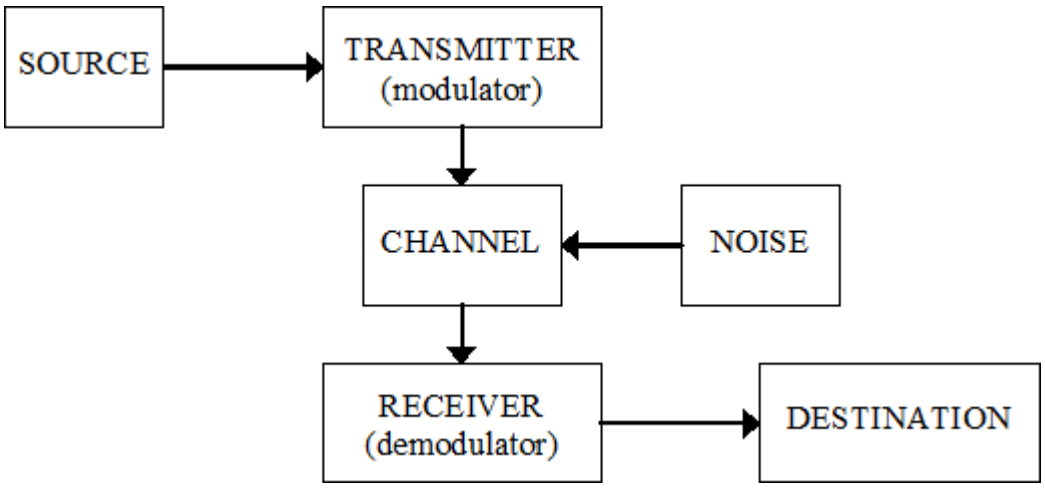
UNIT-I

Introduction to Communication System

Communication is the process by which information is exchanged between individuals through a medium.

Communication can also be defined as the transfer of information from one point in space and time to another point.

The basic block diagram of a communication system is as follows.



- **Transmitter:** Couples the message into the channel using high frequency signals.
- **Channel:** The medium used for transmission of signals
- **Modulation:** It is the process of shifting the frequency spectrum of a signal to a frequency range in which more efficient transmission can be achieved.
- **Receiver:** Restores the signal to its original form.
- **Demodulation:** It is the process of shifting the frequency spectrum back to the original baseband frequency range and reconstructing the original form.

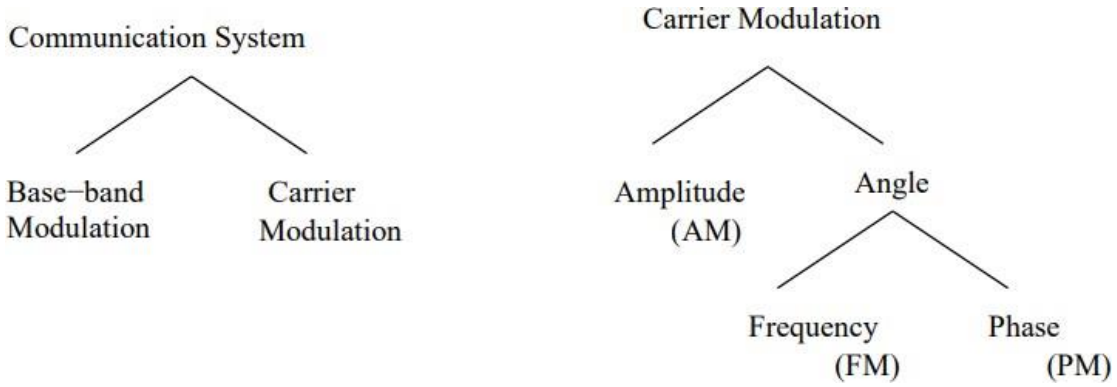
Modulation:

Modulation is a process that causes a shift in the range of frequencies in a signal.

• Signals that occupy the same range of frequencies can be separated.

• Modulation helps in noise immunity, attenuation - depends on the physical medium.

The below figure shows the different kinds of analog modulation schemes that are available



Modulation is operation performed at the transmitter to achieve efficient and reliable information transmission.

For analog modulation, it is frequency translation method caused by changing the appropriate quantity in a carrier signal.

It involves two waveforms:

- A modulating signal/baseband signal – represents the message.
- A carrier signal – depends on type of modulation.

•Once this information is received, the low frequency information must be removed from the high frequency carrier. •This process is known as "Demodulation".

Need for Modulation:

- Baseband signals are incompatible for direct transmission over the medium so, modulation is used to convey (baseband) signals from one place to another.
- Allows frequency translation:
    - Frequency Multiplexing
    - Reduce the antenna height
    - Avoids mixing of signals
    - Narrowbanding
- Efficient transmission
- Reduced noise and interference

Types of Modulation:

Three main types of modulations:

Analog Modulation

- **Amplitude modulation**

Example: Double sideband with carrier (DSB-WC), Double- sideband suppressed carrier (DSB-SC), Single sideband suppressed carrier (SSB-SC), vestigial sideband (VSB)

- **Angle modulation (frequency modulationλ & phase modulation)**

  Example: Narrow band frequency modulation (NBFM), Widebandλ frequency modulation (WBFM), Narrowband phase modulation (NBPM), Wideband phase modulation (NBPM)

Pulse Modulation

- Carrier is a train of pulses
- Example: Pulse Amplitude Modulation (PAM), Pulse width modulation (PWM) , Pulse Position Modulation (PPM)

Digital Modulation

- Modulating signal is analog
  - Example: Pulse Code Modulation (PCM), Delta Modulationλ (DM), Adaptive Delta Modulation (ADM), Differential Pulse Code Modulation (DPCM), Adaptive Differential Pulse Code Modulation (ADPCM) etc.
- Modulating signal is digital (binary modulation)
  - Example: Amplitude shift keying (ASK), frequency Shift Keyingλ (FSK), Phase Shift Keying (PSK) etc

Frequency Division Multiplexing

Multiplexing is the name given to techniques, which allow more than one message to be transferred via the same communication channel. The channel in this context could be a transmission line, *e.g.* a twisted pair or co-axial cable, a radio system or a fibre optic system *etc*.

FDM is derived from AM techniques in which the signals occupy the same physical 'line' but in different frequency bands. Each signal occupies its own specific band of frequencies all the time, *i.e.* the messages share the channel **bandwidth**.

- FDM – messages occupy **narrow** bandwidth – all the time.

Multiplexing requires that the signals be kept apart so that they do not interfere with each other, and thus they can be separated at the receiving end. This is accomplished by separating the signal either in frequency or time. The technique of separating the signals in frequency is referred to as frequency-division multiplexing (FDM), whereas the technique of separating the signals in time is called time-division multiplexing. In this section, we discuss frequency division multiplexing systems, referred hereafter as FDM.

Fig. 1 shows the block diagram of FDM system. As shown in the Fig. 1 , input message signals, assumed to be of the low-pass type are passed through input low-pass filters. This filtering action removes high-frequency components that do not contribute significantly to signal representation but may disturb other message signals that share the common channel. The filtered message signals are then modulated with necessary carrier frequencies with the help of modulators. The most commonly method of modulation in FDM is single sideband modulation, which requires a bandwidth that is approximately equal to that of original message signal. The band pass filters following the modulators are used to restrict the band of each modulated wave to its prescribed range. The outputs of band-pass filters are combined in parallel which form the input to the common channel.

At the receiving end, bandpass filters connected to the common channel separate the message signals on the frequency occupancy basis. Finally, the original message signals are recovered by individual demodulators.

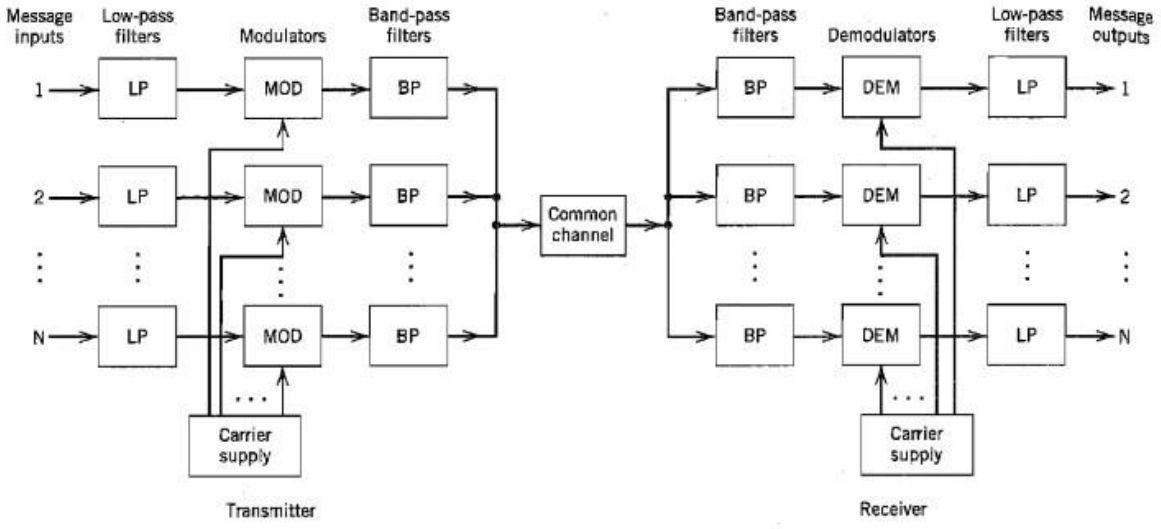


Fig.1. Frequency Division Multiplexing

Amplitude Modulation (AM)

Amplitude Modulation is the process of changing the amplitude of a relatively high frequency carrier signal in accordance with the amplitude of the modulating signal (Information).

The carrier amplitude varied linearly by the modulating signal which usually consists of a range of audio frequencies. The frequency of the carrier is not affected.

- Application of AM - Radio broadcasting, TV pictures (video), facsimile transmission
- Frequency range for AM - 535 kHz – 1600 kHz
- Bandwidth - 10 kHz

Various forms of Amplitude Modulation

• Conventional Amplitude Modulation (Alternatively known as Full AM or Double Sideband Large carrier modulation (DSBLC) /Double Sideband Full Carrier (DSBFC)

• Double Sideband Suppressed carrier (DSBSC) modulation

• Single Sideband (SSB) modulation

• Vestigial Sideband (VSB) modulation

Time Domain and Frequency Domain Description

It is the process where, the amplitude of the carrier is varied proportional to that of the message signal.

Let m (t) be the base-band signal, m (t) $\longleftrightarrow$ M (ω) and c (t) be the carrier, $c(t) = A_c \cos(\omega_c t)$. fc is chosen such that fc >> W, where W is the maximum frequency component of m(t). The amplitude modulated signal is given by

$$s(t) = Ac\,[1 + k_a m(t)]\cos(\omega ct)$$

Fourier Transform on both sides of the above equation

$$S(\omega) = \pi\, Ac/2\,(\delta(\omega - \omega c) + \delta(\omega + \omega c)) + k_a Ac/\,2\,(M(\omega - \omega c) + M(\omega + \omega c))$$

k_a is a constant called amplitude sensitivity.

$k_a m(t) < 1$ and it indicates percentage modulation.

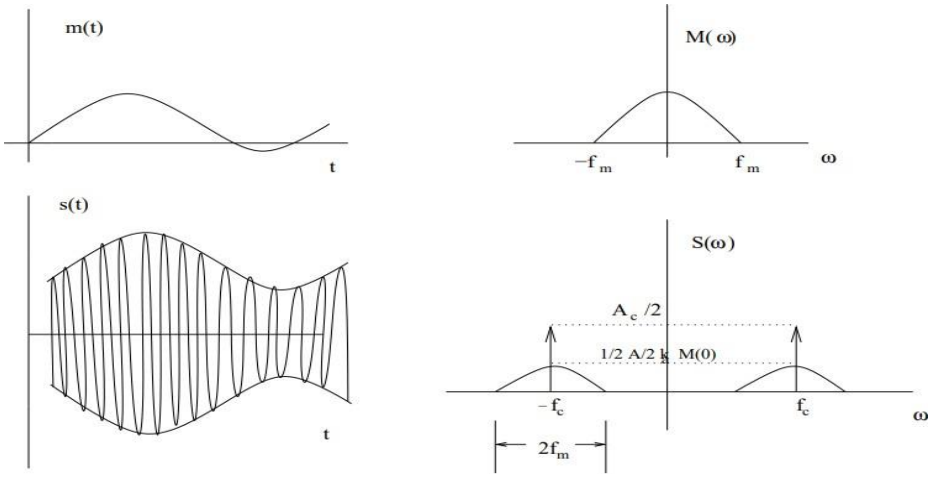


Fig.2. Amplitude modulation in time and frequency domain

Single Tone Modulation:

Consider a modulating wave m(t) that consists of a single tone or single frequency component given by

$$m(t) = A_m \cos(2\pi f_m t) \quad \text{.....(1)}$$

where A_m is peak amplitude of the sinusoidal modulating wave

f_m is the frequency of the sinusoidal modulating wave

Let A_c be the peak amplitude and f_c be the frequency of the high frequency carrier signal. Then the corresponding single-tone AM wave is given by

$$s(t) = A_c[1 + m\cos(2\pi f_m t)]Cos(2\pi f_c t) \quad \text{.....(2)}$$

Let $A_{\max}$ and $A_{\min}$ denote the maximum and minimum values of the envelope of the modulated wave. Then from the above equation (2.12), we get

$$\frac{A_{\max}}{A_{\min}} = \frac{A_c(1+m)}{A_c(1-m)}$$

$$m = \frac{A_{\max} - A_{\min}}{A_{\max} + A_{\min}}$$

Expanding the equation (2), we get

$$s(t) = A_c \cos(2\pi f_c t) + \frac{1}{2} mA_c \cos[2\pi(f_c + f_m)t] + \frac{1}{2} mA_c \cos[2\pi(f_c - f_m)t]$$

The Fourier transform of $s(t)$ is obtained as follows.

$$s(f) = \frac{1}{2} A_c[\delta(f - f_c) + \delta(f + f_c)] + \frac{1}{4} mA_c[\delta(f - f_c - f_m) + \delta(f + f_c + f_m)]$$

$$+ \frac{1}{4} mA_c[\delta(f - f_c + f_m) + \delta(f + f_c - f_m)]$$

Thus the spectrum of an AM wave, for the special case of sinusoidal modulation consists of delta functions at $\pm f_c$, $f_c \pm f_m$, and $-f_c \pm f_m$. The spectrum for positive frequencies is as shown in figure

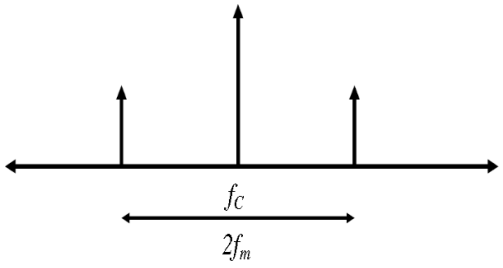


Fig.3. Frequency Domain characteristics of single tone AM

Power relations in AM waves:

Consider the expression for single tone/sinusoidal AM wave

$$s(t) = A_c Cos(2\pi f_c t) + \frac{1}{2} m A_c Cos[2\pi(f_c + f_m)t] + \frac{1}{2} m A_c Cos[2\pi(f_c - f_m)t] \quad \text{.....(1)}$$

This expression contains three components. They are carrier component, upper side band and lower side band. Therefore Average power of the AM wave is sum of these three components.

Therefore the total power in the amplitude modulated wave is given by

$$Pt = \frac{V^2_{car}}{R} + \frac{V^2_{LSB}}{R} + \frac{V^2_{USB}}{R} \quad \text{.....(2)}$$

Where all the voltages are rms values and R is the resistance, in which the power is dissipated.

$$P_C = \frac{V^2_{car}}{R} = \frac{\left(A_c/\sqrt{2}\right)^2}{R} = \frac{A_c^2}{2R}$$

$$P_{LSB} = \frac{V_{LSB}^2}{R} = \left(\frac{mA_c}{2\sqrt{2}}\right)^2 \frac{1}{R} = \frac{m^2 A_c^2}{8R} = \frac{m^2}{4} P_c$$

$$P_{USB} = \frac{V_{USB}^2}{R} = \left(\frac{mA_c}{2\sqrt{2}}\right)^2 \frac{1}{R} = \frac{m^2 A_c^2}{8R} = \frac{m^2}{4} P_c$$

Therefore total average power is given by

$$P_t = P_c + P_{LSB} + P_{USB}$$

$$P_t = P_c + \frac{m^2}{4} P_c + \frac{m^2}{4} P_c$$

$$P_t = P_c\left(1 + \frac{m^2}{4} + \frac{m^2}{4}\right)$$

$$P_t = P_c\left(1 + \frac{m^2}{2}\right) \qquad \text{.....(3)}$$

The ratio of total side band power to the total power in the modulated wave is given by

$$\frac{P_{SB}}{P_t} = \frac{P_c\left(m^2/2\right)}{P_c\left(1 + m^2/2\right)}$$

$$\frac{P_{SB}}{P_t} = \frac{m^2}{2 + m^2} \qquad \text{.....(4)}$$

This ratio is called the efficiency of AM system

Generation of AM waves:

Two basic amplitude modulation principles are discussed. They are square law modulation and switching modulator.

Square Law Modulator

When the output of a device is not directly proportional to input throughout the operation, the device is said to be non-linear. The Input-Output relation of a non-linear device can be expressed as

$$V_0 = a_0 + a_1 V_{in} + a_2 V_{in}^2 + a_3 V_{in}^3 + a_4 V_{in}^4 + \ldots\ldots$$

When the in put is very small, the higher power terms can be neglected. Hence the output is approximately given by $V_0 = a_O + a_1 V_{in} + a_2 V_{in}^2$

When the output is considered up to square of the input, the device is called a square law device and the square law modulator is as shown in the figure 4

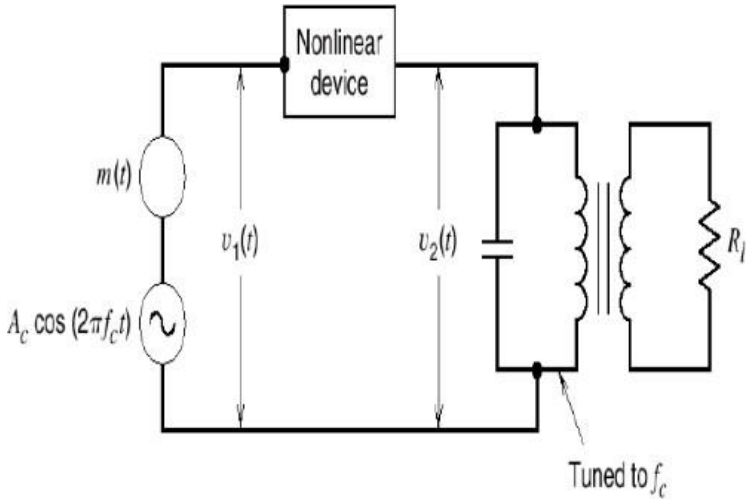


Fig.4. Square Law Modulator

Consider a non-linear device to which a carrier $c(t)=A_c\cos(2\pi f_c t)$ and an information signal m(t) are fed simultaneously as shown in figure 4. The total input to the device at any instant is

$$V_{in} = c(t) + m(t)$$
$$V_{in} = A_c \cos 2\pi f_c t + m(t)$$

As the level of the input is very small, the output can be considered up to square of the input, *i.e.*, $V_0 = a_O + a_1 V_{in} + a_2 V_{in}^2$

$$V_0 = a_0 + a_1[A_c \cos 2\pi f_c t + m(t)] + a_2[A_c \cos 2\pi f_c t + m(t)]^2$$

$$V_0 = a_0 + a_1 A_c \cos 2\pi f_c t + a_1 m(t) + \frac{a_2 A_c^2}{2}(1 + \cos 4\pi f_c t) + a_2[m(t)]^2 + 2a_2 m(t) A_c \cos 2\pi f_c t$$

$$V_0 = a_0 + a_1 A_c \cos 2\pi f_c t + a_1 m(t) + \frac{a_2 A_c^2}{2} \cos 4\pi f_c t + a_2 m^2(t) + 2a_2 m(t) A_c \cos 2\pi f_c t$$

Taking Fourier transform on both sides, we get

$$V_0(f) = (a_0 + \frac{a_2 A_c^2}{2})\delta(f) + \frac{a_1 A_c}{2}[\delta(f - f_c) + \delta(f + f_c)] + a_1 M(f) +$$

$$\frac{a_2 A_c^2}{4}[\delta(f - 2f_c) + \delta(f + 2f_c)] + a_2 M(f) + a_2 A_c[M(f - f_c) + M(f + f_c)]$$

Therefore the square law device output 0 V consists of the dc component at f = 0. The information signal ranging from 0 to W Hz and its second harmonics are signal at f_c and $2f_c$.

Frequency band centered at f_c with a deviation of $\pm W$, Hz.

The required AM signal with a carrier frequency f_c can be separated using a band pass filter at the out put of the square law device. The filter should have a lower cut-off frequency ranging between 2W and (f_c-W) and upper cut-off frequency between (f_c+W) and 2 f_c

Therefore the filter out put is

$$s(t) = a_1 A_c \cos 2\pi f_c t + 2a_2 A_c m(t) \cos 2\pi f_c t$$

$$s(t) = a_1 A_c \left[1 + 2\frac{a_2}{a_1} m(t) \right] \cos 2\pi f_c t$$

If $m(t) = A_m \cos 2\pi f_m t$, we get

$$s(t) = a_1 A_c \left[1 + 2\frac{a_2}{a_1} A_m \cos 2\pi f_m t \right] \cos 2\pi f_c t$$

Comparing this with the standard representation of AM signal,

$$s(t) = A_c [1 + k_a m(t)] \cos(2\pi f_c t)$$

Therefore modulation index of the output signal is given by

$$m = 2\frac{a_2}{a_1} A_m$$

The output AM signal is free from distortion and attenuation only when ($f_c - W$)> 2W or $f_c > 3W$.

Spectrum is as shown below

Switching Modulator

Consider a semiconductor diode used as an ideal switch to which the carrier signal $c(t) = A_c \cos(2\pi f_c t)$ and information signal $m(t)$ are applied simultaneously as shown figure

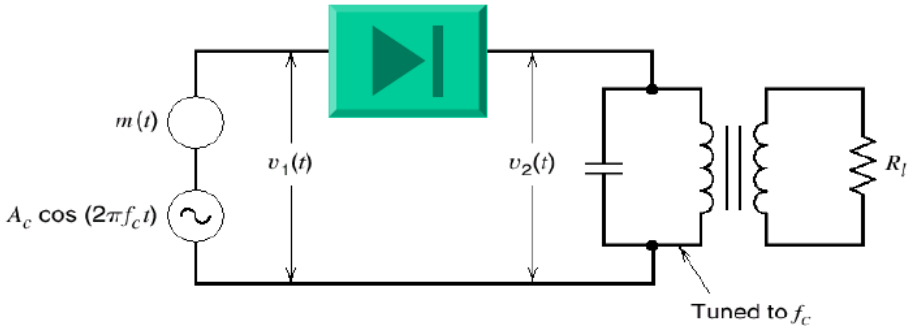


Fig.5. Switching Modulator

The total input for the diode at any instant is given by

$$v_1 = c(t) + m(t)$$

$$v_1 = A_c \cos 2\pi f_c t + m(t)$$

When the peak amplitude of c(t) is maintained more than that of information signal, the operation is assumed to be dependent on only c(t) irrespective of m(t).

When c(t) is positive, v2=v1since the diode is forward biased. Similarly, when c(t) is negative, v2=0 since diode is reverse biased. Based upon above operation, switching response of the diode is periodic rectangular wave with an amplitude unity and is given by

$$p(t) = \frac{1}{2} + \frac{1}{\pi}\sum_{n=-\infty}^{\infty}\frac{(-1)^{n-1}}{2n-1}\cos(2\pi f_c t(2n-1))$$

$$p(t) = \frac{1}{2} + \underset{n=0,1}{\frac{2}{\pi}\cos(2\pi f_c t)} - \underset{n=-1,+2}{\frac{2}{3\pi}\cos(6\pi f_c t)} + -$$

Therefore the diode response V_o is a product of switching response $p(t)$ and input v_1.

$$v_2 = v_1 * p(t)$$

$$V_2 = \left[A_c \cos 2\pi f_c t + m(t)\right]\left[\frac{1}{2} + \frac{2}{\pi}\cos 2\pi f_c t - \frac{2}{3\pi}\cos 6\pi f_c t + - + -\right]$$

Applying the Fourier Transform, we get

$$V_2(f) = \frac{A_c}{4}\left[\delta(f - f_c) + \delta(f + f_{c)}\right] + \frac{M(f)}{2} + \frac{A_c}{\pi}\delta(f)$$
$$+ \frac{A_c}{2\pi}\left[\delta(f - 2f_c) + \delta(f + 2f_c)\right] + \frac{1}{\pi}\left[M(f - f_c) + M(f + f_c)\right]$$
$$- \frac{A_c}{6\pi}\left[\delta(f - 4f_c) + \delta(f + 4f_c)\right] - \frac{A_c}{3\pi}\left[\delta(f - 2f_c) + \delta(f + 2f_c)\right]$$
$$- \frac{1}{3\pi}\left[M(f - 3f_c) + M(f + f_c)\right]$$

The diode output v_2 consists of

a dc component at $f = 0$.

Information signal ranging from 0 to w Hz and infinite number of frequency bands centered at f, $2f_c$, $3f_c$, $4f_c$, ---------

The required AM signal centred at fc can be separated using band pass filter. The lower cut off-frequency for the band pass filter should be between w and fc-w and the upper cut-off frequency between fc+w and 2fc. The filter output is given by the equation

$$S(t) = \frac{A_c}{2}\left[1 + \frac{4}{\pi}\frac{m(t)}{A_c}\right]\cos 2\pi f_c t$$

For a single tone information, let $m(t) = A_m \cos(2\pi f_m t)$

$$S(t) = \frac{A_c}{2}\left[1 + \frac{4}{\pi}\frac{A_m}{A_c}\cos 2\pi f_m t\right]\cos 2\pi f_c t$$

Therefore modulation index, $m = \frac{4}{\pi}\frac{A_m}{A_c}$

The output AM signal is free from distortions and attenuations only when $fc\text{-}w$>w or $fc>2w$.

Detection of AM waves

Demodulation is the process of recovering the information signal (base band) from the incoming modulated signal at the receiver. There are two methods, they are Square law Detector and Envelope Detector

Square Law Detector

Consider a non-linear device to which the AM signal s(t) is applied. When the level of s(t) is very small, output can be considered upto square of the input.

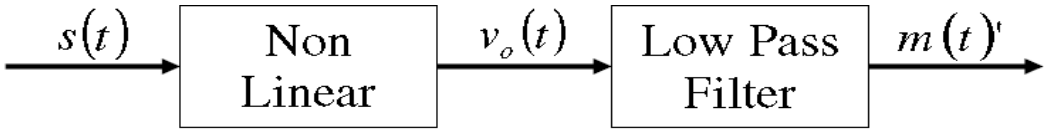


Figure: Demodulation of AM using square law device

Therefore $V_o = a_o + a_1 V_{in} + a_2 V_{in}^2$

If $m(t)$ is the information signal (0-wHz) and $c(t) = A_c \cos(2\pi f_c t)$is the carrier, input AM signal to the non-linear device is given by

$$s(t) = A_c[1 + k_a m(t)]\cos(2\pi f_c t)$$

$$V_o = a_o + a_1 s(t) + a_2 [s(t)]^2$$

$$V_o = a_o + a_1 A_c \cos 2\pi f_c t + a_1 A_c K_a m(t) \cos 2\pi f_c t + a_2 [A_c \cos 2\pi f_c t + A_c k_a m(t) \cos 2\pi f_c t]^2$$

Applying Fourier transform on both sides, we get

$$V_o(f) = \left[a_o + \frac{a_2 A_c^{\,2}}{2} \right] \delta(f) + \frac{a_1 A_c}{2} [\delta(f - f_c) + \delta(f + f_c)]$$

$$+ \frac{a_1 A_c K_a}{2} [M(f - f_c) + M(f + f_c)] + \frac{a_2 A_c^{\,2} K_a^{\,2}}{4} [M(f - 2f_c) + M(f + 2f_c)]$$

$$+ \frac{a_2 A_c^{\,2} K_a^{\,2}}{2} \left[\underset{\pm 2W}{M(f)} \right] + \frac{a_2 A_c^{\,2} K_a^{\,2}}{2} [M(f - 2f_c) + M(f + 2f_c)]$$

$$+ \frac{a_2 A_c^{\,2}}{4} [\delta(f - 2f_c) + \delta(f + 2f_c)] + a_2 A_c^{\,2} K_a [M(f)]$$

The device output consists of a dc component at f =0, information signal ranging from 0-W Hz and its second harmonics and frequency bands centered at fc and 2fc. The required information can be separated using low pass filter with cut off frequency ranging between W and fc-w. The filter output is given by

$$m'(t) = \left(a_o + \frac{a_2 A_c^{\,2}}{2} \right) + a_2 A_c^{\,2} K_a m(t) + \frac{a_2 A_c^{\,2} K_a^{\,2} m^2(t)}{2}$$

DC component + message signal + second harmonic

The dc component (first term) can be eliminated using a coupling capacitor or a transformer. The effect of second harmonics of information signal can be reduced by maintaining its level very low. When *m(t)* is very low, the filter output is given by

$$m^1(t) = a_2 A_c^{\,2} K_a m(t)$$

When the information level is very low, the noise effect increases at the receiver, hence the system clarity is very low using square law demodulator.

Envelope Detector

It is a simple and highly effective system. This method is used in most of the commercial AM radio receivers. An envelope detector is as shown below.

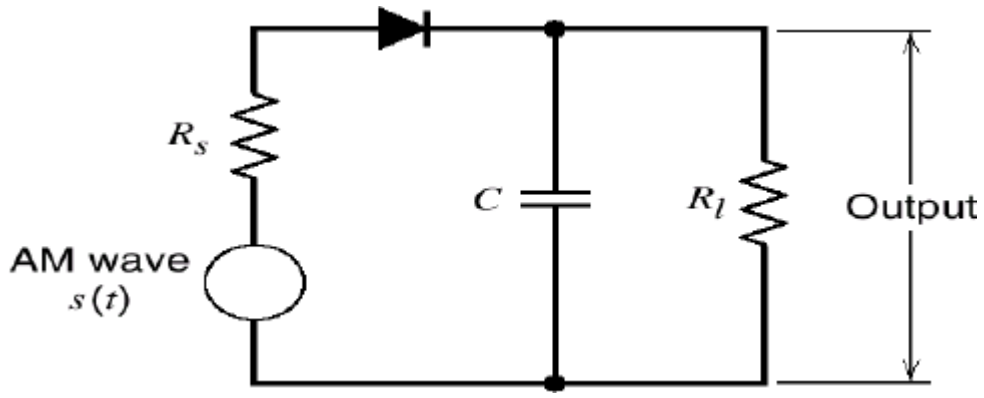


Fig.7. Envelope Detector

During the positive half cycles of the input signals, the diode D is forward biased and the capacitor C charges up rapidly to the peak of the input signal. When the input signal falls below this value, the diode becomes reverse biased and the capacitor C discharges through the load resistor RL.

The discharge process continues until the next positive half cycle. When the input signal becomes greater than the voltage across the capacitor, the diode conducts again and the process is repeated.

The charge time constant (rf+Rs)C must be short compared with the carrier period, the capacitor charges rapidly and there by follows the applied voltage up to the positive peak when the diode is conducting.That is the charging time constant shall satisfy the condition,

$$(r_f + R_s)C << \frac{1}{f_c}$$

On the other hand, the discharging time-constant R_LC must be long enough to ensure that the capacitor discharges slowly through the load resistor R_L between the positive peaks of the carrier wave, but not so long that the capacitor voltage will not discharge at the maximum rate of change of the modulating wave.

That is the discharge time constant shall satisfy the condition,

$$\frac{1}{f_c} << R_L C << \frac{1}{W}$$

Where 'W' is band width of the message signal. The result is that the capacitor voltage or detector output is nearly the same as the envelope of AM wave.

Advantages and Disadvantages of AM:

Advantages of AM:

- Generation and demodulation of AM wave are easy.
- AM systems are cost effective and easy to build.

Disadvantages:

- AM contains unwanted carrier component, hence it requires more transmission power.
- The transmission bandwidth is equal to twice the message bandwidth.

To overcome these limitations, the conventional AM system is modified at the cost of increased system complexity. Therefore, three types of modified AM systems are discussed.

DSBSC (Double Side Band Suppressed Carrier) modulation: In DSBC modulation, the modulated wave consists of only the upper and lower side bands. Transmitted power is saved through the suppression of the carrier wave, but the channel bandwidth requirement is the same as before.

SSBSC (Single Side Band Suppressed Carrier) modulation: The SSBSC modulated wave consists of only the upper side band or lower side band. SSBSC is suited for transmission of voice signals. It is an optimum form of modulation in that it requires the minimum transmission power and minimum channel band width. Disadvantage is increased cost and complexity.

VSB (Vestigial Side Band) modulation: In VSB, one side band is completely passed and just a trace or vestige of the other side band is retained. The required channel bandwidth is therefore in excess of the message bandwidth by an amount equal to the width of the vestigial side band. This method is suitable for the transmission of wide band signals.

DSB-SC MODULATION

DSB-SC Time domain and Frequency domain Description:

DSBSC modulators make use of the multiplying action in which the modulating signal multiplies the carrier wave. In this system, the carrier component is eliminated and both upper and lower side bands are transmitted. As the carrier component is suppressed, the power required for transmission is less than that of AM.

If $m(t)$ is the message signal and $c(t) = A_c \cos(2\pi f_c t)$ is the carrier signal, then DSBSC modulated wave $s(t)$ is given by

$$s(t) = c(t)\, m(t)$$

$$s(t) = A_c \cos(2\pi f_c t) m(t)$$

Consequently, the modulated signal s(t) under goes a phase reversal , whenever the message signal m(t) crosses zero as shown below.

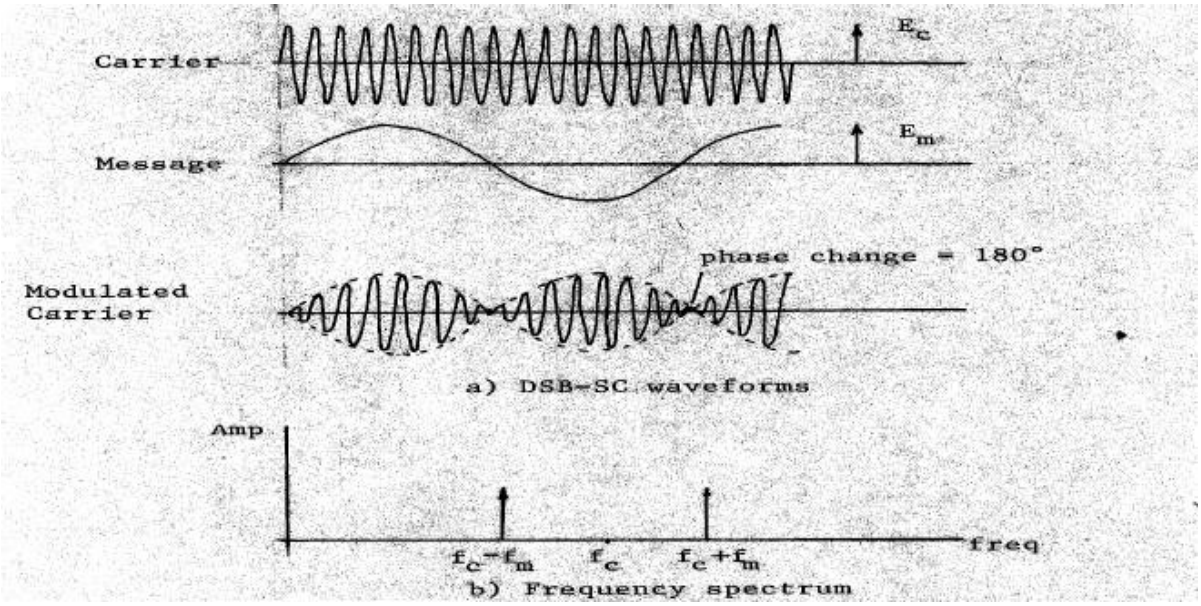


Fig.1. (a) DSB-SC waveform (b) DSB-SC Frequency Spectrum

The envelope of a DSBSC modulated signal is therefore different from the message signal and the Fourier transform of s(t) is given by

$$S(f)=\frac{A_c}{2}[M(f-f_c)+M(f+f_c)]$$

For the case when base band signal *m(t)* is limited to the interval $-W<f<W$ as shown in figure below, we find that the spectrum *S(f)* of the DSBSC wave *s(t)* is as illustrated below. Except for a change in scaling factor, the modulation process simply translates the spectrum of the base band signal by f_c. The transmission bandwidth required by DSBSC modulation is the same as that for AM.

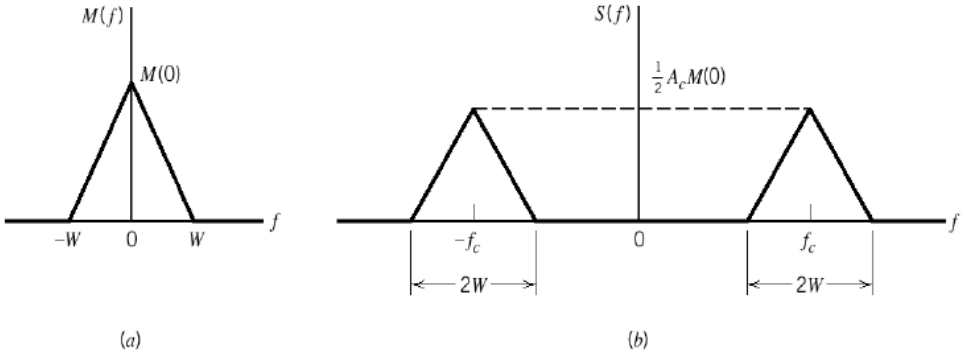


Figure: Message and the corresponding DSBSC spectrum

Generation of DSBSC Waves:

Balanced Modulator (Product Modulator)

A balanced modulator consists of two standard amplitude modulators arranged in a balanced configuration so as to suppress the carrier wave as shown in the following block diagram. It is assumed that the AM modulators are identical, except for the sign reversal of the modulating wave applied to the input of one of them. Thus, the output of the two modulators may be expressed as,

$$s_1(t)=A_c[1+k_a m(t)]\cos(2\pi f_c t)$$
$$s_2(t)=A_c[1-k_a m(t)]\cos(2\pi f_c t)$$

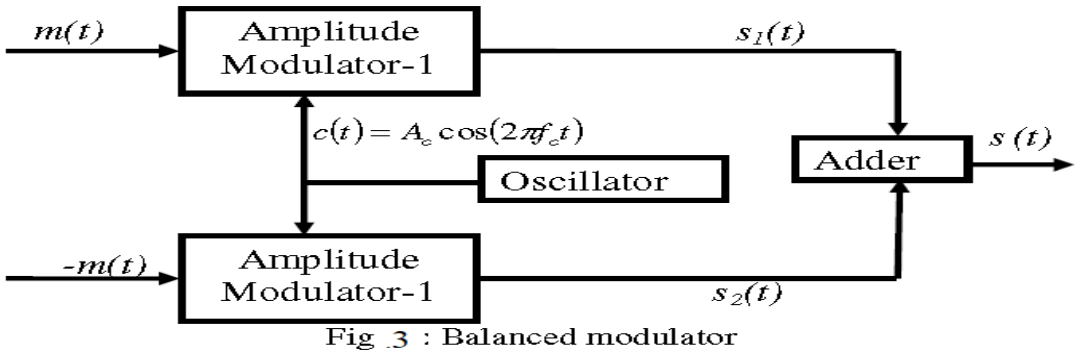


Fig 3 : Balanced modulator

Subtracting $s_2(t)$ from $s_1(t)$,

$$s(t)=s_1(t)-s_2(t)$$
$$s(t)=2k_a m(t)A_c\cos(2\pi f_c t)$$

Hence, except for the scaling factor 2ka, the balanced modulator output is equal to the product of the modulating wave and the carrier.

Ring Modulator

Ring modulator is the most widely used product modulator for generating DSBSC wave and is shown below.

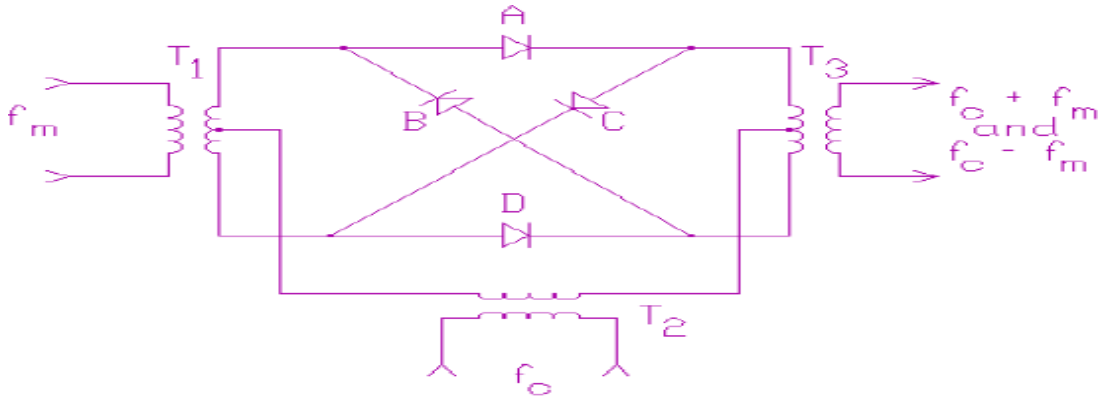


Fig.4 : Ring modulator

The four diodes form a ring in which they all point in the same direction. The diodes are controlled by square wave carrier c(t) of frequency fc, which is applied longitudinally by means of two center-tapped transformers. Assuming the diodes are ideal, when the carrier is positive, the outer diodes D1 and D2 are forward biased where as the inner diodes D3 and D4 are reverse biased, so that the modulator multiplies the base band signal m(t) by c(t). When the carrier is negative, the diodes D1 and D2 are reverse biased and D3 and D4 are forward, and the modulator multiplies the base band signal –m(t) by c(t).

Thus the ring modulator in its ideal form is a product modulator for square wave carrier and the base band signal m(t). The square wave carrier can be expanded using Fourier series as

$$c(t) = \frac{4}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{2n-1}\cos(2\pi f_c t(2n-1))$$

Therefore the ring modulator out put is given by

$$s(t) = m(t)c(t)$$

$$s(t) = m(t)\left[\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{2n-1}\cos(2\pi f_c t(2n-1))\right]$$

From the above equation it is clear that output from the modulator consists entirely of modulation products. If the message signal m(t) is band limited to the frequency band − w < f < w, the output spectrum consists of side bands centred at fc.

Detection of DSB-SC waves:

Coherent Detection:

The message signal m(t) can be uniquely recovered from a DSBSC wave s(t) by first multiplying s(t) with a locally generated sinusoidal wave and then low pass filtering the product as shown.

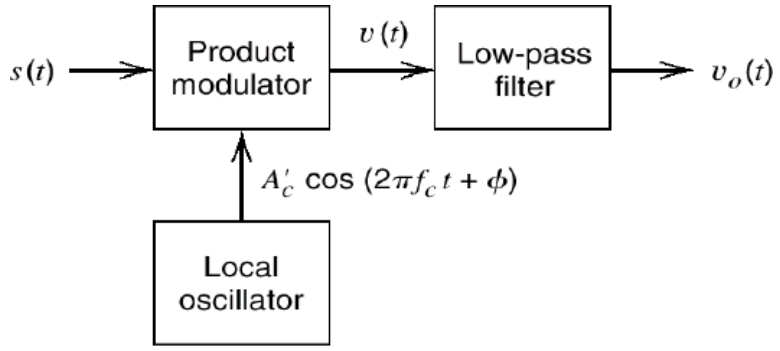


Fig.5 : Coherent detector

It is assumed that the local oscillator signal is exactly coherent or synchronized, in both frequency and phase, with the carrier wave c(t) used in the product modulator to generate s(t). This method of demodulation is known as coherent detection or synchronous detection.

Let $A_c^{\,1}\cos(2\pi f_c t + \phi)$ be the local oscillator signal, and $s(t) = A_c \cos(2\pi f_c t)m(t)$ be the DSBSC wave. Then the product modulator output $v(t)$ is given by

$$v(t) = A_c A_c^{\,1} \cos(2\pi f_c t)\cos(2\pi f_c t + \phi)m(t)$$

$$v(t) = \frac{A_c A_c^{\,1}}{4}\cos(4\pi f_c t + \phi)m(t) + \frac{A_c A_c^{\,1}}{2}\cos(\phi)m(t)$$

The first term in the above expression represents a DSBSC modulated signal with a carrier frequency $2f_c$, and the second term represents the scaled version of message signal. Assuming that the message signal is band limited to the interval $-w < f < w$, the spectrum of $v(t)$ is plotted as shown below.

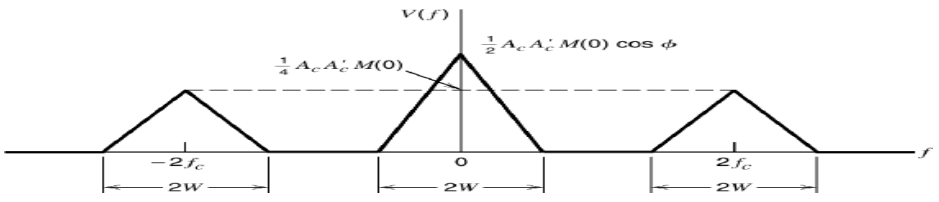


Fig.6.Spectrum of output of the product modulator

From the spectrum, it is clear that the unwanted component (first term in the expression) can be removed by the low-pass filter, provided that the cut-off frequency of the filter is greater than W but less than 2fc-W. The filter output is given by

$$v_o(t) = \frac{A_c A_c^{\,1}}{2}\cos(\phi)m(t)$$

The demodulated signal $v_o(t)$ is therefore proportional to m(t) when the phase error ϕ is constant.

Costas Receiver (Costas Loop):

Costas receiver is a synchronous receiver system, suitable for demodulating DSBSC waves. It consists of two coherent detectors supplied with the same input signal,

that is the incoming DSBSC wave $s(t) = A_c \cos(2\pi f_c t)m(t)$ but with individual local oscillator signals that are in phase quadrature with respect to each other as shown below.

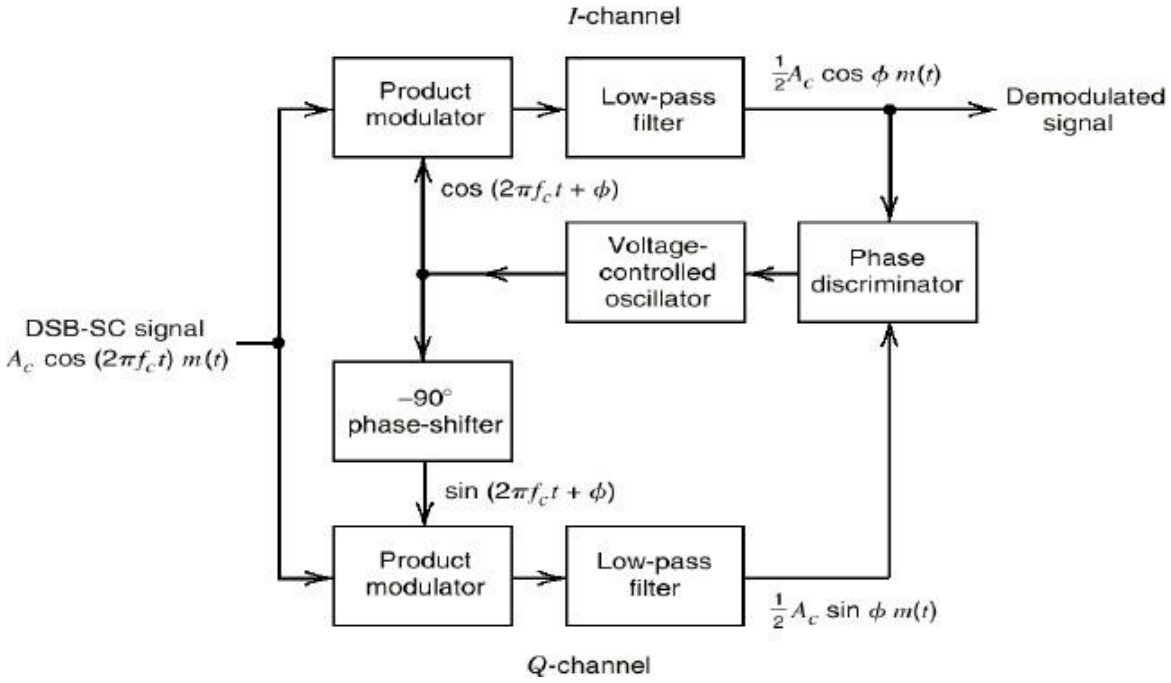


Fig.7. Costas Receiver

The frequency of the local oscillator is adjusted to be the same as the carrier frequency fc. The detector in the upper path is referred to as the in-phase coherent detector or I-channel, and that in the lower path is referred to as the quadrature-phase coherent detector or Q-channel.

These two detector are coupled together to form a negative feedback system designed in such a way as to maintain the local oscillator synchronous with the carrier wave. Suppose

the local oscillator signal is of the same phase as the carrier $c(t) = A_c\cos(2\pi f_c t)$ wave used to generate the incoming DSBSC wave. Then we find that the I-channel output contains the desired demodulated signal m(t), where as the Q-channel output is zero due to quadrature null effect of the Q-channel. Suppose that the local oscillator phase drifts from its proper value by a small angle ϕ radians. The I-channel output will remain essentially unchanged, but there will be some signal appearing at the Q-channel output, which is proportional to $\sin(\phi) \approx \phi$ for small ϕ.

This Q-channel output will have same polarity as the I-channel output for one direction of local oscillator phase drift and opposite polarity for the opposite direction of local oscillator phase drift. Thus by combining the I-channel and Q-channel outputs in a phase discriminator (which consists of a multiplier followed by a LPF), a dc control signal is obtained that automatically corrects for the local phase errors in the voltage-controlled oscillator.

Radio Transmitters

There are two approaches in generating an AM signal. These are known as low and high level modulation. They're easy to identify: A low level AM transmitter performs the process of modulation near the beginning of the transmitter. A high level transmitter performs the modulation step last, at the last or "final" amplifier stage in the transmitter. Each method has advantages and disadvantages, and both are in common use.

Low-Level AM Transmitter:

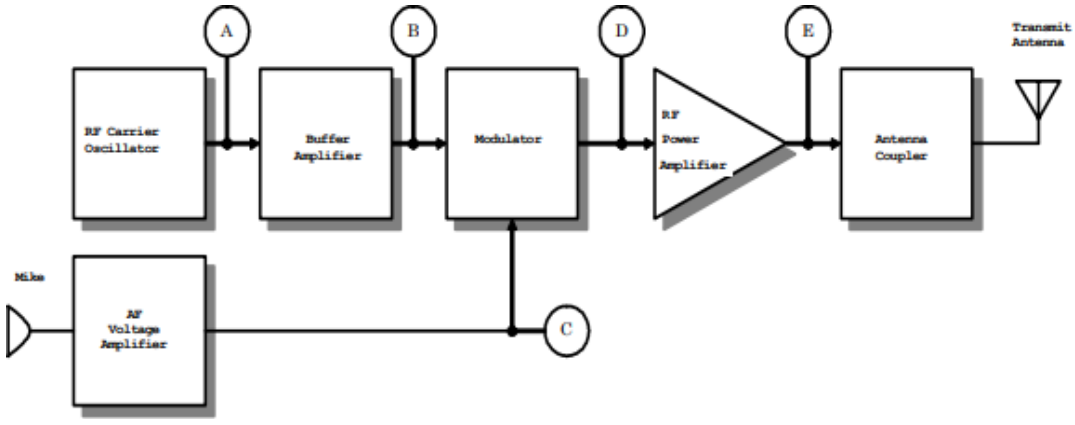


Fig.8. Low-Level AM Transmitter Block Diagram

There are two signal paths in the transmitter, audio frequency (AF) and radio frequency (RF). The RF signal is created in the RF carrier oscillator. At test point A the oscillator's output signal is present. The output of the carrier oscillator is a fairly small AC voltage, perhaps 200 to 400 mV RMS. The oscillator is a critical stage in any transmitter. It must produce an accurate and steady frequency. Every radio station is assigned a different carrier frequency. The dial (or display) of a receiver displays the carrier frequency. If the

oscillator drifts off frequency, the receiver will be unable to receive the transmitted signal without being readjusted. Worse yet, if the oscillator drifts onto the frequency being used by another radio station, interference will occur. Two circuit techniques are commonly used to stabilize the oscillator, buffering and voltage regulation.

The buffer amplifier has something to do with buffering or protecting the oscillator. An oscillator is a little like an engine (with the speed of the engine being similar to the oscillator's frequency). If the load on the engine is increased (the engine is asked to do more work), the engine will respond by slowing down. An oscillator acts in a very similar fashion. If the current drawn from the oscillator's output is increased or decreased, the oscillator may speed up or slow down slightly.

Buffer amplifier is a relatively low-gain amplifier that follows the oscillator. It has a constant input impedance (resistance). Therefore, it always draws the same amount of current from the oscillator. This helps to prevent "pulling" of the oscillator frequency. The buffer amplifier is needed because of what's happening "downstream" of the oscillator. Right after this stage is the modulator. Because the modulator is a nonlinear amplifier, it may not have a constant input resistance -- especially when information is passing into it. But since there is a buffer amplifier between the oscillator and modulator, the oscillator sees a steady load resistance, regardless of what the modulator stage is doing.

Voltage Regulation: An oscillator can also be pulled off frequency if its power supply voltage isn't held constant. In most transmitters, the supply voltage to the oscillator is regulated at a constant value. The regulated voltage value is often between 5 and 9 volts; zener diodes and three-terminal regulator ICs are commonly used voltage regulators. Voltage regulation is especially important when a transmitter is being powered by batteries or an automobile's electrical system. As a battery discharges, its terminal voltage falls. The DC supply voltage in a car can be anywhere between 12 and 16 volts, depending on engine RPM and other electrical load conditions within the vehicle.

Modulator: The stabilized RF carrier signal feeds one input of the modulator stage. The modulator is a variable-gain (nonlinear) amplifier. To work, it must have an RF carrier signal and an AF information signal. In a low-level transmitter, the power levels are low in the oscillator, buffer, and modulator stages; typically, the modulator output is around 10 mW (700 mV RMS into 50 ohms) or less.

AF Voltage Amplifier: In order for the modulator to function, it needs an information signal. A microphone is one way of developing the intelligence signal, however, it only produces a few millivolts of signal. This simply isn't enough to operate the modulator, so a voltage amplifier is used to boost the microphone's signal. The signal level at the output of the AF voltage amplifier is usually at least 1 volt RMS; it is highly dependent upon the transmitter's design. Notice that the AF amplifier in the transmitter is only providing a voltage gain, and not necessarily a current gain for the microphone's signal. The power levels are quite small at the output of this amplifier; a few mW at best.

RF Power Amplifier: At test point D the modulator has created an AM signal by impressing the information signal from test point C onto the stabilized carrier signal from test point B at the buffer amplifier output. This signal (test point D) is a complete AM signal, but has only a few milliwatts of power. The RF power amplifier is normally built with several stages. These stages increase both the voltage and current of the AM signal. We say that power amplification occurs when a circuit provides a current gain. In order to accurately amplify the tiny AM signal from the modulator, the RF power amplifier stages must be linear. You might recall that amplifiers are divided up into "classes," according to the conduction angle of the active device within. Class A and class B amplifiers are considered to be linear amplifiers, so the RF power amplifier stages will normally be constructed using one or both of these type of amplifiers. Therefore, the signal at test point E looks just like that of test point D; it's just much bigger in voltage and current.

Antenna Coupler: The antenna coupler is usually part of the last or final RF power amplifier, and as such, is not really a separate active stage. It performs no amplification, and has no active devices. It performs two important jobs: Impedance matching and filtering. For an RF power amplifier to function correctly, it must be supplied with a load resistance equal to that for which it was designed.

The antenna coupler also acts as a low-pass filter. This filtering reduces the amplitude of harmonic energies that may be present in the power amplifier's output. (All amplifiers generate harmonic distortion, even "linear" ones.) For example, the transmitter may be tuned to operate on 1000 kHz. Because of small nonlinearities in the amplifiers of the transmitter, the transmitter will also produce harmonic energies on 2000 kHz (2nd harmonic), 3000 kHz (3rd harmonic), and so on. Because a low-pass filter passes the fundamental frequency (1000 kHz) and rejects the harmonics, we say that harmonic attenuation has taken place.

High-Level AM Transmitter:

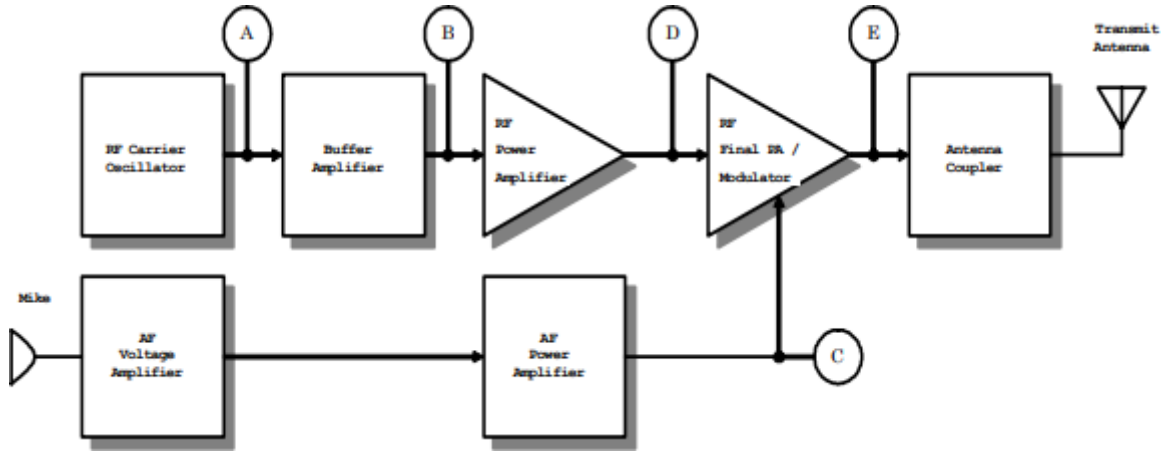


Fig.9. Low-Level AM Transmitter Block Diagram

The high-level transmitter of Figure 9 is very similar to the low-level unit. The RF section begins just like the low-level transmitter; there is an oscillator and buffer amplifier. The difference in the high level transmitter is where the modulation takes place. Instead of

adding modulation immediately after buffering, this type of transmitter amplifies the unmodulated RF carrier signal first. Thus, the signals at points A, B, and D in Figure 9 all look like unmodulated RF carrier waves. The only difference is that they become bigger in voltage and current as they approach test point D.

The modulation process in a high-level transmitter takes place in the last or final power amplifier. Because of this, an additional audio amplifier section is needed. In order to modulate an amplifier that is running at power levels of several watts (or more), comparable power levels of information are required. Thus, an audio power amplifier is required. The final power amplifier does double-duty in a high-level transmitter. First, it provides power gain for the RF carrier signal, just like the RF power amplifier did in the low-level transmitter. In addition to providing power gain, the final PA also performs the task of modulation. The final power amplifier in a high-level transmitter usually operates in class C, which is a highly nonlinear amplifier class.

Comparison:

Low Level Transmitters

- Can produce any kind of modulation; AM, FM, or PM.
- Require linear RF power amplifiers, which reduce DC efficiency and increases production costs.

High Level Transmitters

- Have better DC efficiency than low-level transmitters, and are very well suited for battery operation.
- Are restricted to generating AM modulation only.

Introduction of SSB-SC

Standard AM and DSBSC require transmission bandwidth equal to twice the message bandwidth. In both the cases spectrum contains two side bands of width W Hz, each. But the upper and lower sides are uniquely related to each other by the virtue of their symmetry about the carrier frequency. That is, given the amplitude and phase spectra of either side band, the other can be uniquely determined. Thus if only one side band is transmitted, and if both the carrier and the other side band are suppressed at the transmitter, no information is lost. This kind of modulation is called SSBSC and spectral comparison between DSBSC and SSBSC is shown in the figures 1 and 2.

Figure.1 : Spectrum of the DSBSC wave

Figure .2 : Spectrum of the SSBSC wave

Frequency Domain Description

Consider a message signal $m(t)$ with a spectrum $M(f)$ band limited to the interval $-w < f < w$ as shown in figure 3 , the DSBSC wave obtained by multiplexing $m(t)$ by the carrier wave $c(t) = A_c \cos(2\pi f_c t)$ and is also shown, in figure 4 . The upper side band is represented in duplicate by the frequencies above f_c and those below $-f_c$, and when only upper

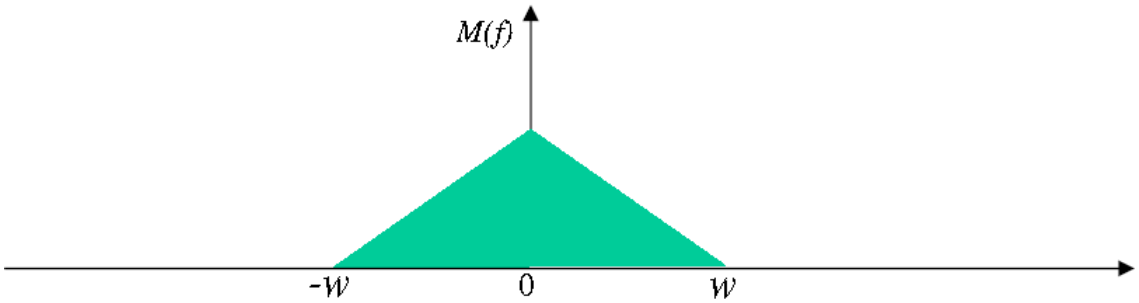


Figure 3. : Spectrum of message wave

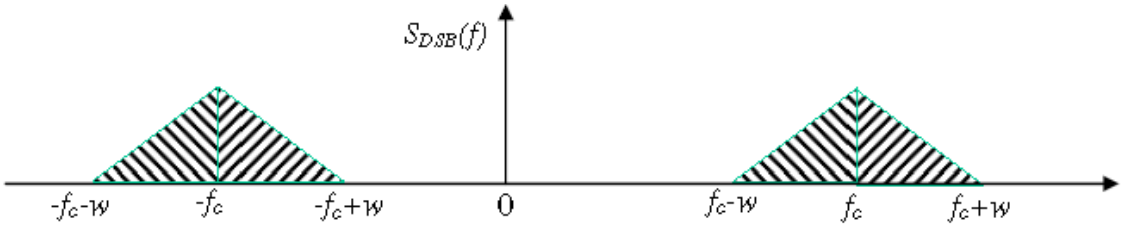


Figure .4 : Spectrum of DSBSC wave

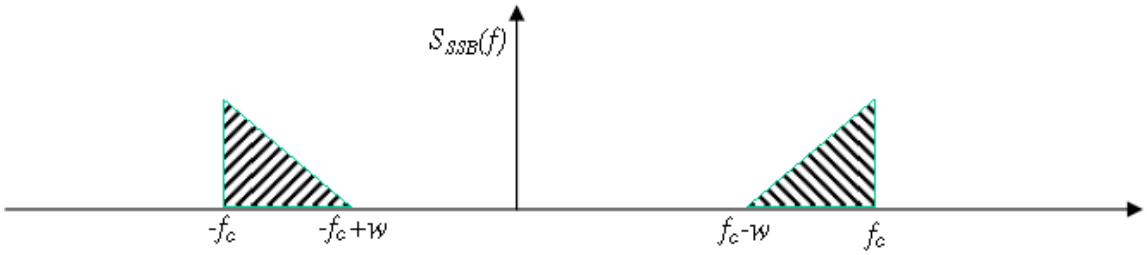


Figure.5 : Spectrum of SSBSC-LSB wave

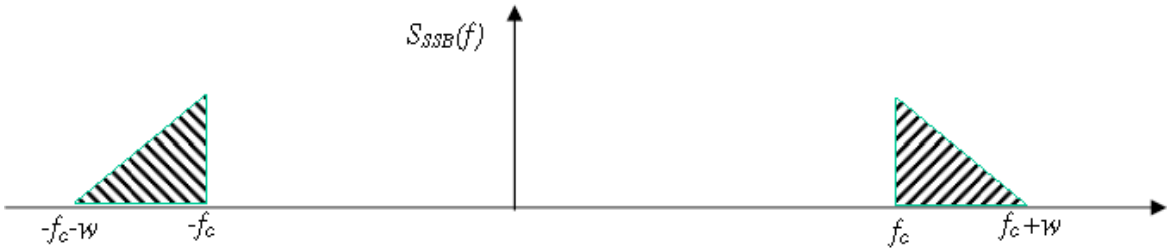


Figure .6 : Spectrum of SSBSC-USB wave

side band is transmitted; the resulting SSB modulated wave has the spectrum shown in figure 1. Similarly, the lower side band is represented in duplicate by the frequencies below fc and those above -fc and when only the lower side band is transmitted, the spectrum of the corresponding SSB modulated wave shown in figure 5.Thus the essential function of the SSB modulation is to translate the spectrum of the modulating wave, either with or without inversion, to a new location in the frequency domain. The advantage of SSB modulation is reduced bandwidth and the elimination of high power carrier wave. The main disadvantage is the cost and complexity of its implementation.

Generation of SSB wave:

Frequency discrimination method

Consider the generation of SSB modulated signal containing the upper side band only. From a practical point of view, the most severe requirement of SSB generation arises from the unwanted sideband, the nearest component of which is separated from the desired side band by twice the lowest frequency component of the message signal. It implies that, for the generation of an SSB wave to be possible, the message spectrum must have an energy gap centered at the origin as shown in figure 7. This requirement is naturally satisfied by voice signals, whose energy gap is about 600Hz wide.

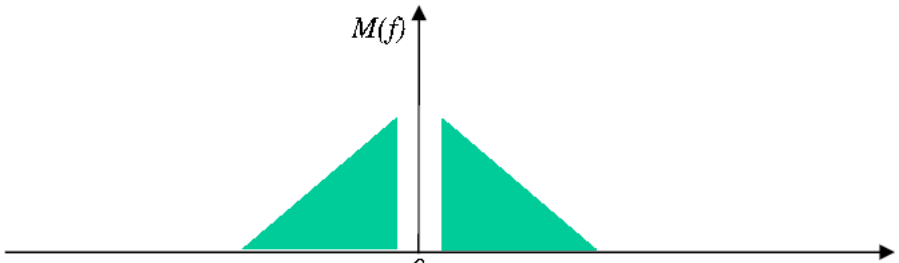


Figure .7 : Message spectrum with energy gap at the origin

The frequency discrimination or filter method of SSB generation consists of a product modulator, which produces DSBSC signal and a band-pass filter to extract the desired side band and reject the other and is shown in the figure 8.

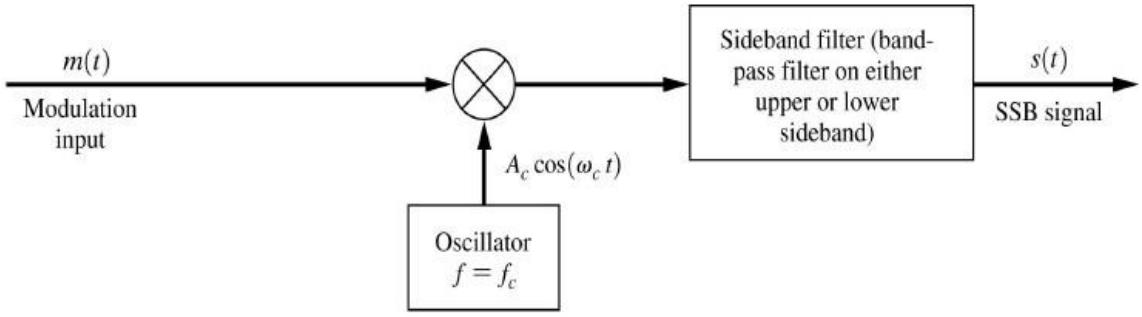


Figure .8 : Frequency discriminator to generate SSBSC wave

Application of this method requires that the message signal satisfies two conditions:

1. The message signal m(t) has no low-frequency content. Example: speech, audio, music.
2. The highest frequency component W of the message signal m(t) is much less than the carrier frequency fc.

Then, under these conditions, the desired side band will appear in a non-overlapping interval in the spectrum in such a way that it may be selected by an appropriate filter.

In designing the band pass filter, the following requirements should be satisfied:

1.The pass band of the filter occupies the same frequency range as the spectrum of the desired SSB modulated wave.

2. The width of the guard band of the filter, separating the pass band from the stop band, where the unwanted sideband of the filter input lies, is twice the lowest frequency component of the message signal.

When it is necessary to generate an SSB modulated wave occupying a frequency band that is much higher than that of the message signal, it becomes very difficult to design an appropriate filter that will pass the desired side band and reject the other. In such a situation it is necessary to resort to a multiple-modulation process so as to ease the filtering

requirement. This approach is illustrated in the following figure 9 involving two stages of modulation.

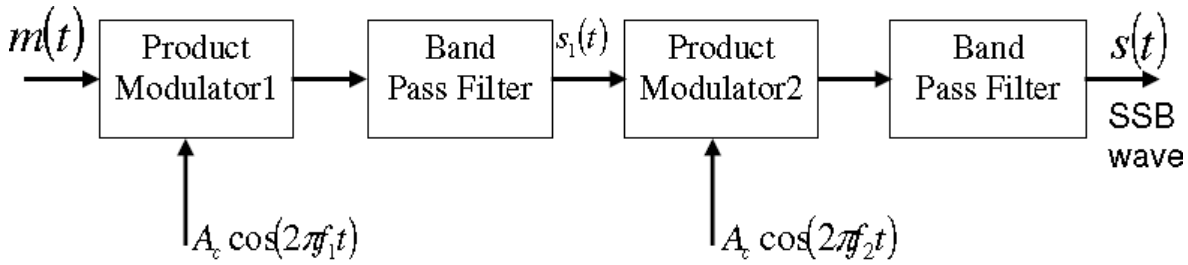


Figure .9 : Two stage frequency discriminator

The SSB modulated wave at the first filter output is used as the modulating wave for the second product modulator, which produces a DSBSC modulated wave with a spectrum that is symmetrically spaced about the second carrier frequency f2. The frequency separation between the side bands of this DSBSC modulated wave is effectively twice the first carrier frequency f1, thereby permitting the second filter to remove the unwanted side band.

Hilbert Transform & its Properties:

The Fourier transform is useful for evaluating the frequency content of an energy signal, or in a limiting case that of a power signal. It provides mathematical basis for analyzing and designing the frequency selective filters for the separation of signals on the basis of their frequency content.Another method of separating the signals is based on phase selectivity, which uses phase shifts between the appropriate signals (components) to achieve the desired separation.
In case of a sinusoidal signal, the simplest phase shift of 180^o is obtained by "Ideal transformer" (polarity reversal). When the phase angles of all the components of a given signal are shifted by 90^o, the resulting function of time is called the "Hilbert transform" of the signal.
Consider an LTI system with transfer function defined by equation 1

$$H(f)=\begin{cases}-j, f>0\\0, f=0\\j, f<0\end{cases} \qquad \text{---------- (1)}$$

and the Signum function given by $\operatorname{sgn}(f)=\begin{cases}1, f>0\\0, f=0\\-1, f<0\end{cases}$

The function $H(f)$ can be expressed using Signum function as given by 2

$$H(f)=-j\operatorname{sgn}(f) \qquad \text{---------- (2)}$$

We know that $1e^{-j\pi/2} = -j$, $1e^{j\pi/2} = j$ and $e^{\pm j\theta} = \cos(\theta) \pm j\sin(\theta)$

Therefore,

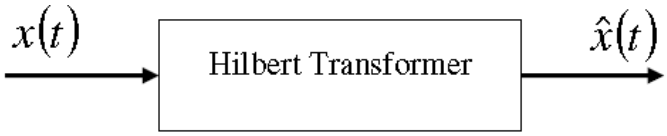
$$H(f) = \begin{cases} 1e^{-j\pi/2}, f > 0 \\ 1e^{j\pi/2}, f < 0 \end{cases}$$

Thus the magnitude $|H(f)| = 1$, for all f, and angle

$$\angle H(f) = \begin{cases} -\pi/2, f > 0 \\ +\pi/2, f < 0 \end{cases}$$

The device which possesses such a property is called Hilbert transformer. Whenever a signal is applied to the Hilbert transformer, the amplitudes of all frequency components of the input signal remain unaffected. It produces a phase shift of -90° for all positive frequencies, while a phase shifts of 90° for all negative frequencies of the signal.

If x(t) is an input signal, then its Hilbert transformer is denoted by x^(t) and shown in the following diagram.



To find impulse response *h(t)* of Hilbert transformer with transfer function *H(f)*.

Consider the relation between Signum function and the unit step function.

$$\text{sgn}(t) = 2u(t) - 1 = x(t),$$

Differentiating both sides with respect to t,

$$\frac{d}{dt}\{x(t)\} = 2\delta(t)$$

Apply Fourier transform on both sides,

$$\text{sgn}(t) \leftrightarrow \frac{2}{j\omega} \longrightarrow \text{sgn}(t) \leftrightarrow \frac{1}{j\pi f}$$

Applying duality property of Fourier transform,

$$-Sgn(f) \leftrightarrow \frac{1}{j\pi t}$$

We have $H(f) = -j\,\mathrm{sgn}(f)$

$$H(f) \leftrightarrow \frac{1}{\pi t}$$

Therefore the impulse response *h(t)* of an Hilbert transformer is given by the equation 3 ,

$$h(t) = \frac{1}{\pi t} \qquad \text{---------- (3)}$$

Now consider any input x(t) to the Hilbert transformer, which is an LTI system. Let the impulse response of the Hilbert transformer is obtained by convolving the input x(t) and impulse response h(t) of the system.

$$\hat{x}(t) = x(t) * h(t)$$

$$\hat{x}(t) = x(t) * \frac{1}{\pi t}$$

$$\hat{x}(t) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{x(\tau)}{(t-\tau)} d\tau \qquad \text{---------- (4)}$$

The equation 3.5 gives the Hilbert transform of *x(t)*.

The inverse Hilbert transform *x(t)* is given by

$$x(t) = \frac{-1}{\pi} \int_{-\infty}^{+\infty} \frac{\hat{x}(\tau)}{(t-\tau)} d\tau \qquad \text{---------- (5)}$$

We have $\hat{x}(t) = x(t) * h(t)$

The Fourier transform $\hat{X}(f)$ of $\hat{x}(t)$ is given by

$$\hat{X}(f) = X(f)H(f)$$

$$\hat{X}(f) = -j\,\mathrm{sgn}(f)X(f) \qquad \text{---------- (6)}$$

Properties:

1. "A signal *x(t)* and its Hilbert transform $\hat{x}(t)$ have the same amplitude spectrum".

The magnitude of *–jsgn(f)* is equal to 1 for all frequencies *f*. Therefore *x(t)* and $\hat{x}(t)$ have the same amplitude spectrum.

That is $\left|\hat{X}(f)\right| = \left|X(f)\right|$ for all f

2. "If $\hat{x}(t)$ is the Hilbert transform of *x(t)*, then the Hilbert transform of $\hat{x}(t)$, is *–x(t)*".

To obtain its Hilbert transform of *x(t)*, *x(t)* is passed through a LTI system with a transfer function equal to *–jsgn(f)*. A double Hilbert transformation is equivalent to passing *x(t)* through a cascade of two such devices. The overall transfer function of such a cascade is equal to

$$[-j\,\mathrm{sgn}(f)]^2 = -1 \quad \text{for all } f$$

The resulting output is *–x(t)*. That is the Hilbert transform of $\hat{x}(t)$ is equal to *–x(t)*.

Time Domain Description:

The time domain description of an SSB wave s(t) in the canonical form is given by the equation 1.

$$s(t) = s_I(t)\cos(2\pi f_c t) - s_Q(t)\sin(2\pi f_c t) \quad \text{---------- (1)}$$

where $S_I(t)$ is the in-phase component of the SSB wave and $S_Q(t)$ is its quadrature component. The in-phase component $S_I(t)$ except for a scaling factor, may be derived from *S(t)* by first multiplying *S(t)* by $\cos(2\pi f_c t)$ and then passing the product through a low-pass filter. Similarly, the quadrature component $S_Q(t)$, except for a scaling factor, may be derived from *s(t)* by first multiplying *s(t)* by $\sin(2\pi f_c t)$ and then passing the product through an identical filter.

The Fourier transformation of $S_I(t)$ and $S_Q(t)$ are related to that of SSB wave as follows, respectively.

$$S_I(f) = \begin{cases} S(f - f_c) + S(f + f_c), -w \le f \le w \\ 0, elsewhere \end{cases} \quad \text{---------- (2)}$$

$$S_Q(f)=\begin{cases} j[S(f-f_c)-S(f+f_c)], -w \le f \le w \\ 0, elsewhere \end{cases} \quad \text{----------} (3)$$

where $-w < f < w$ defines the frequency band occupied by the message signal $m(t)$.

Consider the SSB wave that is obtained by transmitting only the upper side band, shown in figure 10 . Two frequency shifted spectra $S(f-f_c)$ and $S(f+f_c)$ are shown in figure 11 and figure 12 respectively. Therefore, from equations 2 and 3 , it follows that the corresponding spectra of the in- phase component $S_I(t)$ and the quadrature component $S_Q(t)$ are as shown in figure 13 and 14 respectively.

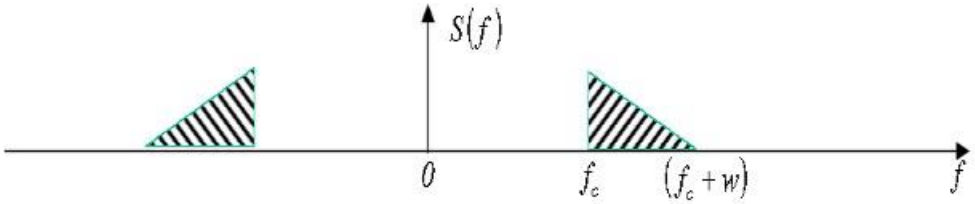


Figure 10 : Spectrum of SSBSC-USB

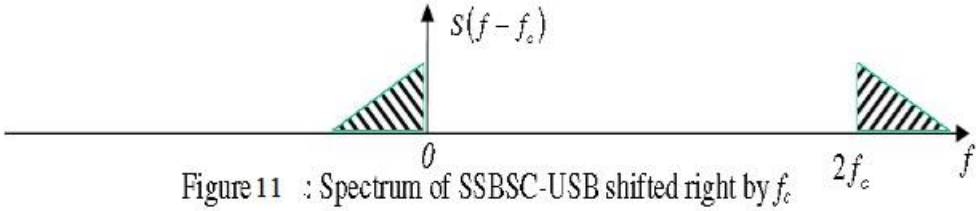


Figure 11 : Spectrum of SSBSC-USB shifted right by f_c

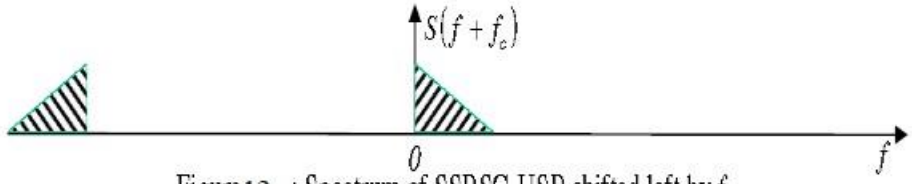


Figure 12 : Spectrum of SSBSC-USB shifted left by f_c

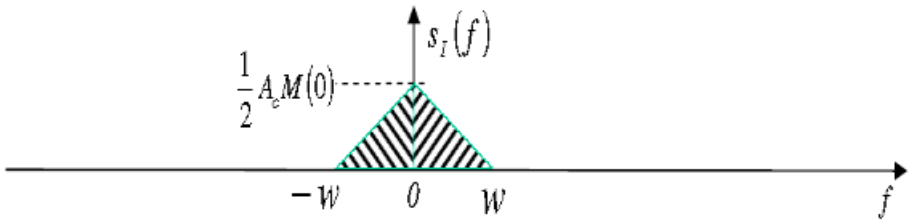


Figure 13 : Spectrum of in-phase component of SSBSC-USB

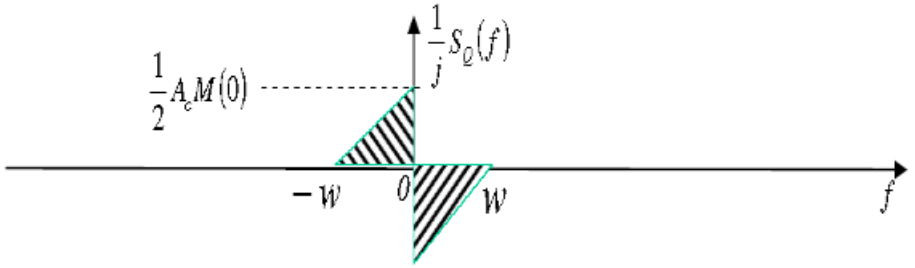


Figure 14 : Spectrum of quadrature component of SSBSC-USB

From the figure 13 , it is found that

$$S_I(f)=\frac{1}{2}A_cM(f)$$

where *M(f)* is the Fourier transform of the message signal *m(t)*. Accordingly in-phase component $S_I(t)$ is defined by equation 4

$$\boxed{s_I(t)=\frac{1}{2}A_cm(t)} \qquad \text{---------- (4)}$$

Now on the basis of figure14 , it is found that

$$S_Q(f)=\begin{cases}\frac{-j}{2}A_cM(f), f>0\\ 0, f=0\\ \frac{j}{2}A_cM(f), f<0\end{cases}$$

$$S_Q(f)=\frac{-j}{2}A_c\,\mathrm{sgn}(f)M(f) \qquad \text{---------- (5)}$$

where $\mathrm{sgn}(f)$ is the Signum function.

But from the discussions on Hilbert transforms, it is shown that

$$-j\,\mathrm{sgn}(f)M(f)=\hat{M}(f) \qquad \text{----------(6)}$$

where $\hat{M}(f)$ is the Fourier transform of the Hilbert transform of *m(t)*. Hence the substituting equation (6) in (5), we get

$$S_Q(f)=\frac{1}{2}A_c\hat{M}(f) \qquad \text{----------(7)}$$

Therefore quadrature component $s_Q(t)$ is defined by equation 8

$$s_Q(t)=\frac{1}{2}A_c\hat{m}(t) \qquad \text{----------(8)}$$

Therefore substituting equations (4) and (8) in equation in (1), we find that canonical representation of an SSB wave *s(t)* obtained by transmitting only the upper side band is given by the equation 9

$$s_U(t)=\frac{1}{2}A_c m(t)\cos(2\pi f_c t)-\frac{1}{2}A_c\hat{m}(t)\sin(2\pi f_c t) \qquad \text{----------(9)}$$

Following the same procedure, we can find the canonical representation for an SSB wave
s(t) obtained by transmitting only the lower side band is given by

$$s_L(t)=\frac{1}{2}A_c m(t)\cos(2\pi f_c t)+\frac{1}{2}A_c\hat{m}(t)\sin(2\pi f_c t) \qquad \text{----------(10)}$$

Phase discrimination method for generating SSB wave:

Time domain description of SSB modulation leads to another method of SSB generation using the equations 9 or 10. The block diagram of phase discriminator is as shown in figure 15.

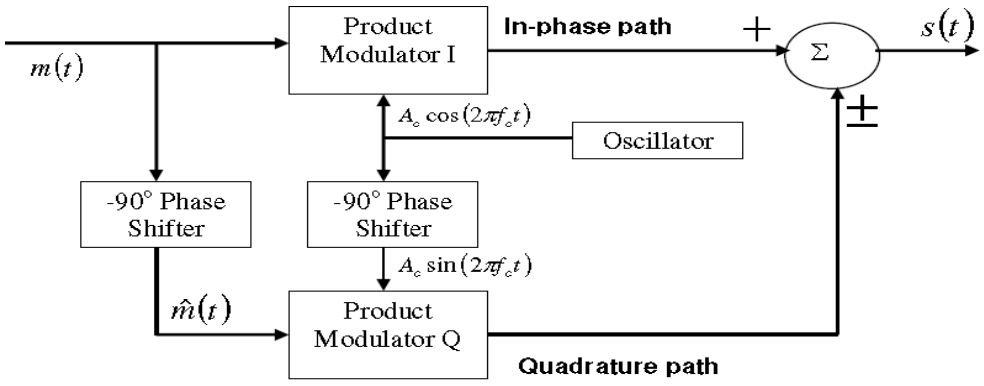


Figure 15 : Block diagram of phase discriminator

The phase discriminator consists of two product modulators I and Q, supplied with carrier waves in-phase quadrature to each other. The incoming base band signal m(t) is applied to product modulator I, producing a DSBSC modulated wave that contains reference phase sidebands symmetrically spaced about carrier frequency fc.

The Hilbert transform mˆ (t) of m (t) is applied to product modulator Q, producing a DSBSC modulated that contains side bands having identical amplitude spectra to those of modulator I, but with phase spectra such that vector addition or subtraction of the two modulator outputs results in cancellation of one set of side bands and reinforcement of the other set.

The use of a plus sign at the summing junction yields an SSB wave with only the lower side band, whereas the use of a minus sign yields an SSB wave with only the upper side band. This modulator circuit is called Hartley modulator.

Demodulation of SSB Waves:

Demodulation of SSBSC wave using coherent detection is as shown in 16 . The SSB wave *s(t)* together with a locally generated carrier $c(t) = A_c^{1} \cos(2\pi f_c t + \phi)$ is applied to a product modulator and then low-pass filtering of the modulator output yields the message signal.

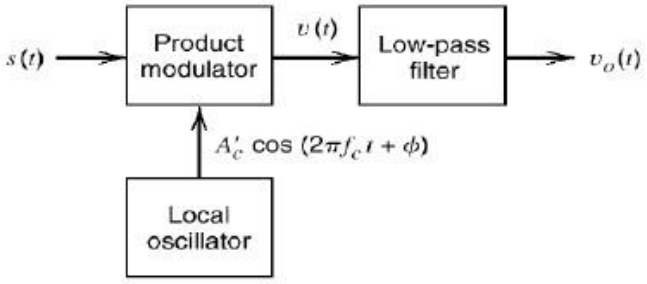


Figure 16 : Block diagram of coherent detector for SSBSC

The product modulator output *v(t)* is given by

$$v(t) = A_c^{1} \cos(2\pi f_c t + \phi) s(t) \quad , \quad \text{Put } \phi = 0$$

$$v(t) = \frac{1}{2} A_c \cos(2\pi f_c t)[m(t)\cos(2\pi f_c t) \pm \hat{m}(t)\sin(2\pi f_c t)]$$

$$v(t) = \frac{1}{4} A_c m(t) + \frac{1}{4} A_c [m(t)\cos(4\pi f_c t) \pm \hat{m}(t)\sin(4\pi f_c t)] \quad \text{.....(1)}$$

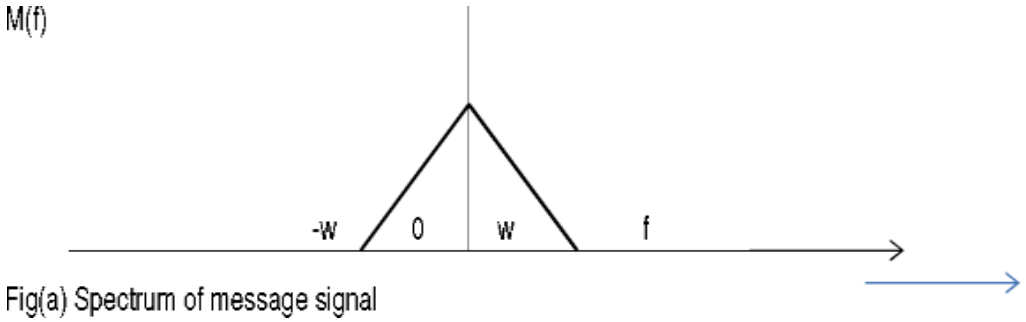
The first term in the above equation 1 is desired message signal. The other term represents an SSB wave with a carrier frequency of $2f_c$ as such; it is an unwanted component, which is removed by low-pass filter.

Introduction to Vestigial Side Band Modulation

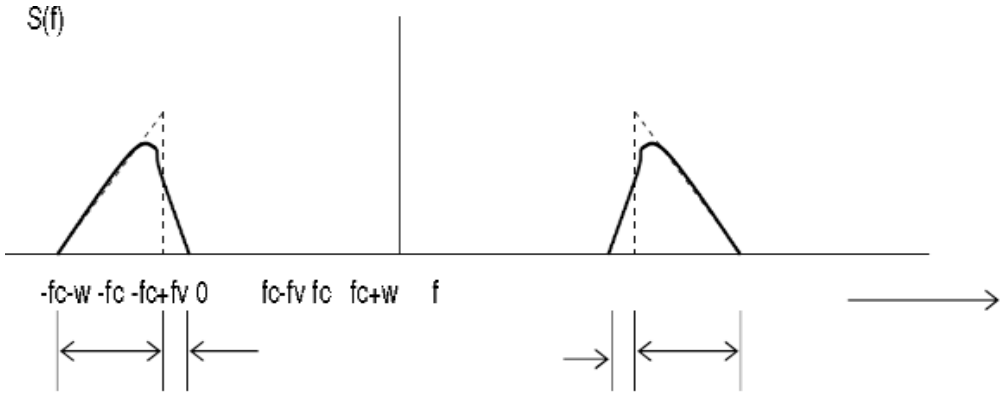
Vestigial sideband is a type of Amplitude modulation in which one side band is completely passed along with trace or tail or vestige of the other side band. VSB is a compromise between SSB and DSBSC modulation. In SSB, we send only one side band, the Bandwidth required to send SSB wave is w. SSB is not appropriate way of modulation when the message signal contains significant components at extremely low frequencies. To overcome this VSB is used.

Frequency Domain Description

The following Fig illustrates the spectrum of VSB modulated wave s (t) with respect to the message m (t) (band limited)



Fig(a) Spectrum of message signal



Fig(b) Spectrum of VSB wave containing vestige of the Lower side band

Assume that the Lower side band is modified into the vestigial side band. The vestige of the lower sideband compensates for the amount removed from the upper sideband. The bandwidth required to send VSB wave is

$B = w+f_v$

Where f_v is the width of the vestigial side band.

Similarly, if Upper side band is modified into the vestigial side band then,

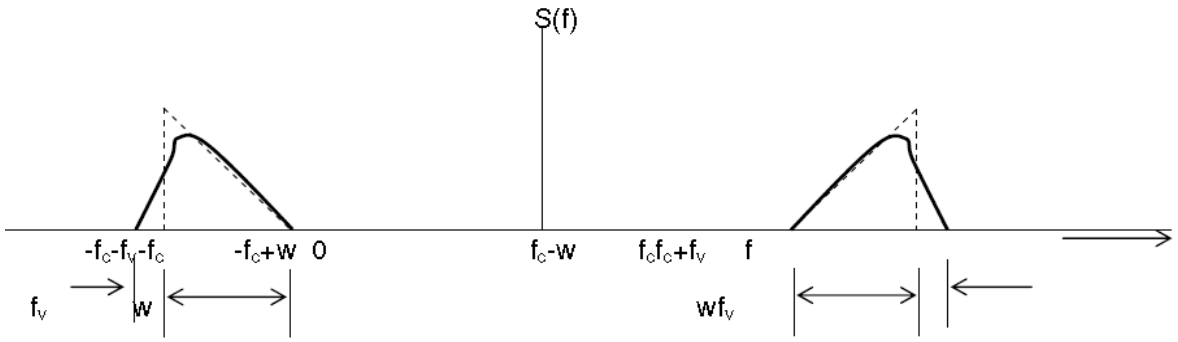


Fig (d) Spectrum of VSB wave containing vestige of the Upper side band

The vestige of the Upper sideband compensates for the amount removed from the Lower sideband. The bandwidth required to send VSB wave is B = w+fv, where fv is the width of the vestigial side band.

Therefore, VSB has the virtue of conserving bandwidth almost as efficiently as SSB modulation, while retaining the excellent low-frequency base band characteristics of DSBSC and it is standard for the transmission of TV signals.

Generation of VSB Modulated Wave

VSB modulated wave is obtained by passing DSBSC through a sideband shaping filter as shown in fig below.

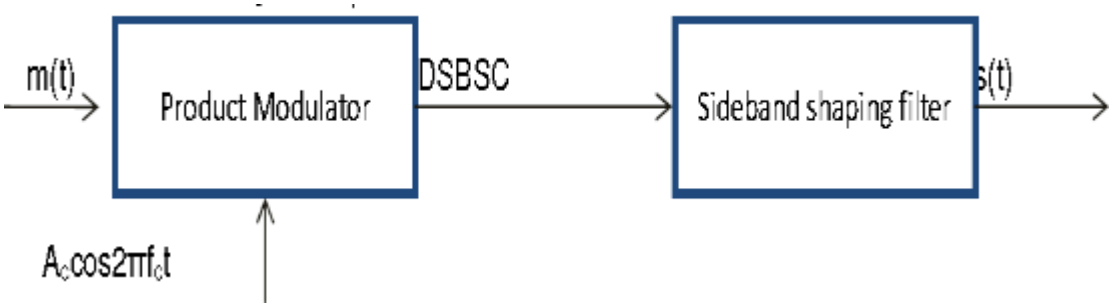


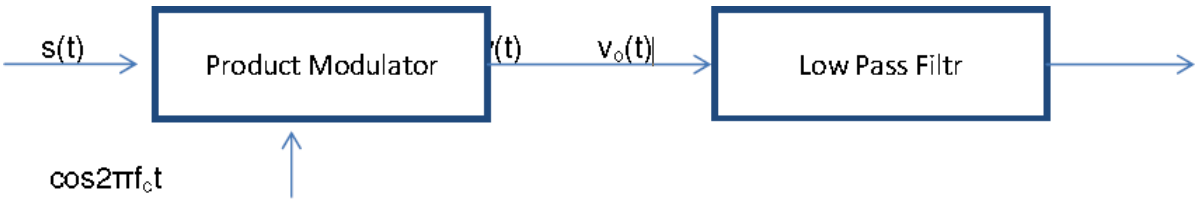
Fig.17. Block Diagram of VSB Modulator

The exact design of this filter depends on the spectrum of the VSB waves. The relation between filter transfer function H (f) and the spectrum of VSB waves is given by

$$S(f) = Ac /2 [M (f - fc) + M(f + fc)]H(f) \text{----------} (1)$$

Where M(f) is the spectrum of Message Signal. Now, we have to determine the specification for the filter transfer function H(f) It can be obtained by passing s(t) to a

coherent detector and determining the necessary condition for undistorted version of the message signal m(t). Thus, s (t) is multiplied by a locally generated sinusoidal wave cos ($2\pi f_c t$) which is synchronous with the carrier wave $A_c\cos(2\pi f_c t)$ in both frequency and phase, as in fig below,



Fig(b). Block diagram of VSB Demodulator

Then, $v(t) = s(t).\cos 2\pi f_c t$ ----------(2)

In frequency domain Eqn (2) becomes,

$V(f) = \frac{1}{2}[S(f - f_c) + S(f + f_c)]$ ----------(3)

Substitution of Eqn (1) in Eqn (3) gives

$V(f) = \frac{1}{2}[A_c/2 [M(f - f_c - f_c) + M(f - f_c + f_c)]H(f - f_c)$

$+ \frac{1}{2}[A_c/2 [M(f + f_c - f_c) + M(f + f_c + f_c)]H(f + f_c)$

$V(f) = \frac{1}{2}[A_c/2 [M(f - 2f_c) + M(f)]H(f - f_c)$

$+ \frac{1}{2}[A_c/2 [M(f) + M(f + 2f_c)]H(f + f_c)$

$V(f) = A_c/4\, M(f)[H(f - f_c) + H(f + f_c)]$

$+ A_c/4 [M(f - 2f_c) H(f - f_c) + M(f + 2f_c) H(f + f_c)]$ ----------(4)

The spectrum of V(f) as shown in fig below,

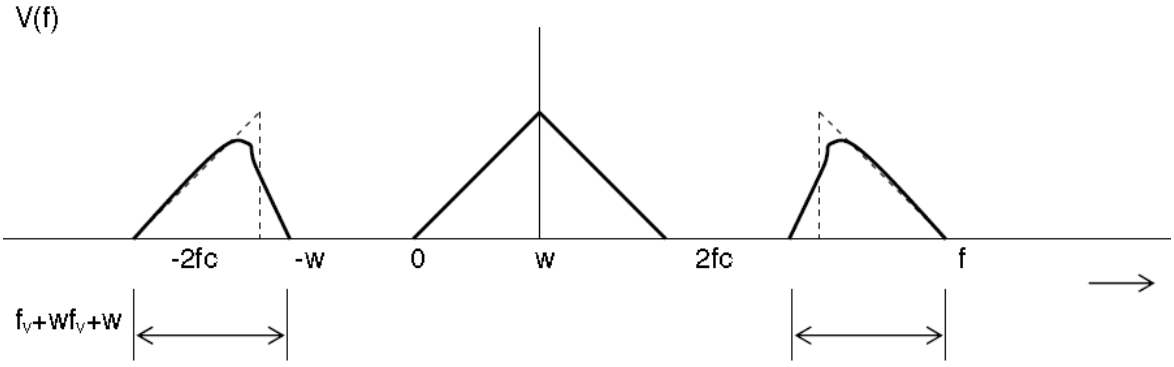


Fig ©. Spectrum of the product modulator output v(t)

Pass v(t) to a Low pass filter to eliminate VSB wave corresponding to $2f_c$.

$V_o(f) = A_c/4\, M(f)[H(f - f_c) + H(f + f_c)]$ ----------(5)

The spectrum of Vo (f) is in fig below,

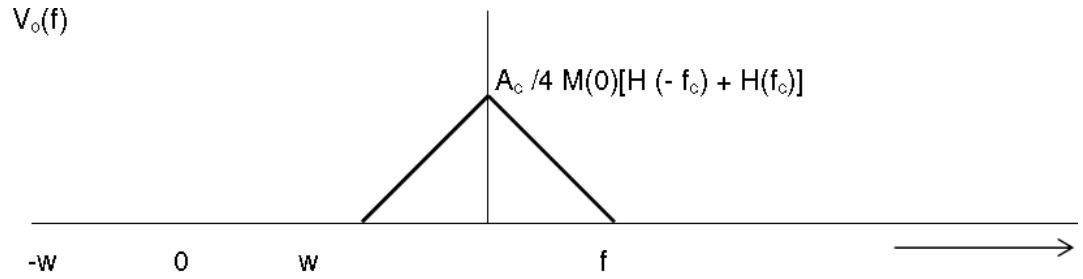


Fig (d). Spectrum of the demodulated Signal $v_o(t)$.

For a distortion less reproduction of the original signal m(t), $V_o(f)$ to be a scaled version of M(f). Therefore, the transfer function H(f) must satisfy the condition

$$H(f - f_c) + H(f + f_c) = 2H(f_c)\text{----------(6)}$$

Where $H(f_c)$ is a constant

Since m(t) is a band limited signal, we need to satisfy eqn (6) in the interval -w≤f≤w. The requirement of eqn (6) is satisfied by using a filter whose transfer function is shown below

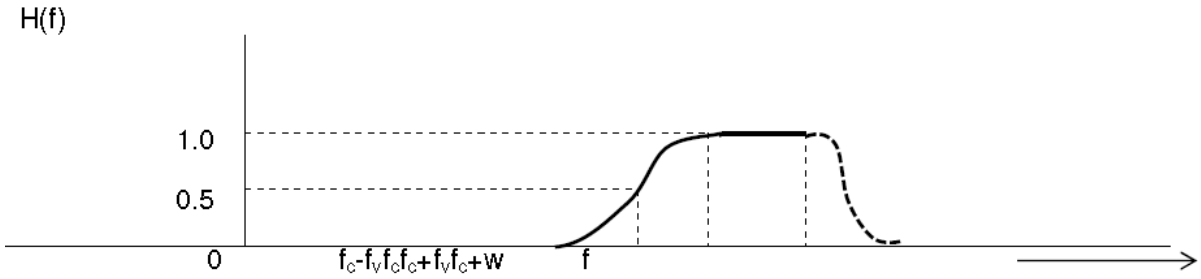


Fig (e) Frequency response of sideband shaping filter

Note: H(f) is Shown for positive frequencies only.

The Response is normalized so that H(f) at f_c is 0.5. Inside this interval f_c-f_v≤f≤f_c+$f_v$response exhibits odd symmetry. i.e., Sum of the values of H(f) at any two frequencies equally displaced above and below is Unity.

Similarly, the transfer function H (f) of the filter for sending Lower sideband along with the vestige of the Upper sideband is shown in fig below,

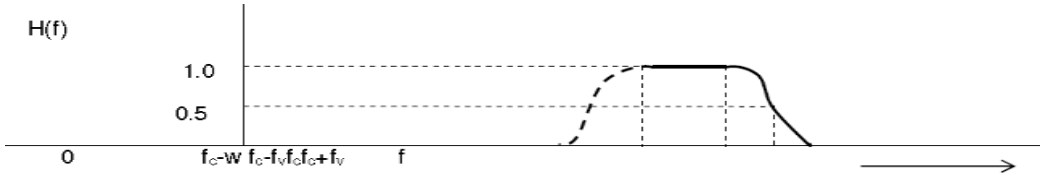


Fig (f) Frequency response of sideband shaping filter

Note: H(f) is Shown for positive frequencies only.

Time Domain Description:

Time domain representation of VSB modulated wave, procedure is similar to SSB Modulated waves. Let s(t) denote a VSB modulated wave and assuming that s(t) containing Upper sideband along with the Vestige of the Lower sideband. VSB modulated wave s(t) is the output from Sideband shaping filter, whose input is DSBSC wave. The filter transfer function H(f) is of the form as in fig below,

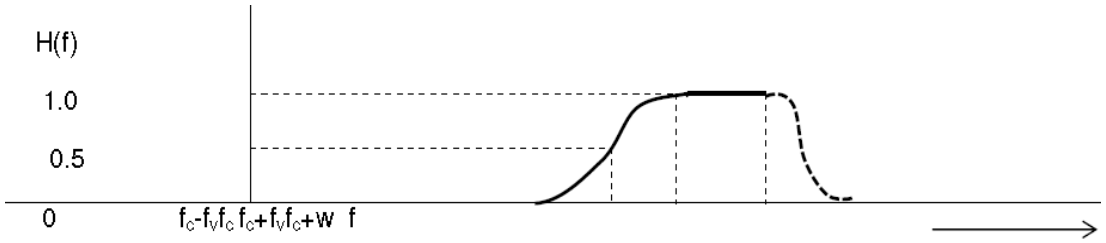


Fig (1) H(f) of sideband shaping filter

The DSBSC Modulated wave is

$S_{DSBSC}(t) = A_c\, m(t) \cos 2\pi f_c t$ ----------(1)

It is a band pass signal and has in-phase component only. Its low pass complex envelope is given by

$\tilde{s}_{DSBSC}(t) = A_c m(t)$ ----------(2)

The VSB modulated wave is a band pass signal.

Let the low pass signal $\tilde{s}(t)$ denote the complex envelope of VSB wave s(t), then

$s(t) = Re[\tilde{s}(t) \exp(j2\pi f_c t)]$ ----------(3)

To determine $\tilde{s}(t)$ we proceed as follows

1. The side band shaping filter transfer function H(f) is replaced by its equivalent complex low pass transfer function denoted by $\tilde{H}(f)$ as shown in fig below

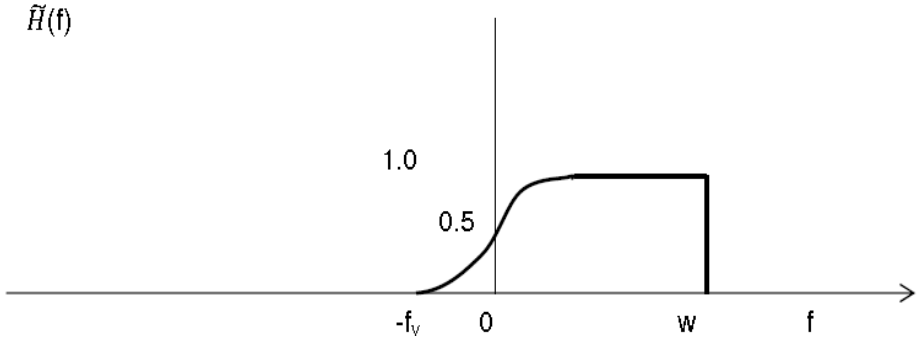


Fig (2) Low pass equivalent to H(f)

We may express $\tilde{H}(f)$ as the difference between two components $\widetilde{Hu}(f)$ and $\widetilde{Hv}(f)$ as

$\tilde{H}(f) = \widetilde{Hu}(f) - \widetilde{Hv}(f)$ ----------(4)

These two components are considered individually as follows

i). The transfer function $\widetilde{Hu}(f)$ pertains to a complex low pass filter equivalent to a band pass filter design to reject the lower side band completely as

$\widetilde{Hu}(f)$

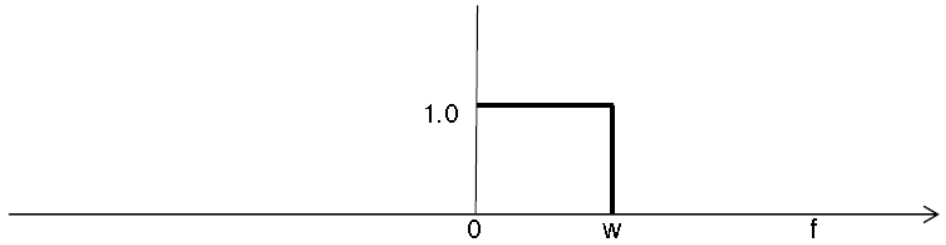


Fig (3) First component of $\tilde{H}(f)$

$$\widetilde{Hu}(f) = \begin{cases} \frac{1}{2}[1+\text{sgn}(f)], & 0<f<w \\ 0, & \text{otherwise} \end{cases} \text{----------(5)}$$

ii). The transfer function $\widetilde{Hv}(f)$ accounts for the generation of vestige and removal of a corresponding portion from the upper side band.

$\widetilde{Hv}(f)$

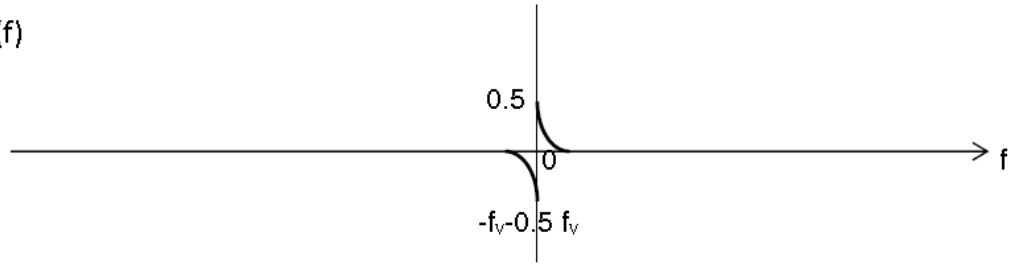


Fig (4) Second component of $\tilde{H}(f)$

Substitute eqn(5) in eqn (4) we get,

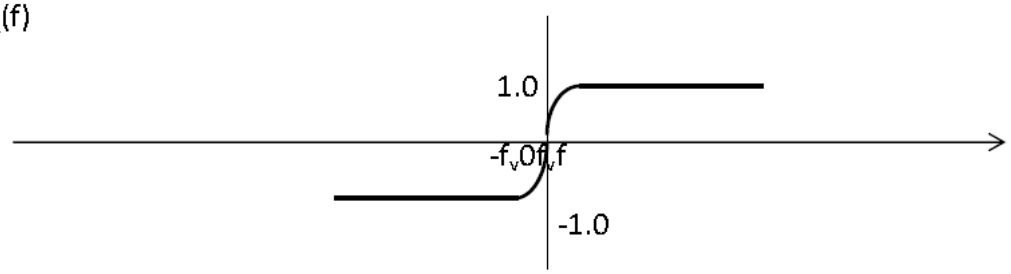
$$\tilde{H}(f) = \begin{cases} \frac{1}{2}[1+\text{sgn}(f) - 2\widetilde{Hv}(f)], & f_v<f<w \\ 0, & \text{otherwise} \end{cases} \text{----------(6)}$$

The sgn(f) and $\widetilde{Hv}(f)$ are both odd functions of frequency, Hence, both they have purely imaginary Inverse Fourier Transform (FT). Accordingly, we may introduce a new transfer function

$H_Q(f) = 1/j[sgn(f) - 2\widetilde{Hv}(f)]$ ----------(7)

It has purely Inverse FT and $h_Q(t)$ denote IFT of $H_Q(f)$

$jH_Q(f)$



Fig(5) Transfer function of the filter $jH_Q(f)$

Rewrite eqn(6) interms of $H_Q(f)$ as

$$\widetilde{H}(f) = \begin{cases} \frac{1}{2}[1+jH_Q(f)], & f_v<f<w \\ 0, & \text{otherwise} \end{cases} \text{----------(8)}$$

2. The DSBSC modulated wave is replaced by its complex envelope as
   $\tilde{S}_{DSBSC}(f) = A_c\, M(f)$ ----------(9)
3. The desired complex envelope $\tilde{s}(t)$ is determined by evaluating IFT of the product $\widetilde{H}(f)\tilde{S}_{DSBSC}(f)$.
   i.e., $\tilde{S}(f) = \widetilde{H}(f)\tilde{S}_{DSBSC}(f)$ ----------(10)
   $\tilde{S}(f) = A_c/2[1+jH_Q(f)]\, M[f]$ ----------(11)
   Take IFT of eqn(11) we get,
   $\tilde{s}(t) = A_c/2[m(t) + jm_Q(t)]$ ----------(12)
   Where $m_Q(t)$ is the response produced by passing the message through a low pass filter of impulse response $h_Q(t)$.
   Substitute eqn(12) in eqn(3) and simplify, we get
   $S(t) = A_c/2\, m(t) \cos 2\pi f_c t - A_c/2\, m_Q(t) \sin 2\pi f_c t$ ----------(13)
   Where $A_c/2\, m(t)$ ----- In-phase component
   $A_c/2\, m_Q(t)$----- Quadrature component

Note:

1. If vestigial side band is increased to full side band, VSB becomes DSCSB ,i.e., mQ(t) = 0.
2. If vestigial side band is reduced to Zero, VSB becomes SSB.
   i.e., $m_Q(t) = \hat{m}(t)$
   Where $\hat{m}(t)$ is the Hilbert transform of m(t).

Similarly If VSB containing a vestige of the Upper sideband, then s(t) is given by

$$S(t) = A_c/2\, m(t) \cos 2\pi f_c t + A_c/2\, m_Q(t) \sin 2\pi f_c t \quad \text{----------(14)}$$

Envelope detection of a VSB Wave plus Carrier

To make demodulation of VSB wave possible by an envelope detector at the receiving end it is necessary to transmit a sizeable carrier together with the modulated wave. The scaled expression of VSB wave by factor k_a with the carrier component $A_c \cos(2\pi f_c t)$ can be given by

$$s(t) = A_c \cos(2\pi f_c t) + \frac{A_c}{2} k_a\, m(t) \cos(2\pi f_c t) - \frac{A_c}{2} k_a\, m_Q(t) \sin(2\pi f_c t)$$

$$= A_c \left[1 + \frac{k_a}{2} m(t)\right] \cos(2\pi f_c t) - \frac{A_c k_a}{2} m_Q(t) \sin(2\pi f_c t) \quad \text{.....(1)}$$

where k_a is the modulation index; it determines the percentage modulation.

When above signal s(t) is passed through the envelope detector, the detector output is given by,

$$a(t) = A_c \left[\left(1 + \frac{k_a}{2} m(t)\right)^2 + \left(\frac{k_a}{2} m_Q(t)\right)^2\right]^{1/2}$$

$$= A_c \left[1 + \frac{k_a}{2} m(t)\right] \left(1 + \left[\frac{\frac{1}{2} k_a\, m_Q(t)}{1 + \frac{1}{2} k_a\, m(t)}\right]^2\right)^{1/2} \quad \text{.....(2)}$$

The detector output is distorted by the quadrature component $m_Q(t)$ as indicated by equation (2).

Methods to reduce distortion

- Distortion can be reduced by reducing percentage modulation, k_a.
- Distortion can be reduced by reducing $m_Q(t)$ by increasing the width of the vestigial sideband.

Comparison of AM Techniques:

Sr. No.	Parameter	Standard AM	SSB	DSBSC	VSB
1	Power	High	Less	Medium	Less than DSBSC but greater than SSB
2	Bandwidth	$2 f_m$	f_m	$2 f_m$	$f_m < B_w < 2 f_m$
3	Carrier supression	No	Yes	Yes	No
4	Receiver complexity	Simple	Complex	Complex	Simple
5	Application	Radio communication	Point to point communication preferred for long distance transmission.	Point to point communication	Television broadcasting
6	Modulation type	Non linear	Linear	Linear	Linear
7	Sideband suppression	No	One sided completely	No	One sideband suppressed partly
8	Transmission efficiency	Minimum	Maximum	Moderate	Moderate

Applications of different AM systems:

- Amplitude Modulation: AM radio, Short wave radio broadcast
- DSB-SC: Data Modems, Color TV's color signals.
- SSB: Telephone
- VSB: TV picture signals

UNIT IV
ANGLE MODULATION

- **Basic concepts**
- **Frequency Modulation**
- **Single tone frequency modulation**
- **Spectrum Analysis of Sinusoidal FM Wave**
- **Narrow band FM, Wide band FM, Constant Average Power**
- **Transmission bandwidth of FM Wave**
- **Generation of FM Waves:**
    - **Indirect FM, Direct FM: Varactor Diode and Reactance Modulator**
- **Detection of FM Waves:**
    - **Balanced Frequency discriminator, Zero crossing detector, Phase locked loop**
- **Comparison of FM & AM**
- **Pre-emphasis & de-emphasis**
- **FM Transmitter block diagram and explanation of each block**

Instantaneous Frequency

The frequency of a cosine function $x(t)$ that is given by

$$x(t) = \cos(\omega_c t + \theta_0)$$

is equal to ω_c since it is a constant with respect to t, and the phase of the cosine is the constant θ_0. The angle of the cosine $\theta(t) = \omega_c t + \theta_0$ is a linear relationship with respect to t (a straight line with slope of ω_c and y–intercept of θ_0). However, for other sinusoidal functions, the frequency may itself be a function of time, and therefore, we should not think in terms of the constant frequency of the sinusoid but in terms of the INSTANTANEOUS frequency of the sinusoid since it is not constant for all t. Consider for example the following sinusoid

$$y(t) = \cos[\theta(t)],$$

where $\theta(t)$ is a function of time. The frequency of $y(t)$ in this case depends on the function of $\theta(t)$ and may itself be a function of time. The instantaneous frequency of $y(t)$ given above is defined as

$$\omega_i(t) = \frac{d\theta(t)}{dt}.$$

As a checkup for this definition, we know that the instantaneous frequency of $x(t)$ is equal to its frequency at all times (since the instantaneous frequency for that function is constant) and is equal to ω_c. Clearly this satisfies the definition of the instantaneous frequency since $\theta(t) = \omega_c t + \theta_0$ and therefore $\omega_i(t) = \omega_c$.

If we know the instantaneous frequency of some sinusoid from $-\infty$ to sometime t, we can find the angle of that sinusoid at time t using

$$\theta(t) = \int_{-\infty}^{t} \omega_i(\alpha) d\alpha.$$

Changing the angle $\theta(t)$ of some sinusoid is the bases for the two types of angle modulation: Phase and Frequency modulation techniques.

Phase Modulation (PM)

In this type of modulation, the phase of the carrier signal is directly changed by the message signal. The phase modulated signal will have the form

$$g_{PM}(t) = A \cdot \cos\left[\omega_c t + k_p m(t)\right],$$

where A is a constant, ω_c is the carrier frequency, $m(t)$ is the message signal, and k_p is a parameter that specifies how much change in the angle occurs for every unit of change of $m(t)$. The phase and instantaneous frequency of this signal are

$$\theta_{PM}(t) = \omega_c t + k_p m(t),$$
$$\omega_i(t) = \omega_c + k_p \frac{dm(t)}{dt} = \omega_c + k_p \dot{m}(t).$$

So, the frequency of a PM signal is proportional to the derivative of the message signal.

Frequency Modulation (FM)

This type of modulation changes the frequency of the carrier (not the phase as in PM) directly with the message signal. The FM modulated signal is

$$g_{FM}(t) = A \cdot \cos\left[\omega_c t + k_f \int_{-\infty}^{t} m(\alpha) d\alpha\right],$$

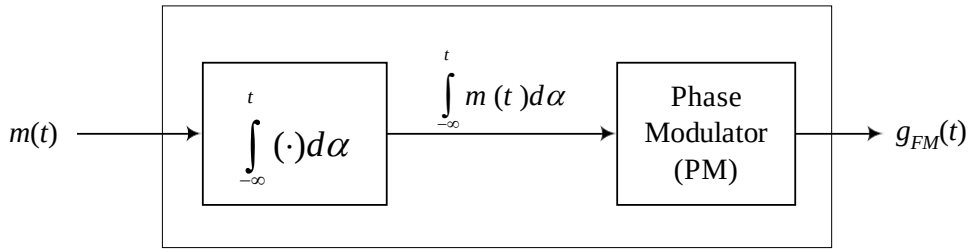
where k_f is a parameter that specifies how much change in the frequency occurs for every unit change of $m(t)$. The phase and instantaneous frequency of this FM are

$$\theta_{FM}(t) = \omega_c t + k_f \int_{-\infty}^{t} m(\alpha) d\alpha,$$
$$\omega_i(t) = \omega_c + k_f \frac{d}{dt}\left[\int_{-\infty}^{t} m(\alpha) d\alpha\right] = \omega_c + k_f m(t).$$

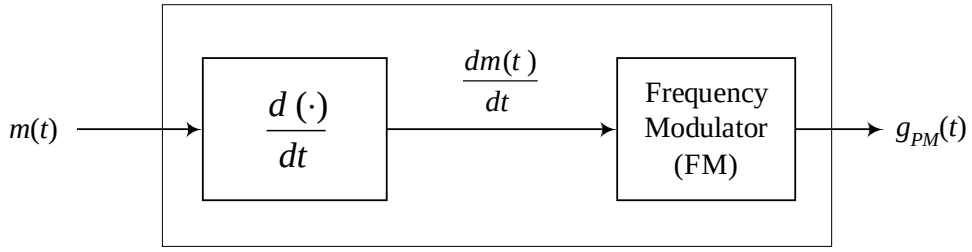
Relation between PM and FM

PM and FM are tightly related to each other. We see from the phase and frequency relations for PM and FM given above that replacing $m(t)$ in the PM signal with $\int_{-\infty}^{t} m(\alpha) d\alpha$ gives an FM signal and replacing $m(t)$ in the FM signal with $\frac{dm(t)}{dt}$ gives a PM signal. This is illustrated in the following block diagrams.

Frequency Modulator (FM)

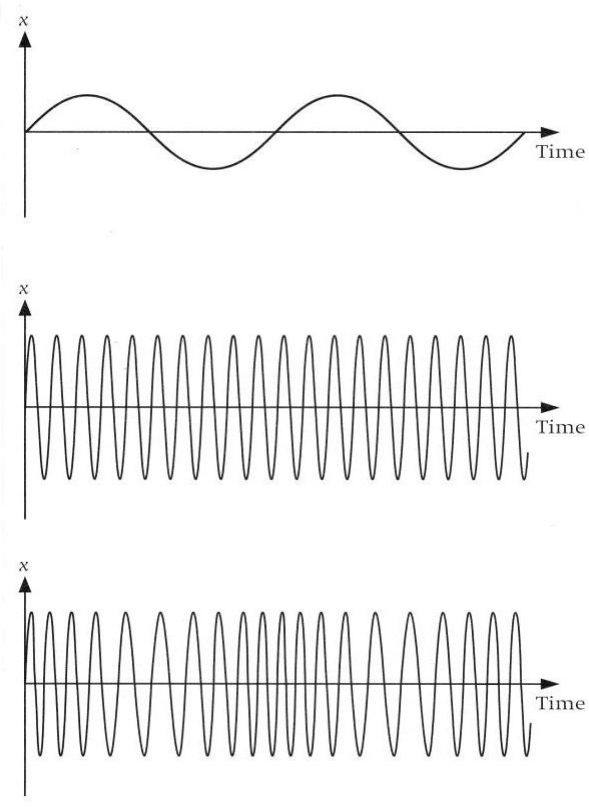


Phase Modulator (PM)



Frequency Modulation

In **F**requency **M**odulation (**FM**) the instantaneous value of the information signal controls the frequency of the carrier wave. This is illustrated in the following diagrams.



Notice that as the information signal increases, the frequency of the carrier increases, and as the information signal decreases, the frequency of the carrier decreases.

The frequency f_i of the information signal controls the rate at which the carrier frequency increases and decreases. As with AM, f_i must be less than f_c. The amplitude of the carrier remains constant throughout this process.

When the information voltage reaches its maximum value then the change in frequency of the carrier will have also reached its maximum deviation above the nominal value. Similarly when the information reaches a minimum the carrier will be at its lowest frequency below the nominal carrier frequency value. When the information signal is zero, then no deviation of the carrier will occur.

The maximum change that can occur to the carrier from its base value f_c is called the frequency deviation, and is given the symbol Δf_c. This sets the dynamic range (i.e. voltage range) of the transmission. The dynamic range is the ratio of the largest and smallest analogue information signals that can be transmitted.

Bandwidth of FM and PM Signals

The bandwidth of the different AM modulation techniques ranges from the bandwidth of the message signal (for SSB) to twice the bandwidth of the message signal (for DSBSC and Full AM). When FM signals were first proposed, it was thought that their bandwidth can be reduced to an arbitrarily small value. Compared to the bandwidth of different AM modulation techniques, this would in theory be a big advantage. It was assumed that a signal with an instantaneous frequency that changes over of range of Δf Hz would have a bandwidth of Δf Hz. When experiments were done, it was discovered that this was not the case. It was discovered that the bandwidth of FM signals for a specific message signal was at least equal to the bandwidth of the corresponding AM signal. In fact, FM signals can be classified into two types: Narrowband and Wideband FM signals depending on the bandwidth of each of these signals

Narrowband FM and PM

The general form of an FM signal that results when modulating a signals $m(t)$ is

$$g_{FM}(t) = A \cdot \cos\left[\omega_c t + k_f \int_{-\infty}^{t} m(\alpha) d\alpha\right].$$

A narrow band FM or PM signal satisfies the condition

$$\left|k_f a(t)\right| \ll 1$$

For FM and

$$\left| k_p \cdot m(t) \right| \ll 1$$

For PM, where

$$a(t) = \int_{-\infty}^{t} m(\alpha) d\alpha ,$$

such that a change in the message signal does not results in a lot of change in the instantaneous frequency of the FM signal.

Now, we can write the above as

$$g_{FM}(t) = A \cdot \cos\left[\omega_c t + k_f a(t)\right].$$

Starting with FM, to evaluate the bandwidth of this signal, we need to expand it using a power series expansion. So, we will define a slightly different signal

$$\hat{g}_{FM}(t) = A \cdot e^{j\{\omega_c t + k_f a(t)\}} = A \cdot e^{j\omega_c t} \cdot e^{jk_f a(t)}.$$

Remember that

$$\hat{g}_{FM}(t) = A \cdot e^{j\{\omega_c t + k_f a(t)\}} = A \cdot \cos\left[\omega_c t + k_f a(t)\right] + jA \cdot \sin\left[\omega_c t + k_f a(t)\right] ,$$

so

$$g_{FM}(t) = \mathrm{Re}\left\{\hat{g}_{FM}(t)\right\}.$$

Now we can expand the term $e^{jk_f a(t)}$ in $\hat{g}_{FM}(t)$, which gives

$$\hat{g}_{FM}(t) = A \cdot e^{j\omega_c t} \cdot \left[1 + jk_f a(t) + \frac{j^2 k_f^2 a^2(t)}{2!} + \frac{j^3 k_f^3 a^3(t)}{3!} + \frac{j^4 k_f^4 a^4(t)}{4!} \cdots \right]$$

$$= A \cdot \left[e^{j\omega_c t} + jk_f a(t) e^{j\omega_c t} - \frac{k_f^2 a^2(t)}{2!} e^{j\omega_c t} - \frac{jk_f^3 a^3(t)}{3!} e^{j\omega_c t} + \frac{k_f^4 a^4(t)}{4!} e^{j\omega_c t} + \cdots \right]$$

Since k_f and $a(t)$ are real ($a(t)$ is real because it is the integral of a real function $m(t)$), and since $\mathrm{Re}\{e^{j\omega c t}\} = \cos(\omega_c t)$ and $\mathrm{Re}\{ je^{j\omega c t}\} = -\sin(\omega_c t)$, then

$$g_{FM}(t) = \mathrm{Re}\left\{\hat{g}_{FM}(t)\right\}$$
$$= A \cdot \left[\cos(\omega_c t) - k_f a(t) \sin(\omega_c t) - \frac{k_f^2 a^2(t)}{2!}\cos(\omega_c t) + \frac{k_f^3 a^3(t)}{3!}\sin(\omega_c t) + \frac{k_f^4 a^4(t)}{4!}\cos(\omega_c t) + \cdots \right]$$

The assumption we made for narrowband FM is ($|k_f a(t)| \ll 1$). This assumption will result in making all the terms with powers of $k_f a(t)$ greater than 1 to be small compared to the first two terms. So, the following is a reasonable approximation for $g_{FM}(t)$

$$\boxed{g_{FM\,(Narrowband)}(t) \approx A \cdot \left[\cos(\omega_c t) - k_f a(t) \sin(\omega_c t)\right]} \qquad \text{when } |k_f a(t)| \ll 1.$$

It must be stressed that the above approximation is only valid for narrowband FM signals that satisfy the condition ($|k_f a(t)| \ll 1$). The above signal is simply the addition (or actually the subtraction) of a cosine (the carrier) with a DSBSC signal (but using a sine as the carrier). The message signal that modulates the DSBSC signal is not $m(t)$ but its integration $a(t)$. One of the properties of the Fourier transform informs us that the bandwidth of a signal $m(t)$ and its integration $a(t)$ (and its derivative too) are the same (verify this). Therefore, the bandwidth of the narrowband FM signal is

$$\boxed{BW_{FM\,(Narrowband)} = BW_{DSBSC} = 2 \cdot BW_{m(t)} =}$$

We will see later that when the condition ($k_f << 1$) is not satisfied, the bandwidth of the FM signal becomes higher that twice the bandwidth of the message signal. Similar relationships hold for PM signals. That is

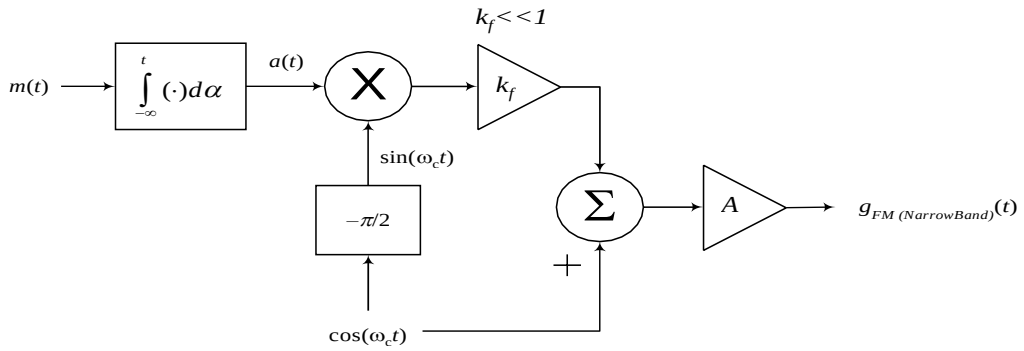
$$\boxed{g_{PM\,(Narrowband)}(t) \approx A \cdot \left[\cos(\omega_c t) - k_p m(t) \sin(\omega_c t)\right]} \qquad \text{when } |k_p \cdot m(t)| \ll 1,$$

and

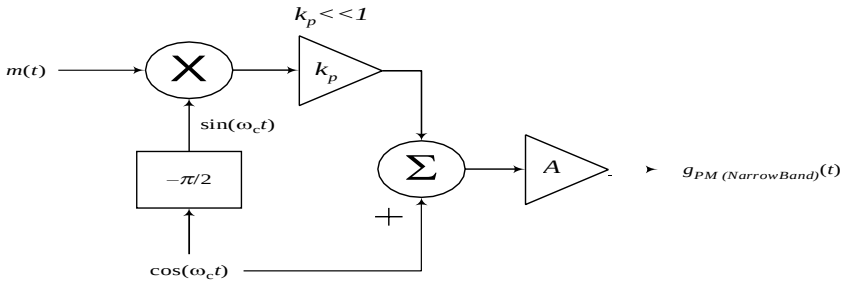
$$\boxed{BW_{PM\,(Narrowband)} = BW_{DSBSC} = 2 \cdot BW_{m(t)} =}$$

Construction of Narrowband Frequency and Phase Modulators

The above approximations for narrowband FM and PM can be easily used to construct modulators for both types of signals



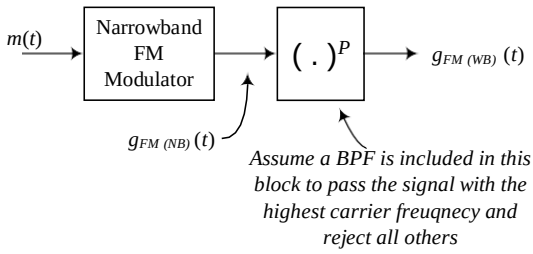
Narrowband FM Modulator



Narrowband PM Modulator

Generation of Wideband FM Signals

Consider the following block diagram



A narrowband FM signal can be generated easily using the block diagram of the narrowband FM modulator that was described in a previous lecture. The narrowband FM modulator generates a narrowband FM signal using simple components such as an integrator (an OpAmp), oscillators, multipliers, and adders. The generated narrowband FM signal can be converted to a wideband FM signal by simply passing it through a non–linear device with power P. Both the carrier frequency and the frequency deviation Δf of the narrowband signal

are increased by a factor *P*. Sometimes, the desired increase in the carrier frequency and the desired increase in Δf are different. In this case, we increase Δf to the desired value and use a frequency shifter (multiplication by a sinusoid followed by a BPF) to change the carrier frequency to the desired value.

SINGLE-TONE FREQUENCY MODULATION

Time-Domain Expression

Since the FM wave is a nonlinear function of the modulating wave, the frequency modulation is a nonlinear process. The analysis of nonlinear process is the difficult task. In this section, we will study single-tone frequency modulation in detail to simplify the analysis and to get thorough understanding about FM.

Let us consider a single-tone sinusoidal message signal defined by

$$n(t) = A_n \cos(2nf_n t) \qquad (5.13)$$

The instantaneous frequency from Eq. (5.8) is then

$$f(t) = f_c + k_f A_n \cos(2nf_n t) = f_c + \Delta f \cos(2nf_n t) \qquad (5.14)$$

where

$$\Delta f = k_f A_n$$

$$\theta(t) = 2\pi f_c t + 2\pi k_f \int_0^t A_m \cos(2\pi f_m t)dt$$

$$= 2\pi f_c t + 2\pi k_f \frac{A_m}{2\pi f_m} \sin(2\pi f_m t)$$

$$= 2\pi f_c t + \frac{k_f A_m}{f_m} \sin(2\pi f_m t)$$

$$= 2\pi f_c t + \frac{\Delta f}{f_m} \sin(2\pi f_m t)$$

$$\therefore \theta(t) = 2\pi f_c t + \beta_f \sin(2\pi f_m t)$$

Where

$$\beta_f = \frac{\Delta f}{f_m} = \frac{k_f A_m}{f_m}$$

is the modulation index of the FM wave. Therefore, the single-tone FM wave is expressed by

$$s_{FM}(t) = A_c \cos[2\pi f_c t + \beta_f \sin(2\pi f_n t)] \quad (5.18)$$

This is the desired time-domain expression of the single-tone FM wave

Similarly, **single-tone phase modulated wave** may be determined from Eq.as

$$s_{PM}(t) = A_c \cos[2\pi f_c t + k_p A_n \cos(2\pi f_n t)]$$

$$\text{or, } s_{PM}(t) = A_c \cos[2\pi f_c t + \beta_p \cos(2\pi f_n t)] \quad (5.19)$$

where

$$\beta_p = k_p A_n \quad (5.20)$$

is the modulation index of the single-tone phase modulated wave.
The frequency deviation of the single-tone PM wave is

$$s_{FM}(t) = A_c \cos\left[2\pi f_c t + \beta_f \sin(2\pi f_m t)\right]$$

Spectral Analysis of Single-Tone FM Wave

The above Eq. can be rewritten as

$$s_{FM}(t) = \mathrm{Re}\{A_c e^{j2nf_c t} e^{j\beta \sin(2nf_n t)}\}$$

For simplicity, the modulation index of FM has been considered as β instead of β_f afterward. Since $\sin(2nf_n t)$ is periodic with fundamental period $T = 1/f_n$, the complex expontial $e^{j\beta \sin(2nf_n t)}$ is also periodic with the same fundamental period. Therefore, this complex exponential can be expanded in Fourier series representation as

$$e^{j\beta \sin(2\pi f_m t)} = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n f_m t}$$

where the Fourier series coefficients c_n are obtained as

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} e^{j\beta \sin(2\pi f_m t)} e^{-j2\pi n f_m t} dt \tag{5.24}$$

Let $2\pi f_m t = x$, then Eq. (5.24) reduces to

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j\beta \sin(x)} e^{-jnx} dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j(\beta \sin(x) - nx)} dx \tag{5.25}$$

The integral on the right-hand side is known as the nth order Bessel function of the first kind and is denoted by $J_n(\beta)$. Therefore, $c_n = J_n(\beta)$ and Eq. (4.23) can be written as

$$e^{j\beta \sin(2\pi f_m t)} = \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j2\pi n f_m t} \tag{5.26}$$

By substituting Eq. (5.26) in Eq. (5.22), we get

$$\begin{aligned} s_{FM}(t) &= \mathrm{Re}\left\{ A_c \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j2\pi n f_m t}\, e^{j2\pi f_c t} \right\} \\ &= A_c \sum_{n=-\infty}^{\infty} J_n(\beta) \cos[2\pi(f_c + n f_m)t] \end{aligned} \tag{5.27}$$

Taking Fourier transform of Eq. (5.27), we get

$$S(f) = \frac{1}{2} A_c \sum_{n=-\infty}^{\infty} J_n(\beta)[\delta(f - f_c - n f_m) + \delta(f + f_c + n f_m)] \tag{5.28}$$

From the spectral analysis we see that there is a carrier component and a number of side-frequencies around the carrier frequency at $\pm n f_m$.

The Bessel function may be expanded in a power series given by

$$J_n(\beta) = \sum_{k=0}^{\infty} \frac{(-1)^k \left(\frac{1}{2}\beta\right)^{n+2k}}{k!\,(k+n)!} \tag{5.29}$$

Plots of Bessel function $J_n(\beta)$ versus modulation index β for $n = 0, 1, 2, 3, 4$ are shown in Figure 5.3.

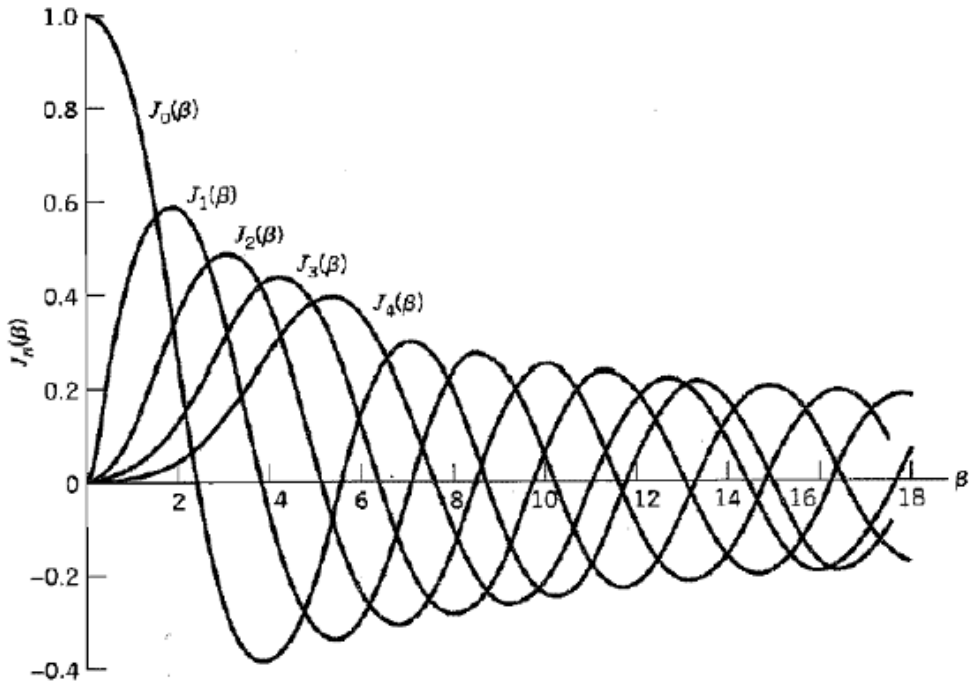


Figure 5.3 Plot of Bessel function as a function of modulation index.

Figure 5.3 shows that for any fixed value of n, the magnitude of $J_n(\beta)$ decreases as β increases. One property of Bessel function is that

$$J_{-n}(\beta) = \begin{cases} J_n(\beta), & n \text{ even} \\ -J_n(\beta), & n \text{ odd} \end{cases} \tag{5.30}$$

One more property of Bessel function is that

$$\sum_{n=-\infty}^{\infty} J_n^2(\beta) = 1 \tag{5.31}$$

(iii) The average power of the FM wave remains constant. To prove this, let us determine the average power of Eq. (5.27) which is equal to

$$P = \frac{1}{2}A_c^2 \sum_{n=-\infty}^{\infty} J_n^2(\beta)$$

Using Eq. (5.31), the average power P is now

$$P = \frac{1}{2}A_c^2$$

TRANSMISSION BANDWIDTH OF FM WAVE

The transmission bandwidth of an FM wave depends on the modulation index þ. The modulation index, on the other hand, depends on the modulating amplitude and modulating frequency. It is almost impossible to determine the exact bandwidth of the FM wave. Rather, we use a rule-of-thumb expression for determining the FM bandwidth.

For single-tone frequency modulation, the approximated bandwidth is determined by the expression

$$B = 2(\Delta f + f_m) = 2(\beta + 1)f_m = 2\Delta f\left(1 + \frac{1}{\beta}\right)$$

This expression is regarded as the Carson's rule. The FM bandwidth determined by this rule accommodates at least 98 % of the total power.

For an arbitrary message signal n(t) with bandwidth or maximum frequency W, the bandwidth of the corresponding FM wave may be determined by Carson's rule as

$$B = 2(\Delta f + W) = 2(D + 1)W = 2\Delta f\left(1 + \frac{1}{D}\right)$$

GENERATION OF FM WAVES

FM waves are normally generated by two methods: indirect method and direct method.

Indirect Method (Armstrong Method) of FM Generation

In this method, narrow-band FM wave is generated first by using phase modulator and then the wideband FM with desired frequency deviation is obtained by using frequency multipliers.

$$s(t) = A_c \cos\left[2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right]$$

$$\text{or, } s(t) = A_c \cos[2\pi f_c t + \emptyset(t)]$$

$$\emptyset(t) = 2\pi k_f \int_0^t m(t)dt$$

$$s(t) = A_c \cos(2\pi f_c t)\cos[\emptyset(t)] - A_c \sin(2\pi f_c t)\sin[\emptyset(t)]$$

The above eq is the expression for narrow band FM wave

In this case $\cos[\emptyset(t)] \approx 1$ and $\sin[\emptyset(t)] \approx \emptyset(t)$

$$s(t) = A_c \cos(2\pi f_c t) - A_c \sin(2\pi f_c t)\emptyset(t)$$

$$\text{or, } s(t) = A_c \cos(2\pi f_c t) - 2\pi A_c k_f \sin(2\pi f_c t) \int_0^t m(t)dt$$

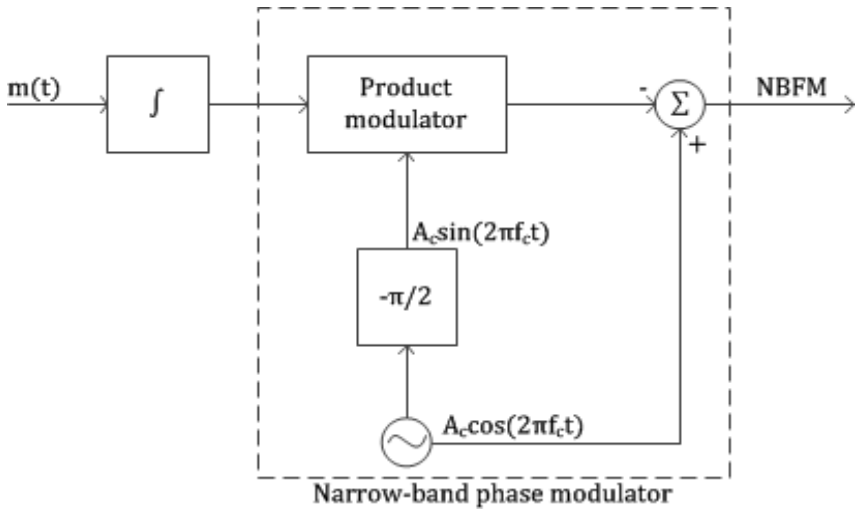


Fig: Narrowband FM Generator

The frequency deviation Δf is very small in narrow-band FM wave. To produce wideband FM, we have to increase the value of Δf to a desired level. This is achieved by means of one or multiple frequency multipliers. A frequency multiplier consists of a nonlinear device and a bandpass filter. The n^{th} order nonlinear device produces a dc component and n number of frequency modulated waves with carrier frequencies f_c, $2f_c$, … nf_c and frequency deviations Δf, $2\Delta f$, … $n\Delta f$, respectively. If we want an FM wave with frequency deviation of $6\Delta f$, then we may use a 6^{th} order nonlinear device or one 2^{nd} order and one 3^{rd} order nonlinear devices in cascade followed by a bandpass filter centered at $6f_c$. Normally, we may require very high value of frequency deviation. This automatically increases the carrier frequency by the same factor which may be higher than the required carrier frequency. We may shift the carrier frequency to the desired level by using mixer which does not change the frequency deviation.

The narrowband FM has some distortion due to the approximation made in deriving the expression of narrowband FM from the general expression. This produces some amplitude modulation in the narrowband FM which is removed by using a limiter in frequency multiplier.

Direct Method of FM Generation

In this method, the instantaneous frequency f(t) of the carrier signal c(t) is varied directly with the instantaneous value of the modulating signal n(t). For this, an oscillator is used in which any one of the reactive components (either C or L) of the resonant network of the oscillator is varied linearly with n(t). We can use a varactor diode or a varicap as a voltage-variable capacitor whose capacitance solely depends on the reverse-bias voltage applied across it. To vary such capacitance linearly with n(t), we have to reverse-bias the diode by the fixed DC voltage and operate within a small linear portion of the capacitance-voltage characteristic curve. The unmodulated fixed capacitance C_0 is linearly varied by n(t) such that the resultant capacitance becomes

$$C(t) = C_0 - kn(t)$$

where the constant k is the sensitivity of the varactor diode (measured in capacitance per volt).

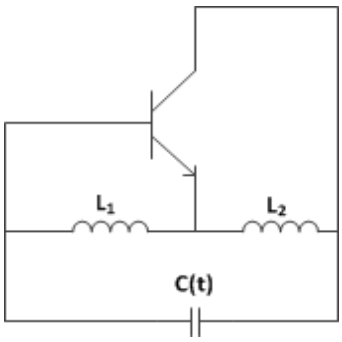


fig: Hartley oscillator for FM generation

The above figure shows the simplified diagram of the Hartley oscillator in which is implemented the above discussed scheme. The frequency of oscillation for such an oscillator is given

$$f(t) = \frac{1}{2\pi\sqrt{(L_1 + L_2)C(t)}}$$

$$f(t) = \frac{1}{2\pi\sqrt{(L_1 + L_2)(C_0 - km(t))}}$$

$$= \frac{1}{2\pi\sqrt{(L_1 + L_2)C_0}} \frac{1}{\sqrt{1 - \frac{km(t)}{C_0}}}$$

$$\text{or,} \qquad f(t) = f_c \left(1 - \frac{km(t)}{C_0}\right)^{-1/2}$$

where f_c is the unmodulated frequency of oscillation. Assuming,

$$\frac{km(t)}{C_0} \ll 1$$

we have from binomial expansion,

$$\left(1-\frac{km(t)}{C_0}\right)^{-1/2} \approx 1+\frac{km(t)}{2C_0}$$

$$f(t) \approx f_c\left(1+\frac{km(t)}{2C_0}\right)$$

$$= f_c + \frac{kf_c m(t)}{2C_0}$$

$$\text{or, } f(t) = f_c + k_f m(t)$$

$$k_f = \frac{kf_c}{2C_0}$$

is the frequency sensitivity of the modulator. The Eq. (5.42) is the required expression for the instantaneous frequency of an FM wave. In this way, we can generate an FM wave by direct method.

Direct FM may be generated also by a device in which the inductance of the resonant circuit is linearly varied by a modulating signal n(t); in this case the modulating signal being the current.

The main advantage of the direct method is that it produces sufficiently high frequency deviation, thus requiring little frequency multiplication. But, it has poor frequency stability. A feedback scheme is used to stabilize the frequency in which the output frequency is compared with the constant frequency generated by highly stable crystal oscillator and the error signal is feedback to stabilize the frequency.

DEMODULATION OF FM WAVES

The process to extract the message signal from a frequency modulated wave is known as frequency demodulation. As the information in an FM wave is contained in its instantaneous frequency, the frequency demodulator has the task of changing frequency variations to amplitude variations. Frequency demodulation method is generally categorized into two types: direct method and indirect method. Under direct method category, we will discuss about limiter discriminator method and under indirect method, phase-locked loop (PLL) will be discussed.

Limiter Discriminator Method

Recalling the expression of FM signal,

$$s(t) = A_c \cos\left[2\pi f_c t + 2\pi k_f \int_0^t n(t)dt\right]$$

In this method, extraction of n(t) from the above equation involves the three steps: amplitude limit, discrimination, and envelope detection.

A. Amplitude Limit

During propagation of the FM signal from transmitter to receiver the amplitude of the FM wave (supposed to be constant) may undergo changes due to fading and noise. Therefore, before further processing, the amplitude of the FM signal is limited to reduce the effect of fading and noise by using limiter as discussed in the section 5.9. The amplitude limitation will not affect the message signal as the amplitude of FM does not carry any information of the message signal.

B. Discrimination/ Differentiation

In this step we differentiate the FM signal as given by

$$\frac{ds(t)}{dt} = \frac{d}{dt}\left\{A_c \cos\left[2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right]\right\}$$

$$= \frac{d\left\{A_c \cos\left[2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right]\right\}}{d\left\{2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right\}} \frac{d\left\{2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right\}}{dt}$$

$$= -A_c\left[2\pi f_c + 2\pi k_f m(t)\right] \sin\left[2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right]$$

Here both the amplitude and frequency of this signal are modulated.

In this case, the differentiator is nothing but a circuit that converts change in frequency into corresponding change in voltage or current as shown in Fig. 5.11. The ideal differentiator has transfer function

$$H(jw) = j2nf$$

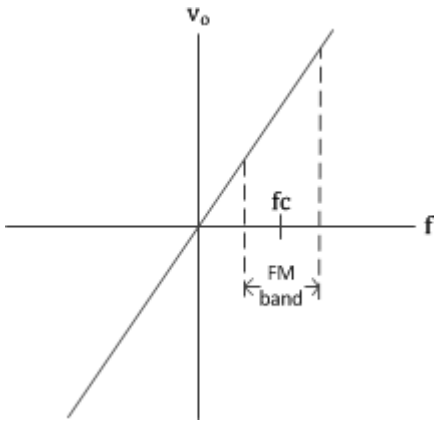


Figure : Transfer function of ideal differentiator.

Instead of ideal differentiator, any circuit can be used whose frequency response is linear for some band in positive slope. This method is known as slope detection. For this, linear segment with positive slope of RC high pass filter or LC tank circuit can be used. Figure 5.13 shows the use of an LC circuit as a differentiator. The drawback is the limited linear portion in the

slope of the tank circuit. This is not suitable for wideband FM where the peak frequency deviation is high.

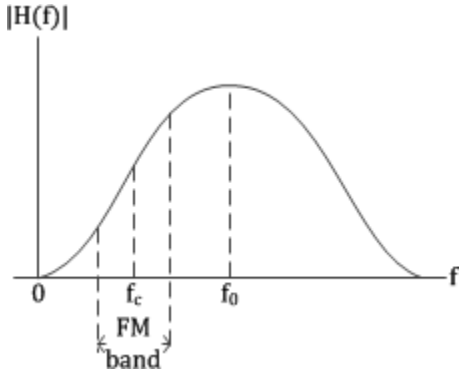


Figure : Use of LC tank circuit as a differentiator.

A better solution is the ratio or balanced slope detector in which two tank circuits tuned at $f_c + \Delta f$ and $f_c - \Delta f$ are used to extend the linear portion as shown in below figure.

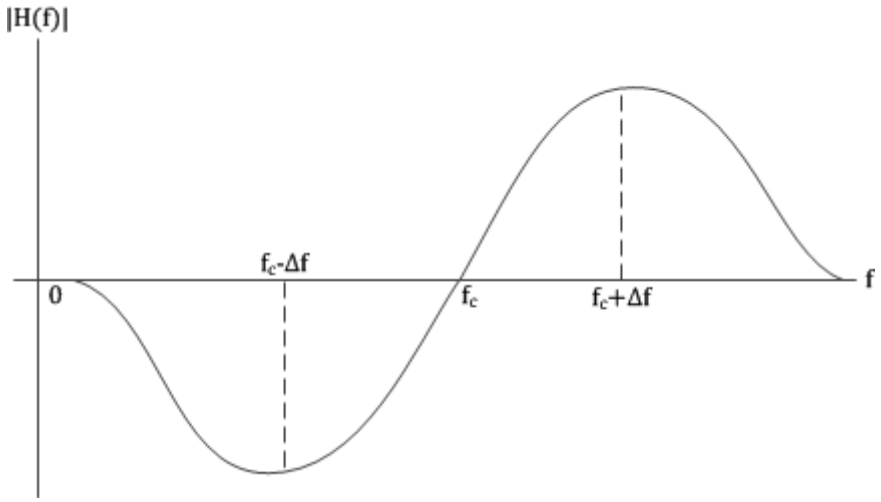


Figure : Frequency response of balanced slope detector.

Another detector called Foster-seely discriminator eliminates two tank circuits but still offer the same linear as the ratio detector.

C. Envelope Detection

The third step is to send the differentiated signal to the envelope detector to recover the message signal.

Phase-Locked Loop (PLL) as FM Demodulator

A PLL consists of a multiplier, a loop filter, and a VCO connected together to form a feedback loop as shown in Fig. 5.15. Let the input signal be an FM wave as defined by

$$s(t) = A_c \cos[2\pi f_c t + \emptyset_1(t)]$$

$$\emptyset_1(t) = 2\pi k_f \int_0^t m(t)dt$$

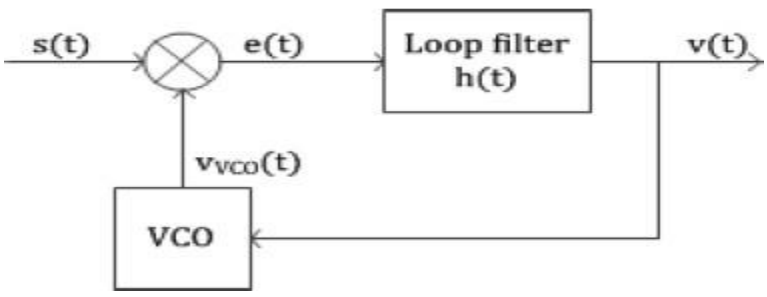


Fig: PLL Demodulator

Let the VCO output be defined by

$$v_{VCO}(t) = A_v \sin[2nf_c t + \emptyset_2(t)]$$

where

$$\emptyset_2(t) = 2nk_v \int_0^t v(t)dt$$

Here k_v is the frequency sensitivity of the VCO measured in hertz per volt. The multiplication of $s(t)$ and $v_{VCO}(t)$ results

$$s(t)v_{VCO}(t) = A_c \cos[2\pi f_c t + \emptyset_1(t)] A_v \sin[2\pi f_c t + \emptyset_2(t)]$$

$$= \frac{A_c A_v}{2} \sin[4\pi f_c t + \emptyset_1(t) + \emptyset_2(t)] + \frac{A_c A_v}{2} \sin[\emptyset_2(t) - \emptyset_1(t)]$$

The high-frequency component is removed by the low-pass filtering of the loop filter. Therefore, the input signal to the loop filter can be considered as

$$e(t) = \frac{A_c A_v}{2} \sin[\emptyset_2(t) - \emptyset_1(t)]$$

The difference $\emptyset_2(t) - \emptyset_1(t) = \emptyset_e(t)$ constitutes the phase error. Let us assume that the PLL is in phase lock so that the phase error is very small. Then,

$$\sin[\emptyset_2(t) - \emptyset_1(t)] \approx \emptyset_2(t) - \emptyset_1(t)$$

$$\emptyset_e(t) = 2\pi k_v \int_0^t v(t)dt - \emptyset_1(t)$$

$$e(t) = \frac{A_c A_v}{2} \emptyset_e(t)$$

Differentiating Eq. (5.48) with respect to time, we get

$$\frac{d\emptyset_e(t)}{dt} = 2\pi k_v v(t) - \frac{d\emptyset_1(t)}{dt}$$

Since

$$v(t) = e(t) * h(t) = \frac{A_c A_v}{2} [\emptyset_e(t) * h(t)]$$

Eq. (5.50) becomes

$$\frac{d\emptyset_e(t)}{dt} = 2\pi k_v \frac{A_c A_v}{2} [\emptyset_e(t) * h(t)] - \frac{d\emptyset_1(t)}{dt}$$

$$or, \pi k_v A_c A_v [\emptyset_e(t) * h(t)] - \frac{d\emptyset_e(t)}{dt} = \frac{d\emptyset_1(t)}{dt}$$

Taking Fourier transform of Eq. (5.52), we get

$$\pi k_v A_c A_v \emptyset_e(f) H(f) - j2\pi f \emptyset_e(f) = j2\pi f \emptyset_1(f)$$

$$or, \emptyset_e(f) = \frac{j2\pi f}{\pi k_v A_c A_v H(f) - j2\pi f} \emptyset_1(f)$$

$$or, \emptyset_e(f) = \frac{1}{\frac{\pi k_v A_c A_v}{j2\pi f} H(f) - 1} \emptyset_1(f)$$

Fourier transform of Eq. (5.51) is

$$V(f) = \frac{A_c A_v}{2} \emptyset_e(f) H(f)$$

Substituting Eq. (5.53) into (5.54), we get

$$V(f) = \frac{A_c A_v}{2} \frac{1}{\frac{\pi k_v A_c A_v}{j2\pi f} H(f) - 1} \emptyset_1(f) H(f)$$

We design $H(f)$ such that

$$\left| \frac{\pi k_v A_c A_v}{j2\pi f} H(f) \right| \gg 1$$

in the frequency band $|f| < W$ of the message signal.

78

$$\therefore V(f) = \frac{A_c A_v}{2} \frac{1}{\frac{\pi k_v A_c A_v}{j2\pi f} H(f)} \emptyset_1(f) H(f)$$

$$or, V(f) = \frac{1}{2\pi k_v} j2\pi f \emptyset_1(f)$$

Taking inverse Fourier transform of Eq. (4.56), we get

$$v(t) = \frac{1}{2\pi k_v} \frac{d\emptyset_1(t)}{dt}$$

$$= \frac{1}{2\pi k_v} \frac{d}{dt} \left\{ 2\pi k_f \int_0^t m(t) dt \right\}$$

$$= \frac{1}{2\pi k_v} 2\pi k_f m(t)$$

$$\therefore v(t) = \frac{k_f}{k_v} m(t)$$

Since the control voltage of the VCO is proportional to the message signal, v(t) is the demodulated signal.

We observe that the output of the loop filter with frequency response H(f) is the desired message signal. Hence the bandwidth of H(f) should be the same as the bandwidth W of the message signal. Consequently, the noise at the output of the loop filter is also limited to the bandwidth W. On the other hand, the output from the VCO is a wideband FM signal with an instantaneous frequency that follows the instantaneous frequency of the received FM signal.

PREEMPHASIS AND DEEMPHASIS NETWORKS

In FM, the noise increases linearly with frequency. By this, the higher frequency components of message signal are badly affected by the noise. To solve this problem, we can use a preemphasis filter of transfer function $H_p(f)$ at the transmitter to boost the higher frequency components before modulation. Similarly, at the receiver, the deemphasis filter of transfer function $H_d(f)$can be used after demodulator to attenuate the higher frequency components thereby restoring the original message signal.

The preemphasis network and its frequency response are shown in Figure 5.19 (a) and (b) respectively. Similarly, the counter part for deemphasis network is shown in Figure 5.20.

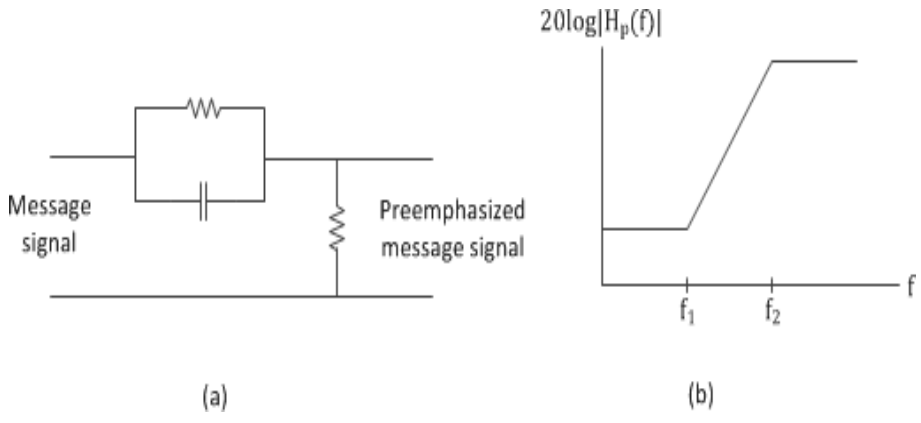


Figure ;(a) Preemphasis network. (b) Frequency response of preemphasis network.

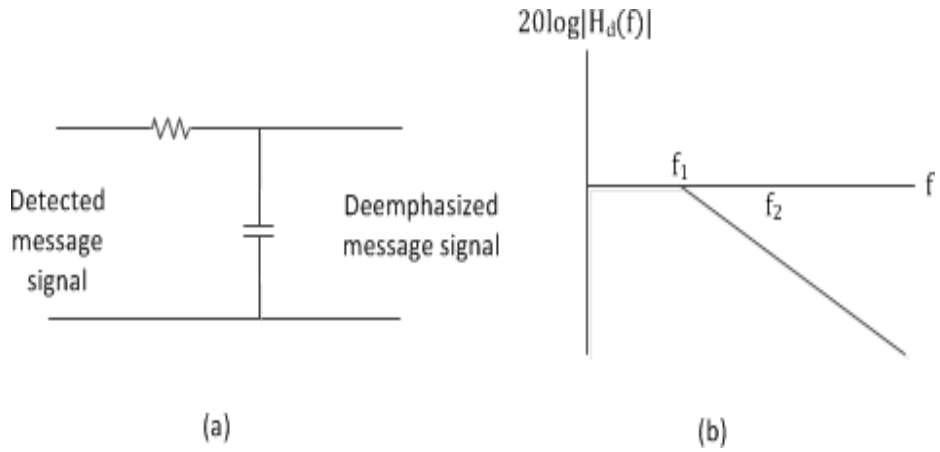


Figure (a) Deemphasis network. (b) Frequency response of Deemphasis network.

In FM broadcasting, f_1 and f_2 are normally chosen to be 2.1 kHz and 30 kHz respectively.

The frequency response of preemphasis network is

$$H_p(f) = \left(\frac{w_2}{w_1}\right)\frac{jw + w_1}{jw + w_2}$$

Here, $w = 2\pi f$ and $w_1 = 2\pi f_1$. For $w \ll w_1$,

$$H_p(f) \approx 1$$

And for $w_1 \ll w \ll w_2$,

$$H_p(f) \approx \frac{j2\pi f}{w_1}$$

So, the amplitude of frequency components less than 2.1 kHz are left unchanged and greater than that are increased proportional to f.

The frequency response of deemphasis network is

$$H_d(f) = \frac{w_1}{j2\pi f + w_1}$$

For $w \ll w_2$,

$$H_p(f) \approx \frac{j2\pi f + w_1}{w_1}$$

such that

$$H_p(f)H_d(f) \approx 1$$

over the baseband of 0 to 15 KHz.

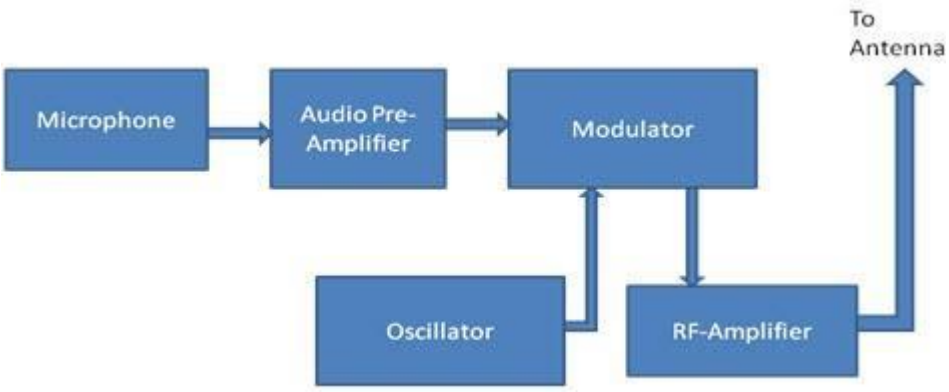
Comparison of AM and FM:

S.NO	AMPLITUDE MODULATION	FREQUENCY MODULATION
1.	Band width is very small which is one of the biggest advantage	It requires much wider channel (7 to 15 times) as compared to AM.
2.	The amplitude of AM signal varies depending on modulation index.	The amplitude of FM signal is constant and independent of depth of the modulation.
3.	Area of reception is large	The are of reception is small since it is limited to line of sight.
4.	Transmitters are relatively simple & cheap.	Transmitters are complex and hence expensive.
5.	The average power in modulated wave is greater than carrier power. This added power is provided by modulating source.	The average power in frequency modulated wave is same as contained in un-modulated wave.
6.	More susceptible to noise interference and has low signal to noise ratio, it is more difficult to eliminate effects of noise.	Noise can be easily minimized amplitude variations can be eliminated by using limiter.
7.	it is not possible to operate without interference.	it is possible to operate several independent transmitters on same frequency.
8.	The maximum value of modulation index = 1, other wise over-modulation would result in distortions.	No restriction is placed on modulation index.

FM Transmitter

The FM transmitter is a single transistor circuit. In the telecommunication, the frequency modulation (FM)transfers the information by varying the frequency of carrier wave according to the message signal. Generally, the FM transmitter uses VHF radio frequencies of 87.5 to 108.0 MHz to transmit & receive the FM signal. This transmitter accomplishes the most excellent range with less power. The performance and working of the wireless audio transmitter circuit is depends on the induction coil & variable capacitor. This article will explain about the working of the FM transmitter circuit with its applications.

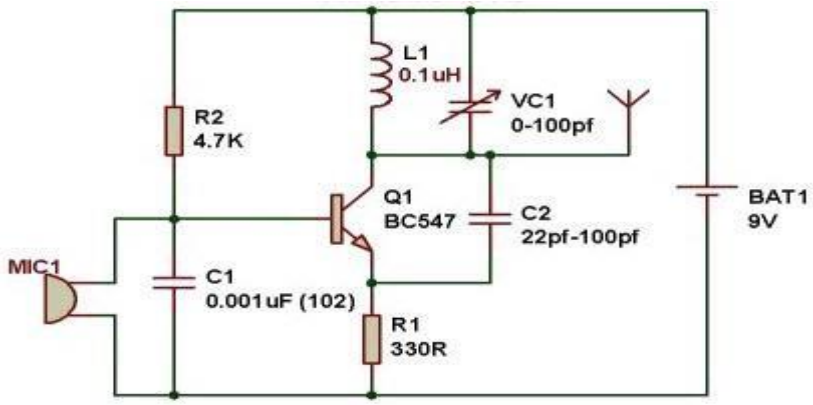
The FM transmitter is a low power transmitter and it uses FM waves for transmitting the sound, this transmitter transmits the audio signals through the carrier wave by the difference of frequency. The carrier wave frequency is equivalent to the audio signal of the amplitude and the FM transmitter produce VHF band of 88 to 108MHZ.Plese follow the below link for: Know all About Power Amplifiers for FM Transmitter



Block Diagram of FM Transmitter

Working of FM Transmitter Circuit

The following circuit diagram shows the FM transmitter circuit and the required electrical and electronic components for this circuit is the power supply of 9V, resistor, capacitor, trimmer capacitor, inductor, mic, transmitter, and antenna. Let us consider the microphone to understand the sound signals and inside the mic there is a presence of capacitive sensor. It produces according to the vibration to the change of air pressure and the AC signal.



FM Transmitter circuit

The formation of the oscillating tank circuit can be done through the transistor of 2N3904 by using the inductor and variable capacitor. The transistor used in this circuit is an NPN transistor used for general purpose amplification. If the current is passed at the inductor L1 and variable capacitor then the tank circuit will oscillate at the resonant carrier frequency of the FM modulation. The negative feedback will be the capacitor C2 to the oscillating tank circuit.

To generate the radio frequency carrier waves the FM transmitter circuit requires an oscillator. The tank circuit is derived from the LC circuit to store the energy for oscillations.

The input audio signal from the mic penetrated to the base of the transistor, which modulates the LC tank circuit carrier frequency in FM format. The variable capacitor is used to change the resonant frequency for fine modification to the FM frequency band. The modulated signal from the antenna is radiated as radio waves at the FM frequency band and the antenna is nothing but copper wire of 20cm long and 24 gauge. In this circuit the length of the antenna should be significant and here you can use the 25-27 inches long copper wire of the antenna.

Application of Fm Transmitter

- The FM transmitters are used in the homes like sound systems in halls to fill the sound with the audio source.
- These are also used in the cars and fitness centers.
- The correctional facilities have used in the FM transmitters to reduce the prison noise in common areas.

Advantages of the FM Transmitters

- The FM transmitters are easy to use and the price is low
- The efficiency of the transmitter is very high
- It has a large operating range
- This transmitter will reject the noise signal from an amplitude variation.

UNIT II
NOISE

- Noise in communication System,
- White Noise
- Narrowband Noise –In phase and Quadrature phase components
- Noise Bandwidth
- Noise Figure
- Noise Temperature
- Noise in DSB& SSB System
- Noise in AM System
- Noise in Angle Modulation System
- Threshold effect in Angle Modulation System

Noise in communication system

A signal may be contaminated along the path by noise. Noise may be defined as any unwanted introduction of energy into the desired signal. In radio receivers, noise may produce "hiss" in the loudspeaker output. Noise is random and unpredictable.

Noise is produced both external and internal to the system. External noise includes atmospheric noise (e.g., from lightning), galactic noise (thermal radiation from cosmic bodies), and industrial noise (e.g., from motors, ignition). We can minimize or eliminate external noise by proper system design. On the other hand, internal noise is generated inside the system. It is resulted due to random motion of charged particles in resistors, conductors, and electronic devices. With proper system design, it can be minimized but never can be eliminated. This is the major constraint in the rate of telecommunications.

- Noise is unwanted signal that affects wanted signal
- Noise is random signal that exists in communication systems

Effect of noise

- Degrades system performance (Analog and digital)
- Receiver cannot distinguish signal from noise
- Efficiency of communication system reduces

Types of noise

- Thermal noise/white noise/Johnson noise or fluctuation noise
- Shot noise
- Noise temperature
- Quantization noise

Noise temperature

Equivalent noise temperature is not the physical temperature of amplifier, but a theoretical construct, that is an equivalent temperature that produces that amount of noise power

$T_e = (F - 1)$

White noise

One of the very important random processes is the *white noise* process. Noises in many practical situations are approximated by the white noise process. Most importantly, the white noise plays an important role in modelling of WSS signals.

A white noise process $\{W(t)\}$ is a random process that has constant power spectral density at all frequencies. Thus

$$S_W(\omega) = \frac{N_0}{2} \qquad -\infty < \omega < \infty$$

where N_0 is a real constant and called the *intensity* of the white noise. The corresponding autocorrelation function is given by

$R_W(\tau) = \frac{N}{2}\delta(\tau)$ where $\delta(\tau)$ is the Dirac delta.

The average power of white noise

$$P_{avg} = EW^2(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{N}{2}d\omega \to \infty$$

The autocorrelation function and the PSD of a white noise process is shown in Figure 1 below.

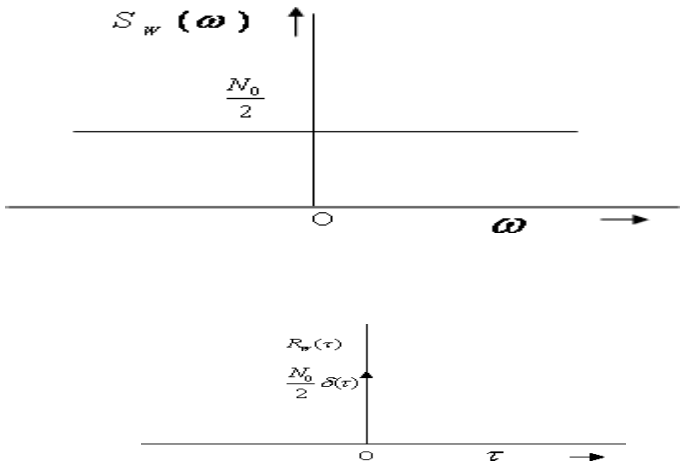


fig: auto correlation and psd of white noise

NARROWBAND NOISE (NBN)

In most communication systems, we are often dealing with band-pass filtering of signals. Wideband noise will be shaped into band limited noise. If the bandwidth of the band limited noise is relatively small compared to the carrier frequency, we refer to this as *narrowband noise*.

the narrowband noise is expressed as as

$$n(t) = x(t)\cos 2\pi f_C t - y(t)\sin 2\pi f_C t$$

where f_C is the carrier frequency within the band occupied by the noise. $x(t)$ and $y(t)$ are known as the *quadrature components* of the noise $n(t)$. The Hibert transform of

$n(t)$ is

Proof.

The Fourier transform of $n(t)$ is

$$N(f) = \frac{1}{2}X(f - f_c) + \frac{1}{2}X(f + f_c) + \frac{1}{2}jY(f - f_c) - \frac{1}{2}jY(f + f_c)$$

Let $\hat{N}(f)$ be the Fourier transform of $\hat{n}(t)$. In the frequency domain, $\hat{N}(f) = N(f)[-j\ \mathrm{sgn}(f)]$. We simply multiply all positive frequency components of $N(f)$ by $-j$ and all negative frequency components of $N(f)$ by j. Thus

$$\hat{n}(t) = H[n(t)] = x(t)\sin 2\pi f_c t + y(t)\cos 2\pi f_c t$$

$$\hat{N}(f) = -j\frac{1}{2}X(f - f_c) + j\frac{1}{2}X(f + f_c) - j\frac{1}{2}jY(f - f_c) - j\frac{1}{2}jY(f + f_c)$$

$$\hat{N}(f) = -j\frac{1}{2}X(f - f_c) + j\frac{1}{2}X(f + f_c) + \frac{1}{2}Y(f - f_c) + \frac{1}{2}Y(f + f_c)$$

and the inverse Fourier transform of $\hat{N}(f)$ is

$$\hat{n}(t) = x(t)\sin 2\pi f_c t + y(t)\cos 2\pi f_c t$$

The quadrature components x(t) and $y(t)$ can now be derived from equations

$x(t) = n(t)\cos 2\pi f_c t + \hat{n}(t)\sin 2\pi f_c t$ \

and

$y(t) = n(t)\cos 2\pi f_c t - \hat{n}(t)\sin 2\pi f_c t$

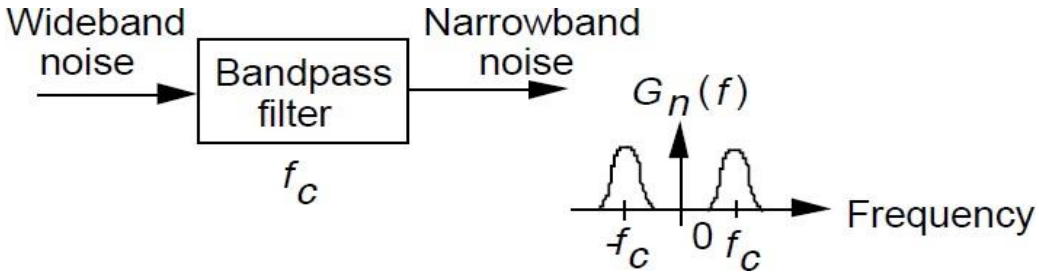


Fig: generation of narrow band noise

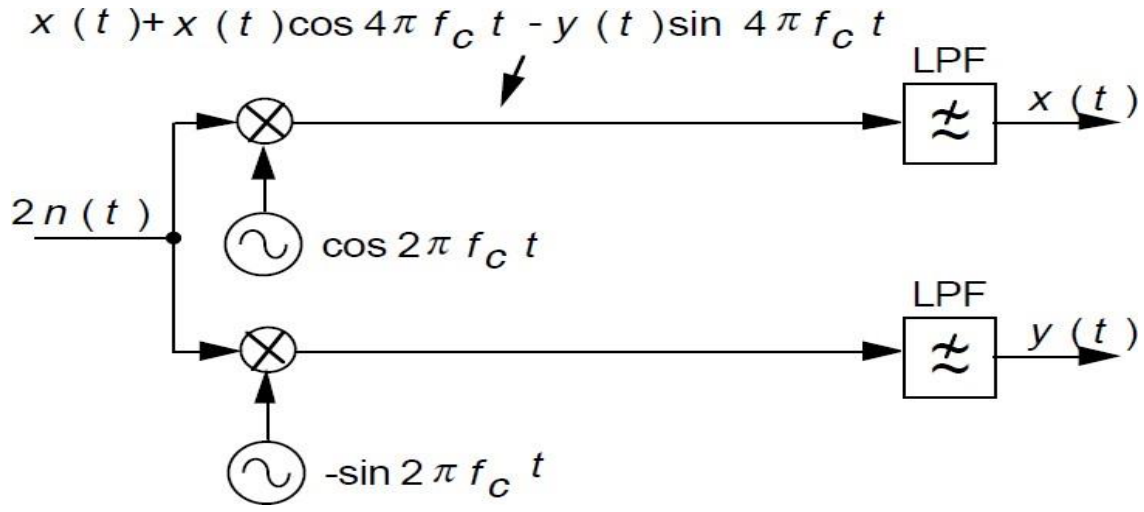


Fig: Generation of quadrature components of $n(t)$.

- Filters at the receiver have enough bandwidth to pass the desired signal but not too big to pass excess noise.
- Narrowband (NB) fc center frequency is much bigger that the bandwidth.
- Noise at the output of such filters is called narrowband noise (NBN).
- NBN has spectral concentrated about some mid-band frequency f_C
- The sample function of such NBN $n(t)$ appears as a sine wave of frequency f_C which modulates slowly in amplitude and phase

Input signal-to-noise ratio (SNR_I): is the ratio of the average power of modulated signal $s(t)$ to the average power of the filtered noise.

Output signal-to-noise ratio (SNR_O): is the ratio of the average power of demodulated message to the average power of the noise, both measured at the receiver output.

Channel signal-to-noise ratio (SNR_C): is the ratio of the average power of modulated signal $s(t)$ to the average power of the noise in the message bandwidth, both measured at the receiver input.

Noise figure

The Noise figure is the amount of noise power added by the electronic circuitry in the receiver to the thermal noise power from the input of the receiver. The thermal noise at the input to the receiver passes through to the demodulator. This noise is present in the receive channel and cannot be removed. The noise figure of circuits in the receiver such as amplifiers and mixers, adds additional noise to the receive channel. This raises the noise floor at the demodulator.

$$Noise\ Figure = \frac{Signal\ to\ noise\ ratio\ at\ input}{Signal\ to\ noise\ ratio\ at\ output}$$

Noise Bandwidth

A filter's equivalent noise bandwidth (ENBW) is defined as the bandwidth of a perfect rectangular filter that passes the same amount of power as the cumulative bandwidth of the channel selective filters in the receiver. At this point we would like to know the noise floor in our receiver, i.e. the noise power in the receiver intermediate frequency (IF) filter bandwidth that comes from kTB. Since the units of kTB are Watts/ Hz, calculate the noise floor in the channel bandwidth by multiplying the noise power in a 1 Hz bandwidth by the overall equivalent noise bandwidth in Hz.

NOISE IN DSB-SC SYSTEM:

Let the transmitted signal is

$$u(t) = A_c m(t)\cos(2\pi f_c t)$$

The received signal at the output of the receiver noise- limiting filter : Sum of this signal and filtered noise .A filtered noise process can be expressed in terms of its in-phase and quadrature components as

$$n(t) = A(t)\cos[2\pi f_c t + \theta(t)] = A(t)\cos\theta(t)\cos(2\pi f_c t) - A(t)\sin\theta(t)\sin(2\pi f_c t)$$
$$= n_c(t)\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t)$$

where $n_c(t)$ is in-phase component and $n_s(t)$ is quadrature component

Received signal (Adding the filtered noise to the modulated signal)

$$r(t) = u(t) + n(t)$$
$$= A_c m(t)\cos(2\pi f_c t) + n_c(t)\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t)$$

Demodulate the received signal by first multiplying r(t) by a locally generated sinusoid $\cos(2\pi f_c t + \phi)$, where ϕ is the phase of the sinusoid.Then passing the product signal through an ideal lowpass filter having a bandwidth W.

The multiplication of r(t) with $\cos(2\pi f_c t + \phi)$ yields

$$r(t)\cos(2\pi f_c t + \phi)$$
$$= u(t)\cos(2\pi f_c t + \phi) + n(t)\cos(2\pi f_c t + \phi)$$
$$= A_c m(t)\cos(2\pi f_c t)\cos(2\pi f_c t + \phi)$$
$$+ n_c(t)\cos(2\pi f_c t)\cos(2\pi f_c t + \phi) - n_s(t)\sin(2\pi f_c t)\cos(2\pi f_c t + \phi)$$
$$= \tfrac{1}{2} A_c m(t)\cos(\phi) + \tfrac{1}{2} A_c m(t)\cos(4\pi f_c t + \phi)$$
$$+ \tfrac{1}{2}\left[n_c(t)\cos(\phi) + n_s(t)\sin(\phi)\right] + \tfrac{1}{2}\left[n_c(t)\cos(4\pi f_c t + \phi) - n_s(t)\sin(4\pi f_c t + \phi)\right]$$

The low pass filter rejects the double frequency components and passes only the low pass components.

$$y(t) = \tfrac{1}{2} A_c m(t)\cos(\phi) + \tfrac{1}{2}\left[n_c(t)\cos(\phi) + n_s(t)\sin(\phi)\right]$$

the effect of a phase difference between the received carrier and a locally generated carrier at the receiveris a drop equal to $\cos^2(\phi)$ in the received signal power.

Phase-locked loop

The effect of a phase-locked loop is to generate phase of the received carrier at the receiver.

If a phase-locked loop is employed, then $\phi = 0$ and the demodulator is called a coherent or synchronous demodulator.

In our analysis in this section, we assume that we are employing a coherent demodulator.

With this assumption, we assume that $\phi = 0$

$$y(t) = \tfrac{1}{2}\left[A_c m(t) + n_c(t)\right]$$

Therefore, at the receiver output, the message signal and the noise components are additive and we are able to define a meaningful SNR. The message signal power is given by

$$P_o = \frac{A_c^2}{4} P_M$$

Power P_M is the content of the messagesignal

The noise power is given by

$$P_{n_0} = \frac{1}{4} P_{n_c} = \frac{1}{4} P_n$$

The power content of n(t) can be found by noting that it is the result of passing $n_w(t)$ through a filter with bandwidth B_c.Therefore, the power spectral density of n(t) is given by

$$S_n(f) = \begin{cases} \frac{N_0}{2} & |f - f_c| < W \\ 0 & otherwise \end{cases}$$

The noise power is

$$P_n = \int_{-\infty}^{\infty} S_n(f) df = \frac{N_0}{2} \times 4W = 2WN_0$$

Now we can find the output SNR as

$$\left(\frac{S}{N}\right)_0 = \frac{P_0}{P_{n_0}} = \frac{\frac{A_c^2}{4} P_M}{\frac{1}{4} 2WN_0} = \frac{A_c^2 P_M}{2WN_0}$$

In this case, the received signal power, given by

$$P_R = A_c^2 P_M / 2.$$

The output SNR for DSB-SC AM may be expressed as

$$\left(\frac{S}{N}\right)_{0_{DSB}} = \frac{P_R}{N_0 W}$$

which is identical to baseband SNR.

In DSB-SC AM, the output SNR is the same as the SNR for a baseband system. DSB-SC AM does not provide any SNR improvement over a simple baseband communication system.

NOISE IN SSB-SC SYSTEM:

Let SSB modulated signal is

$$u(t) = A_c m(t) \cos(2\pi f_c t) \mp A_c \hat{m}(t) \sin(2\pi f_c t)$$

Input to the demodulator

$$
\begin{aligned}
r(t) &= A_c m(t)\cos(2\pi f_c t) \mp A_c \hat{m}(t)\sin(2\pi f_c t) + n(t) \\
&= A_c m(t)\cos(2\pi f_c t) \mp A_c \hat{m}(t)\sin(2\pi f_c t) + n_c(t)\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t) \\
&= \left[A_c m(t) + n_c(t)\right]\cos(2\pi f_c t) + \left[\mp A_c \hat{m}(t) - n_s(t)\right]\sin(2\pi f_c t)
\end{aligned}
$$

Assumption : Demodulation with an ideal phase reference.

Hence, the output of the lowpass filter is the in-phase component (with a coefficient of ½) of the precedingsignal.

Parallel to our discussion of DSB, we have

$$P_o = \frac{A_c^2}{4} P_M$$

$$P_{n_0} = \frac{1}{4} P_{n_c} = \frac{1}{4} P_n$$

$$P_n = \int_{-\infty}^{\infty} S_n(f)\,df = \frac{N_0}{2} \times 2W = WN_0$$

$$\left(\frac{S}{N}\right)_0 = \frac{P_0}{P_{n_0}} = \frac{A_c^2 P_M}{WN_0}$$

$$P_R = P_U = A_c^2 P_M$$

$$\left(\frac{S}{N}\right)_{0_{SSB}} = \frac{P_R}{N_0 W} = \left(\frac{S}{N}\right)_b$$

The signal-to-noise ratio in an SSB system is equivalent to that of a DSB system.

Noise in Conventional AM

DSB AM signal : $u(t) = A_c[1 + am_n(t)]\cos(2\pi f_c t)$

Received signal at the input to the demodulator

$$
\begin{aligned}
r(t) &= A_c[1 + am_n(t)]\cos(2\pi f_c t) + n(t) \\
&= A_c[1 + am_n(t)]\cos(2\pi f_c t) + n_c(t)\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t) \\
&= \left[A_c[1 + am_n(t)] + n_c(t)\right]\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t)
\end{aligned}
$$

Where *a* is the modulation index

$m_n(t)$ is normalized so that its minimum value is *-1*

If a synchronous demodulator is employed, the situation is basically similar to the DSB case, except that we have $1 + am_n(t)$ instead of $m(t)$.

$$y(t) = \tfrac{1}{2}\left[A_c a m_n(t) + n_c(t)\right]$$

Received signal power

$$P_R = \frac{A_c^2}{2}\left[1 + a^2 P_{M_n}\right]$$

- Assumed that the message process is zero mean.

Now we can derive the output SNR as

$$\left(\frac{S}{N}\right)_{0_{AM}} = \frac{\frac{1}{4}A_c^2 a^2 P_{M_n}}{\frac{1}{4}P_{n_c}} = \frac{A_c^2 a^2 P_{M_n}}{2N_0 W} = \frac{a^2 P_{M_n}}{1 + a^2 P_{M_n}} \frac{\frac{A_c^2}{2}\left[1 + a^2 P_{M_n}\right]}{N_0 W}$$

$$= \frac{a^2 P_{M_n}}{1 + a^2 P_{M_n}} \frac{P_R}{N_0 W} = \frac{a^2 P_{M_n}}{1 + a^2 P_{M_n}} \left(\frac{S}{N}\right)_b = \eta \left(\frac{S}{N}\right)_b$$

- η denotes the modulation efficiency
- Since $a^2 P_{M_n} < 1 + a^2 P_{M_n}$, the SNR in conventional AM is always smaller than the SNR in a baseband system.

- In practical applications, the modulation index a is in the range of 0.8-0.9.
- Power content of the normalized message process depends on the message source.
- Speech signals : Large dynamic range, P_M is about 0.1.
- The overall loss in SNR, when compared to a baseband system, is a factor of 0.075 or equivalent to a loss of 11 dB.

The reason for this loss is that a large part of the transmitter power is used to send the carrier component of the modulated signal and not the desired signal. To analyze the envelope-detector performance in the presence of noise, we must use certain approximations.

This is a result of the nonlinear structure of an envelope detector, which makes an exact analysis difficult

In this case, the demodulator detects the envelope of the received signal and the noise process.

The input to the envelope detector is

$$r(t) = \left[A_c[1 + a m_n(t)] + n_c(t)\right]\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t)$$

Therefore, the envelope of *r (t)* is given by

$$V_r(t) = \sqrt{\left[A_c[1 + a m_n(t)] + n_c(t)\right]^2 + n_s^2(t)}$$

Now we assume that the signal component in *r (t)* is much stronger than the noise component. Then

$$P\left(n_c(t) << A_c[1+am_n(t)]\right) \approx 1$$

Therefore, we have a high probability that

$$V_r(t) \approx A_c[1+am_n(t)]+n_c(t)$$

After removing the DC component, we obtain

$$y(t) = A_c am_n(t) + n_c(t)$$

which is basically the same as y*(t)* for the synchronous demodulation without the ½ coefficient.

This coefficient, of course, has no effect on the final SNR. So we conclude that, under the assumption of high SNR at the receiver input, the performance of synchronous and envelope demodulators is the same.

However, if the preceding assumption is not true, that is, if we assume that, at the receiver input, the noise power is much stronger than the signal power, Then

$$\begin{aligned} V_r(t) &= \sqrt{\left[A_c[1+am_n(t)]+n_c(t)\right]^2 + n_s^2(t)} \\ &= \sqrt{A_c^2[1+am_n(t)]^2 + n_c^2(t) + n_s^2(t) + 2A_c n_c(t)[1+am_n(t)]} \\ &\xrightarrow{a} \sqrt{\left(n_c^2(t)+n_s^2(t)\right)\left[1+\frac{2A_c n_c(t)}{n_c^2(t)+n_s^2(t)}\left(1+am_n(t)\right)\right]} \\ &\xrightarrow{b} V_n(t)\left[1+\frac{A_c n_c(t)}{V_n^2(t)}\left(1+am_n(t)\right)\right] \\ &= V_n(t)+\frac{A_c n_c(t)}{V_n(t)}\left(1+am_n(t)\right) \end{aligned}$$

(a) : $A_c^2[1+am_n(t)]^2$ is small compared with the other components

(b) : $\sqrt{n_c^2(t)+n_s^2(t)} = V_n(t)$:the envelope of the noise process

Use the approximation

$\sqrt{1+\varepsilon} \approx 1+\frac{\varepsilon}{2}$, *for small* ε, where $\varepsilon = \frac{2A_c n_c(t)}{n_c^2(t)+n_s^2(t)}\left(1+am_n(t)\right)$

Then $$V_r(t) = V_n(t) + \frac{A_c n_c(t)}{V_n(t)}\left(1 + a m_n(t)\right)$$

We observe that, at the demodulator output, the signal and the noise components are no longer additive. In fact, the signal component is multiplied by noise and is no longer distinguishable. In this case, no meaningful SNR can be defined. We say that this system is operating below the threshold. The subject of threshold and its effect on the performance of a communication system will be covered in more detail when we discuss the noise performance in angle modulation.

Effect of threshold in angle modulation system:

FM THRESHOLD EFFECT FM threshold is usually defined as a Carrier-to-Noise ratio at which demodulated Signal-to-Noise ratio falls 1dB below the linear relationship . This is the effect produced in an FM receiver when noise limits the desired information signal. It occurs at about 10 dB, as earlier stated in 5 the introduction, which is at a point where the FM signal-to-Noise improvement is measured. Below the FM threshold point, the noise signal (whose amplitude and phase are randomly varying) may instantaneously have amplitude greater than that of the wanted signal. When this happens, the noise will produce a sudden change in the phase of the FM demodulator output. In an audio system, this sudden phase change makes a "click". In video applications the term "click noise" is used to describe short horizontal black and white lines that appear randomly over a picture

An important aspect of analogue FM satellite systems is FM threshold effect. In FM systems where the signal level is well above noise received carrier-to-noise ratio and demodulated signal-to-noise ratio are related by:

$$S/N = 3\beta^2 C/N \qquad \text{Eqn 9}$$

where
S/N = signal-to-noise ratio at output of FM demodulator
β = FM deviation ratio or modulation index = (Δf/B)
C/N = carrier-to-noise ratio at input of FM demodulator
Δf = peak deviation
B = basebandwidth of signal being modulated

The expression however does not apply when the carrier-to-noise ratio decreases below a certain point. Below this critical point the signal-to-noise ratio decreases significantly. This is

known as the FM threshold effect (FM threshold is usually defined as the carrier-to-noise ratio at which the demodulated signal-to-noise ratio fall 1 dB below the linear relationship given in Eqn 9. It generally is considered to occur at about 10 dB).

Below the FM threshold point the noise signal (whose amplitude and phase are randomly varying), may instantaneously have an amplitude greater than that of the wanted signal. When this happens the noise will produce a sudden change in the phase of the FM demodulator output. In an audio system this sudden phase change makes a "click". In video applications the term "click noise" is used to describe short horizontal black and white lines that appear randomly over a picture, because satellite communications systems are power limited they usually operate with only a small design margin above the FM threshold point (perhaps a few dB). Because of this circuit designers have tried to devise techniques to delay the onset of the FM threshold effect. These devices are generally known as FM threshold extension demodulators. Techniques such as FM feedback, phase locked loops and frequency locked loops are used to achieve this effect. By such techniques the onset of FM threshold effects can be delayed till the C/N ratio is around 7 dB.

Noise in Angle Modulated Systems

Like AM, noise performance of angle modulated systems is characterized by parameter γ

$$\gamma_{FM} = \frac{3}{2}\beta^2$$

If it is compared with AM

$$\frac{\gamma_{FM}}{\gamma_{AM}} = \frac{1}{2}\left(\frac{\omega_{FM}}{\omega_{AM}}\right)^2$$

Note: if bandwidth ratio is increased by a factor 2.Then $\frac{\gamma_{FM}}{\gamma_{AM}}$ increases by a factor 4

This exchange of bandwidth and noise performance is an important feature of FM

$$\text{Figure of merit } (\gamma) = \frac{SNR_O}{SNR_C}$$

CW- Modulation System	SNR_O	SNR_C	Figure of merit	Figure of merit (single tone)
DSB-SC	$\frac{C^2 A_c^2 P}{2WN_0}$	$\frac{C^2 A_c^2 P}{2WN_0}$	1	1
SSB	$\frac{C^2 A_c^2 P}{4WN_0}$	$\frac{C^2 A_c^2 P}{4WN_0}$	1	1
AM	$\frac{A_c^2 k_a^2 P}{2WN_0}$	$\frac{A_c^2 (1+k_a^2 P)}{2WN_0}$	$\approx \frac{k_a^2 P}{1+k_a^2 P} < 1$	$\frac{\mu^2}{2+\mu^2}$
FM	$\frac{3A_c^2 k_f^2 P}{2N_0 W^3}$	$\frac{A_c^2}{2WN_0}$	$\frac{3k_f^2 P}{W^2}$	$\frac{3}{2}\beta^2$

P is the average power of the message signal.

C^2 is a constant that ensures that the ration is dimensionless.

W is the message bandwidth.

A_c is the amplitude of the carrier signal.

k_a is the amplitude sensitivity of the modulator.

$\mu = k_a A_m$ and A_m is the amplitude sinusoidal wave

$\beta = \frac{\Delta f}{W}$ is the modulation index.

k_f is the frequency sensitivity of the modulator.

Δf is the frequency deviation.

UNIT 5 - PULSE ANALOG MODULATION

Sampling Theorem:

All pulse modulation scheme undergoes sampling process. Sampling of low frequency(LF) signal is achieved using a pulse train. Sampling process provides samples of the message signal. Sampling rate of sampling process must be proper to get original signal back. Sampling theorem defines the sampling rate of sampling process in order to recover the message signal. The solution to sampling rate was provided by Shannon.

Basically there are two types of message signal, such as-

(i) Low-pass (baseband) signal,

(ii) Band-pass (passband) signal.

- **Sampling rate for Low-Pass Signal:--**

  Sampling theorem states that if g(t) being a lowpass signal of finite energy and is band limited to W Hz, then the signal can be completely described by and recovered from its sampled values taken at a rate of 2W samples or more per

second.

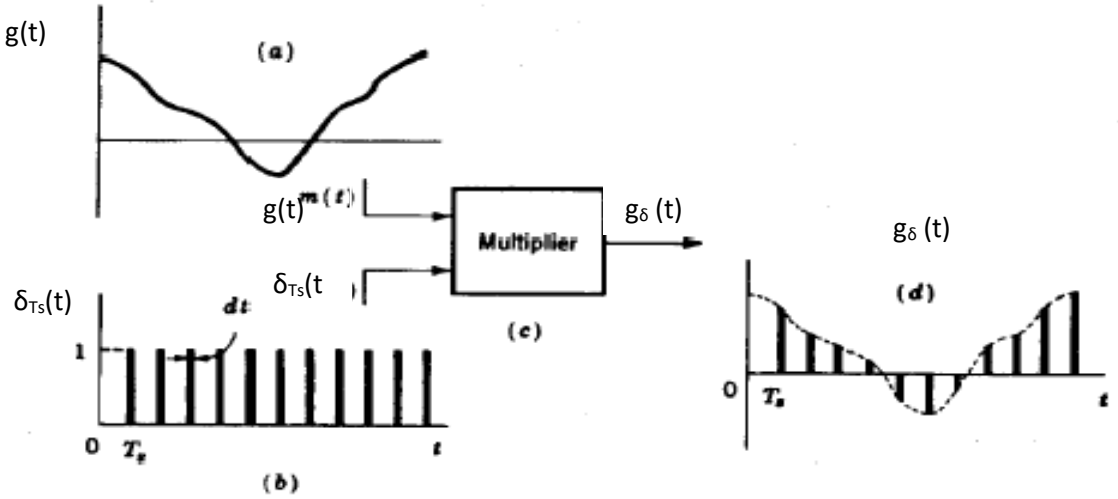


Fig. 1.1 Representation of sampling process.

Thus the time period of sampled signal must be, $Ts \leq 1/(2W)$.

Considering a signal g(t) as shown be a low pass signal where fourier transform of g(t),

$G(f) = 0$, for $f > W$

$=$ finite, for $f \leq W$.

Ideally, we can get sampled values of g(t) at a regular time interval of time T_s if we multiply a train of pulses δ_{Ts} to g(t) as shown.

The product signal [$g_\delta(t)$], ie, the sampled values can be written as,

$$g_\delta(t) = g(t)\ \delta_{Ts}(t) \qquad (1.1a)$$

or, $$g_\delta(t) = g(t) \qquad (1.1b)$$

If we denote $g(nT_s)$ as the weights of low pass signal at sampled interval, then we can write,

$$g_\delta(t) = \qquad (1.2)$$

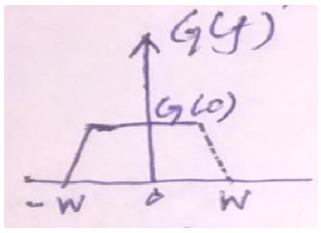
Taking the fourier transform of equation 1.2, we get

$$G_r(f) = G(f)$$

Or, $$Gr(f) = 1/T_s$$

or, $$Gr(f) = 1/T_s \qquad (1.3)$$

Now, we can draw graphically the frequency components of both the original signal and the sampled signal as follows,



ig. 1.1a Spectrum of original signal.

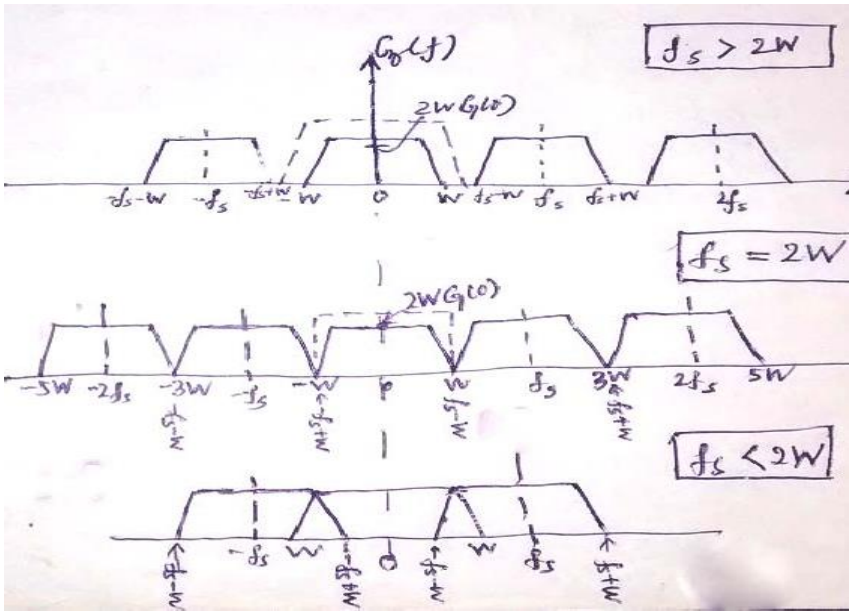


Fig. 1.1b Spectrum of Sampled signal.

Note:- The process of uniformly sampling a baseband signal in time domain results in a periodic spectrum in the frequency domain with a period, $f_s = 1/T_s$, where T_s is the sampling period in time domain and $\leq 1/2W$.

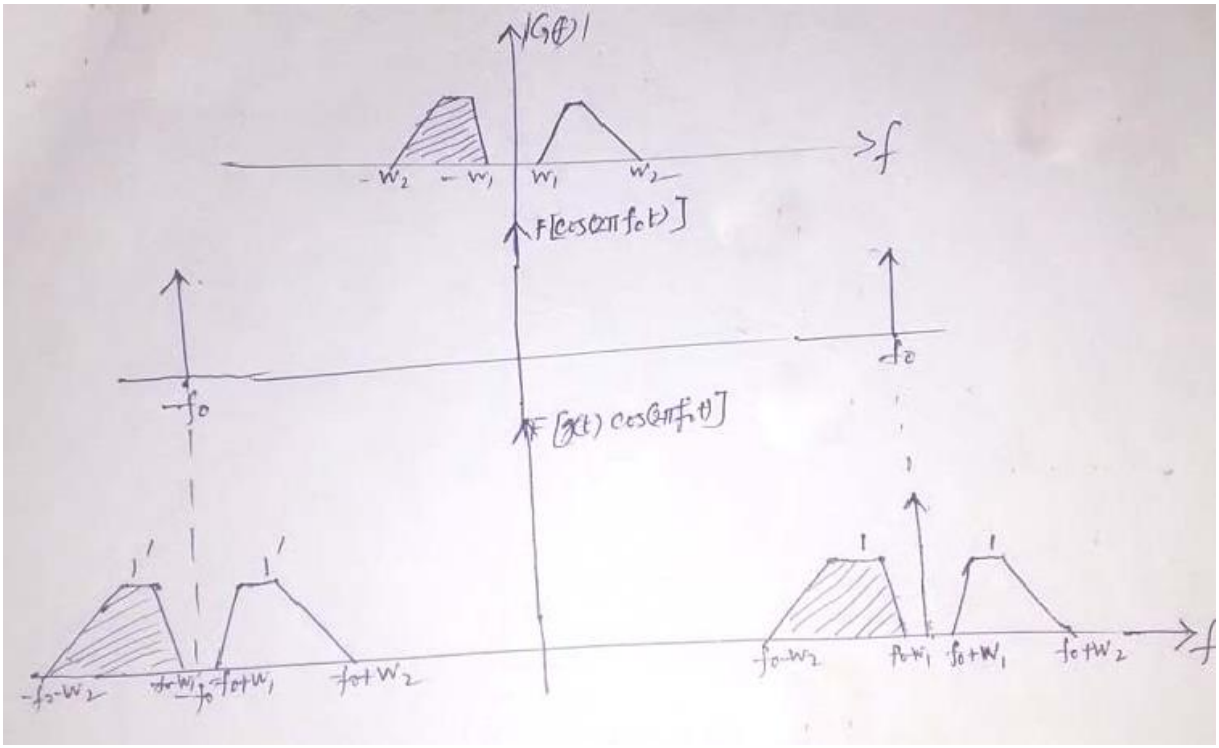


Fig. 1.1c Spectrum of baseband, carrier and modulated carrier signal.

- **Sampling of Bandpas Signal:**
  If the spectral range of a signal extends from 10 MHz to 10.1 MHz, the signal may ne recovered from samples taken at a frequency f_s=2{10.1 - 10} = 0.2 MHz.
  The sampling signal $\delta_{Ts}(t)$ is periodic. So,

$$\delta_{Ts} = dt/ds + 2.dt/ds(\cos 2\pi t/Ts + \cos(2.2\ \pi t/T_s) + \cos(3.2\ \pi t/T_s) + \ldots\ldots)$$
$$= f_s dt + 2f_s dt(\cos 2\pi f_{st} + \cos(2\pi.2f_s t) + \cos 2\pi.3f_s t + \ldots\ldots)$$

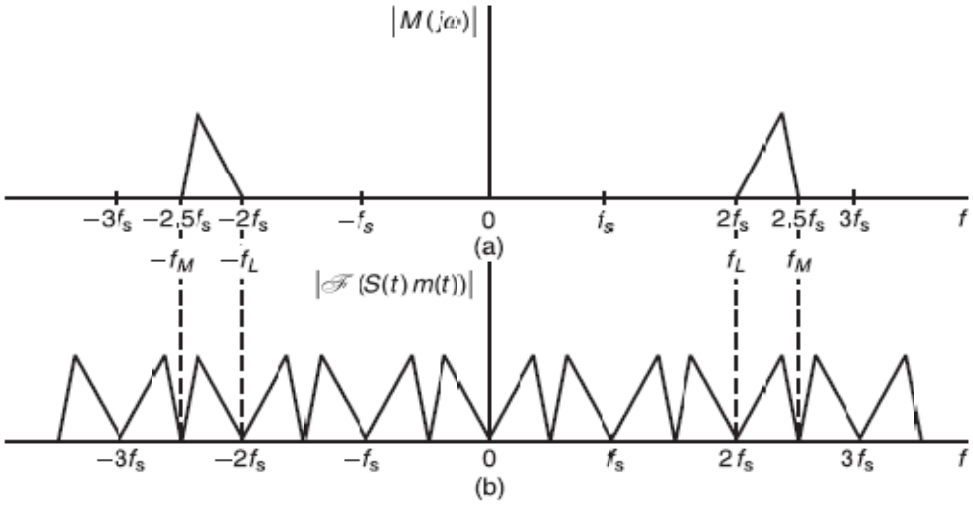


Fig. 1.2 Spectrum of bandpass and its sampled version signal

In fig. 1.2 the spectrum of g(t) extends over the first half of the frequency interval between harmonics of the sampling frequency, that is, from $2f_s$ to $2.5f_s$. As a result, there is no spectrum overlap, and signal recovery is possible. It may also be seen from the figure that if the spectral range of g(t) extends over the second half of the interval from 2.5 f_s to $3f_s$, there would similarly be no overlap. Suppose, however that the spectrum of g(t) were confined neither to the first half nor to the second half of the interval between sampling frequency harmonics. In such a case, there would be overlap between the spectrum patterns, and signal recovery would not be possible. Hence the minimum sampling frequency allowable is $f_s=2(f_M - f_L)$ provided that either f_M or f_L is a harmonic of f_s.

If neither f_M nor f_L is a harmonic of f_S, a more general analysis is required. In fig 1.3a, we have reproduced the spectral pattern of fig 1.2. The positive frequency part and negative frequency part of the spectrum are called PS and NS respectively. Let us , for simplicity, consider separately PS and NS and the manner in which they are shifted due to the sampling and let us consider initially what constraints must be imposed so that we cause no overlay over, say, PS. The product of g(t) and the dc component of the sampling waveform leaves PS unmoved, which will be considered to reproduce the original signal. If we select the minimum value of $f_s=2(f_m - f_L) = 2B$, then the shifted Ps patterns will not overlap

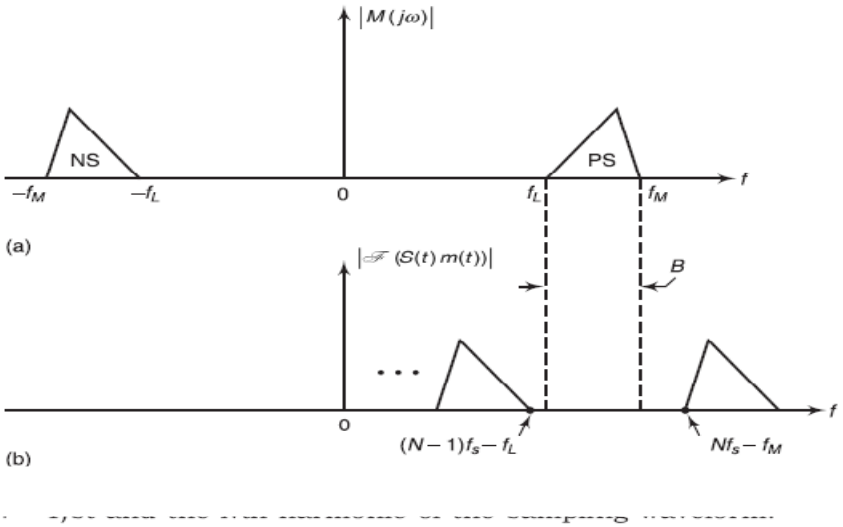


Fig. 1.3 (a) Spectrum of the bandpass signal (b) Spectrum of NS shifted by the (N-1)st and the Nth harmonic of the sampling waveform.

PS. The NS will also generate a series of shifted patterns to the left and to the right. The left shiftings can not cause an overlap with unmoved PS. However, the right shifting of NS might cause an overlap and these right shifting of NS are the only possible source of such overlap over PS. Shown in fig. 1.3b, are the right shifted patterns of NS due to the (N-1)th and Nth harmonics of the sampling waveform. It is clear that to avoid overlap it is necessary that,

$$(N-1)f_s - f_L \le f_L \qquad (1.4a)$$

and,

$$Nf_s - f_M \ge f_M \qquad (1.4b)$$

So that, with $B = f_M - f_L$, we have

$$(N - 1)f_s \leq 2(f_M - B) \quad (1.4c)$$

and, $$Nf_s \geq 2f_M \quad (1.4d)$$

If we let $k = f_M/B$, eqn. (1.4c) & (1.4d) become

$$f_S \leq 2B(K-1)/(N-1) \quad (1.4e)$$

and, $$fs \leq 2B(K/N) \quad (1.4f)$$

In which $k \geq N$, since $fs \geq 2B$. Eqn. (1.4e) and (1.4f) establish the constraint which must be observed to avoid an overlap on PS. It is clear from the symmetry of the initial spectrum and the symmetry of the shiftings required that this same constraint assumes the there will be no overlap on NS. Eqn.(1.4e) and (1.4f) has been plotted in fig. 1.4 for several values of N.

Let us take a case where f_L=2.5 KHz and f_M=3.5 KHz. So, B=1 KHz and $K=f_M/$ = 3.5. On the plot of fig. 1.4 line for k=3.5 has been erected vertically. For this value of k if f_s = 2B, then overlapping occurs. If f_s is increased in the range of 3.5 to 5 KHz, then no overlap occurs corresponding to N=2. If f_s is 7B or more then no overlap occurs.

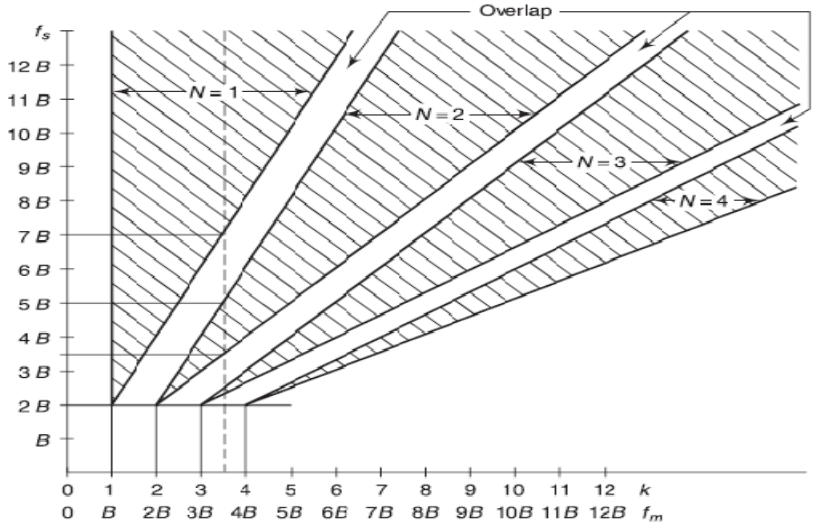


Fig. 1.4 The shaded region are the regions where the constraints eqn. (1.4e) and (1.4f) are satisfied.

From this discussion, we can write bandpass sampling theorem as follows---A bandpass signal with highest frequency f_H and bandwidth B, can be recovered from its samples through bandpasss filtering by sampling it with frequency f_s=2 f_H/k, where k is the largest integer not exceeding f_H/B. All frequencies higher than f_s but below $2f_H$(lower limit from low pass sampling theorem) may or may not be useful for bandpass sampling depending on overlap of shifted spectrums.

m(t) – low pass signal band limits to f_M.
s(t) – impulse train

$s(t) = \Delta t/Ts + 2.\ \Delta t/Ts(\cos 2\pi t/Ts + \cos(2.2\ \pi t/T_s) + \cos(3.2\ \pi t/T_s) + \ldots\ldots)$
$= \Delta t.fs + 2.\ \Delta t.fs(\cos 2\pi.fs.t + \cos(2.2\ \pi.fs.t) + \cos(3.2\ \pi.fs.t) + \ldots\ldots)$ (1.4g)

Product of m(t) and s(t) si the sampled m(t) ie, $m_s(t)$
$m_s(t) = m(t).s(t)$
$= \Delta t/Ts.m(t) + \Delta t/Ts[2.m(t)\cos 2\pi.fs.t + 2.m(t).\cos(2\pi.2.fs.t) + 2.m(t).\cos(2\ \pi.3.fs.t) + \ldots\ldots]$ (1.4h)

By using a low pass filter(ideal) with cut-off frequency at f_m then $\Delta t/Ts.m(t)$ will be passed so the m(t) can be recovered from the sample.

Band pass m(t) with lower frequency 'f_L' & upper frequency 'f_H', $f_H - f_L = B$. The minimum sampling frequency allowable is $f_s = 2(f_H - f_L)$ provided that either f_H or f_L is a harmonic of f_s.

A bandpass signal with highest frequency f_H and bandwidth B, can be recovered from its samples through bandpass filtering by sampling it with frequency $f_s = 2.f_H/k$, where k is the largest integer not exceeding f_H/B. All frequencies higher than f_s but below $2.f_H$(lower limit from low pass sampling theorem) may or may not be useful for bandpass sampling depending on overlap of shifted spectrum.

Eg. Let us say, f_L=2.5 KHz and f_H=3.5 KHz.
So, B=1 KHz, $k=f_M / B$ =3.5.
Selecting fs = 2B = 2 KHz cause overlap.
If k is taken as 3 then f_s = 2*3.5 kHz/3 = 7/3 kHz cause no overlap.
If k is taken as 2 then f_s = 2*3.5 KHz/2 = 3.5 KHz cause no overlap.

- **Aliasing Effect:-**

From the spectrum of $G_s(f)$ we can filter out one of the spectrum, say $-W<f<W$, using a low pass filter and can reconstruct the time domain representation of it after doing inverse fourier transform of the spectrum. This is possible only when $f_s >= 2W$.
But when $f_s < 2W$, ie, $T_s > 1/2W$, then there will be overlap of adjacent spectrums. Here high frequency part of 1st spectrum interfere with low frequency part of 2nd spectrum. This phenomenon is the aliasing effect. In such a case the original signal g(t) cannot be recovered exactly from its sampled values $g_s(t)$.

➢ Signal Reconstruction :

The process of reconstructing a continuous time signal g(t)[bandlimited to W Hz] from its samples is also known as interpolation. This is done by passing the sampled signal through an ideal low pass filter of bandwidth W Hz. As seen from eqn. 1.4, the sampled signal contains a component $1/T_s\ G(f)$, and to recover G(f)[or g(t)], the sampled signal must be passed through on ideal low-pass filter of bandwidth W hz and gain T_s.

Thus the reconstruction(or interpolating) filter transfer function is,

$$H(f) = T_s \, \text{rect}(f/2W) \tag{1.5}$$

The interpolation process here is expressed in the frequency domain as a filtering operation.

Let the signal interpolating (reconstruction) filter impulse response be h(t). Thus, if we were to pass the sampled signal $g_r(t)$ through this filter, its response would be g(t).

Let us now consider a very simple interpolating filter whose impulse response is $\text{rect}(t/T_s)$, as shown in fig. 1.5. This is a gate pulse of unit height, cantered at the origin, and of width T_s(the sampling interval). Each sample in $g_\delta(t)$, being an impulse generates a gate pulse of the height equal to the strength of the sample. For instance the kth sample is an impulse of strength $g(kT_s)$ located at $t=kT_s$, and can be expressed as $g(kT_s)\ \delta(t-kT_s)$. When this impulse passes thorugh the filter, it generates and ouput of $g(kT_s)\ \text{rect}(t/T_s)$. This is a gate pulse of height $g(kT_s)$, centred at $t=kT_s$(shown shaded in fig. 1.5).

Each sample in $g_\delta(t)$ will generate a corresponding gate pulse resulting in an output,

$$y(t) = \sum_k g(k..Ts)\ \text{rect}\left(\frac{t}{Ts}\right) \tag{1.6}$$

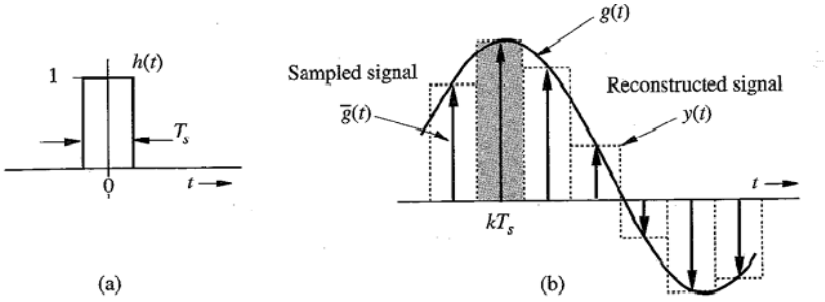


Fig. 1.5 Simple interpolation using zero-order hold circuit

The filter output is a staircase approximation of g(t), shown dotted in fig. 1.5b. This filter thus provides a crude form of interpolation.

The transfer function of this filter H(f) is the fourier transform of the impulse response $\text{rect}(t/T_s)$. Assuming the Nyquist sampling rate, ie, $T_s = 1/2W$,

$$W(t) = \text{rec}(t/Ts) = \text{rect}(2Wt)$$

and, $$H(f) = T_s.\text{sinc}(\pi.f.Ts) = 1/(2W).\text{sinc}(\pi f/2W) \tag{1.7}$$

The amplitude response $|H(f)|$ for this filter shown in fig. 1.6, explains the reason for the crudeness of this interpolation. This filter is also known as the zero order hold filter, is a poor approximation of the ideal low pass filter(as shown double shaded in fig. 1.6).

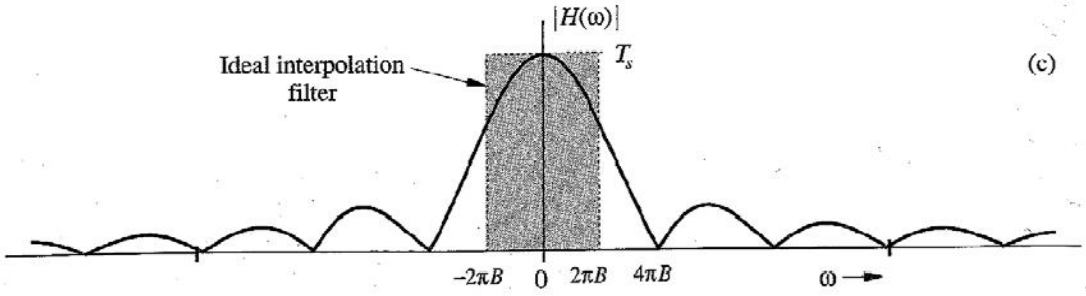


Fig. 1.6 Amplitude response of interpolation filter.

We can improve on the zero order hold filter by using the first order hold filter, which results in a linear interpolation instead of the staircase interpolation. The linear interpolator, whose impulse response is a triangular pulse $\Delta(t/2T_s)$, results in an interpolation in which successive sample tops are connected by straight line segments. The ideal interpolation filter transfer function found in eqn. 1.5 is shown in fig. 1.7a. The impulse response of this filter, the inverse fourier transform of H(f) is,

$$h(t) = 2.W.Ts.sinc(Wt),$$

Assuming the Nyquist sampling rate, ie, $2WT_s = 1$, then

$$h(t) = sinc(Wt) \tag{1.8}$$

This h(t) is shown in fig. 1.7b.

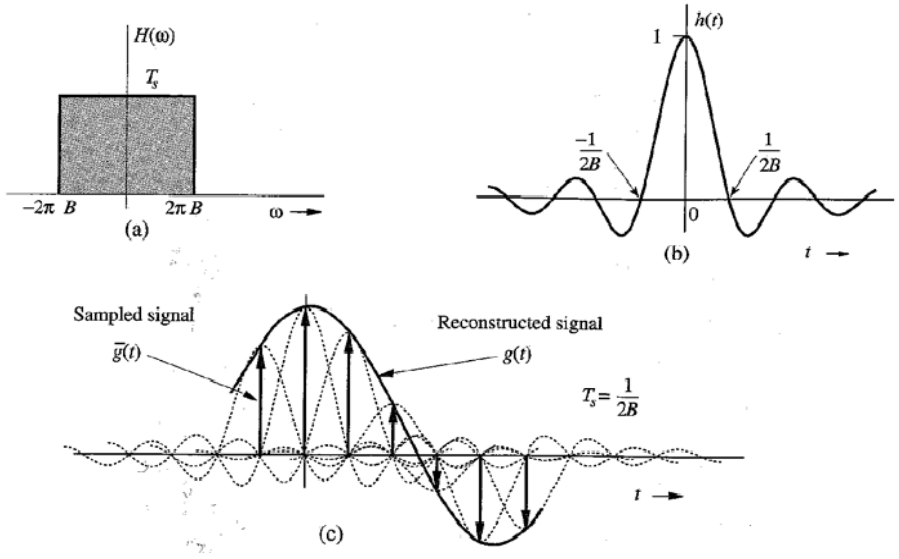


Fig. 1.7 Ideal interpolation.

The very interesting fact we observe is that, h(t) = 0 at all Nyquist sampling instants(t = ±n/2W) except at t=0. When the sampled signal $g_\delta(t)$ is applied at the input of this filter, the output is g(t). Each sample in $g_\delta(t)$, being an impulse, generates a sine pulse of height equal to the strength of the sample, as shown fig. 1.7c.

The process is identical to that shown in fig. 1.7b, except that h(t) is a sine pulse instead of gate pulse. Addition of the sine pulses generated by all the samples results in g(t). The k^{th} sample of the input $g_\delta(t)$ is the impulse $g(kT_s)\delta(t-kT_s)$; the filter output of this impulse is $g(kT_s)h(t-kT_s)$. Hence, the filter output to $g_\delta(t)$, which is g(t), can now be expressed as a sum.

$$g(t) = \sum_k g(k.Ts)\, h(t — KTs)$$

$$= \sum_k g(k.Ts)\, \text{sinc}[W(t — KTs)] \quad (1.9a)$$

$$= \sum_k g(k.Ts)\, \text{sinc}[Wt — K/2] \quad (1.9b)$$

Eqn. 1.9 is the interpolation formula, which yields values of g(t) between samples as a weighted sum of all the sample values.

➢ Practical Difficulties:

If a signal is sampled at the Nyquist rate $f_s = 2W$ hz, the spectrum $G_\delta(f)$ without any gap between successive cycles.. To recover g(t) from $g_\delta(t)$, we need to pass the sampled signal $g_\delta(t)$ through an ideal low pass filter. Such filter is unrealizable; it can be closely approximated only with infinite time delay in the response. This means that we can recover the signal g(t) from its samples with infinite time delay. A practical solution to this problem is to sample the signal at a rate higher h=than the Nyquist rate($f_s > 2W$). This yields $G_\delta(f)$, consisting of repetition of G(t) with a finite band gap between successive cycles. We can now recover G(g) from $G_\delta(f)$ from $G_\delta(f)$ using a low pass filter with a gradual cut-off characteristics. But even in this case, the filter gain is required to be zero beyond the first cycle of G(f). By Paley-Wiener criterion, it is also impossible to realize even this filter. The only advantage in this case is that the required filter can be closely approximated with a smaller time delay.

This indicated that it is impossible in practice to recover a band limited signal $g_\delta(t)$ exactly from its samples even if sampling rate is higher than the Nyquist rate. However as the sampling rate increases, the recovered signal approaches the desired signal more closely.

➢ **The Treachery of Aliasing:**

There is another fundamental practical difficulty in reconstructing a signal from its samples. The sampling theorem was proved on the assumption that the signal g(t) is bandlimited. All practical signals are time limited, ie, they are of finite duration width. A signal cannot be time-limited and band-limited simultaneously. If a signal is time limited, it cannot be band limited and vice-versa(but it can be simultaneously non time limited and non band limited). This means that all practical signals which are time limited are non band limited; they have infinite bandwidth and the spectrum $G_\delta(f)$ consists of overlapping cycles of G(f) repeating every f_s hz(the sampling frequency) as shown in fig. 1.8.

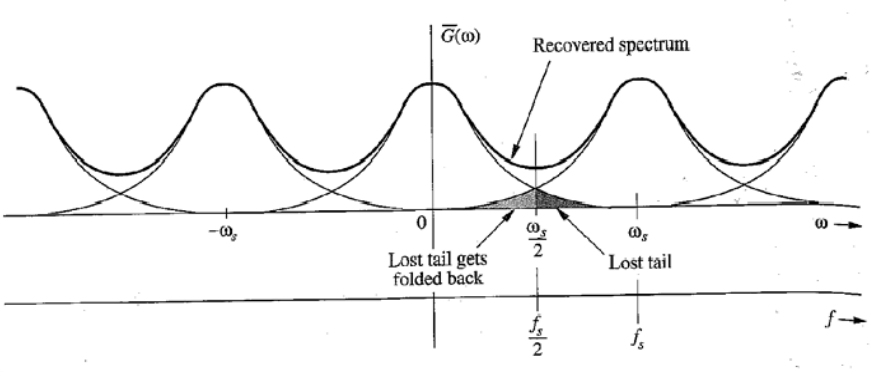


Fig. 1.8 Aliasing effect

Because of the overlapping tails, $G_\delta(f)$ no longer has complete information about G(f) and it is no longer possible even theoretically to recover g(t) from the sampled signal $g_\delta(t)$. If the sampled signal is passed through and ideal low pass filter the output is not G(f) but a version of G(f) distorted as a result of two separate causes:

1. The loss of the tail of G(f) beyond $|f| > f_s/2$ Hz.
2. The reappearance of this tail inverted or folded onto the spectrum.

The spectra cross at frequency $f_s/2 = 1/2T_s$ Hz, is called the folding frequency. The spectrum, therefore, folds onto itself at the folding frequency. In fig. 1.8, the components of frequencies above $f_s/2$ reappear as components of frequencies below $f_s/2$. This tail inversion, known as spectral folding or aliasing is shown shaded in fig. 1.8. In this process of aliasing, we are not only losing all the components of frequencies above $f_s/2$ Hz, but these very components reappear(aliased) as lower frequency components also as in fig. 1.8.

➢ A Solution: The Antialiasing Filter

The potential defectors are all the frequency components beyond $f_s/2 = 1/2T_s$ Hz. We should eliminate (suppress) these components from g(t) before sampling g(t). This way, we lose only the components beyond the folding frequency $f_s/2$ Hz. These components now cannot reappear to corrupt the components with frequencies below the folding frequency. This suppression of higher frequencies can be accomplished by an ideal low pass filtr of bandwidth $f_s/2$ hz. This filter is called the antialiasing filter. This antialiasing operation must be performed before the signal is sampled.

The antialiasing filter, being an ideal filter, is unrealizable. In practice we use a steep cut off filter which leaves a sharply attenuated residual spectrum beyond the folding frequency $f_\delta/2$.

Even using antialiasing filter, the original signal may not be recovered if $T_s > 1/2W$, ie, $f_s < 2W$. For this case also aliasing will occur. To avoid this sampling frequency f_s should be always greater than or atleast equal to 2W, where W is the highest frequency component available in information signal.

➢ Some Applications of the Sampling Theorem:

In the field of digital communication the transmission of a continuous time message is replaced by the transmission of a sequence of numbers. These open doors to many new techniques of communicating continuous time signals by pulse trains. The continuous time signal g(t) is sampled, and samples values are used to modify certain parameters of a periodic pulse train. As per these parameters, we have pulse amplitude modulation (PAM), pulse width modulation (PWM) and pulse position modulation (PPM). In all these cases instead of transmitting g(t), we transmit the corresponding pulse modulated signal. One advantage of using pulse modulation is that it permits the simultaneous transmission of several signals on a time sharing basis-time division multiplexing (TDM) which is the dual of FDM.

➢ Pulse Amplitude Modulation(PAM) :

In PAM, the amplitude of regularly spaced rectangular pulses vary with the instantaneous sample value of a continuous message signal in one to one fashion.

$$V_{PAM}(t) = \Sigma_{n=-\infty}^{\infty k}[1 + Ka.\, g(n.\, Ts)]\delta(t - n.\, Ts)$$

Where $g(nT_S)$ represents the nth sample of the message signal g(t), T_s is the sampling time, k_a is a constant called the amplitude sensitivity(or modulation index of PAM) and $\delta_{TS}(t)$ demotes the pulse train. 'k_a' is chosen so as to maintain a single polarity, ie, $\{1+k_a g(nT_s)\} > 0$ for all values of $g(nT_s)$.

Different forms of pulse analog modulation (PAM, PWM & PPM) are illustrated below:-

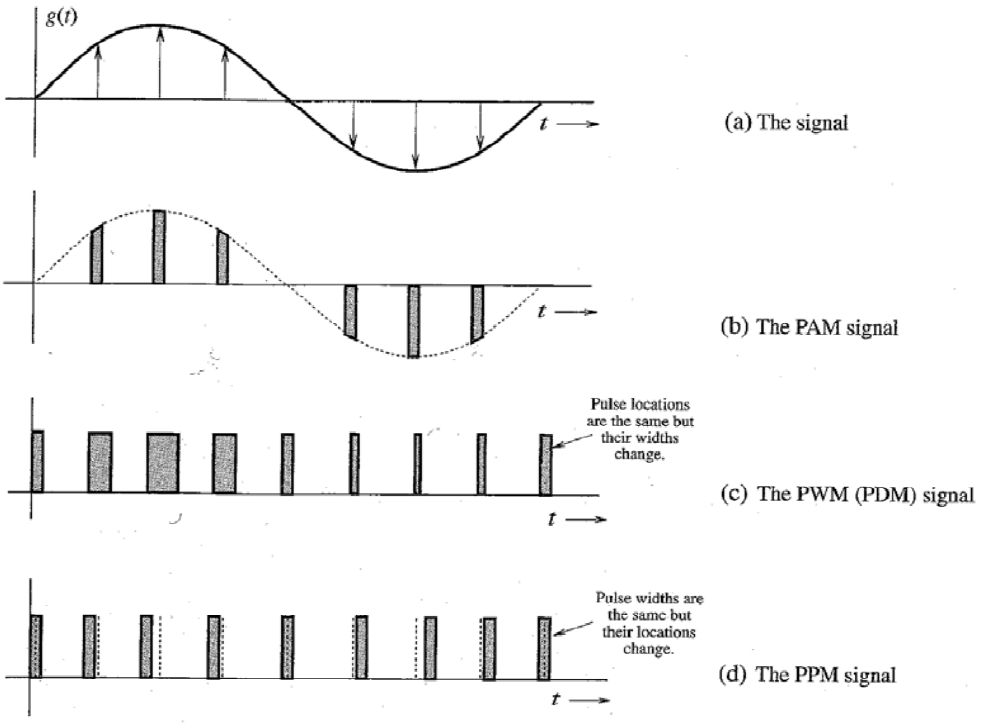


Fig. 1.9 Pulse modulated signals.

Transmission BW in PAM

We know $\tau << T_s \leq 1/2W$

Considering 'ON' and 'OFF' time of PAM it is velar the maximum frequency of PAM is $f_{max} = 1/2\tau$.

So transmission BW $\geq f_{max} = 1/2\tau >> W$.

Noise performance of PAM is never better than the baseband signal transmission.

However we need PAM for message processing for a TDM system, from which PCM can be easily generated or other form of pulse modulation can be generated.

Be it single or multi user system the detection should be done in synchronism. So synchronization between transmitter and receiver is an important requirement.

➢ Pusle Width Modulation(PWM):

In pulse width modulation, the instantaneous sample values of the message signal are used to vary the duration of the individual pulses. This form of modulation is also referred to as pulse duration modulation (PDM) or pulse length modulation (PLM).

Here the modulating wave may vary the time of occurrence of leading edge, the trailing edge or both edges of the pulse.

Disadvantage – In PWM, long pulses (more width) expand considerable power during the pulse transmission while bearing no additional information.

$$V_{PWM} = P(t - n.T_s) = \delta(t - n.Ts) \quad \text{for } nTs \in t \in (nTs + kn.g(nTs))$$
$$= 0 \quad \text{for } [nT_s + k_w.g(nT_S)] \leq t \leq (n+1)T_s$$

➢ Generation of PWM and PPM waves:

The figure below depicts the generation of PWM and PPM waves. Hence for the PWM wave the trailing edge is varied according to the sample value of the message.

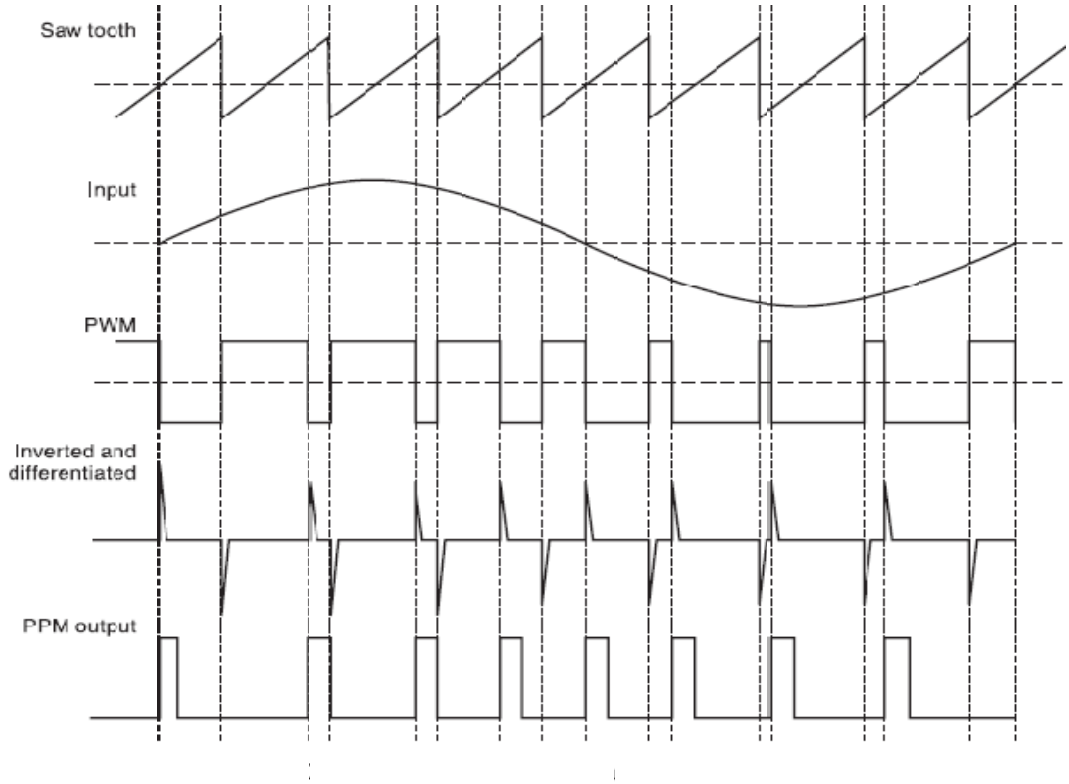


Fig. 1.10 Principle of PWM and PPM generation.

The saw tooth generator generates the sawtooth signal of frequency f_s ($f_s = 1/T_s$). If sawtooth waveform is reversed, then leading edge of the pulse will be varied with samples of the signal and if the sawtooth waveform is replaced by a triangular waveform then both the edged will vary according to samples.

PPM waveform is generated when PWM wave is used as the trigger input to a monostable multivibrator. The monostable multivibrator is triggered on the falling (trailing edge) of PWM. The output of monostable is then switches to positive saturation value and remain there for a fixed period and then goes low. Thus a pulse is generated which occurs at a time which occurs at a time which depends upon the amplitude of the sampled value.

Demodulation of PWM waves

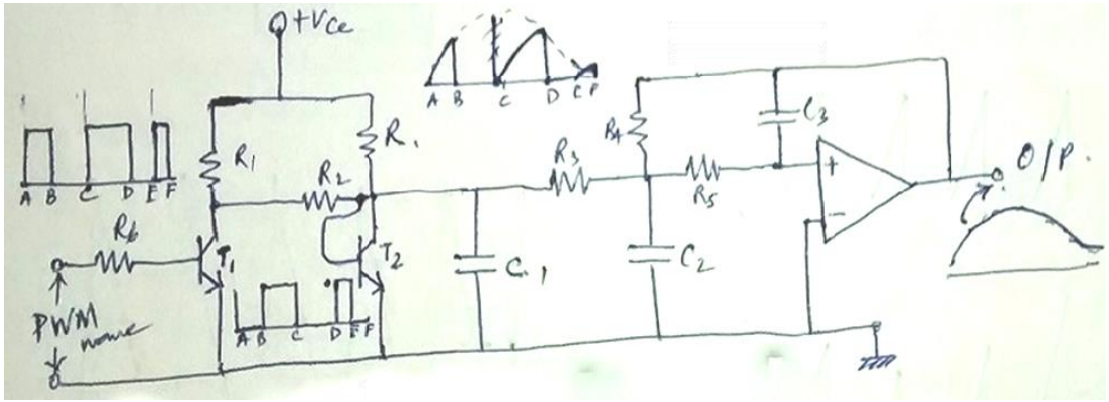


Fig. 1.10a A PWM demodulator circuit.

Here transistor T1 acts as an inverter. Hence when transfer is off capacitor C1 will chase through R as when it is 'on' C1 discharges quickly through T1 as the resistance in the path is very small. This produces a sawtooth wave at the output of T2. This sawtooth wave when passed through an op-amp with 2nd order LPF produces the desired wave at the output.

Demodulation of PPM waves:

Since in PPM the gaps in between pulses contains information, so during the gaps, say OA, BC and DE the transfer T remains off and capacitor the capacitor C gets charged. The voltage across the capacitor depends on time of charging as the value of R and C. Rest of the operation is same as above.

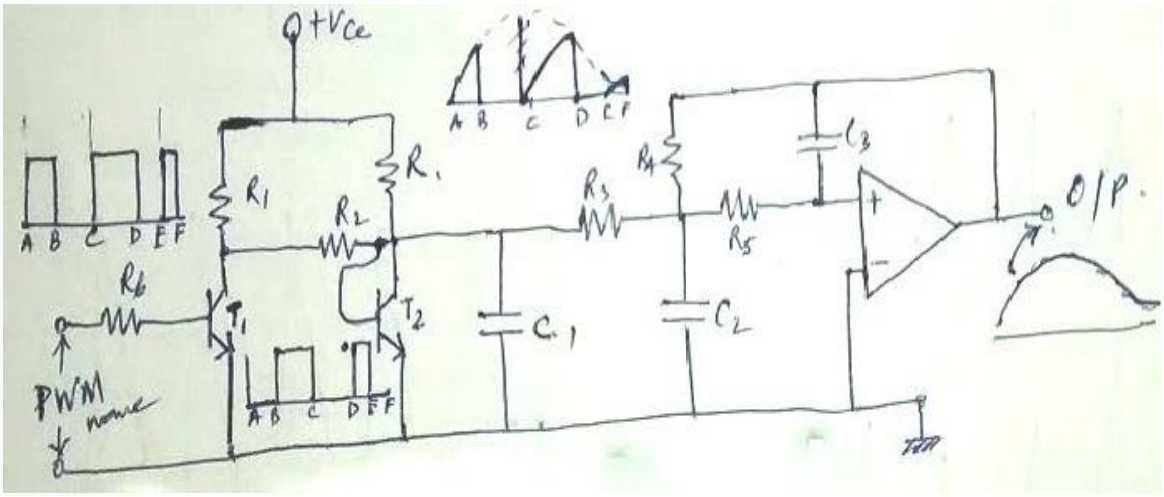


Fig. 1.10b A PPM demoddator circuit.

UNIT 6- PULSE DIGITAL MODULATION

PCM is the most useful and widely used of all the pulse modulations mentioned. Basically, PCM is a method of converting an analog signal into a digital signal (A/D conversion).An analog signal is characterized by the fact that its amplitude can take on any value over a continuous range. This means that it can take on an infinite number of values. On the other hand, digital signal amplitude can take on only a finite number of values. An analog signal can be converted into a digital signal by means of sampling and quantizing, that is, rounding off its value to one of the closest permissible numbers (or quantized levels) as shown in fig 2.1.

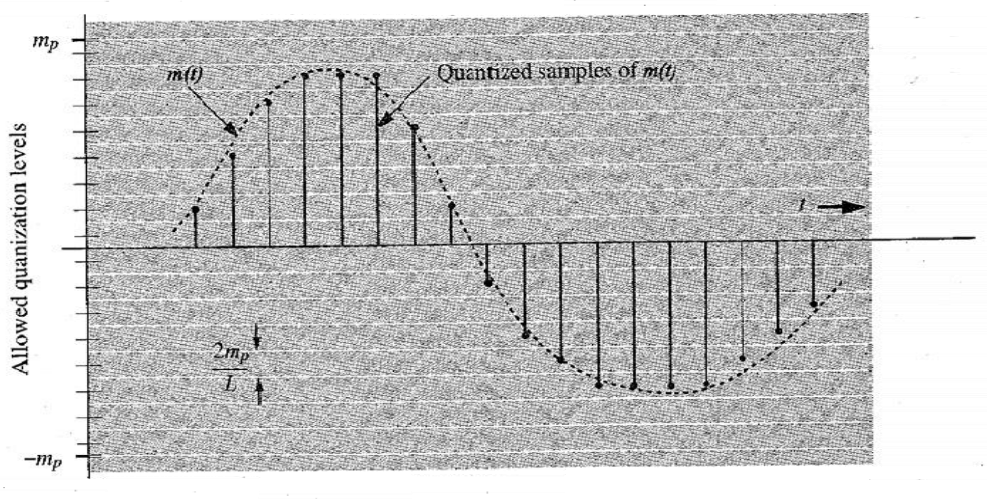


Fig. 2.1 Quantization of a sampled analog signal.

Quantization is of two types:--uniform and non-uniform quantization.

- ## Uniform Quantization :--

  Amplitude quantizing is the task of mapping samples of a continuous amplitude waveform to a finite set of amplitudes. The hardware that performs the mapping is the analog-to-digital converter (ADC or A-to-D). The amplitude quantizing occurs after the sample-and-hold operation. The simplest quantizer to visualize performs an instantaneous mapping from each continuous input sample level to one of the preassigned equally spaced output levels. Quantizers that exhibit equally spaced increments between possible quantized output levels are called uniform or linear quantizers.

  Possible instantaneous input-output characteristics are easily visualized by a simple staircase graph consisting of risers and treads of the types shown in Fig 2.2. Fig 2.2 a, b, and d show quantizers with uniform quantizing steps, while fig 2.2c is a quantizer with nonuniform quantizing steps.

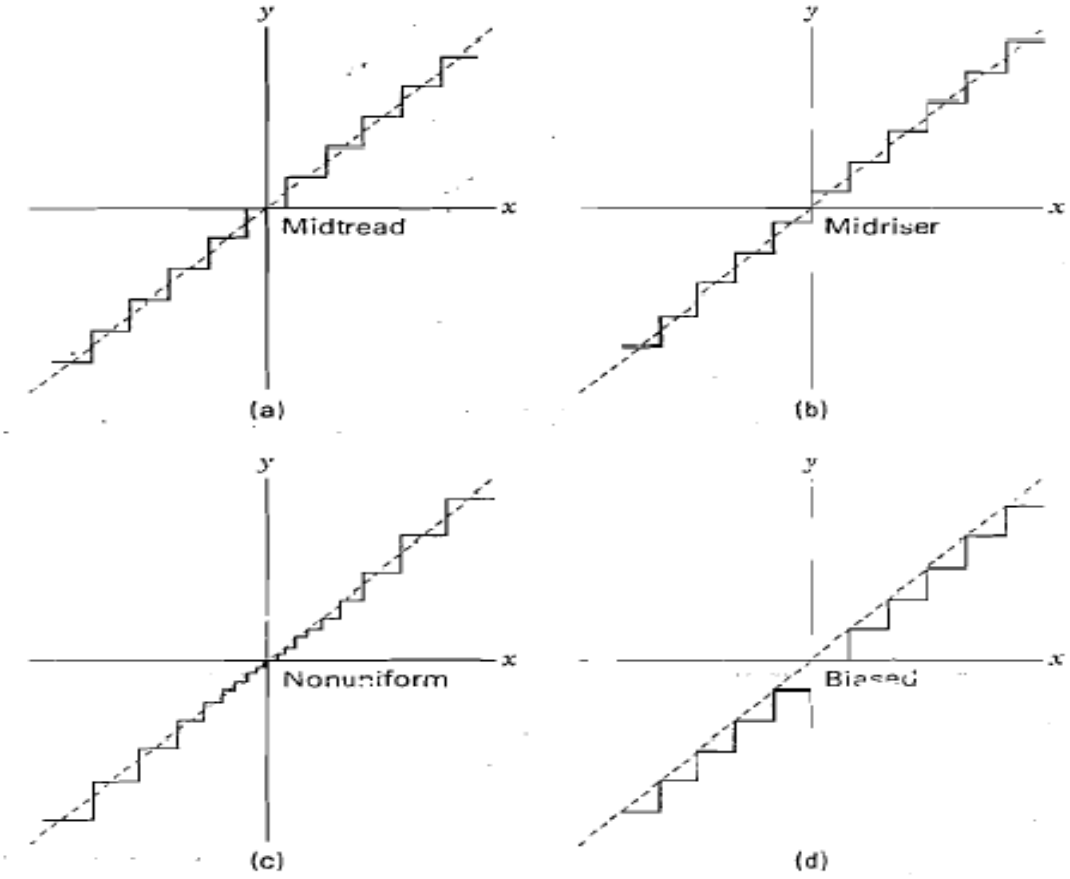


Fig. 2.2 Various quantizers transfer functions.

- Non Uniform Quantization:

For many classes of signals the uniform quantization is not efficient, for example, in speech communication it is found(statistically) that smaller amplitudes predominate in speech and that larger amplitudes are relatively rare. The uniform quantizing scheme is thus wasteful for speech signals; many of the quantizing levels are rarely used. An efficient scheme is to employ a non uniform quantizing method in which smaller steps for small amplitudes are used.

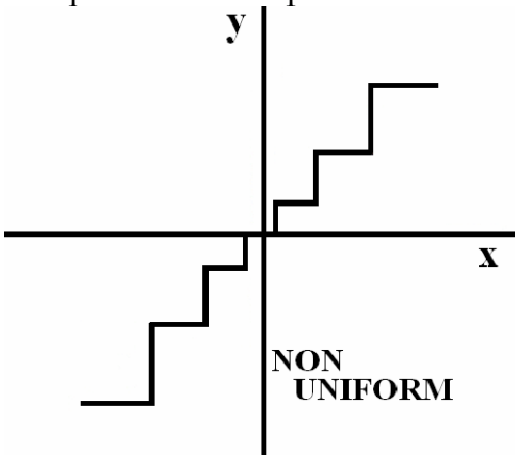


Fig. 2.3. Non-uniform quantization

The same result can be achieved by first compressing the signal samples and then using a uniform quantizing. The input-output characteristics of a compressor are shown in below fig. 2.4

The same result can be achieved by first compressing the signal samples and then using a uniform quantizing. The input output characteristics of a compressor are shown in fig. The horizontal axis is the normalized input signal (ie, g/g_p), and the vertical axis is the output signal y. The compressor maps input signal increment Δg, into larger increment Δy for small signal input signals and small increments for larger input signals. Hence, by applying the compressed signals to a uniform quantizer a given interval Δg contains a larger no. of steps (or smaller step-size) when g is small.

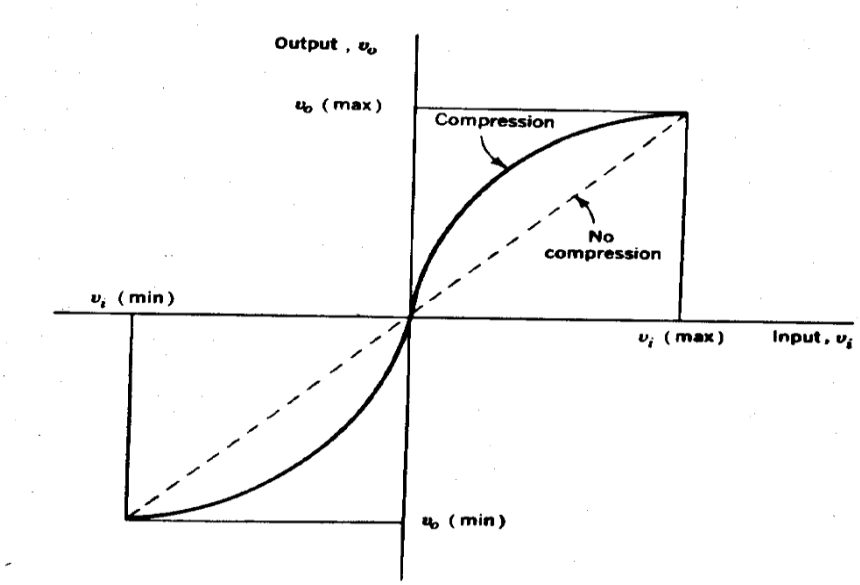


Fig. 2.4 Characteristics of Compressor.

A particular form of compression law that is used in practice (in North America and Japan) in the so called μ law (μ law compressor), defined by

$y = \ln(1 + \mu|g/g_p|)/\ln(1 + \mu).\text{sgn}(g)$ for $|g/g_p| \le 1$ where, μ is a +ve constant and sgn(g) is a signum function.

Another compression law popular in Europe is the so A-law, defined by,

$$y = A/(1+\ln A).(g/g_p) \quad \text{for } 0 \le g/g_p \le 1/A$$
$$= (1+\ln A|g/g_p|/(1+\ln A)).\text{sgn}(g) \quad \text{for } 1/A \le |g/g_p| \le 1 \qquad (2.1)$$

The values of μ & A are selected to obtain a nearby constant output signal to quantizing noise ratio over an input signal power dynamic range of 40 dB.
To restore the signal samples to their correct relative level, an expander with a characteristic complementary to that of compressor is used in the receiver. The combination of compression and expansion is called companding.

- **Encoding:-**

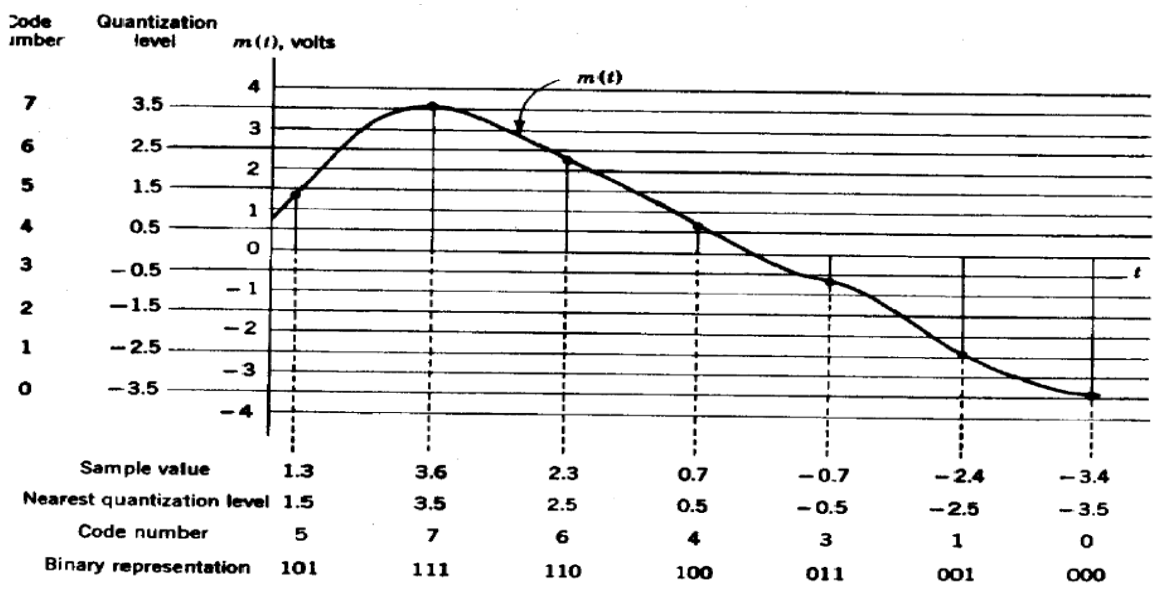


Fig. 2.5 Representation of each sample by its quantized value and binary representation.

A signal g(t) bandlimited to B hz is sampled by a periodic pulse train $P_{Ts}(t)$ made up of a rectangular pulse of width 1/8B seconds (cantered at origin), amplitude 1 unit repeating at the Nyquist rate(2B pulses per second. Show that the sampled signal is given by,

$$g^{-}(t) = ¼.g(t) + \sum_{n=1}^{œ} (\frac{2}{nn}.\sin(nn/4)g(t)\cos n.ws.t) \qquad (2.2)$$

➢ Quantizing Noise or Quantizing Error :

We assume that the amplitude of g(t) is confined to the range($-g_p$, g_p). This range is divided into L no. of equal segments. Each segment is having step size Δ, given by,

$$\Delta = 2.g_p/L \qquad (2.3)$$

A sample amplitude value is approximated by the mid-points of the interval in which it lies. The input-output characteristic of a midrise uniform quantizer is shown in fig.

The difference between the input and output signals of the quantizer becomes the quantizing error or quantizing noise.

It is apparent that with a random input signal, the quantizing error 'q_e' varies randomly within the interval,

$$-\Delta/2 \le q_e \le \Delta/2 \qquad (2.4)$$

Assuming that the error is equally likely to lie anywhere in the range $(-\Delta/2, \Delta/2)$, the mean square quantizing error $\langle q_e^2 \rangle$ in given by,

$$\langle q_e^2 \rangle = 1/\Delta \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} q_e^2 dq_e = \Delta^2/12 \qquad (2.5)$$

Substituting eqn.(2.3) in eqn.(2.5) we get,

$$\langle q_e^2 \rangle = g_p^2/(3L^2)$$

$$S_i = \langle g^2(t) \rangle = \int_{-g_p}^{g_p} g^2(t) \cdot \frac{1}{2} g.dg = g_p^2/3$$

➢ Transmission Bandwidth and the output SNR :

For binary PCM, we assign a distinct group of 'n' binary digits(bits) to each of the L quantization levels. Because a sequence of n binary digits can be arranged in 2^n distinct patterns,

$$L = 2^n \text{ or } n = \log_2 L \qquad (2.6)$$

Each quantized sample is thus, encoded into 'n' bits. Because a signal g(t) bandlimited to W Hz requires a minimum of 2W samples second, we require a total of 2nW bits per second(bps), ie, 2nW pieces of information per second. Because a unit bandwidth (1 Hz) can transmit a maximum of two pieces of information per second, we require a minimum channel of bandwidth B_T Hz, given by,

$$B_T = n.W \quad \text{Hz} \qquad (2.7)$$

This is the theoretical minimum transmission bandwidth required to transmit the PCM signal. We shall see that for practical reasons we may use transmission bandwidth higher than as in eqn.(2.7).

$$\text{Quantizing Noise} = N_o = \langle q_e^2 \rangle = g_p^2/(3.L^2) \qquad (2.8)$$

Assuming the pulse detection error at the receiver is negligible, the reconstructed signal $\hat{g}(t)$ at the receiver output is,

$$\hat{g}(t) = g(t) + q_e(t) \qquad (2.9)$$

The desired signal at the output is g(t), and the (quantizing) noise is $q_e(t)$. Since the power of the message signal g(t) is $\langle g^2(t) \rangle$, then

$$S_0 = \langle g^2(t) \rangle \qquad (2.10)$$

So, $SNR = S_o/N_o = \langle g^2(t)\rangle / (g^2_p/(3.L^2)) = 3L^2 \langle g^2(t)\rangle / g^2_p$ (2.11)

$S_o/N_o(dB) = 10.\log(3L^2) \langle g^2(t)\rangle / g^2_p$

Signal to noise ration can be written as,

$$S_o/N_o = 3.2^{(2n)} \langle g^2(t)\rangle / g^2_p \quad (2.12)$$

$$= C(2)^{2n} \quad (2.13)$$

Where,

$C = 3.\langle g^2(t) \rangle / g^2_p$ (uncompressed case, as in eqn.(2.12))

$= 3/[\ln(1+\mu)]^2$ (compressed case)

For a μ-law compander, the output SNR is,

$S_o/N_o = 3.l^2/[\ln(1+\mu)]^2$ $\mu^2 >> g^2_p/\langle g^2(t)\rangle$

Substituting eqn.(2.7) in eqn.(2.12), we find

$$S_o/N_o = C(2)^{2.B_T/W} \quad (2.14)$$

From eqn.(2.14), it is observed that SNR increases almost exponentially with the transmission bandwidth B_T. This trade-off SNR with bandwidth is attractive and come close to the upper theoretical limit. A small increase in bandwidth yields a large benefit in terms of SNR. This trade relationship is clearly seen by rewriting eqn.(2.14) using decibel scale as,

S_0/N_0 (dB) $= 10.\log(S_o/N_o)$

$= 10\log(C2^{2n})$

$= 10\log C + 20\log 2$

$= (\alpha + 6n)$ dB (2.15)

Where, $\alpha = 10\log C$. This shows that increasing n by 1, quadruples the output SNR(6 dB increase). Thus if we increase 'n' from 8 to 9, the SNR quadruples, but the transmission bandwidth increases only from 32 to 36 Khz(an increase of only 12.5%). This shows that in PCM, SNR can be controlled by transmission bandwidth. We shall see later that frequency and phase modulation also do this. But it requires a doubling of the bandwidth to quadruple the SNR. In this respect, PCM is strikingly superior to FM or PM.

➢ Digital Multiplexer :--

This is a device which multiplexers or combines several low bit rate signals to form one high bit rate signal to be transmitted over a high frequency medium. Because of the medium is time shared by various incoming signals, this is a case of time-division

multiplexing (TDM. The signals from various incoming channels may be such diverse nature as digitized voice signal (PCM), a computer output, telemetry data, a digital facsimile and so on. The bit rates of the various tributaries (channels) need not be the same.

Multiplexing can be done on a bit-by-bit basis(known as bit or digit interleaving) or on a word-by-word basis(known as byte or word interleaving). The third category is interleaving channel having different bit rate.

T1 carrier system:-- The input to the (fast) 13-bit ADC comes from an analog multiplexer. The digital processor compresses the digital value according to μ-law.

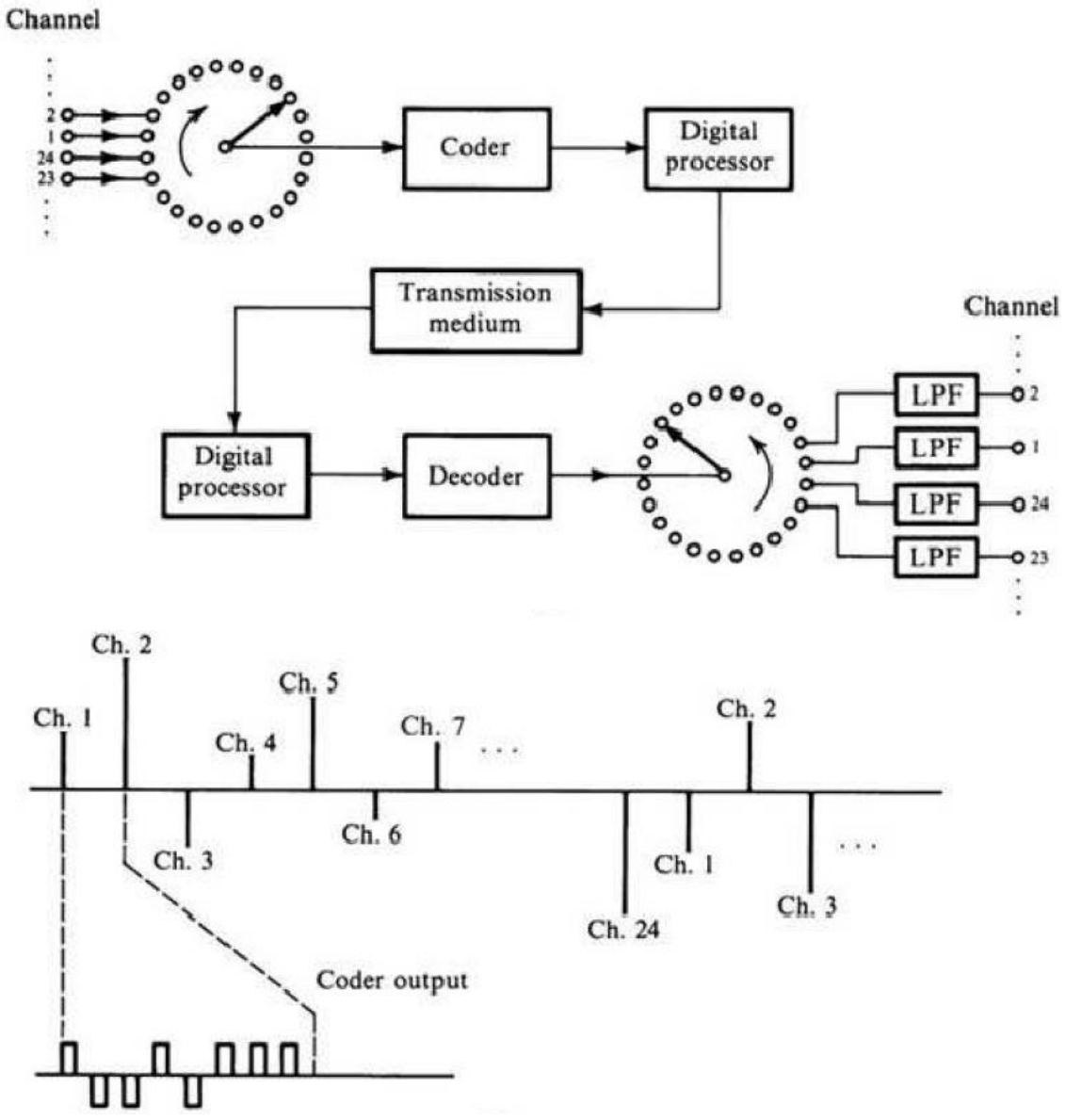


Fig. 2.6 T-1 carrier system.

The 8-bit compressed voice values are sent consecutively, MSB first. The samples of all 24 inputs comprise a frame. Most serial communications transmits data LSB first ("little endian").

➢ Synchronizing & Signalling :

Binary code words corresponding to samples of each of the 24 channels are multiplexed in a sequence as shown in fig 2.7. A segment containing one codeword (corresponding to one sample) from each of the 24 channels is called a frame. Each frame has 24x8 = 192 information bits. Because the sampling rate is 8000 samples per second, each frame takes 125 μs. At the receiver it is necessary to be sure where each frame begins in order to separate information bits separately. For this purpose, s framing bit is added at the beginning of each frame. This makes a total of 193 bits per frame. Framing bits are chosen so that a sequence of framing bits, one at the beginning of each frame, forms a special pattern that is unlikely to be formed in a speech channel.

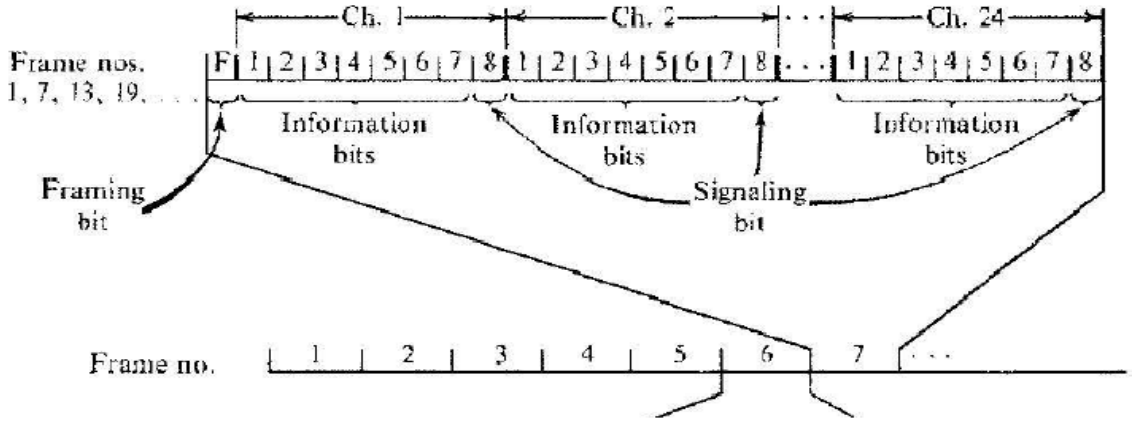


Fig. 2.7 T-1 frame.

The sequence formed by the first bit from each frame is examined by the logic of the receiving terminal. If this sequence does not follow the given coded pattern (framing bit pattern), then a synchronization lost is detected and the next position is examined to determine whether it is actually the framing bit. It takes about 0.4 to 6 ms to detect and about 50 ms (in the worst possible case) to reframe.

In addition to information and framing bits we need to transmit signalling bits corresponding to dialling pulses, as well as telephone on-hook/off-hook signals. When channels developed by this system are used to transmit signals between telephone switching systems, the switches must be able to communicate with each other to use the channels effectively. Since all eight bits are now used for transmission instead of the seven bits used in the earlier version, the signalling channel provided by the eighth bit is no longer available. Since only a rather low speed signalling channel is required, rather than create extra time slots for this information we use one information bit(the least significant bit) of every sixth sample of a signal to transmit this information. This means every sixth sample of each voice signal will have a possible error corresponding to the least significant digit. Every sixth frame, therefore, has 7x24 = 168 information bits, 24 signalling bits and 1 framing bit. In all the remaining frames, there are 192 information bits

and 1 framing bit. This technique is called 75/6 bit encoding and the signalling channel so derived is called robbed-bit signalling. The slight SNR degradation suffered by impairing one out of six frame is considered to be an acceptable penalty. The signalling bits for each signal occur at a rate of 8000/6 = 1333 bits/sec.

In such above case detection of boundary of frames is important. A new framing structure called the super frame was developed to take care of this. The framing bits are transmitted at the 8 kbps rate as before (earlier case) and occupy the first bit of each frame. The framing bits form a special pattern which repeats in twelve frames: 100011011100. The pattern thus allows the identification of frame boundaries as before, but also allows the determination of the locations of the sixth and twelfth frames within the superframe. Since two signalling frames are used so two specific job can be initiated. The odd numbered frames are used for frame and sample synchronization and the even numbered frames are used to identify the A & B channel signalling frames(frames 6 & 12).

A new superframe structure called the extended superframe (ESF) format was introduced during 1970s to take advantage of the reduced framing bandwidth requirement. An ESG is 24 frames in length and carries signalling bits in the eighth bit of each channel in frames 6, 12, 18 and 24. Sixteen state signalling is thus possible. Out of 24 framing bits 4^{th}, 8^{th}, 12^{th}, 16^{th}, 20^{th} and 24^{th}(2 kbps) are used for frame synchronization and have a bit sequence 001011. Framing bits 1, 5, 9, 13, 17 and 21(2 kbps) are for error detection code. 12 remaining bits are for management purpose and called as facility data link(FDL). The function of signalling is also the common channel interoffice signalling (CCIS).

Differential Pulse Code Modulation :

In analog messages we can make a good guess about a sample value from a knowledge of the past sample values. In other words, the sample values are not independent and there is a great deal of redundancy in the Nyquist samples. Proper exploitation of this redundancy leads to encoding a signal with a lesser number of bits. Consider a simple scheme where instead of transmitting the sample values, we transmit the difference between the successive sample values.

If g[k] is the kth sample instead of transmitting g[k], we transmit the difference d[k] = g[k] – g[k-1]. At the receiver, knowing the d[k] and the sample value g[k-1], we can construct g[k]. Thus form the knowledge of the difference d[k], we can reconstruct g[k] iteratively at he receiver. Now the difference between successive samples is generally much smaller than the sample values. Thus peak amplitude, g_p of the transmitted values is reduced considerably. Because the quantization interval $\Delta = g_p/L$, for a given L(or n) this reduces the quantization interval Δ. Thus, reducing the quantization noise which is given by $\Delta^2/12$.

This means that for a given n(or transmission bandwidth), we can increase the SNR or for a given SNR we can reduce n(or transmission bandwidth).

We can improve upon scheme by estimating the value of the kth sample g[k] from knowledge of the previous sample values. If this estimate is $\hat{g}[k]$, then we transmit the difference (prediction error) d[k] = g[k] – $\hat{g}[k]$. At the receiver also we determine the estimate $\hat{g}[k]$ from the previous sample values and then generate g[k] by adding the received d[k] to the estimate $\hat{g}[k]$. Thus we reconstruct the samples at the receiver iteratively. If our prediction is worthful the predicted value $\hat{g}[k]$will be close to g[k] and their difference (prediction error) d[k] will be even smaller than the difference between the successive samples. Consequently this scheme known as the differential PCM(DPCM) is superior to that described in the previous paragraph which is a special case of DPCM, where the estimate of a sample value is taken as the previous sample value, ie, $\hat{g}[k]$=g[k-1].

Consider for example a signal g(t) which has derivative of all orders at 't'. Using Taylor series for this signal, we can express $g(t+T_s)$ s,

$$g(t+T_s) = g(t) + T_s.\dot{g}(t) + T^2_s/2!\,\ddot{g}(t) + \ldots\ldots \quad (2.16)$$

$$= g(t) + T_s.\dot{g}(t) \qquad \text{for small } T_s. \quad (2.17)$$

So from eqn.(2.16) it is clear a future signal can be predicted from the present signal and its all derivatives. Even if we know the first derivative we can predict the approximated signal.

Let us denote the kth sample of g(t0 by g[k], ie, $g[kT_s]$ =g[k] and $g(kT_s \pm T_s) = g[k \pm 1]$ and so on. Setting $t=kT_s$ in eqn.(2.17) and recognizing $g(kT_s) \approx [g(kT_s) - g(kT_s - T_s)]/T_s$.

We obtain,

$$g[k+1] \approx g[k] + T_s[\{g[k] - g[k-1\}/T_s]]$$

$$= 2g[k] - g[k-1] \quad (2.18)$$

This shows that we can find a crude prediction of the (k+1)th sample from two previous samples. The approximation in eqn.(2.17) improves as we add more terms in the series on the right hand side. To determine the higher order derivatives in the series, we require more samples in the past. The larger the member of past samples we use, the better will be the prediction. Thus, in general we can express the prediction formula as,

$$g[k] \approx a_1 g[k-1] + a_2 g[k-2] + \ldots\ldots\ldots + a_N g[k-N] \quad (2.19)$$

The right hand side of eqn.(2.19), is , $\hat{g}$[k, the predicted value of g[k]. Thus,

$$\hat{g}[k] = a_1 g[k-1] + a_2 g[k-2] + \ldots\ldots\ldots + a_N g[k-N] \quad (2.20)$$

This is the eqn. of an Nth order predictor. Larger n would result in better prediction in general. The output of this filter (predictor) is $\hat{g}[k]$, the predicted value of g[k]. the input is the previous samples g[k-1], g[k-2],……,g[k-n], although it is customary to say that the input is g[k] and the output is $\hat{g}[k]$.

Eqn.(2.20) reduces to $\hat{g}[k] = g[k-1]$ for the 1st order predictor. This is similar to eqn.(2.17). This means $a_1 = 1$ and the 1at order predictor is a simple time delay.

The predictor described in eqn.(2.20) is called a linear predictor. It is basically a transversal filter(a tapped delay line), where the tap gains are set equal to the prediction coefficients as shown in fig. 2.8.

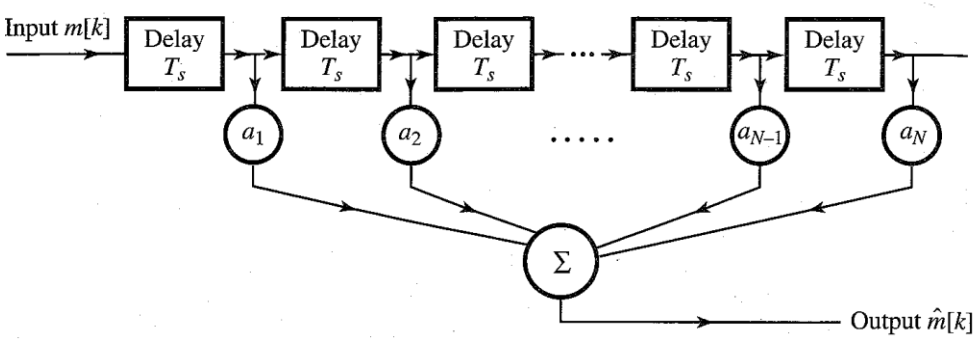


Fig. 2.8 Transversal filter(tapped delay line) used as a liner predictor

➢ Analysis of DPCM :

As mentioned earlier, in DPCM we transmit not the present sample g[k] but d[k] (the difference between g[k] and its predicted value $\hat{g}[k]$). At the receiver, we generate $\hat{g}[k]$ from the past sample values to which the received d[k] is added to generate g[k]. There is, however, one difficulty in this scheme. At the receiver, instead of the past samples g[k-1], g[k-2],……. as well as d[k], we have their quantized versions $g_p[k-1]$, $g_p[k-2]$,….. Hence, we cannot determine $\hat{g}[k]$. We can only determine $g_p[k]$, the estimate of the quantized sample $g_q[k]$ in terms of the quantized samples $g_q[k-1]$, $g_q[k-2]$,……. This will increase the error in reconstruction. In such a case, a better strategy is to determine $\hat{g}_q[k]$, the estimate of $g_q[k]$(instead of g[k]), at the transmitter also from the quantized samples $g_q[k-1]$, $g_q[k-2]$,……. The difference $d[k] = g[k] - g_q[k-2]$,……. is now transmitted using PCM. At the receiver we can generate $\hat{g}_q[k]$, and from the received d[k], we can reconstruct $g_q[k]$.

Fig 2.9 shows a DPCM transmitter. We shall soon see that the predictor input is $g_q[k]$. Naturally its output is $\hat{g}_q[k]$, the predicted value of $g_q[k]$. The difference,

$$d[k] = g[k] - \hat{g}_q[k] \qquad (2.21)$$

is quantized to yield

$$d_q[k] = d[k] + q[k] \tag{2.22}$$

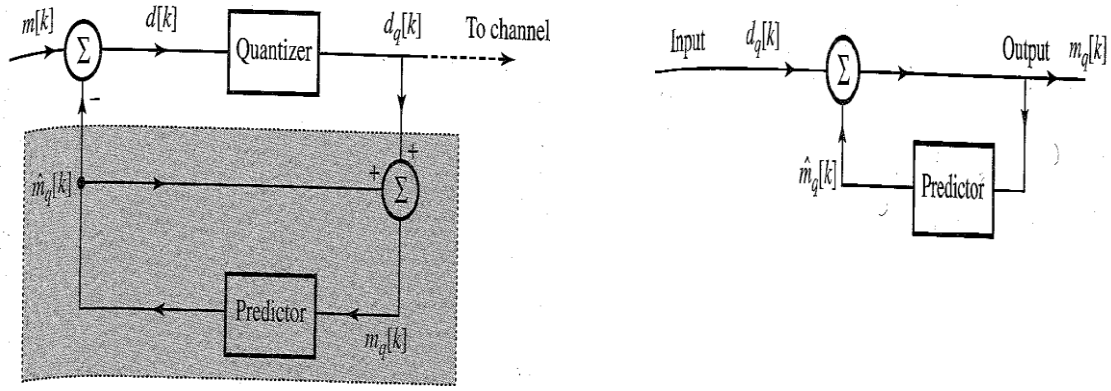


Fig. 2.9 DPCM system—Tansmitter and Receiver

In eqn.(2.22) q[k] is the quantization error. The predictor output $\hat{g}_q[k]$ is fed back to its input so that the predictor input $g_q[k]$ is,

$$\begin{aligned} g_q[k] &= \hat{g}_q[k] + d_q[k] \\ &= g[k] - d[k] + d_q[k] \\ &= g[k] + q[k] \end{aligned} \tag{2.23}$$

This shows that $g_q[k]$ is a quantized version of g[k]. The predictor input is indeed $g_q[k]$ as assumed. The quantized signal $d_q[k]$ is now transmitted over the channel. The receiver shown in fig 2.9 is identical to the shaded portion of the transmitter. The input in both cases is also the same, viz., $d_q[k]$. Therefore, the predictor output must be $\hat{g}_q[k]$ (the same as the predictor output at the transmitter). Hence, the receiver output (which is the predictor input) is also the same, viz., $g_q[k] = g[k] + q[k]$, as found in eqn.(2.23). This shows that we are able to receive the desired signal g[k] plus the quantization noise q[k]. This is the quantization noise associated with the difference signal d[k], which is much smaller than g[k]. The received samples are decoded and passed through a low pass filter of D/A conversion.

- **SNR Improvement :**

To determine the improvement in DPCM over PCM, let g_p and d_p be the peak amplitudes of g(t) and d(t). If we use the same value of 'L' in both cases, the quantization step Δ in DPCM is reduced by the factor g_p/d_p. Because the quantization noise power is $\Delta^2/12$, the quantization noise in DPCM reduced by the factor $(g_p/d_p)^2$ and the SNR increases by the same factor. Moreover, the signal power is proportional to its peak value squared (assuming other statistical properties invariant). Therefore, G_p(SNR improvement due to prediction) is

$$G_p = P_g/P_d \tag{2.24}$$

Where P_g and P_d are the powers of g(t) and d(t) respectively. In terms of dB units, this means that the SNR increases by $10\log(P_m/P_d)$ dB.
For PCM,

$$(S_0/N_0) = \alpha + Gn \qquad \text{where, } \alpha = 10\log C \qquad (2.25)$$

In case of PCM the value of α is higher by $10\log(P_g/P_d)$ dB. A second order predictor processor for speech signals can provide the SNR improvement of around 5.6 dB. In practice, the SNR improvement may be as high as 25 dB. Alternately, for the same SNR, the bit rate for DPCM could be lower than that for PCM by 3 to 4 bits per sample. Thus telephone systems using DPCM can often operate at 32 kbits/s or even 24 kbits/s.

➢ Delta Modulation:

Sample correlation used in DPCM is further exploited in delta modulation(DM) by oversampling(typically 4 times the Nyquist rate) the baseband signal. This increases the correlation between adjacent samples, which results in a small prediction error that can be encoded using only one bit (L=2) for quantization of the $g[k] - \hat{g}_q[k]$. In comparison to PCM even DPCM, it us very simple and inexpensive method of A/D conversion. A 1-bit code word in DM makes word framing unnecessary at the transmitter and the receiver. This strategy allows us to use fewer bits per sample for encoding a baseband signal.

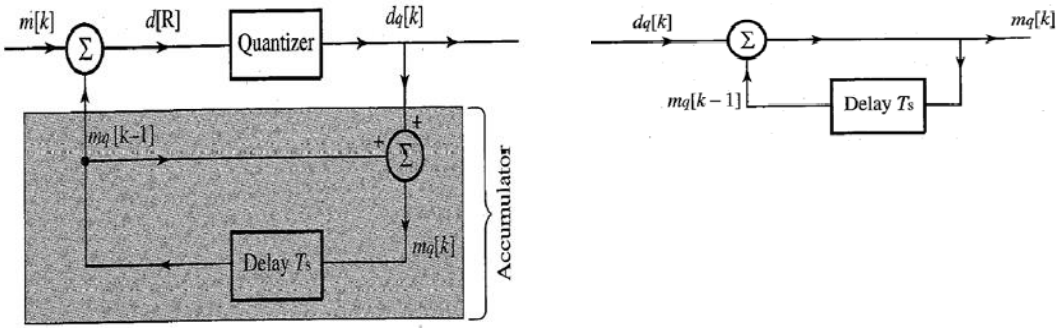


Fig. 2.10 Delta Modulation is a special case of DPCM

In DM, we use a first order predictor which as seen earlier is just a time delay of T_s(the sampling interval). Thus, the DM transmitter (modulator) and the receiver (demodulator) are identical to those of the DPCM in fig2.9 with a time delay for the predictor as shown in fig 2.10. From this figure, we obtain,

$$\hat{g}_q[k] = g_q[k-1] + d_q[k] \qquad (2.26)$$

Hence, $$g_q[k-1] = g_q[k-2] + d_q[k-1] \qquad (2.27)$$

Substituting eqn.(2.27) into eqn.(2.26) yields

$$g_q[k] = g_q[k-2] + d_q[k] + d_q[k-1] \quad (2.28)$$

Proceeding iteratively in this manner and assuming zero initial condition, ie, $g_q[0] = 0$, yields,

$$g_q[k] = \sum_{g=0}^{k} dq[g] \quad (2.29)$$

This shows that the receiver(demodulator) is just an accumulator(adder). If the output $d_{q=}[k]$ is represented by an integrator because its output is the sum of the strengths of the input impulses(sum of the areas under the impulses). We may also replace the feedback portion of the modulator (which is identical to the demodulator) by an integrator. The demodulator output is $g_p[k]$, which when passed through a low pass filter yields the desired signal reconstructed from the quantized samples.

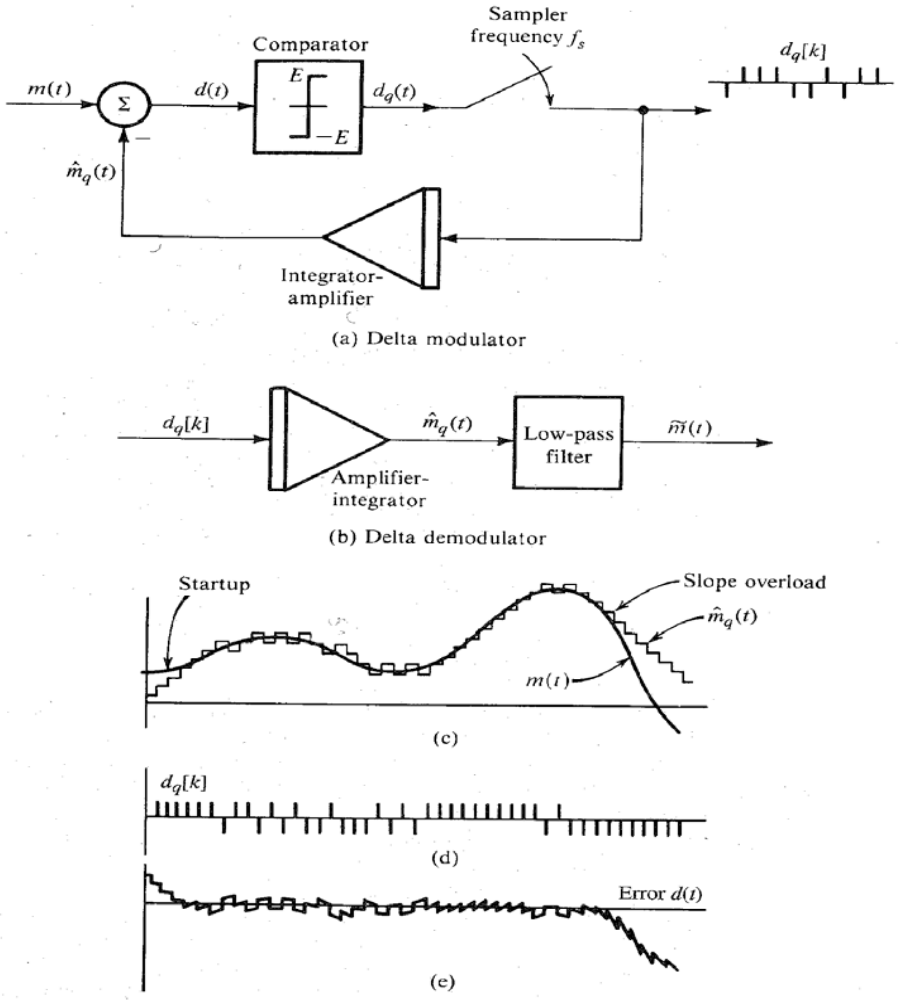


Fig. 2.11 Delta Modulation

Fig 2.11 shows a practical implementation of the delta modulator and demodulator. As discussed earlier, the first order predictor is replaced by a low cost integrator circuit (such as and RC integrator). The modulator consists of a comparator and a sampler in the direct path and an integrator amplifier n the feedback path. Let us see how this delta modulator works.

The analog signal g(t) is compared with the feedback signal(which served as a predicted signal) $\hat{g}_q[k]$. The error signal d(t) = g(t) – $\hat{g}_q[k]$ is applied to a comparator. If d(t) is +ve, the comparator output is a constant signal of amplitude E, and if d(t) is –ve, the comparator output is –E. Thus, the difference is a binary signal [L = 2] that is needed to generate a 1-bit DPCM. The comparator output is sampled by a sampler at a rate of f_s samples per second. The sampler thus produces a train of narrow pulses $d_q[k]$ with a positive pulse when $g(t) > \hat{g}_q[k]$ and a negative pulse when $g(t) < \hat{g}_q[k]$. The pulse train $d_q(t)$ is the delta modulated pulse train. The modulated signal $d_q(t)$ is amplified and integrated in the feedback path to generate $\hat{g}_q[k]$ which tries to follow g(t).

To understand how this works we note that each pulse in $d_q[k]$ at the input of the integrator gives rise to a step function (positive or negative depending on pulse polarity) in $\hat{g}_q[k]$. If, eg, $g(t) > \hat{g}_q[k]$, a positive pulse is generated in $d_q[k]$, which gives rise to a positive step in $\hat{g}_q[k]$, trying to equalize $\hat{g}_q[k]$ to g(t) in small steps at every sampling instant as shown in fig 2.11. It can be seen that $\hat{g}_q[k]$ is a kind of staircase approximation of g(t). The demodulator at the receiver consists of an amplifier integrator (identical to that in the feedback path of the modulator) followed by a low pass filter.

- **DM transmits the derivative of g(t)**

  In DM, the modulated signal carries information not about the signal samples but about the difference between successive samples. If the difference is positive or negative a positive or negative pulse (respectively) is generated in the modulated signal $d_q[k]$. Basically, therefore, DM carries the information about the derivative of g(t) and , hence, the name delta modulation. This can also be seen from the face that integration of the delta modulated signal yields $g_q(t)$, which is an approximation of g(t).

- **Threshold of coding and overloading**

  Threshold and overloading effects can be clearly seen in fig 2.11c. Variation in g(t) smaller than the step value(threshold coding) are lost in DM. Moreover, if g(t) changes too fast ie, $\hat{g}_q[k]$ is too high, $\hat{g}_q[k]$ cannot follow g(t), and overloading occurs. This is the so called slope overload which gives rise to slope overload noise. This noise is one of the basic limiting factors in the performance of DM. We should expect slope overload rather than amplitude overload in DM, because DM basically carries the information about $\hat{g}_q[k]$. The granular nature of the output signal gives rise to the granular noise similar to the quantization noise. The slope overload noise can be reduced by increasing the step size Δ. This unfortunately increases granular noise. There is an

optimum value of Δ, which yields the best compromise giving the minimum overall noise. This optimum value of Δ depends on the sampling frequency f_s and the nature of the signal.

The slope overload occurs when $g_q[k]$ cannot follow g(t). During the sampling interval Ts, $g_q[k]$ is capable of changing by Δ, where Δ is the height of the step Hence, the maximum slope that $g_q[k]$ can follow is $\Delta/T_{s=}$, or Δf_s, where f_s is the sampling frequency. Hence, no overload occurs if

$$|g^{\cdot}(t)| < \Delta f_s \tag{2.30}$$

Consider the case of a single tone modulation,

ie, $g(t) = A.\cos(wt)$

The condition for no overload is

$$|g^{\cdot}(t)|_{max} = wA < \Delta f_s \tag{2.31}$$

Hence, the maximum amplitude 'A_{max}' of this signal that can be tolerated without overload is given by

$$A_{max} = \Delta f_s / W \tag{2.32}$$

The overload amplitude of the modulating signal is inversely proportional to the frequency W. For higher modulating frequencies, the overload occurs for smaller amplitudes. For voice signals, which contain all frequency components up to(say) 4 KHz, calculating A_{max} by using W = 2.pi.4000 in eqn.(2.32) will give an overly conservative value. It has been shown by De Jager that 'A_{max}' for voice signals can be calculated by using W_r = 2.pi.800 in eqn.(2.32),

$$[A_{max}]_{voice} \approx \Delta f_s / w_r \tag{2.33}$$

Thus, the maximum voice signal amplitude 'A_{max}' that can be used without causing slope overload in DM is the same as the maximum amplitude of a sinusoidal signal of reference frequency f_r(f_r = 800 Hz) that can be used without causing slope overload in the same system.

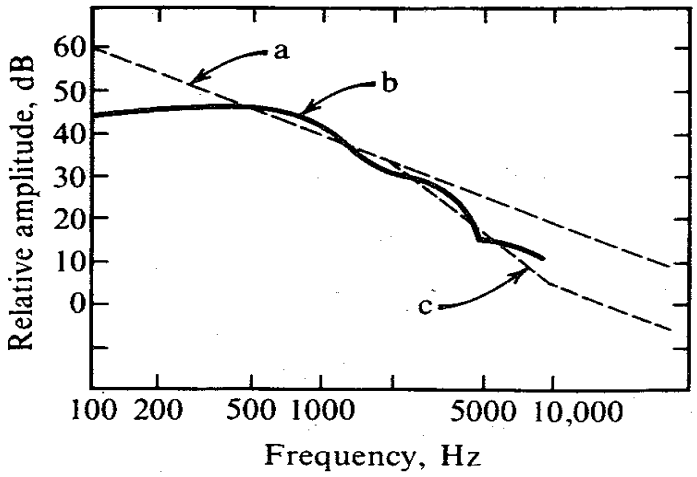


Fig. 2.12 Voice Signal Spectrum

Fortunately, the voice spectrum (as well as the TV video signal) also decays with frequency and closely follows the overload characteristics (curve c, fig 2.11). For this reason, DM is well suited for voice (and TV) signals. Actually, the voice signal spectrum (curve b) decrease as 1/W upto 2000 Hz, land beyond this frequency, it decreases as $1/W^2$. Hence, a better match between the voice spectrum and the overload characteristics is achieved by using a single integration up to 2000 Hz and a double interaction beyond 2000 Hz. Such a circuit (the double integration) is fast responding, but has a tendency to instability, which can be reduced by using some lower order prediction along with double integration. The double integrator can be built by placing in cascade tow low pass RC integrators with the time constant $R_1C_1 = 1/2000.pi$ and $R_2C_2 = 1/4000.pi$, respectively. This result in single integration from 100 Hz to 2000 Hz and double integration beyond 2000 Hz.

- **Adaptive Delta Modulation**

The DM discussed so far suffers from one serious disadvantage. The dynamic range of amplitudes is too small because of the threshold and overload effects discussed earlier. To correct this problem, some type of signal compression is necessary. In DM a suitable method appears to be the adaptation of the step value 'Δ' according to the level of the input signal derivative. For example in fig.2.11c when the signal g(t) is falling rapidly, slope overload occurs. If we can increase the step size during this period, this could be avoided. On the other hand, if the slope of g(t) is small, a reduction of step size will reduce the threshold level as well as the granular noise. The slope overload causes dq[k] to have several pulses of same polarity in succession. This call for increased step size. Similarly, pulses in dq[k] alternating continuously in polarity indicates small amplitude variations, requiring a reduction in step size. This results in a much larger dynamic range for DM.

- **Output SNR**

The error d(t) caused by the granular noise in DM, (excluding slope overload), lies in the range (-Δ,Δ) , where Δ is the step height in $g_q(t)$. The situation is similar to that encountered in PCM, where the quantization error amplitude was in the range from –Δ/2 to Δ/2. The quantization noise is,

$$\langle q_e^2 \rangle = 1/\Delta \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} q_e^2 dq_e = \Delta^2/12 \qquad (2.34)$$

Similarly the granular noise power $\langle g_n^2 \rangle$ is

$$\langle g_n^2 \rangle = 1/(2A) \int_{-A}^{A} g_n^2 . dg_n = A^3/3 \qquad (2.35)$$

The granular noise PSD has continuous spectrum, with most of the power in the frequency range extending well beyond the sampling frequency 'fs'. At the output, most of this will be suppressed by the baseband filter of bandwidth W. Hence the granular noise power N_0 will be well below that indicated in equation (18). To compute N_0 we shall assume that PSD of the quantization noise is uniform and concentrated in the band of 0 to fs Hz. This assumption has been verified experimentally. Because the total power $\Delta^3/3$ is uniformly spread over the bandwidth f_s, the power within the baseband W is

$$N_0 = (\Delta^3/3)W/f_s = \Delta^2.W/(3f_s) \qquad (2.36)$$

The output signal power is $S_0 = \langle g^2(t) \rangle$. Assuming no slope overload distortion

$$S_0/N_0 = 3.f_s \langle g^2(t) \rangle / (\Delta^2.W) \qquad (2.37)$$

If g_p is the peak signal amplitude, then eqn. (2.33) an be written as,

$$g_p = \Delta f_s / W_r$$

& $$S_0/N_0 = 3.f_s^3 \langle g^2(t) \rangle / (W_r^2.W.g_p^2) \qquad (2.38)$$

Because we need to transmit f_s pulses per second, the minimum transmission bandwidth $B_T = f_s/2$. Also for voice signals, W=4000 and W_r =2.pi.800 =1600.pi. Hence,

$$S_0/N_0 = [3.(2B_T)^3 \langle g^2(t) \rangle]/[1600x1600.\pi^2 W g_p^2]$$
$$= 150/\pi^2.(B_T/W)^3.\langle g^2(t) \rangle / g_p^2 \qquad (2.39)$$

Thus the output SNR varies as the cube of the bandwidth expansion ratio B_T/W. This result is derived for the single integration case. For double integration DM, Greefkes and De Jager have shown that,

$$S_0/N_0 = 5.34(B_T/W)^5 \langle g^2(t) \rangle / g_p^2 \qquad (2.40)$$

It should be remembered that these results are valid only for voice signals. In all the preceding developments, we have ignored the pulse detection error at the receiver.

- **Comparison With PCM**

The SNR in DM varies as a power of B_T/W, being proportional to $(B_T/W)^3$ for single integration and $(B_T/W)^5$ for double integration. In PCM on the other hand the SNR varies exponentially with B_T/W. Whatever the initial value, the exponential will always outrun the power variation. Clearly for higher values of B_T/W, PCM is expected to be superior to DM. The output SNR for voice signals as a function of the bandwidth expansion ratio B_T/W is plotted in fig. for tone modulation, for which $\langle g^2 \rangle / g_p^2 = 0.5$. The transmission band is assumed to be the theoretical minimum bandwidth for DM as well as PCM. It is clear that DM with double integration has a performance superior to companded PCM(which is the practical case) for lower valued of $B_T/W = 10$. In practice, the crossover value is lower than 10, usually between 6 & 7(f_s = 50 kbits/s). This is true only for voice and TV signals, for which DM is ideally suited. For other types of signals, DM does not comparable as well with PCM. Because the DM signal is digital signal, it has all the advantages of digital system, such as the use of regenerative repeaters and other advantages as mentioned earlier. As far as detection of errors are concerned, DM is more immune to this kind of error than PCM, where weight of the detection error depends on the digit location; thus for n=8, the error in the first digit is 128 times as large as the error in the last digit.

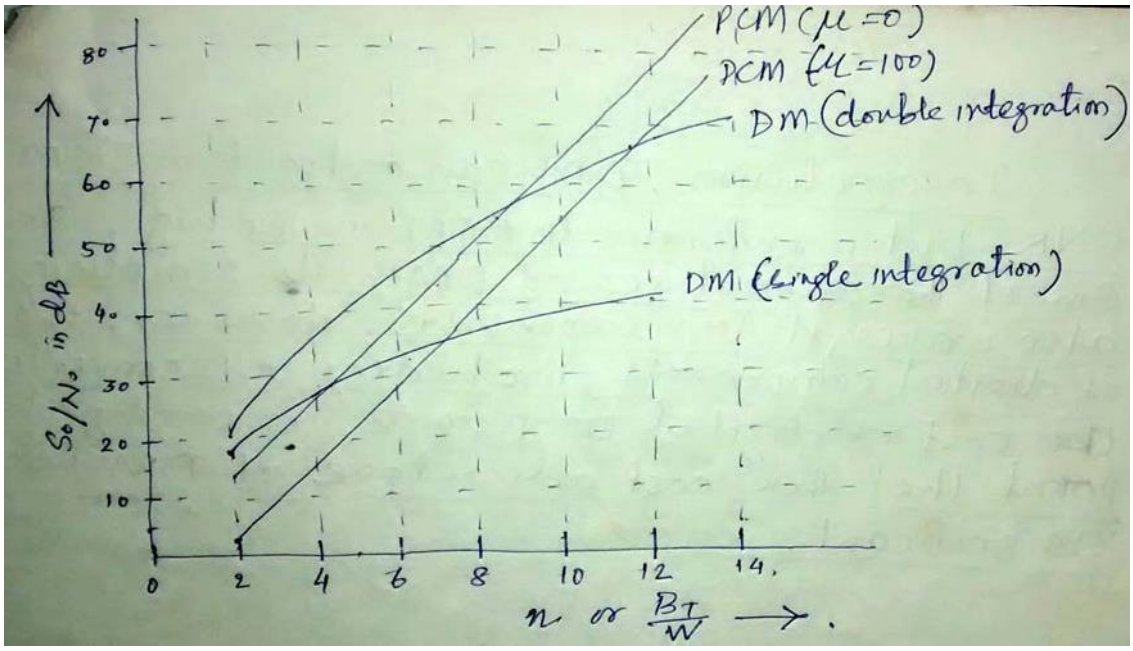


Fig. 2.21a Comparison of DM and PCM.

For DM, on the other hand, each digit has equal importance. Experiments have shown that an error probability 'Pe' on the order of 10^{-1} does not affect the

intelligibility of voice signals in DM, where as 'Pe' as low as 10^{-4} can cause serious error, leading to threshold in PCM. For multiplexing several channels, however, DM suffers from the fact that each channel requires its own coder and decoder, whereas for PCM, one coder and one decoder are shared by each channel. But his very fact of an individual coder and decoder for each channel also permits more flexibility in DM. On the route between terminals, it is easy to drop one or more channels and insert other incoming channels. For PCM, such operations can be performed at the terminals. This is particularly attractive for rural areas with low population density and where population grows progressively. The individual coder-decoder also avoids cross-talk, thus alleviating the stringent design requirements in the multiplexing circuits in PCM.

In conclusion, DM can outperform PCM at low SNR, but is inferior to PCM in the high SNR case. One of the advantages of DM is its simplicity, which also makes it less expensive. However, the cost of digital components, including A/D converters, ie, coming down to the point that the cost advantage of DM becomes insignificant.

➢ Noise in PCM and DM

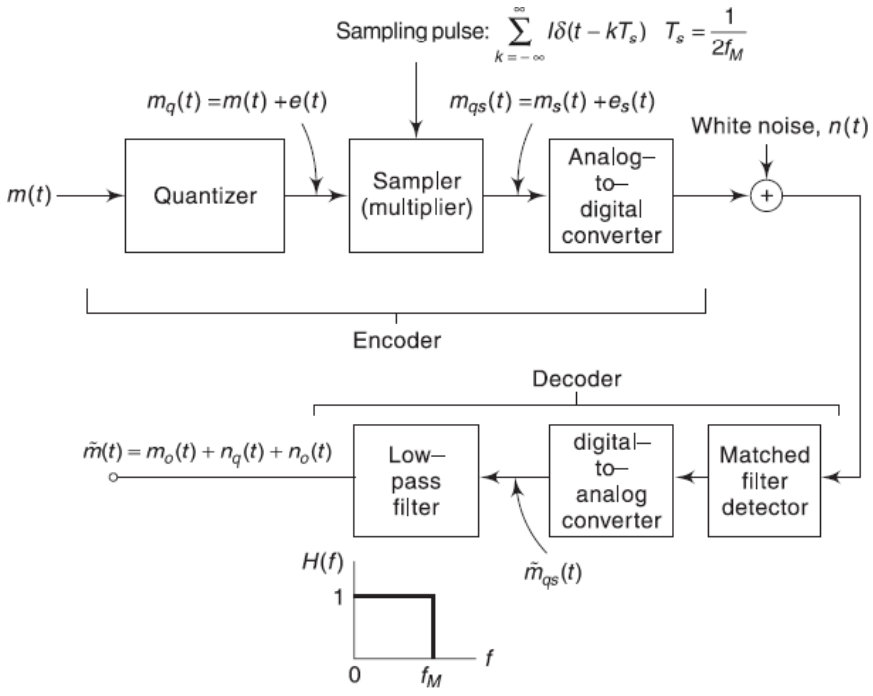


Fig. 2.13 A binary PCM encoder-decoder.

In the above figure m(t) is same as g(t). The baseband signal g(t) is quantized, giving rise to quantized signal $g_q(t)$, where

$$g_q(t) = g(t) + e(t) \qquad (2.41)$$

($e(t)$ is same as $q_e(t)$ as discussed earlier).

The sampling interval is $T_s=1/2f_m$, where f_m is the frequency to which the signal g(t) is bandlimited.

The sampling pulses considered here are narrow enough so that the sampling may be considered as instantaneous. With such instantaneous sampling, the sampled signal may be reconstructed exactly by passing the sequence of samples through a low pass filter with cut off frequency of f_m. Now as a matter of mathematical convenience, we shall represent each sampling pulse as an impulse. The area of such an impulse is called its strength, and an impulse of strength I is written as $I\delta(t)$.
The sampling impulse train is therefore s(t), given by,

$$s(t) = I\sum_{-\infty}^{\infty} \delta(t - k)\, T_c \qquad (2.42)$$

Where, $T_s = 1/(2.f_m)$

From equation 1 and 2 , the quantized signal $g_q(t)$ after sampling becomes $g_{qs}(t)$, written as,

$$g_{qs}(t) = g(t)I\sum_{k=-\infty}^{\infty} \delta(t - kT_c) + e(t)I\sum_{k=-\infty}^{\infty} \delta(t - kT_c) \qquad (2.43a)$$

$$= g_s(t) + e_s(t) \qquad (2.43b)$$

The binary output of the A/D converter is transmitted over a communication channel and arrives at the receiver contaminated as a result of the addition of white thermal noise W(t). Transmission may be direct as indicated in fig.2.13, or the binary output signal may be used to modulate a carrier as in PSK or FSK.
In any event the received signal is detected by a matched filter to minimize errors in determining each binary bit and thereafter passed on to a D/A converter. The output of a D/A converter is called $g_{qs}(t)$. In the absence of thermal noise and assuming unity gain from the input to the A/D converter to the output of the D/A converter, we should have $g\sim_{qs}(t) = g_{qs}(t)$. Finally the signal $g\sim_{qs}(t)$ is passed through the low pass baseband filter. At the output of the filter we find a signal $g_0(t)$ which aside from a possible difference in amplitude has exactly the waveform of the original baseband signal g(t). This output signal however in accompanied by a noise waveform $W_q(t)$ due to thermal noise.

- **<u>Calculation of Quantization Noise</u>**

Let us calculate the output power due to the quantization noise in the PCM system as in fig.2.14 ignoring the effect of thermal noise.
The sampled quantization error waveform, as given by eq^n (2.43b),

$$e_s(t) = e(t)I\sum_{k=-\infty}^{\infty} g(t - kT_c) \qquad (2.44)$$

It is to be noted that if the sampling rate is selected to be the nyquist rate for the baseband signal g(t) the sampling rate will be inadequate to allow reconstruction of the error signal e(t) from its sample $e_s(t)$. In fi.2 the quantization levels are separated by amount Δ. We observe that e(t) executes a complete cycle and exhibits an abrupt discontinuity every time g(t) makes an excursion of amount Δ. Hence spectral range of e(t) extends for beyond the band limit f_m of g(t).

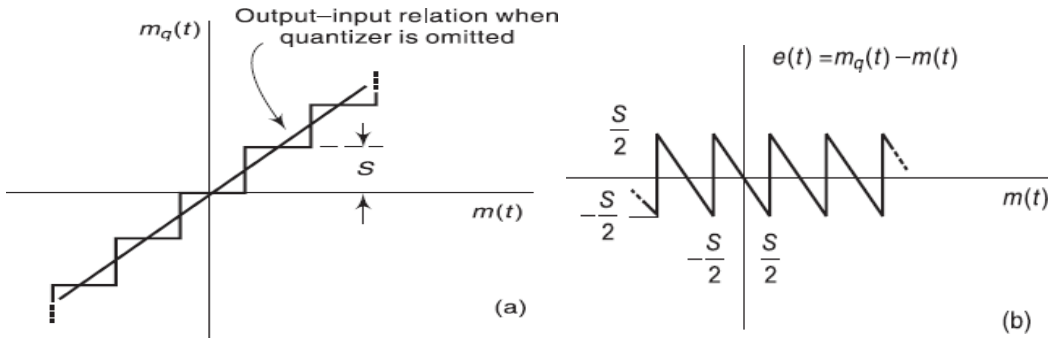


Fig. 2.14 Plot of $m_q(t)$ and e(t) as a function of m(t).

To find the quantization noise output power N_q, we require the PSD of the sampled quantization error $e_s(t)$ given in eq^n (2.44).

Since $\delta(t-kT_s) = 0$ except when $t=kT_s$ $e_s(t)$ may be written as,

$$e_s(t) = I.\sum_{k=-\infty}^{\infty} e(kT_c)\delta(t - kT_c) \qquad (2.45)$$

The waveform of eq^n (2.45) consists of a sequence of impulses of area=A=$e(kT_s)$ I occurring at intervals T_s. The quantity $e(kT_s)$ is the quantization error at sampling time and is a random variable.
The PSD $G_{es}(f)$ of the sampled quantization error is,

$$G_{e_s}(f) = \frac{I^2}{T_s}\overline{e^2(kT_s)} \qquad (2.46)$$

and,

$$e^2(t) = e^2(kT_s) = \frac{S^2}{12}$$

For a step size of Δ the quantization error is

$$e^2(t) = \Delta^2/12 \qquad (2.47)$$

Equation 6 involves $\langle e^2(kT_s)\rangle$ rather than $\langle e^2(t)\rangle$. However since the probability density of e(t) does not depend on time the variance of e(t) is equal to the variance of e(t= kT_s) .

Thus, $\langle e^2(t)\rangle = \langle e^2(kT_s)\rangle = \Delta^2/12$ (2.48)
From eqn. (2.46) and eqn. (2.49) we have,

$$G_{es}(f) = I^2\Delta^2/(T_s.12) \qquad (2.49)$$

Finally the quantization noise N_q is, from eqn. (2.50),

$$N_q = \int_{-f_M}^{f_M} G_{e_s}(f)\,df = \frac{I^2}{T_s}\frac{S^2}{12}2f_M = \frac{I^2}{T_s^2}\frac{S^2}{12} \qquad (2.50)$$

[take 'S' as 'Δ']

- **The Output Signal Power**

The sampled signal which appears at the input to the baseband filter shown in fig.2.14 is given by $g_s(t)$ in eq^n(2.43) as.

$$g_s(t) = g(t).I.\sum_{k=-\infty}^{\infty} \delta(t - kT_c) \qquad (2.51)$$

Since the impulse train is periodic it can be represented by a fourier series. Because the impulses have strength I and are separated by a time T_s, the first term in Fourier series is the dc component which is $1/T_s$. Hence the signal $g_0(t)$ at the output of the baseband filter is

$$g_0(t) = I/T_s.g(t) \qquad (2.52)$$

Since $T_s=1/2f_m$, other terms in the series of equation 11 lie outside the passband of the filter. The normalised signal output power is from eq^n (2.52),

$$\overline{g_0^2(t)} = I^2/T^2 .\overline{g^2(t)} \qquad (2.53)$$

We can now express $\overline{g^2(t)}$ in terms of the number M of quantization levels and the step size Δ. To do this we can say that the signal can vary from $-m\Delta/2$ to $m\Delta/2$, i.e we assume that the instantaneous value of g(t) may fall anywhere in its allowable range of '$m\Delta$' volts with equal likelihood. Then the probability density of the instantaneous value of g in f(g) given by,

$$f(g) = 1/(M\Delta)$$

The variance σ^2 of g(t), ie, $\overline{g^2(t)}$ is,

$$\overline{g^2(t)} = \int_{-\frac{M\Delta}{2}}^{\frac{M\Delta}{2}} g^2 f(g)dg = M^2.\Delta^2/12 \qquad (2.54)$$

Hence from eqn. (2.53), the output signal power is

$$S_0 = \overline{g_0^2(t)} = I^2/T^2 . M^2.\Delta^2/12 \qquad (2.55)$$

From eqn.(2.50) and (2.55) we find the signal to quantization noise ratio is

$$S_o /N_q = M^2 = (2^N)^2 \qquad (2.56)$$

where, N is the number of binary digits needed to assign individual binary code designations to the M quantization levels.

- **The Effects of Thermal Noise**

The effect of additive thermal noise is to calculate the matched filter detector of fig.2.14 to make an occasional error in determining whether a binary 1 or binary 0 was transmitted. If the thermal noise is white and Gaussian the probability of such an error depends on the ratio E_b/η. Where E_b is signal energy transmitted during a bit and $\eta/2$ is the two sided power spectral density of the noise. The probability depends also on the type of modulation employed.

Rather typically, PCM system operate with error probabilities which are small enough so that we may ignore the likelihood that more than a single bit error will occur with in a single word. For example, if the error probability $P_e=10^{-5}$ and a word of 8 bits we would expect on the average that 1 word would be in error for every 12500 word transmitted. Indeed the probability of two words being transmitted in error in the same 8 bit word is $28*10^{-10}$.

Let us assume that a code word used to identify a quantization level has N bits. We assume further that the assignment of code words to levels is in the order of numerical significance of the word. Thus we assign 00. 00 to the most negative level to the next higher level until the most positive level is assigned the codeword 1 1. 1 1.

An error which occurs in the least significant bit of the code word corresponds to an incorrect determination by amount 'Δ' in the quantized value $g_s(t)$ of the sampled signal. An error in the next higher significant bit corresponds to an error 2Δ; in the next higher, 4Δ, etc.
Let us call the error δg_s. Then assuming that an error may occur with equal likelihood inany bit of the word, the variance of the error is,

$$<\delta g^2_s> = 1/N.[\ \Delta^2 + (2\Delta)^2 + (4\Delta)^2 + \ldots\ldots\ldots + (2^{N-1}\Delta)^2]$$
$$= \Delta^2/N.[\ 1^2 + (2)^2 + (4)^2 + \ldots\ldots\ldots + (2^{N-1})^2] \qquad (2.57)$$

The sum of the geometric progression in eqn.(2.57),

$$<\delta g^2_s> = \Delta^2/N.2^{(2N-1)}/(2^2-1) = 2^{2N}.\Delta^2/(3N), \quad \text{for } N \geq 2 \qquad (2.58a)$$

The preceding discussion indicates that the effect of thermal noise errors may be taken into account by adding at the input to the A/D converter in fig. 2.14, an error voltage δg_s , and by detecting the white noise source and the matched filter. We have assumed unity gain from the input to the A/D converter to the output of the D.A converter. Thus the same error voltage appears at the input to the lowpass baseband filter. The results of a succession of errors is a train of impulses, each of strength $I(\delta g_s)$. These impulses are of random amplitude and of random time of occurrence.

A thermal noise error impulse occurs on each occasion when a word is in error. With P_e the probability of a bit error, the mean separation between bits which are in errors is $1/ P_e$.
With N bits per word , the mean separation between words which are in error is $1/N\ P_e$ words. Words are separated in time by the sampling interval T_s. Hence the mean time between words which are in error is T, given by

$$T = \frac{T_s}{NP_e} \qquad (2.58b)$$

The power spectral density of the thermal noise error impulses train is, using eqn.(2.58a) and(2.58b),

$$G_{th}(f) = I^2/T < \delta.g_s^2 > = NP_eI^2/T_s <\delta.g_s^2> \qquad (2.59)$$

using eqn.(2.58a), we have

$$G_{th}(f) = 2^{2N}\Delta^2 P_e I^2/(3T^2_e) \quad (2.60)$$

Finally, the output power due to the thermal error noise is,

$$N_{th} = \int_{-f_N}^{f_N} G_{th}(f)df = 2^{2N}.\Delta^2 P_e I^2/(3.T_s^2) \quad (2.61)$$

- **Output Signal To Noise Ratio in PCM**

The output SNR including both quantization and thermal noise , is found by combining equation 10,16 and 23. The result is

$$\frac{S_o}{N_o} = \frac{S_o}{N_q + N_{th}} = \frac{(I^2/T_s^2)(M^2S^2/12)}{(I^2/T_s^2)(S^2/12) + (I^2/T_s^2)(P_e 2^{2N} S^2/3)}$$

[replace 'S' by 'Δ'; S is same as Δ]

$$= \frac{2^{2N}}{1 + 4P_e\, 2^{2N}} \quad (2.62)$$

In PSK(or for direct transmission) we have,

$$(P_e)_{PSK} = \frac{1}{2}\operatorname{erfc}\sqrt{\frac{E_b}{\eta}} \quad (2.63)$$

Where, E_b is the signal energy of a bit and η/2 is the two sided thermal noise power spectral density. Also, for coherent reception of FSK we have,

$$(P_e)_{FSK} = \frac{1}{2}\operatorname{erfc}\sqrt{0.6\frac{E_b}{\eta}} \quad (2.64)$$

To calculate E_b, we note that if a sample is taken at intervals of T_s and the code word of N bit occupies the total interval between samples, then a bit has a duration T_s/N. If the received signal power is S_i , energy associated with a single bit is

$$E_b = S_i \frac{T_s}{N} = S_i \frac{1}{2 f_M N} \quad (2.65)$$

Combining eqns. (2.62), (2.63) & (2.65), we find,

$$\left(\frac{S_o}{N_o}\right)_{PSK} = \frac{2^{2N}}{1 + 2^{2N+1}\operatorname{erfc}\sqrt{(1/2N)(S_i/\eta f_M)}} \quad (2.66)$$

using eqn.(2.64) in place of eqn.(2.63), we have

$$\left(\frac{S_o}{N_o}\right)_{FSK} = \frac{2^{2N}}{1 + 2^{2N+1}\operatorname{erfc}\sqrt{(0.3/N)(S_i/\eta f_M)}} \quad (2.67)$$

Note that for $S_i/\eta f_M \gg 1$ and $N = 8$

$$\left(\frac{S_o}{N_o}\right)_{\text{PSK, FSK}} = 10 \log (2^{16}) = 48 \text{ dB} \tag{2.68}$$

From fig. we find both the PCM system exhibit threshold, FSK threshold occurring at a $S_i/\eta f_m$ which is 2.2 dB greater than that for PSK. Experimentally, the onset of threshold in PCM is marked by an abrupt increase in a crackling noise analogous to the clicking noise heard below threshold in analogue FM systems.

- **Delta Modulation:**

A delta modulation system including a thermal noise source is shown in fig.2.15. The impulse generator applies the modulator a continuous sequence of impulses $p_i(t)$ of time separation τ. The modulator output is a sequence of pulses $P_0(t)$ whose polarity depends on the polarity of the difference signal $\delta(t)=g(t) - g\tilde{}(t)$, where $g\tilde{}(t)$ is the integrator output. We assume that the integrator has been adjusted so that its response to an input impulse of strength I is a step size Δ; i.e. $g\tilde{}(t) = (\Delta/I)\int P_0(t)dt$.

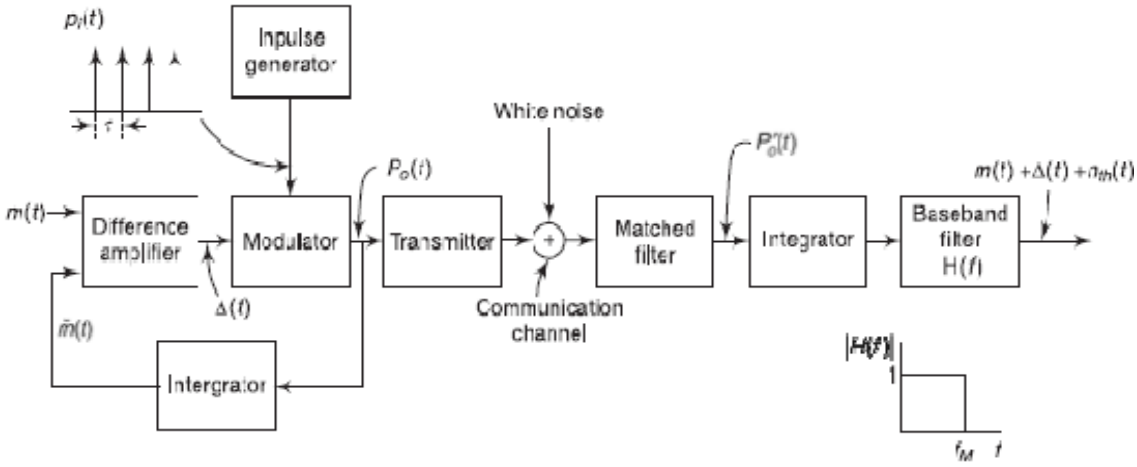


Fig. 2.15 A delta modulation system.

A typical impulse train $P_0(t)$ is shown in fig.2.16(a). Before transmission, the impulse waveform will be converted to the two level waveform of fig.2.16(b). Since this latter waveform has much greater power than a train of narrow pulses. This conversion is

accomplished by the block in fig.2.15 marked "transmitter". The transmitter in principle need be nothing more complicated than a bistable multivibrator. We may readily

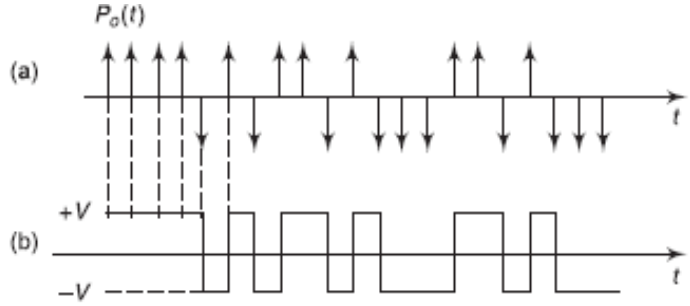


Fig.2.16 (a) A typical impulse train $p_0(t)$appearing at the modulator output in previous fig. (b) The two-level signal transmitted over the communication channel.

arrange that two positive impulses set the flip-flop into one of its stable states, while the negative impulses reset the flip-flop to its other stable state. The binary waveform of fig.2.16(b) will be transmitted directly or used to modulate as a carrier in FSK or PSK. After detection by the matched filter shown in fig.2.15, the binary waveform will be reconverted to a sequence of impulses $P_0'(t)$. In the absence of thermal noise $P_0'(t)=P_0(t)$, and the signal $\tilde{g}(t)$ is recovered at the receiver by passing $P_0'(t)$ through an integrator. We assume that transmitter and receiver integrators are identical and that the input to each consists of a train of impulses of strength +I or -I. Hence in the absence of thermal noise , the output of both the integrators are identical.

➢ Quantization Noise in Delta Modulation

Here in fig. 2.17 $\tilde{g}(t)$ in the delta modulator approximation to g(t). Fig 2.17 shows the error waveform δ(t) given by,

$$\delta(t) = g(t) - \tilde{g}(t) \qquad (2.69)$$

This error waveform is the source for quantization noise.

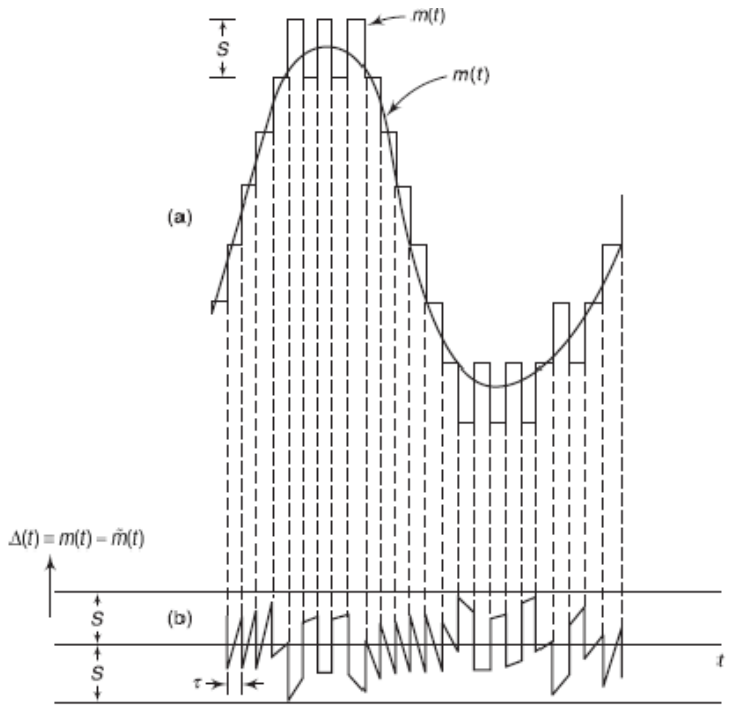


Fig. 2.17 The estimate $\tilde{g}(t)$ and error $\Delta(t)$ when g(t) is sinusoidal.

We observe that, as long as slope overloading is avoided, the error $\delta(t)$ is always less than the step size Δ. We shall assume that $\delta(t)$ takes on all values between $-\Delta$ and $+\Delta$ with equal likelihood. So we can assume the probability $\delta(t)$ is,

$$f(\delta) = 1/(2\Delta), \qquad -\Delta \leq \delta(t) \leq \Delta \tag{2.70}$$

The normalization power of the waveform $\delta(t)$ is then,

$$<[\delta(t)]^2> = \int_{-A}^{A} f(\delta)\, \delta^2 d\delta = \Delta^2/3 \tag{2.71}$$

Our interest is in estimating how much of this power will pass through a baseband filter. For this purpose we need to know something about the PSD of $\delta(t)$.

In fig. 2.17 the period of the sinusoidal waveform g(t) i.e. T has been selected so that T is an integral multiple of step duration τ. We then observe that the $\delta(t)$ is periodic with fundamental period T, and is of course, rich in harmonics. Suppose, however, that the period T is charged very slightly by amount δT. Then the fundamental period of $\delta(t)$ will not be T but will be instead $T * \tau/\delta T$ corresponding to a fundamental frequency near zero as δT tends to 0. And again, of course $\delta(t)$ will be rich in harmonics. Hence, in the general case, especially with g(t) a random signal, it is reasonable to assume that $\delta(t)$ has a spectrum which extends continuously over a frequency which begins near zero.

To get some idea of the upper frequency range of the spectrum of the waveform $\delta(t)$. Let us contemplate passing $\delta(t)$ through a LPF of adjustable cutoff frequency. Suppose

that initially the cutoff frequency is high enough so that $\delta(t)$ may pass with nominally no distortion. As we lower the cutoff frequency, the first type of distortion we would note is that the abrupt discontinuities in the waveform would exhibit finite rise and fall times. Such is the case since it is the abrupt changes which contribute the high frequency power content of the signal. To keep the distortion within reasonable limits, let us arrange that the rise time be rather smaller than the interval τ. To satisfy this condition we require the filter cutoff frequency f_c be of the order of $f_c=1/\tau$, since the transmitted bit rate $f_b=1/\tau$, $f_c=f_b$ as expected.

We now have made it appear reasonable, by a rather heuristic arguments that the spectrum of $\delta(t)$ extends rather continuously from nominally zero to $f_c = f_b$. We shall assume further that over this range the spectrum is white. It has indeed been established experimentally that the spectrum of $\delta(t)$ is approximately white over the frequency range indicated.

We may now finally calculate the quantization noise that will appear at the output of a baseband filter of cutoff frequency f_m. Since the quantization noise power in a frequency range f_b is $\Delta^3/3$ as given by equation 32, the output noise power in the baseband frequency range f_m is

$$N_q = \frac{S^2}{3}\frac{f_M}{f_b} = \frac{S^2 f_M}{3 f_b} \quad \text{[replace 'S' with 'Δ']} \qquad (2.72)$$

We may note also, in passing, that the two-sided power spectral density of $\delta(t)$ is,

$$G_\delta(f) = \Delta^2/(3.2.f_b) = \Delta^2/(6.f_b), \qquad -f_b \le f \le f_b \qquad (2.73)$$

➢ The Output Signal Power

In PCM, the signal power is determined by the step size and the number of quantization levels. Thus, with step size Δ and M levels, the signal could make excursion only between $-M\Delta/2$ and $M\Delta/2$. In delta modulation there is no similar restriction on the amplitude of the signal waveform, because the number of levels is not fixed. On the other hand, in delta modulation there is a limitation on the slope of the signal wave form which must be observed if slope overload is to be avoided. If however, the signal waveform changes slowly , there is normally no limit to the signal power which may be transmitted.

Let us consider a worst case for delta modulation. We assume that the signal power is concentrated at the upper end of the baseband. Specifically let the signal be,

$$g(t) = A.\sin(w_m t)$$

With 'A' the amplitude and $\omega_m = 2\pi f_m$, where f_m is the upper limit of the baseband frequency range. Then the output signal power

$$S_0(t) = \overline{g^2(t)} = A^2/2 \qquad (2.74a)$$

The maximum slope of g(t) is $\omega_m A$. The maximum average slope of the delta modulator approximation $\tilde{g}(t)$ is $\Delta/\tau = \Delta f_b$, where Δ is step size and f_b the bit rate. The limiting value of 'A' just before the onset of slope overload is, therefore given by the condition,

$$w_M \cdot A = \Delta f_b \qquad (2.74b)$$

From eqns.(2.74a) and (2.74b), we have that the maximum power which may be transmitted in,

$$S_0 = \Delta^2 f^2_b/(2w^2_M) \qquad (2.75)$$

The condition specified in equation 37 is unduly severe. A design procedure, more often employed, is to select the Δf_b product to be equal to the rms value of the slope g(t). In this case the output signal power can be increased above the value given in equation 38.

➢ Output Signal to Quantization Noise Ratio for Delta Modulation

The output signal to quantization noise ratio for delta modulation is found by dividing eqn.(2.75) by eqn.(2.72). The result is

$$\frac{S_o}{N_q} = \frac{5}{8\pi^2}\left(\frac{f_b}{f_M}\right)^3 \cong \frac{3}{80}\left(\frac{f_b}{f_M}\right)^3 \qquad (2.76)$$

It is of interest to note that when our heuristic analysis is replaced by a rigorous analysis, it is found that eqn. 39 continues to apply, except with a factor 3/80 replaced by 3/64, corresponding to a difference of less than 1dB.

The dependence of S_0/N_q on the product f_b/f_m should be anticipated. For suppose that the signal amplitude were adjusted to the point of slope overload, if now, say, f_m were increased by some order to continue to avoid overload.

Let us now make a comparison of the performance of PCM and DM in the matter of the ratio S_0/N_q. We observe that the transmitted signals in DM and in PCM are of the same waveform, a binary pulse train. In PCM a voltage level, corresponding to a single bit persists for the time duration allocated to one bit of codeword. With sampling at the Nyquist rate $1/2f_m$ s , and with N bits per code word , the PCM bit rate is $f_b=2f_mN$. In DM, a voltage corresponding to a single bit is held for a duration τ which is the interval between samples. Thus the DM system operates at a bit rate $f_b=1/\tau$.

If the communication channel is of limited bandwidth, then there is a possibility of interference in either DM or PCM. Whether such inter-symbol interference occurs in DM depends on the ratio of f_b to the bandwidth of the channel and similarly in PCM on the ratio of f_b to the channel bandwidth. For a fixed channel bandwidth, if inter-symbol

interference is to be equal in the two cases, DM or PCM , we require that both systems operate at the same bit rate or

$$f_b = f'_b = 2f_m N \qquad (2.77)$$

Combining eq 17 and 40 for PCM yields

$$S_0/N_q = 2^{2N} = 2^{fb/fm} \qquad (2.78)$$

Combining eq 39 and 40 for delta modulation yields

$$S_0/N_q = N^3 (3/\pi^2) \qquad (2.79)$$

Comparing equation 41 with 42 , we observe that for a fixed channel bandwidth the performance of DM is always poorer than PCM. For example if a channel is adequate to accommodate code words in PCM with N=8, equation 41 gives S_0/N_q = 48dB. The same channel used for DM would, from equation 42 yield S_0/N_q =22dB.

- **Comparison of DM and PCM for Voice**

when signal to be transmitted is the waveform generated by voice, the comparison between DM and PCM is overly pessimistic against DM. For as appears in the discussion leading to equation 37, in our concern to avoid slope overload under any possible circumstances, we have allowed for the very worst possible case. We have provided for the possibility that all the signal power might be concentrated at the angular frequency ω_m which is the upper edge of the signal bandwidth. Such is certainly not the case for voice. Actually for speech a bandwidth f_m = 3200Hz is adequate and the voice spectrum has a pronounced peak at 800Hz = $f_m/4$. If we replace ω_m by $\omega_m/4$ in eqn. (2.74b) we have,

$$w_M .A/4 = \Delta.f_b$$

The amplitude 'A' will now be four times larger than before and the allowed signal power before slope overload will be increased by a factor of it(12dB). Correspondingly, equation 39 now becomes,

$$S_0/N_q = 6/\pi^2.(f_b/f_M)^3 = 0.6(f_b/f_M)^3 = 5N^3 \qquad (2.80)$$

It may be readily verified that for $(f_b/f_m) \leq 8$ the signal to noise ratio for DM , SNR(δ), given by eqn.(2.80) is larger than SNR(PCM) given by eq^n (2.78). At about (f_b/f_m) = 4 the ratio SNR(DM)/ SNR(PCM) has maximum value 2.4 corresponding to 3.8db advantage. Thus if we allow f_m = 4KHz for voice, then to avail ourselves of this maximum advantage offered by DM we would take f_b = 16KHz.

In our derivation of the SNR in PCM we assumed that at all times the signal is strong enough to range widely through its allowable excursion. As a matter of fact, we specifically assumed that the distribution function f(g) for the instantaneous signal value g(t) was uniform throughout the allowable signal range. As a matter of practice, such would hardly be the case. The commercial PCM systems using companding, are designed so that the SNR remains at about 30dB over a 40dB range of signal power. In

short while eq^n (2.78) predicts a continuous increase in SNR(PCM) with increase in f_b/f_m, this result is for uncompanded PCM and in practice SNR(PCM) is approximately constant at 30dB. The linear DM discussed above has a dynamic range of 15dB. In order to widen this dynamic range to 40dB one employs adaptive DM(ADM), which yields advantages similar to the companding of PCM. When adaptive DM is employed, the SNR is comparable to the SNR of companded PCM. Today the satellite business system employs ADM operating at 32kb/s rather than companded PCM which operates at 64kb/s thereby providing twice as many voice channels in a given frequency band.

- **The Effect of Thermal Noise in DM**

When thermal noise is present, the matched filter in the receiver will occasionally make an error in determining the polarity of the transmitted waveform. Whenever such an error occurs , the received impulse stream $P_0'(t)$ will exhibit an impulse of incorrect polarity. The received impulse stream is then

$$P_0'(t)=P_0(t) + P_{th}(t) \qquad (2.81)$$

In which $P_{th}(t)$ is the error impulse stream due to thermal noise. If the strength of the individual impulses is I, then each impulse in P_{th} is of strength 2I and occurs only at each error. The factor of two results from the fact that an error reverses the polarity of the impulse.

The thermal error noise appears as a stream of impulses of of random time of occurrence and of strength ±2I. The average time of separation between these impulses is τ/P_e, where P_e is the bit error probability and τ is the time duration of a bit. The PSD of thermal noise impulses is

$$G_{pth}(f) = \frac{P_e}{c}(2I)^2 \qquad (2.82)$$

Now the integrators (assumed identical in both the DM transmitter and receiver) as having the property that when the input is an impulse of strength the output is a step of amplitude Δ is

$$\begin{aligned} F\{\Delta u(t)\} &= \Delta/j\omega \quad ; \ \omega\neq 0 \\ &= \Delta\pi\delta(\omega) \quad ; \ \omega=0 \end{aligned} \qquad (2.83)$$

We may ignore the dc component in the transform since such dc components will not be transmitted through the baseband filter. Hence we may take the transfer function of the integrator to be $H_i(f)$ given by

$$H_i(f) = \frac{\Delta}{I}\frac{1}{jm} \quad ; \quad \omega \neq 0 - \qquad (2.84)$$

And $$|H_i(f)|^2 = \left(\frac{\Delta}{I}\right)^2 \frac{1}{m^2} \quad ; \omega=0 \qquad (2.85)$$

From equation 46 and 49 we find that the PSD of the thermal noise at the input to the baseband filter is $G_{th}(f)$ given by

$$G_{th}(f) = |H_i(f)|^2 G_{pth}(f) = \frac{4\Delta^2 Pe}{cm^2} \qquad (2.86)$$

It would now appear that to find the thermal noise output, we need not to integrate $G_{th}(f)$ over the passband of the baseband filter. During integration we have extended the range of integration from $-f_m$ through f=0 to $+f_m$, even though we recognised that baseband filter does not pass dc and eventually has a low frequency cutoff f_1. However in other cases the PSD of the noise near f=0 is not inordinately large in comparison with the density throughout the baseband range generally. Hence, it as is normally the case, $f_1 << f_m$, the procedure is certainly justified as a good approximation. We observe however that in the

present case [eq^n (2.86)], $G_{th}(f) \to \infty$ at $\omega \to 0$, and more importantly that the integral of $G_{th}(f)$, over a range which include $\omega \to 0$,is infinite. Let us then explicitly take account of the low frequency cutoff f_1 of the baseband filter. The thermal noise output is using eq^n (2.86) with $\omega = 2\pi f$ and since $f_b = \frac{1}{\tau}$,

$$N_{th} = \frac{\otimes^2 P_e}{\pi^2 \tau}\left[\int_{-f_m}^{-f_1} \frac{df}{f^2} + \int_{f_1}^{f_m} \frac{df}{f^2}\right] \quad (2.87)$$

$$= \frac{2\otimes^2 P_e}{\pi^2 \tau}\left[\frac{1}{f_1} - \frac{1}{f_m}\right] \quad (2.88)$$

$$= \frac{2\otimes^2 P_e}{\pi^2 \tau f_1} = \frac{2\otimes^2 P_e f_b}{\pi^2 f_1} \quad (2.89)$$

If $f_1 << f_m$, unlike the situation encountered in all other earlier cases, the thermal noise output in delta modulation depends upon the low frequency cutoff rather than the higher frequency limit of the baseband range. In many application such as voice encoder where the voice signal is typically band limited from 300 to 3200 H_z, the use of band pass output filter(f_1=300H_z) is common place.

➢ Output Signal-to-Noise ratio in DM

The o/p SNR is obtained by combining eq^n (2.72), (2.80) and (2.89), the result is

$$\frac{S_0}{N_0} = \frac{S_0}{N_q + N_{th}} = \frac{(2\otimes^2/\pi)(f_b/f_m)^2}{(\otimes^2 f_m/3f_b) + (2\otimes^2 P_e f_b/\pi^2 f_1)} \quad (2.90)$$

Which may be written as

$$\frac{S_0}{N_0} = \frac{0.6(f_b/f_m)^3}{1 + 0.6P_e(f_b^2/f_m f_1)} \quad (2.91)$$

If transmission is direct or by means of PSK,

$$P_e = \frac{1}{2} erfc\sqrt{E_s/\eta} \quad (2.92)$$

Where E_s is the signal energy is a bit, is related to the received signal power S_i

By $E_s = S_i T_b = S_i/f_b$ (2.93)

Combining eq^n (2.91), (2.92) and (2.93), we have

$$\frac{S_0}{N_0} = \frac{0.6(f_b/f_m)^3}{1+[0.3\,f_b^2/f_m f_1]\operatorname{erfc}\sqrt{S_i/\eta f_b}} \qquad (2.94)$$

➢ Comparison of PCM and DM

We can now compare the output signal SNR I PCM and DM by comparing eq^n(2.66)and (2.94). To ensure that the communications channels bandwidth required is same in the two cases, we use the condition, given in eq^n(2.77), that $2N = f_b/f_m$. Then eq^n(2.66) can be written as

$$\frac{S_0}{N_0} = \frac{2^{\frac{f_b}{f_m}}}{1+2(2^{\frac{f_b}{f_m}})\operatorname{erfc}\sqrt{S_i/\eta f_b}} \qquad (2.95)$$

Eq^n (2.95) and (2.94) are compared in fig.2.18 for N=8(f_b(DM)=48 Kb/s) : to obtain the thermal performance of the delta modulator system, we assume voice transmission where f_m=300 H_z and f_1 = 300 H_z .

Thus f_b/f_m =16 (2.96)
And f_m/f_1 = 10 (2.97)

Let us compare the ratios S_0/N_0 for PCM and DM for case of voice transmission. We assume that f_m=3000 H_z, f_1 = $2Nf_m$= 48 x 10^3 H_z. Using these numbers and resulting that the probability of an error in a bit as $P_{eb} = \frac{1}{2}\operatorname{erfc}\sqrt{S_i/5f_b}$ we have from eq^n (2.94) & (2.95) the result for DM is,

$$\left(\frac{S_0}{N_0}\right)_{DM} = \frac{2457.6}{1+768\,erfc\sqrt{S_i/\eta f_b}} = \frac{2457.6}{1+1536P_e} \qquad (2.98)$$

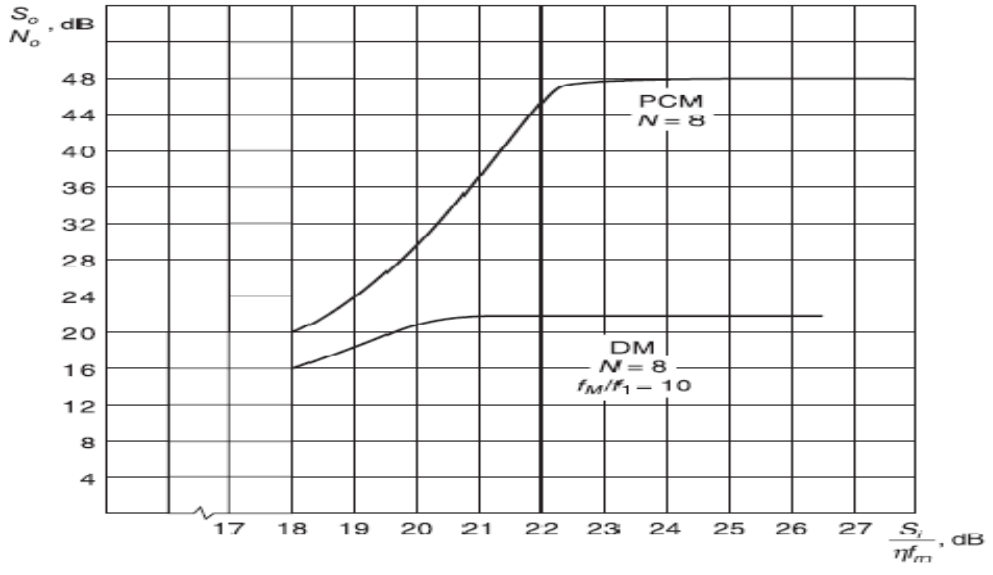


Fig.2.18 A comparison of PCM & DM

And for PCM

$$\left(\frac{S_0}{N_0}\right)_{PCM} = \frac{65,536}{1+131,072 erfc\sqrt{S_i/\eta f_b}} = \frac{65536}{1+262144 P_e} \qquad (2.99)$$

When the probability of bit error is very small, the PCM system is seen to have higher output SNR than the DM system. Indeed the o/p SNR for PCM system is 48 dB and only about 33 dB for DM system. However, an o/p SNR of 30 dB is all that is required in a communication system. Indeed if commanded PCM is employed the o/p SNR will decrease by about 12 dB to 36 dB for PCM system. Thus eq^n (2.99) indicates that the output SNR is higher for PCM system, the output SNR. In practice, can we consider as being comparable.

With regard to the threshold, we see that when $P_e \sim 10^{-6}$ the PCM system has reached threshold with the DM system reaches threshold when $P_e \sim 10^{-4}$. In practice, we find that our ear does not detect threshold P_e is about 10^{-4} for PCM and 10^{-2} for DM and ADM. Some ADM systems can actually produce understandable speech at error rates as high as 10^{-1}-. Fig.2.18 shows a comparison of PCM and DM for N=8 and $f_m/f_1 = 10$.

UNIT 7

Microwave communications

Introduction:-

Satellites offer a number of features not readily available with other means of communications. Because very large areas of the earth are visible from a satellite, the satellite can form the star point of a communications net, simultaneously linking many users who may be widely separated geographically. The same feature enables satellites to provide communications links to remote communities in sparsely populated areas that are difficult to access by other means. Of course, satellite signals ignore political boundaries as well as geographic ones, which may or may not be a desirable feature.

Satellites are also used for remote sensing, examples being the detection of water pollution and the monitoring and reporting of2 Chapter One weather conditions. Some of these remote sensing satellites also form a vital link in search and rescue operations for downed aircraft and the like. Satellites are specifically made for telecommunication purpose. They are used for mobile applications such as communication to ships, vehicles, planes, hand-held terminals and for TV and radio broadcasting. They are responsible for providing these services to an assigned region (area) on the earth. The power and bandwidth of these satellites depend upon the preferred size of the footprint, complexity of the traffic control protocol schemes and the cost of ground stations. A satellite works most efficiently when the transmissions are focused with a desired area. When the area is focused, then the emissions do not go outside that designated area and thus minimizing the interference to the other systems. This leads more efficient spectrum usage.

Satellite's antenna patterns play an important role and must be designed to best cover the designated geographical area (which is generally irregular in shape). Satellites should be designed by keeping in mind its usability for short and long term effects throughout its life time. The earth station should be in a position to control the satellite if it drifts from its orbit it is subjected to any kind of drag from the external forces.

History of Satellite Communications

The first artificial satellite used solely to further advances in global communications was a balloon named Echo 1. Echo 1 was the world's first artificial communications satellite capable of relaying signals to other points on Earth. The first American satellite to relay communications

was Project SCORE in 1958, which used a tape recorder to store and forward voice messages. It was used to send a Christmas greeting to the world from U.S. President Dwight D. Eisenhower. NASA launched the Echo satellite in 1960; the 100-foot (30 m) aluminised PET film balloon served as a passive reflector for radio communications. Courier 1B, built by Philco, also launched in 1960, was the world's first active repeater satellite. The first communications satellite was Sputnik 1. Put into orbit by the Soviet Union on October 4, 1957, it was equipped with an onboard radio-transmitter that worked on two frequencies: 20.005 and 40.002 MHz. Sputnik 1 was launched as a step in the exploration of space and rocket development. While incredibly important it was not placed in orbit for the purpose of sending data from one point on earth to another. And it was the first artificial satellite in the steps leading to today's satellite communications. Telstar was the second active, direct relay communications satellite. Belonging to AT&T as part of a multi-national agreement between AT&T, Bell Telephone Laboratories, NASA, the British General Post Office, and the French National PTT (Post Office) to develop satellite communications, it was launched by NASA from Cape Canaveral on July 10, 1962, the first privately sponsored space launch. Relay 1 was launched on December 13, 1962, and became the first satellite to broadcast across the Pacific on November 22, 1963.Satellites (spacecraft) orbiting the earth follow the same laws that govern the motion of the planets around the sun. From early times much has been learned about planetary motion through careful observations. Johannes Kepler (1571–1630) was able to derive empirically three laws describing planetary motion. Later, in 1665, Sir Isaac Newton (1642–1727) derived Kepler's laws from his own laws of mechanics and developed the theory of gravitation.

Kepler's Laws of Planetary Motion

Kepler's First Law:- *Kepler's first law* states that the path followed by a satellite around the primary will be an ellipse. An ellipse has two focal points shown as *F1* and *F2* in Fig.1.

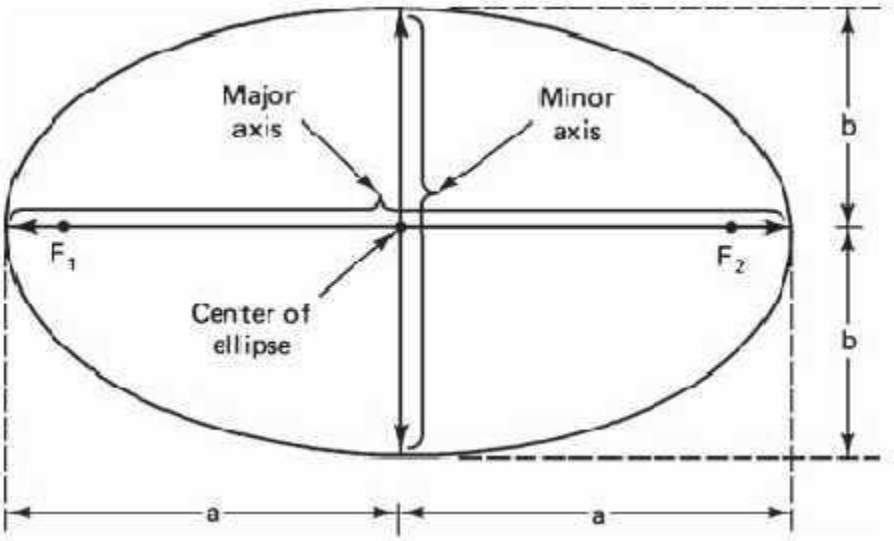


Fig.2.1 The foci *F1* and *F2,* the semimajor axis *a,* and the semiminor axis *b* of an ellips

The foci *F* The eccentricity and the semimajor axis are two of the orbital parameters specified for satellites (spacecraft) orbiting the earth. For an elliptical orbit, $0<e<1$. When $e = 0$, the orbit becomes circular.

Kepler's Second Law:- *Kepler's second law* states that, for equal time intervals, a satellite will sweep out equal areas in its orbital plane, focused at the barycenter. The center of mass of the two-body system, termed the *barycenter,* is always centered on one of the foci.

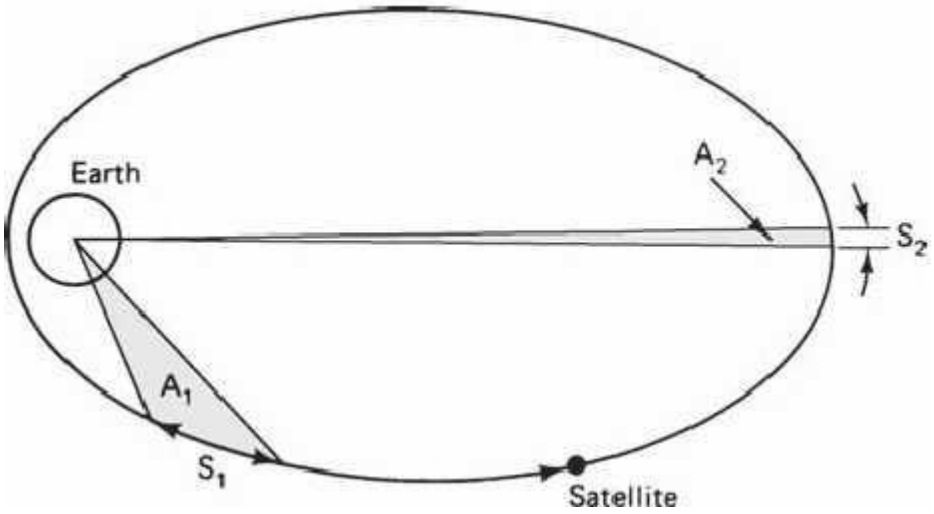


Figure 2.2. Kepler's second law. The areas *A*1 and *A*2 swept out in unit time are equal.

Kepler's Third Law:- *Kepler's third law* states that the square of the periodic time of orbit is proportional to the cube of the mean distance between the two bodies. The mean distance is equal to the semimajor axis *a*. For the

artificial satellites orbiting the earth, Kepler's third law can be written as follows

$$a^3 = \frac{\mu}{n^2} \qquad \ldots (1)$$

where n is the mean motion of the satellite in radians per second and μ is the earth's geocentric gravitational constant.

$$\mu = 3.986005 \times 10^{14} m^3 / s^3 \qquad \ldots (2)$$

The importance of Kepler's third law is that it shows there is a fixed relationship between period and semimajor axis.

Satellite Orbits

There are many different satellite orbits that can be used. The ones that receive the most attention are the geostationary orbit used as they are stationary above a particular point on the Earth. The orbit that is chosen for a satellite depends upon its application. These orbits are given in table 1.Geostationary or geosynchronous earth orbit (GEO)

A satellite in a geostationary orbit appears to be stationary with respect to the earth, hence the name *geostationary*. GEO satellites are synchronous with respect to earth. Looking from a fixed point from Earth, these satellites appear to be stationary. These satellites are placed in the space in such a way that only three satellites are sufficient to provide connection throughout the surface of the Earth. GEO satellite travels eastward at the same rotational speed as the earth in circular orbit with zero inclination.

A geostationary orbit is useful for communications because ground antennas can be aimed at the satellite without their having to track the satellite's motion. This is relatively inexpensive. In applications that require a large number of ground antennas, such as DirectTVdistribution, the savings in ground equipment can more than outweigh the cost and complexity of placing a satellite into orbit.

Table: 1

STELLITE ORBIT NAME	ORBIT	SATELLITE ORBIT ALTITUDE (KM ABOVE EARTH'S SURFACE)	APPLICATION
Low Earth Orbit	LEO	200 - 1200	Satellite phones, Navstar or Global Positioning (GPS) system
Medium Earth Orbit	MEO	1200 - 35790	High-speed telephone signals
Geosynchronous Orbit	GSO	35790	Satellite Television
Geostationary Orbit	GEO	35790	Direct broadcast television

Low Earth Orbit (LEO) satellites

A low Earth orbit (LEO) typically is a circular orbit about 200 kilometres (120 mi) above the earth's surface and, correspondingly, a period (time to revolve around the earth) of about 90 minutes. Because of their low altitude, these satellites are only visible from within a radius of

roughly 1000 kilometers from the sub-satellite point. In addition, satellites in low earth orbit change their position relative to the ground position quickly. So even for local applications, a large number of satellites are needed if the mission requires uninterrupted connectivity. s. LEO systems try to ensure a high elevation for every spot on earth to provide a high qualitycommunication link. Each LEO satellite will only be visible from the earth for around ten minutes.

Low-Earth-orbiting satellites are less expensive to launch into orbit than geostationary satellites and, due to proximity to the ground, do not require as high signal strength (Recall that signal strength falls off as the square of the distance from the source, so the effect is dramatic). Thus there is a trade off between the number of satellites and their cost. In addition, there are important differences in the onboard and ground equipment needed to support the two types of missions. One general problem of LEOs is the short lifetime of about five to eight years due to atmospheric drag and radiation from the inner Van Allen belt1.

Medium Earth Orbit (MEO) satellites

A MEO satellite is in orbit somewhere between 8,000 km and 18,000 km above the earth's surface. MEO satellites are similar to LEO satellites in functionality. MEO satellites are visible for much longer periods of time than LEO satellites, usually between 2 to 8 hours. MEO satellites have a larger coverage area than LEO satellites. A MEO satellite's longer duration of visibility and wider footprint means fewer satellites are needed in a MEO network than a LEO network. One disadvantage is that a MEO satellite's distance gives it a longer time delay and weaker signal than a LEO satellite, though not as bad as a GEO satellite. Due to the larger distance to the earth, delay increases to about 70–80 ms. so these satellites need higher transmit power and special antennas for smaller footprints.

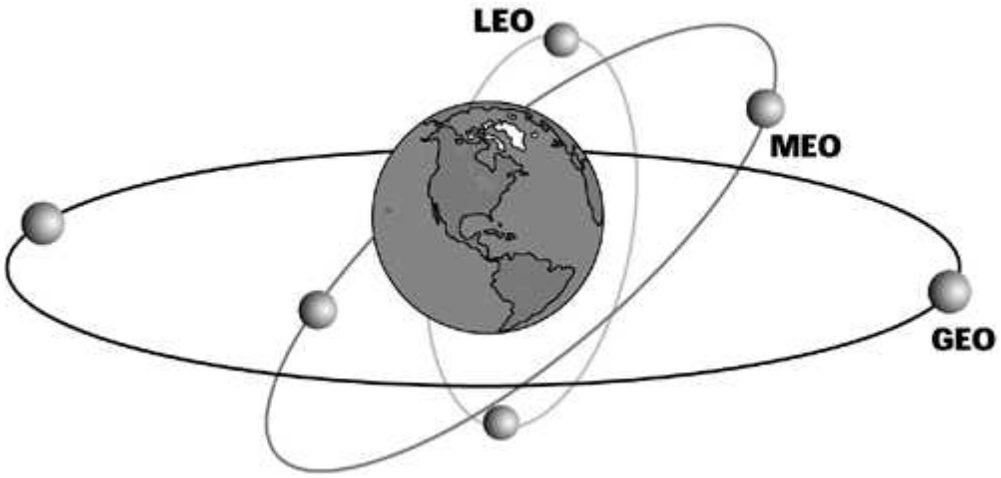


Fig. 2.3 Satellite Orbits

Spacing and Frequency Allocation

Allocating frequencies to satellite services is a complicated process which requires international coordination and planning. This is carried out under the supervision of the *International Telecommunication Union* (ITU). This frequency allocation is done based on different areas. So this world is divided into three areas.

Area 1:- : Europe, Africa, Soviet Union, and Mongolia

Area 2: North and South America and Greenland

Area 3: Asia (excluding area 1 areas), Australia, and the south-west Pacific

Within these regions, frequency bands are allocated to various satellite services, although a given service may be allocated different frequency bands in different regions. Some of the services provided by satellites are:

- Fixed satellite service *(FSS)*

  The FSS provides links for existing telephone networks as well as for transmitting television signals to cable companies for distribution over cable systems. Broadcasting satellite services are intended mainly for direct broadcast to the home, sometimes referred to as *direct broadcast satellite* (DBS) service [in Europe it may be known as *direct-to-home* (DTH) service]. Mobile satellite services would include land mobile, maritime mobile, and aeronautical mobile. Navigational satellite services include *global positioning systems* (GPS), and satellites intended for the meteorological services often provide a search and rescue service.

  TABLE 2: ITU Frequency Band Designations

Band number	Symbols	Frequency range (lower limit exclusive, upper limit inclusive)
4	VLF	3–30 kHz
5	LF	30–300 kHz
6	MF	300–3000 kHz
7	HF	3–30 MHz
8	VHF	30–300 MHz
9	UHF	300–3000 MHz
10	SHF	3–30 GHz
11	EHF	30–300 GHz
12		300–3000 GHz

TABLE 3: Frequency Band Designations

Frequency range, (GHz)	Band designation
0.1–0.3	VHF
0.3–1.0	UHF
1.0–2.0	L
2.0–4.0	S
4.0–8.0	C
8.0–12.0	X
12.0–18.0	Ku
18.0–27.0	K
27.0–40.0	Ka
40.0–75	V
75–110	W
110–300	mm
300–3000	μm

- Broadcasting satellite service *(BSS)*

  Provides Direct Broadcast to homes. E.g. Live Cricket matches etc.

- Mobile satellite services
  - Land Mobile
  - Maritime Mobile
  - Aeronautical mobile
- Navigational satellite services

- Include Global Positioning systems

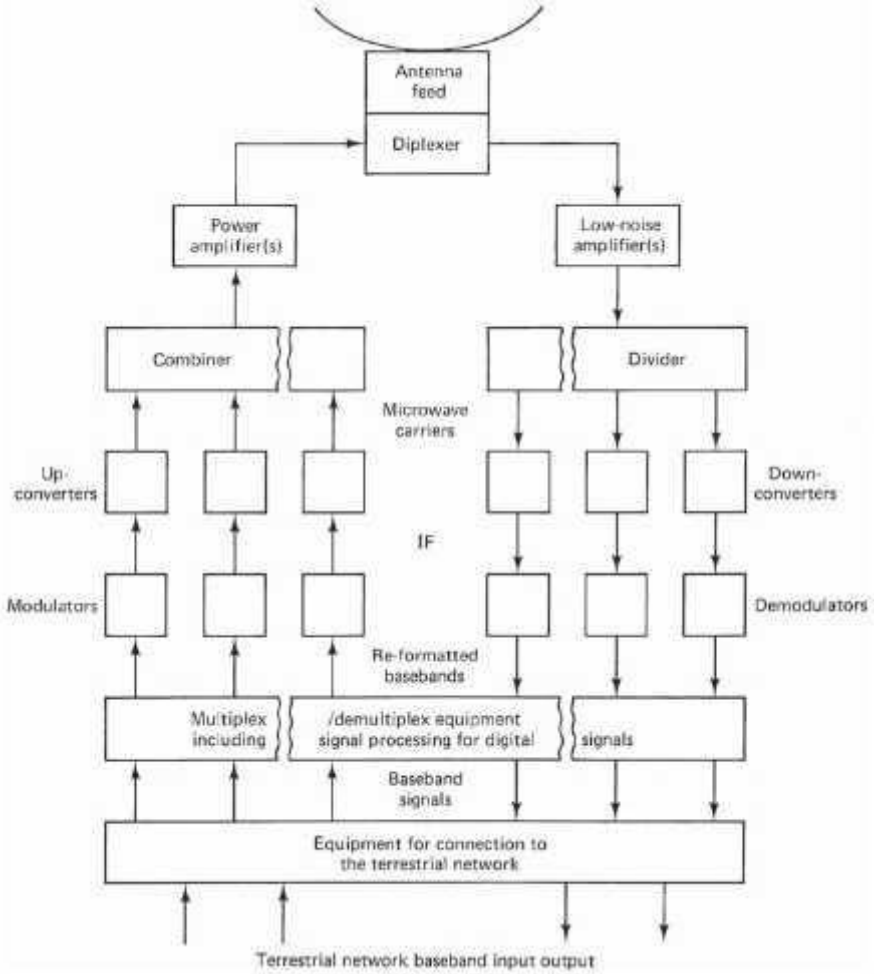
- Meteorological satellite services
  - They are often used to perform Search and Rescue service.

Earth Station

The earth segment of a satellite communications system consists of the transmit and receive earth stations. The station's antenna functions in both, the transmit and receive modes, but at different frequencies.

An earth station is generally made up of a multiplexor, a modem, up and downconverters, a high power amplifier (HPA) and a low noiseamplifier (LNA). Almost all transmission to satellites is d igital, and the digital data streams are combined in a multiplexor and fed to a modemthat modula tes a carrier frequency in the 50 to 180 MHz range. An upconverter bumps the carrier into the gi gahertz range, which goes to the HPA and antenna.

For receiving, the LNA boosts the signals to the downconverter, which lowers the freque ncy and sends itto the modem. The modemdemodulates the carrier, and the digital output goes to the demultiplexing device and then to its destinations. See earth station on board vessel and base station. A detailed block diagram is shown in fig. 3.2.



and dish.

Figur e 3.2:- Block diagram of a transmit-receive earth station

Satellite Sub-systems

A satellite communications system can be broadly divided into two segments—a ground segment and a space segment. The space segment will obviously include the satellites, but it also includes the ground facilities needed to keep the satellites operational, these being referred to as the *tracking, telemetry, and command* (TT&C) facilities. In many networks it is common practice to employ a ground station solely for the purpose of TT&C.

In a communications satellite, the equipment which provides the connecting link between the satellite's transmit and receive antennas is referred to as the *transponder*. The transponder forms one of the main sections of the payload, the other being the antenna subsystems.

PAYLOAD:- The payload comprises of a Repeater and Antenna subsystem and performs the primary function of communication.

1- **REPEATER:-** It is a device that receives a signal and retransmits it to a higher level and/or higher power onto the other side of the obstruction so that the signal can cover longer distance.

2- **Transparent Repeater:-** It only translates the uplink frequency to an appropriate downlink frequency. It does so without processing the baseband signal. The main element of a typical transparent repeater is a single beam satellite. Signals from antenna and the feed system are fed into the low-noise amplifier through a bandpass filter.

3- **Regenerative Repeater :-** A repeater, designed for digital transmission, in which digital signals are amplified, reshaped, retimed, and retransmitted. Regenerative Repeater can also be called as a device which regenerates incoming digital signals and then retransmits these signals on an outgoing circuit.

4- **Antennas :-** The function of an antenna of a space craft is to receive signals and transmit signals to the ground stations located within the coverage area of the satellite. The choice of the antenna system is therefore governed by the size and shape of the coverage area. Consequently, there is also a limit to the minimum size of the antenna footprint.

Satellite System Link Models

System Link Budget calculations basically relate two quantities, the transmit power and the receive power, and show in detail how the difference between these two powers is accounted for. Link-power budget calculations also need the additional losses and noise factor which is incorporated with the transmitted and the received signals. Along with losses, this unit also discusses the system noise parameters. Various components of the system add to the noise in the signal that has to be transmitted.

EQUIVALENT ISOTROPIC RADIATED POWER

The key parameter in link-power budget calculations is the equivalent isotropic radiated power factor, commonly denoted as EIRP. Is the amount of power that a theoretical isotropic antenna (which evenly distributes power in all directions) would emit to produce the peak power density observed in the direction of maximum antenna gain. EIRP can be defined as the power input to one end of the transmission link and the problem to find the power received at the other end.

$$EIRP = G\ Ps$$

Where,
G - Gain of the Transmitting antenna and G is in decibels.
Ps- Power of the sender (transmitter) and is calculated in watts.

$$[EIRP] = [G] + [Ps]\ dBW$$

TRANSMISSION LOSSES:-

As EIRP is thought of as power input of one end to the power received at the other, the problem here is to find the power which is received at the other end. Some losses that occur in the transmitting – receiving process are constant and their values can be pre – determined.

Free-Space Transmission Losses (FSL)

This loss is due to the spreading of the signal in space. Going back to the power flux density equation

$$\psi_m = P_s / 4\pi r^2$$

The power that is delivered to a matched receiver is the power flux density. It is multiplied by the effective aperture of the receiving antenna. Hence, the received power is:

$$P_R = \psi_M A_{eff}$$

$$EIRP \lambda^2 G$$

$$= \frac{R}{4\pi r^2}$$

Where
r- distance between transmitter and receiver, G_R - power gain at the receiver
In decibels, the above equation becomes:

$\left(4\pi r \right)^2$

$$[P_R] = [EIRP] + [G_R] - 10\log\left(\frac{4\pi r}{\lambda}\right)^2$$

$$[FSL] = 10\log\left(\frac{4\pi r}{\lambda}\right)^2$$

$$[P_R] = [EIRP] + [G_R] - [FSL]$$

Feeder Losses (RFL):- This loss is due to the connection between the satellite receiver device and the receiver antenna is improper. Losses here occur is connecting wave guides, filers and couplers. The receiver feeder loss values are added to free space loss.

3.4.2.1 Antenna Misalignment Losses (AML):- To attain a good communication link, the earth station‟s antenna and the communicating satellite‟s antenna must face each other in such a way that the maximum gain is attained.

3.4.2.1 Fixed Atmospheric (AA) and Ionospheric losses (PL):-The gases present in the atmosphere absorb the signals. This kind of loss is usually of a fraction of decibel in quantity. Along with the absorption losses, the ionosphere introduces a good amount of depolarization of signal which results in loss of signal.

Link Equations

The EIRP can be considered as the input power to a transmission link. Due to the above discussed losses, the power at the receiver that is the output can be considered as a simple calculation of EIRP– losses.

Losses = [FSL] + [RFL] + [AML] + [AA] + [PL]

The received power that is P

$$[P_R] = [EIRP] + [G_R] - [Losses]$$

Where;

$[P_R]$ - Received power in dB, [EIRP] - equivalent isotropic radiated power in dBW.

$[G_R]$- Isotropic power gain at the receiver and its value is in dB.

[FSL]-Free-space transmission loss in dB.

[RFL] -Receiver feeder loss in dB.

[AA] -Atmospheric absorption loss in dB.

[AML] -Antenna misalignment loss in dB.

[PL] - Depolarization loss in dB.

UNIT 8
OPTICAL COMMUNICATION SYSTEMS

An **optical fiber** (or **optical fibre**) is a flexible, transparent fiber made of extruded glass (silica) or plastic, slightly thicker than a human hair. It can function as a waveguide, or "light pipe", to transmit light between the two ends of the fiber.The field of applied science and engineering concerned with the design and application of optical fibers is known as **fiber optics**.

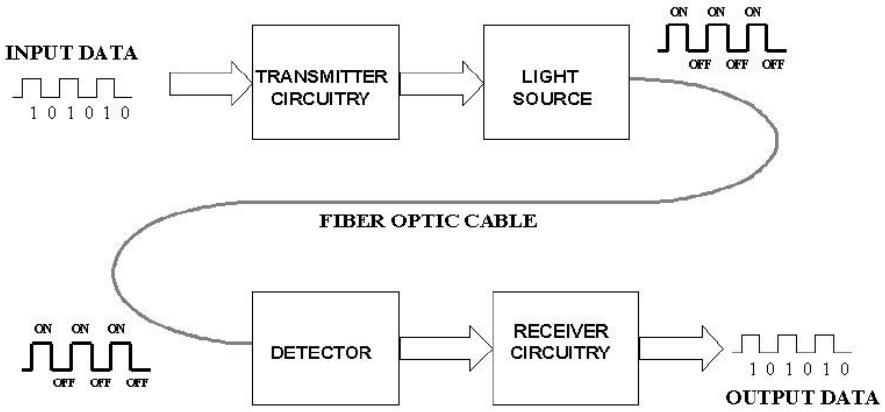
Optical fibers are widely used in fiber-optic communications, where they permit transmission over longer distances and at higher bandwidths (data rates) than wire cables. Fibers are used instead of metal wires because signals travel along them with less loss and are also immune to electromagnetic interference. Fibers are also used for illumination, and are wrapped in bundles so that they may be used to carry images, thus allowing viewing in confined spaces. Specially designed fibers are used for a variety of other applications, including sensors and fiber lasers.

Optical fibers typically include a transparent core surrounded by a transparent cladding material with a lower index of refraction. Light is kept in the core by total internal reflection. This causes the fiber to act as a waveguide. Fibers that support many propagation paths or transverse modes are called multi-mode fibers (MMF), while those that only support a single mode are called single-mode fibers (SMF). Multi-mode fibers generally have a wider core diameter, and are used for short-distance communication links and for applications where high power must be transmitted. Single-mode fibers are used for most communication links longer than 1,000 meters (3,300 ft).

How a Fiber Optic Communication Works?

Unlike copper wire based transmission where the transmission entirely depends on electrical signals passing through the cable, the fiber optics transmission involves transmission of signals in the form of light from one point to the other. Furthermore, a fiber optic communication network consists of transmitting and receiving circuitry, a light source and detector devices like the ones shown in the figure.

When the input data, in the form of electrical signals, is given to the transmitter circuitry, it converts them into light signal with the help of a light source. This source is of LED whose amplitude, frequency and phases must remain stable and free from fluctuation in order to have efficient transmission. The light beam from the source is carried by a fiber optic cable to the destination circuitry wherein the information is transmitted back to the electrical signal by a receiver circuit.



The Receiver circuit consists of a photo detector along with an appropriate electronic circuit, which is capable of measuring magnitude, frequency and phase of the optic field. This type of communication uses the wave lengths near to the infrared band that are just above the visible range. Both LED and Laser can be used as light sources based on the application.

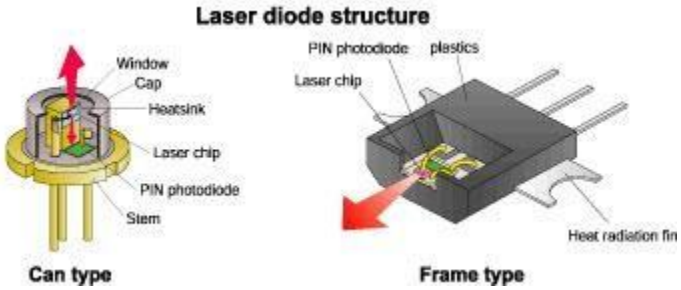
3 Basic Elements of a Fiber Optic Communication System

There are three main basic elements of fiber optic communication system. They are

1. Compact Light Source
2. Low loss Optical Fiber
3. Photo Detector

Accessories like connectors, switches, couplers, multiplexing devices, amplifiers and splices are also essential elements in this communication system.

1. Compact Light Source



Laser Diodes

Depending on the applications like local area networks and the long haul communication systems, the light source requirements vary. The requirements of the sources include power, speed, spectral line width, noise, ruggedness, cost, temperature, and so on. Two components are used as light sources: light emitting diodes (LED's) and laser diodes.

The light emitting diodes are used for short distances and low data rate applications due to their low bandwidth and power capabilities. Two such LEDs structures include Surface and Edge Emitting Systems. The surface emitting diodes are simple in design and are reliable, but due to its broader line width and modulation frequency limitation edge emitting diode are mostly used. Edge emitting diodes have high power and narrower line width capabilities.

For longer distances and high data rate transmission, Laser Diodes are preferred due to its high power, high speed and narrower spectral line width characteristics. But these are inherently non-linear and more sensitive to temperature variations.

LED Versus Laser

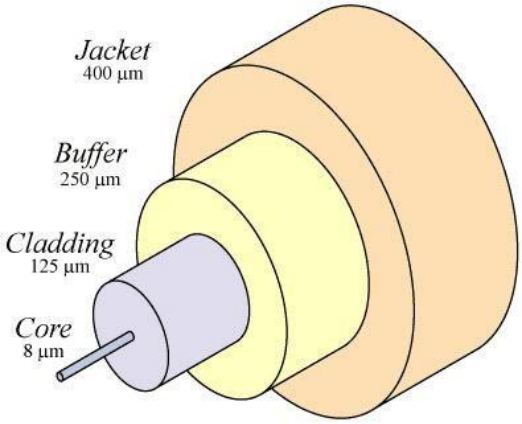
Characteristic	LED	Laser
Output power	Lower	Higher
Spectral width	Wider	Narrower
Numerical aperture	Larger	Smaller
Speed	Slower	Faster
Cost	Less	More
Ease of operation	Easier	More difficult

LED vs Laser Diodes

Nowadays many improvements and advancements have made these sources more reliable. A few of such comparisons of these two sources are given below. Both these sources are modulated using either direct or external modulation techniques.

2. Low Loss Optical Fiber

Optical fiber is a cable, which is also known as cylindrical dielectric waveguide made of low loss material. An optical fiber also considers the parameters like the environment in which it is operating, the tensile strength, durability and rigidity. The Fiber optic cable is made of high quality extruded glass (si) or plastic, and it is flexible. The diameter of the fiber optic cable is in between 0.25 to 0.5mm (slightly thicker than a human hair).



A Fiber Optic Cable consists of four parts.

- Core
- Cladding
- Buffer
- Jacket

Core

The core of a fiber cable is a cylinder of plastic that runs all along the fiber cable's length, and offers protection by cladding. The diameter of the core depends on the application used. Due to internal reflection, the light travelling within the core reflects from the core, the cladding boundary. The core cross section needs to be a circular one for most of the applications.

Cladding

Cladding is an outer optical material that protects the core. The main function of the cladding is that it reflects the light back into the core. When light enters through the core (dense material) into the cladding(less dense material), it changes its angle, and then reflects back to the core.

Buffer

The main function of the buffer is to protect the fiber from damage and thousands of optical fibers arranged in hundreds of optical cables. These bundles are protected by the cable's outer covering that is called jacket.

JACKET

Fiber optic cable's jackets are available in different colors that can easily make us recognize the exact color of the cable we are dealing with. The color yellow clearly signifies a single mode cable, and orange color indicates multimode.

2 Types of Optical Fibers

Single-Mode Fibers: Single mode fibers are used to transmit one signal per fiber; these fibers are used in telephone and television sets. Single mode fibers have small cores.

Multi-Mode Fibers: Multimode fibers are used to transmit many signals per fiber; these signals are used in computer and local area networks that have larger cores.

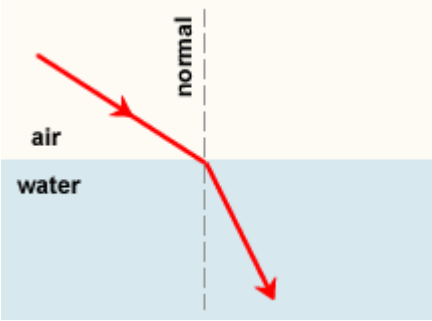
3. Photo Detectors

The purpose of photo detectors is to convert the light signal back to an electrical signal. Two types of photo detectors are mainly used for optical receiver in optical communication system: PN photo diode and avalanche photo diode. Depending on the application's wavelengths, the material composition of these devices vary. These materials include silicon, germanium, InGaAs, etc.

Basic optical laws

Refraction of light

As a light ray passes from one transparent medium to another, it changes direction; this phenomenon is called refraction of light. How much that light ray changes its direction depends on the refractive index of the mediums.



Refractive Index

Refractive index is the speed of light in a vacuum (abbreviated **c**, **c**=299,792.458km/second) divided by the speed of light in a material (abbreviated **v**). Refractive index measures how much a material refracts light. Refractive index of a material, abbreviated as **n**, is defined as

n=c/v

Snell's Law

In 1621, a Dutch physicist named Willebrord Snell derived the relationship between the different angles of light as it passes from one transparent medium to another. When light passes from one transparent material to another, it bends according to Snell's law which is defined as:

$$n_1 sin(\vartheta_1) = n_2 sin(\vartheta_2)$$

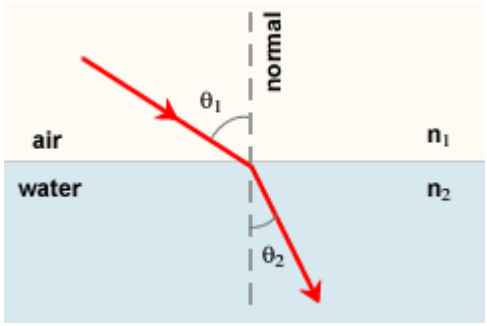
where:

$\boldsymbol{n_1}$ is the refractive index of the medium the light is leaving

$\boldsymbol{\theta_1}$ is the incident angle between the light beam and the normal (normal is 90° to the interface between two materials)

$\boldsymbol{n_2}$ is the refractive index of the material the light is entering

$\boldsymbol{\theta_2}$ is the refractive angle between the light ray and the normal

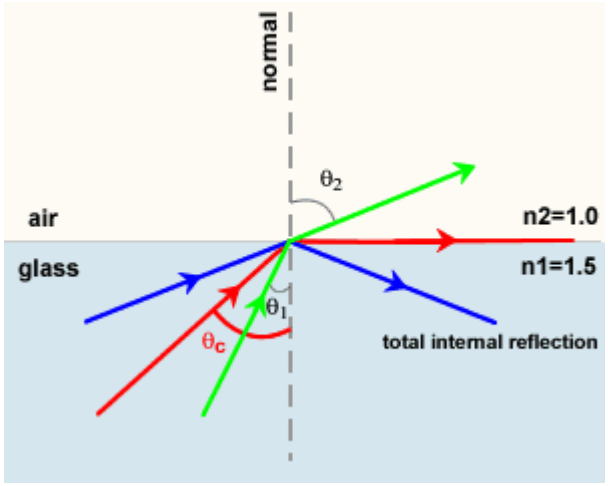


Note:

For the case of $0_1 = 0°$ (i.e., a ray perpendicular to the interface) the solution is $0_2 = 0°$ regardless of the values of n_1 and n_2. That means a ray entering a medium perpendicular to the surface is never bent.

The above is also valid for light going from a dense (higher n) to a less dense (lower n) material; the symmetry of Snell's law shows that the same ray paths are applicable in opposite direction.

Total Internal Reflection



When a light ray crosses an interface into a medium with a higher refractive index, it bends towards the normal. Conversely, light traveling cross an interface from a higher refractive index medium to a lower refractive index medium will bend away from the normal.

This has an interesting implication: at some angle, known as the **critical angle $\mathbf{0_c}$**, light traveling from a higher refractive index medium to a lower refractive index medium will be refracted at 90°; in other words, refracted along the interface.

If the light hits the interface at any angle larger than this critical angle, it will not pass through to the second medium at all. Instead, all of it will be reflected back into the first medium, a process known as **total internal reflection**.

The critical angle can be calculated from Snell's law, putting in an angle of 90° for the angle of the refracted ray $\mathbf{0_2}$. This gives $\mathbf{0_1}$:

$$\mathbf{\theta_1 = \arcsin[(n_2/n_1)*\sin(\theta_2)]}$$

Since

$\mathbf{0_2}$ = 90°

So

$\mathbf{\sin(0_2)}$ = 1

Then

$$0_c = 0_1 = \arcsin(n_2/n_1)$$

For example, with light trying to emerge from glass with n_1=1.5 into air (n_2 =1), the critical angle θ_c is arcsin(1/1.5), or 41.8°.

For any angle of incidence larger than the critical angle, Snell's law will not be able to be solved for the angle of refraction, because it will show that the refracted angle has a sine larger than 1, which is not possible. In that case all the light is totally reflected off the interface, obeying the law of reflection.

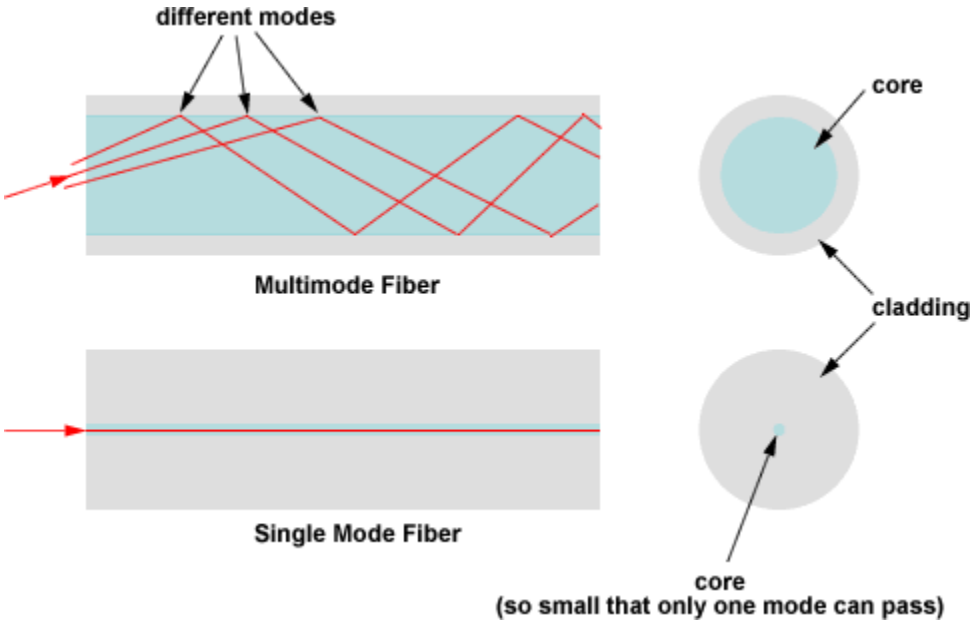
Optical Fiber Mode

What is Fiber Mode?

An optical fiber guides light waves in distinct patterns called *modes*. Mode describes the distribution of light energy across the fiber. The precise patterns depend on the wavelength of light transmitted and on the variation in refractive index that shapes the core. In essence, the variations in refractive index create boundary conditions that shape how light waves travel through the fiber, like the walls of a tunnel affect how sounds echo inside.

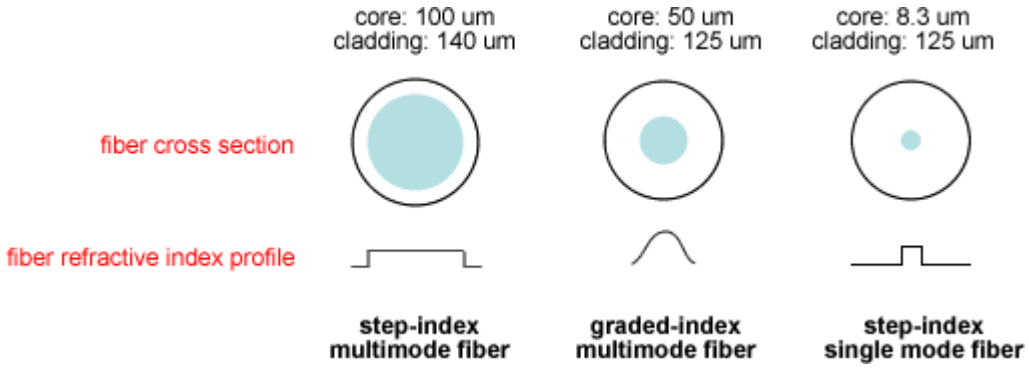
We can take a look at large-core step-index fibers. Light rays enter the fiber at a range of angles, and rays at different angles can all stably travel down the length of the fiber as long as they hit the core-cladding interface at an angle larger than critical angle. These rays are different modes.

Fibers that carry more than one mode at a specific light wavelength are called multimode fibers. Some fibers have very small diameter core that they can carry only one mode which travels as a straight line at the center of the core. These fibers are single mode fibers. This is illustrated in the following picture.



Optical Fiber Index Profile

Index profile is the refractive index distribution across the core and the cladding of a fiber. Some optical fiber has a step index profile, in which the core has one uniformly distributed index and the cladding has a lower uniformly distributed index. Other optical fiber has a graded index profile, in which refractive index varies gradually as a function of radial distance from the fiber center. Graded-index profiles include power-law index profiles and parabolic index profiles. The following figure shows some common types of index profiles for single mode and multimode fibers.



Multimode Fibers

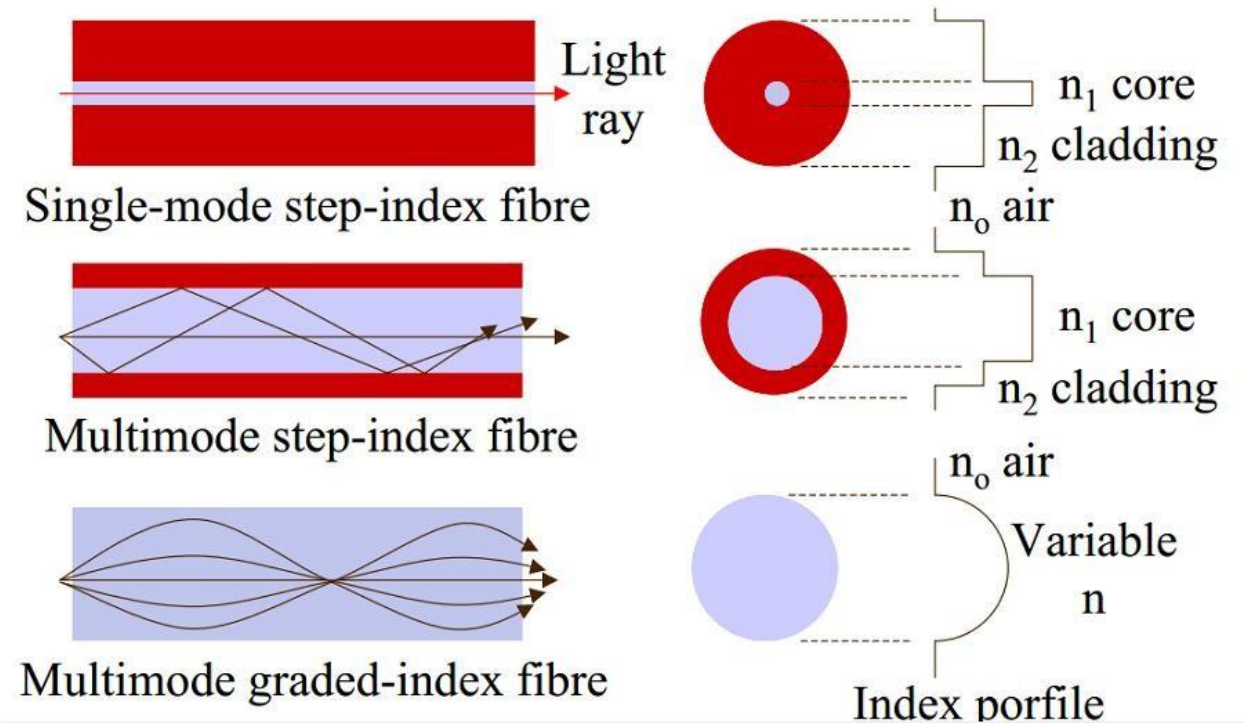
As their name implies, multimode fibers propagate more than one mode. Multimode fibers can propagate over 100 modes. The number of modes propagated depends on the core size and numerical aperture (NA).

As the core size and NA increase, the number of modes increases. Typical values of fiber core size and NA are 50 to 100 micrometer and 0.20 to 0.29, respectively.

Single Mode Fibers

The core size of single mode fibers is small. The core size (diameter) is typically around 8 to 10 micrometers. A fiber core of this size allows only the fundamental or lowest order mode to propagate around a 1300 nanometer (nm) wavelength. Single mode fibers propagate only one mode, because the core size approaches the operational wavelength. The value of the normalized frequency parameter (V) relates core size with mode propagation.

In single mode fibers, V is less than or equal to 2.405. When V = 2.405, single mode fibers propagate the fundamental mode down the fiber core, while highorder modes are lost in the cladding. For low V values (<1.0), most of the power is propagated in the cladding material. Power transmitted by the cladding is easily lost at fiber bends. The value of V should remain near the 2.405 level.



Multimode Step Index Fiber

Core diameter range from 50-1000 m .Light propagate in many different ray paths, or modes, hence the name multimode Index of refraction is same all across the core of the fiber Bandwidth range 20-30 MHz . Multimode Graded Index Fiber The index of refraction across the core is gradually changed from a maximum at the center to a minimum near the edges, hence the name "Graded Index" Bandwidth ranges from 100MHz-Km to 1GHz-Km

Pulse dispersion in a step index optical fiber is given by

$$\text{pulse dispersion} = \frac{\triangle\, n_1\, \ell}{c}$$

where

$\triangle$is the difference in refractive indices of core and cladding.

$n_1$is the refractive index of core

$\ell$is the length of the optical fiber under observation

$c = 3 \times 10^8\,\text{ms}^{-1}$

Graded-Index Multimode Fiber

Contains a core in which the refractive index diminishes gradually from the center axis out toward the cladding. The higher refractive index at the center makes the light rays moving down the axis advance more slowly than those near the cladding. Due to the graded index, light in the core curves helically rather than zigzag off the cladding, reducing its travel distance. The shortened path and the higher speed allow light at the periphery to arrive at a receiver at about the same time as the slow but straight rays in the core axis. The result: digital pulse suffers less dispersion. This type of fiber is best suited for local-area networks.

Pulse dispersion in a graded index optical fiber is given by

$$\text{Pulse dispersion} = \frac{k\delta n \; n_1 \; l}{c},$$

where

δn is the difference in refractive indices of core and cladding,

n_1 is the refractive index of the cladding,

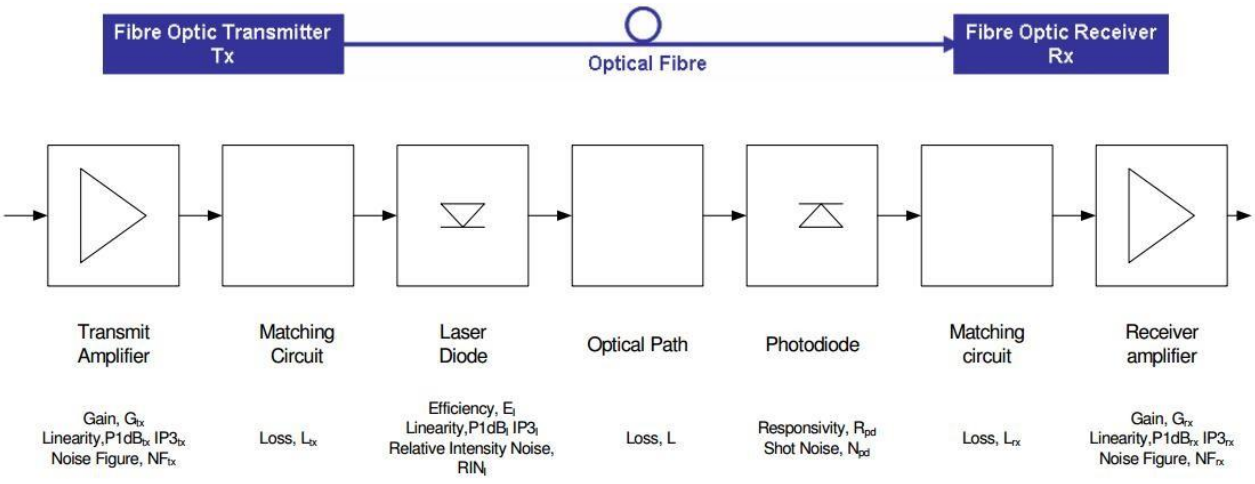
l is the length of the fiber taken for observing the pulse dispersion,

$c \approx 3 \times 10^8 \ \mathrm{m/s}$ is the speed of light, and

k is the constant of graded index profile.

Fibre Optic Link Budget

The FOL budget provides the design engineer with quantitative performance information about the FOL.It is determined by computing the FOL power budget and overall link gain.



Fibre Optic Power Budget

The FOL power budget (PB) is simply the difference between the maximum and minimum signals that the FOL can transport.

Fibre Optic Link Gain

FOL link gain is a summation of gains and losses derived from the different elements of the FOL as shown in above figure . Gains and losses attributed to the Tx, Rx, optical fibre and connectors, as well as any additional in-line components such as splitters, multiplexers, splices etc, must be taken into accounts when computing the linkloss budget.

In the case of a simple point-to-point link described in Above figure , and resistively matched (50 ohms) components,

the link gain (G) is expressed as:-

$G = T + R - 2LO$ (1)

Where T is the gain of the Tx, R is the gain of theRx, and LO is the insertion loss attributed to the fibre link. Note the factor of two in this last optical term, meaning that for each dB optical loss there is a corresponding 2dB RF loss.

To calculate LO the following information is needed.

Standard Corning SMF28 single mode fibre has an insertion loss 0.2dB/km at 1310nm and 0.15dB/km at 1550nm. Optical connectors such as FC/APC typically have an insertion loss of 0.25dB. Optical splices introduce a further 0.25dB loss. Refer to TIA 568 standard forInterfacility and Premise cable specifications.

Output Noise Power

The output noise power of an analogue FOL must alsobe considered when quantifying the overall link budget. The measured output noise power is defined as:-

Output Noise Power = ONF + 10log10 (BW)

Where ONF (Optical Noise Floor) is the noise output of the link on its own, defined in a bandwidth of 1Hz,and BW is the bandwidth of the service transported over fibre. In a real installation, the NF, or Noise Figure is used to define the noise performance of the fibre optic link and is related to the output noise floor as follows:

ONF = -174dBm + NF + G (3)

-174dBm, is the noise contribution from an ideal 1ohm resistive load at zero degrees Kelvin.

The measured output noise power is given as:-

= -174dBm + NF + G + 10log10 (MBW)

ATTENUATION ON OPTICAL FIBER

The signal on optical attenuates due to following mechanisms.

1. Intrinsic loss in the fiber material.
2. Scattering due to micro irregularities inside the fiber.
3. Micro-bending losses due to micro-deformation of the fiber.
4. Bending or radiation losses on the fiber.

The first two losses are intrinsically present in any fiber and the last two depend on the environment in which the fiber is laid.

Material Loss

(a) Due to impurities: The material loss is due to the impurities present in glass used for making fibers. Inspite of best purification efforts, there are always impurities like Fe, Ni, Co, Al which are present in the fiber material. The Fig. shows attenuation due to various molecules inside glass as a function of wavelength. It can be noted from the figure that the material loss due to impurities reduces substantially beyond about 1200nm wavelength.

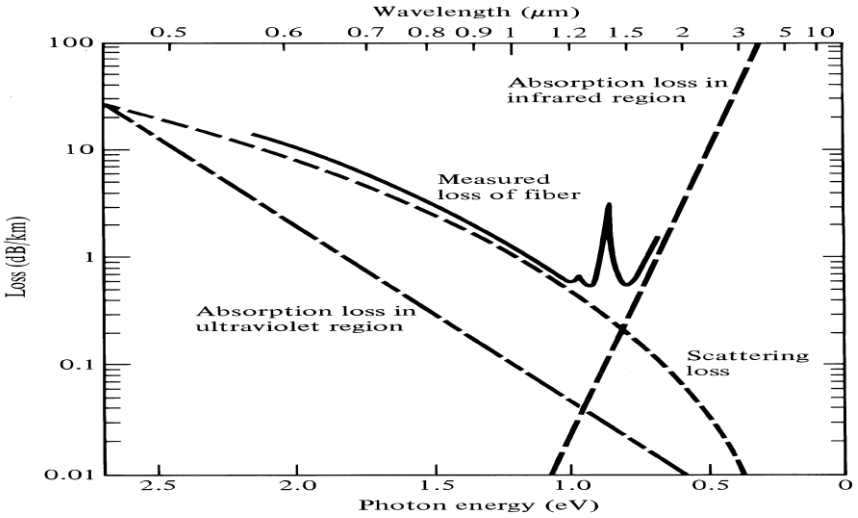
(b)Due to OH molecule: In addition, the OH molecule diffuses in the material and causes absorption of light. The OH molecule has main absorption peak somewhere in the deep infra-red wavelength region. However, it shows substantial loss in the range of 1000 to 2000nm.

(b) Due to infra-red absorption : Glass intrinsically is a good infra-red absorber. As we increase the wavelength the infra-red loss increases rapidly.

SCATTERING LOSS

The scattering loss is due to the non-uniformity of the refractive index inside the core of the fiber. The refractive index of an optical fiber has fluctuation of the order of 10^{-4} over spatial scales much smaller than the optical wavelength. These fluctuations act as scattering centres for the light passing through the fiber. The process is, **Rayleigh Scattering** . A very tiny fraction of light gets scattered and therefore contributes to the loss.

The Rayleigh scattering is a very strong function of the wavelength. The scattering loss varies as Z^{-4}.
This loss therefore rapidly reduces as the wavelength increases. For each doubling of the wavelength, the scattering loss reduces by a factor of 16. It is then clear that the scattering loss at 1550nm is about factor of 16 lower than that at 800nm.
The following Fig. shows the infrared, scattering and the total loss as a function of wavelength.

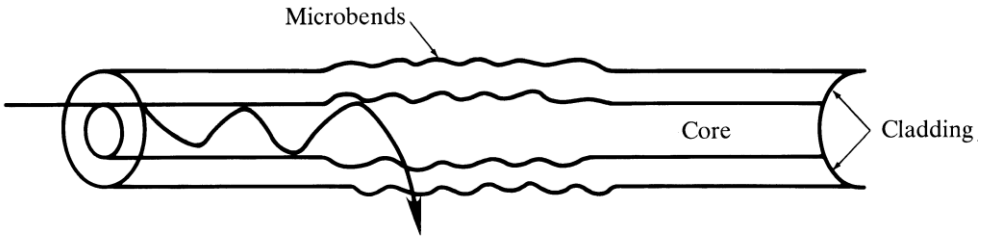


It is interesting to see that in the presence of various losses, there is a natural window in the optical spectrum where the loss is as low as 0.2-0.3dB/Km. This window is from 1200nm to 1600nm.

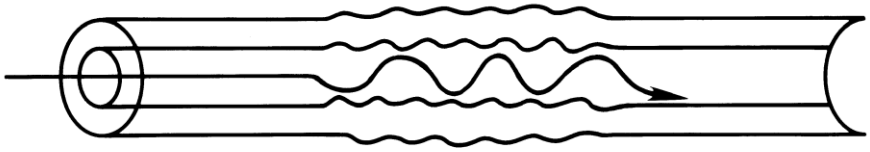
There is a local attenuation peak around 1400nm which is due to OH absorption. The low-loss window therefore is divided into sub-windows, one around 1300nm and other around 1550nm. In fact these are the windows which are the II and III generation windows of optical communication.

MICRO-BENDING LOSSES

While commissioning the optical fiber is subjected to micro-bending as shown in Fig.



Power loss from higher-order modes



Power coupling to higher-order modes

The analysis of micro-bends is a rather complex task. However, just for basic understanding of how the loss takes place due to micro-bending, we use following arguments.

In a fiber without micro-bends the light is guided by total internal reflection (ITR) at the core-cladding boundary. The rays which are guided inside the fiber has incident angle greater than the critical angle at the core-cladding interface. In the presence of micro-bends however, the direction of the local normal to the core-cladding interface deviates and therefore the rays may not have angle of incidence greater than the critical angle and consequently will be leaked out.

A part of the propagating optical energy therefore leaks out due to micro-bends.

Depending upon the roughness of the surface through which the fiber passes, the micro-bending loss varies.

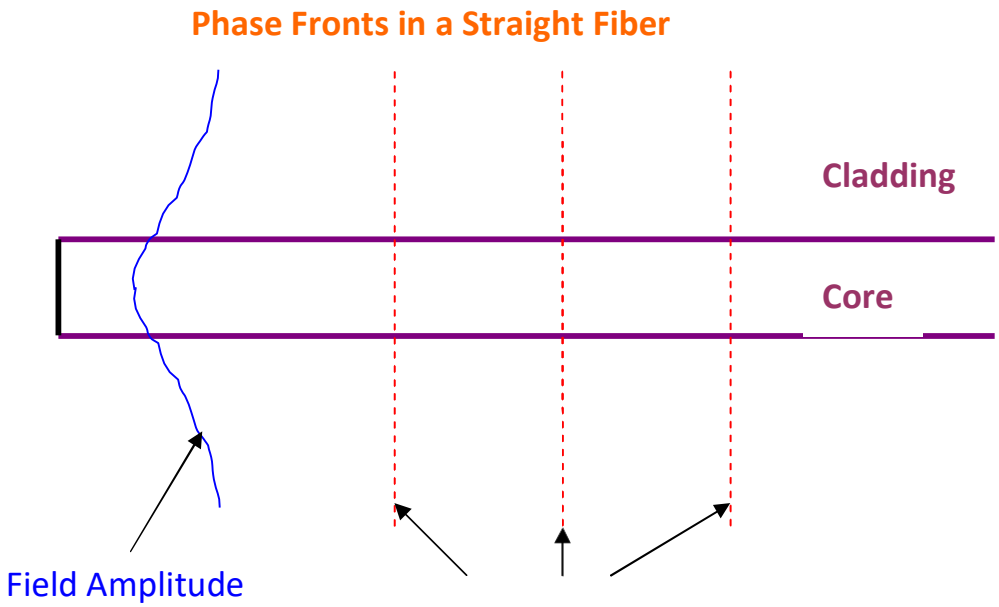
Typically the micro-bends increase the fiber loss by 0.1-0.2 dB/Km.

RADIATION OR BENDING LOSS

While laying the fiber the fiber may undergo a slow bend. In micro-bend the bending is on micron scale, whereas in a slow bend the bending is on cm scale. A typical example of a slow bend is a formation of optical fiber loop.

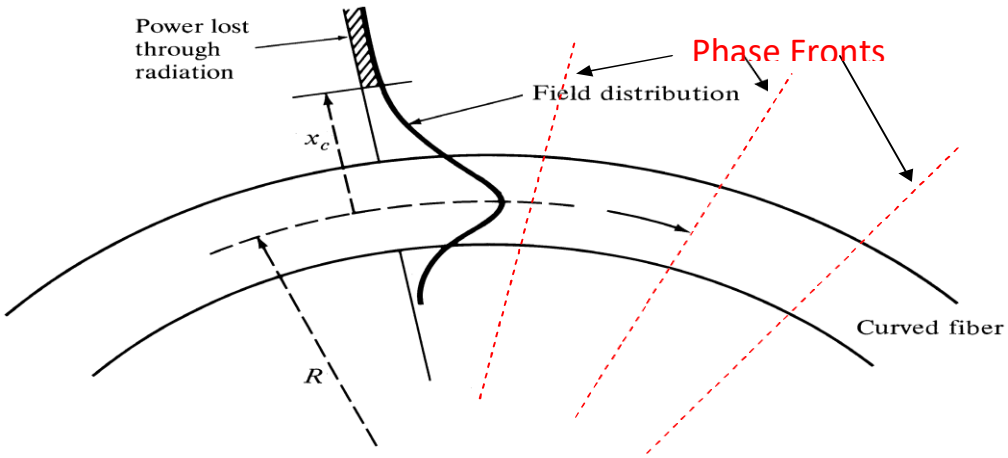
The loss mechanism due to bending loss can be well understood using modal propagation model.

As we have seen, the light inside a fiber propagates in the form of modes. The modal fields decay inside the cladding away from the core cladding interface. Theoretically the field in the cladding is finite no matter how far away we are from the core-cladding interface. Now look at the amplitude and phase distribution for the fibers which are straight and which are bent over an circular arc as shown in Fig.



It can be noted that for the straight the phase fronts are parallel and each point on the phase front travels with the same phase velocity.

Phase Fronts for a Bent Fiber



However, as soon the fiber is bent (no matter how gently) the phase fronts are no more parallel. The phase fronts move like a fan pivoted to the center of curvature of the bent fiber (see Fig.). Every point on the phase front consequently does not move with same velocity. The velocity increases as we move radially outwards the velocity of the phase front increases. Very quickly we reach to a distance x_c from the fiber where the velocity tries to become greater than the velocity of light in the cladding medium.

Since the velocity of energy can not be greater than velocity of light, the energy associated with the modal field beyond x_c gets detached from the mode and radiates away. This is called the bending or the radiation loss.

Phase fronts

Following important things can be noted about the bending loss.

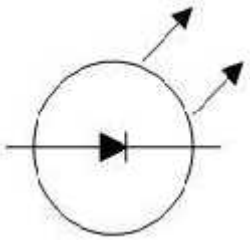
1. The radiation loss is present in every bent fiber no matter how gentle the bend is.
2. Radiation loss depends upon how much is the energy beyond x_c.
3. For a given modal field distribution if x_c reduces, the radiation loss increases. The x_c reduces as the radius of curvature of the bent fiber reduces, that is the fiber is sharply bent.
4. The number of modes therefore reduces in a multimode fiber in presence of bends.

Light Emitting Diodes:-

INTRODUCTION

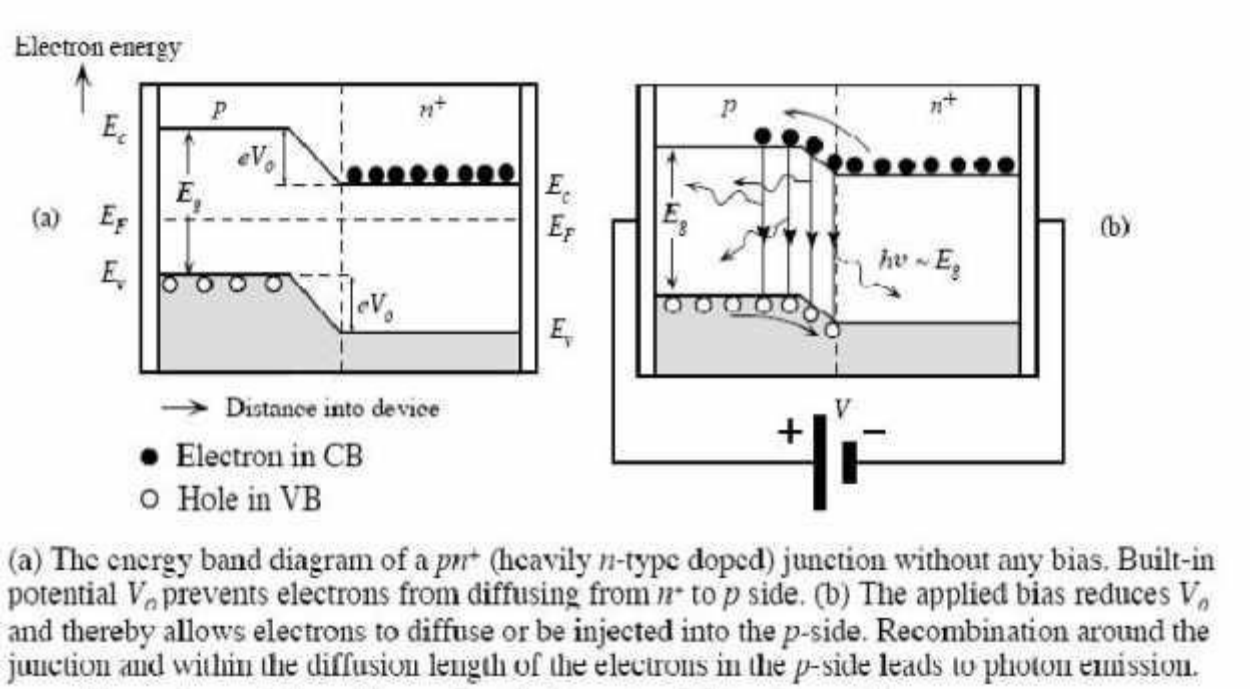
Over the past 25 years the light-emitting diode (LED) has grown from a laboratory curiosity to a broadly used light source for signaling applications . In 1992 LED production reached a level of approximately 25 billion chips , and $2 . 5 billion worth of LED-based components were shipped to original equipment manufacturers .
This article covers light-emitting diodes from the basic light-generation processes todescriptions of LED products . First , we will deal with light-generation mechanisms and light extraction . Four major types of device structures—from simple grown or dif fused homojunctions to complex double heterojunction devices are discussed next , followed by a description of the commercially important semiconductors used for LEDs , from the pioneering GaAsP system to the AlGaInP system that is currently revolutionizing LED technology . Then processes used to fabricate LED chips are explained—the growth of GaAs and GaP substrates ; the major techniques used for growing the epitixal material in which the light-generation processes occur ; and the steps required to create LED chips up to the point of assembly . Next the important topics of quality and reliability—in particular , chip degradation and package-related failure mechanisms—will be addressed . Finally , LED-based products , such as indicator lamps , numeric and alphanumeric displays , optocouplers , fiber-optic transmitters , and sensors , are described . This article covers the mainstream structures , materials , processes , and applications in use today . It does not cover certain advanced structures , such as quantum well or strained layer devices , The reader is also referred to for current information on edge-emitting LEDs , whose fabrication and use are similar to lasers .

Schematic:



Theory:

A Light emitting diode (LED) is essentially a pn junction diode. When carriers are injected across a forward-biased junction, it emits incoherent light. Most of the commercial LEDs are realized using a highly doped n and a p Junction.



(a) The energy band diagram of a pn^+ (heavily n-type doped) junction without any bias. Built-in potential V_0 prevents electrons from diffusing from n^+ to p side. (b) The applied bias reduces V_0 and thereby allows electrons to diffuse or be injected into the p-side. Recombination around the junction and within the diffusion length of the electrons in the p-side leads to photon emission.

Figure 1: p-n+ Junction under Unbiased and biased conditions

To understand the principle, let's consider an unbiased pn+ junction (Figure1 shows the pn+ energy band diagram). The depletion region extends mainly into the p-side. There is a potential barrier from Ec on the n-side to the Ec on the p-side, called the built-in voltage, V0. This potential barrier prevents the excess free electrons on the n+ side from diffusing into the p side. When a Voltage V is applied across the junction, the built-in potential is reduced from V0 to V0 – V. This allows the electrons from the n+ side to get injected into the p-side. Since electrons are the minority carriers in the p-side, this process is called minority carrier injection. But the hole injection from the p side to n+ side is very less and so the current is primarily due to the flow of electrons into the p-side. These electrons injected into the p-side recombine with the holes. This recombination results in spontaneous emission of photons (light). This effect is called injection electroluminescence. These photons should be allowed to escape from the device without being reabsorbed.

The recombination can be classified into the following two kinds
• Direct recombination
• Indirect recombination

Direct Recombination:
In direct band gap materials, the minimum energy of the conduction band lies directly above the maximum energy of the valence band in momentum space energy . In this material, free electrons at the bottom of the conduction band can recombine directly with free holes at the top of the valence band, as the momentum of the two particles is the same. This transition from conduction band to valence band involves photon emission (takes care of the principle of energy conservation). This is known as direct recombination. Direct recombination occurs spontaneously. GaAs is an example of a direct band-gap material.

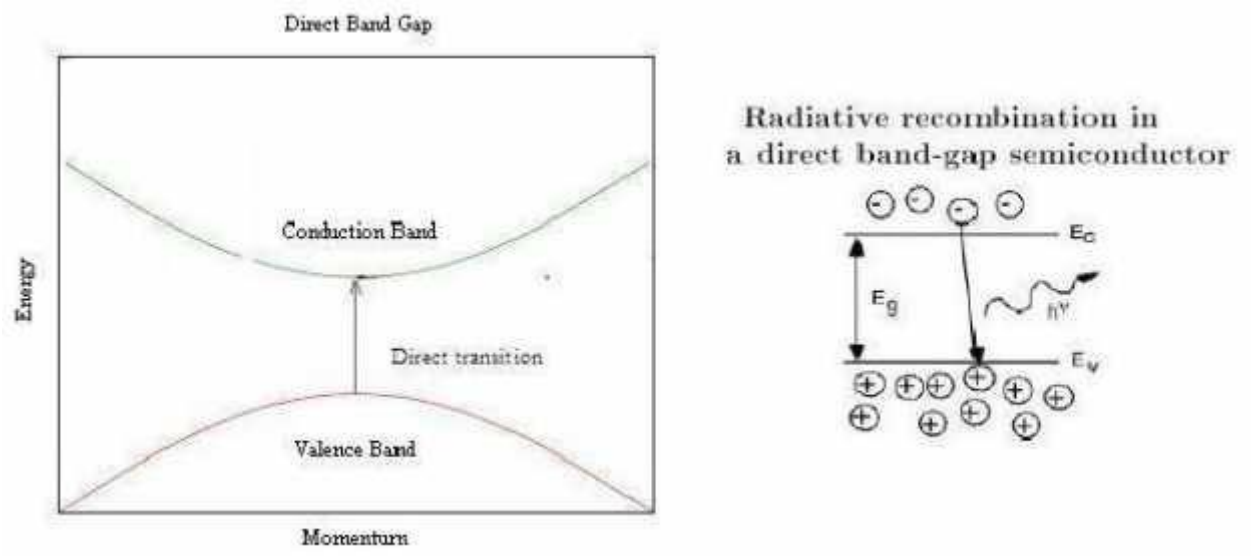


Figure 2: Direct Bandgap and Direct Recombination

Indirect Recombination:

In the indirect band gap materials, the minimum energy in the conduction band is shifted by a k-vector relative to the valence band. The k-vector difference represents a difference in momentum. Due to this difference in momentum, the probability of direct electronhole recombination is less.

In these materials, additional dopants(impurities) are added which form very shallow donor states. These donor states capture the free electrons locally; provides the necessary momentum shift for recombination. These donor states serve as the recombination centers. This is called Indirect (non-radiative) Recombination.

Nitrogen serves as a recombination center in GaAsP. In this case it creates a donor state, when SiC is doped with Al, it recombination takes place through an acceptor level. when SiC is doped with Al, it recombination takes place through an acceptor level.

The indirect recombination should satisfy both conservation energy, and momentum.

Thus besides a photon emission, phonon emission or absorption has to take place.

GaP is an example of an indirect band-gap material.

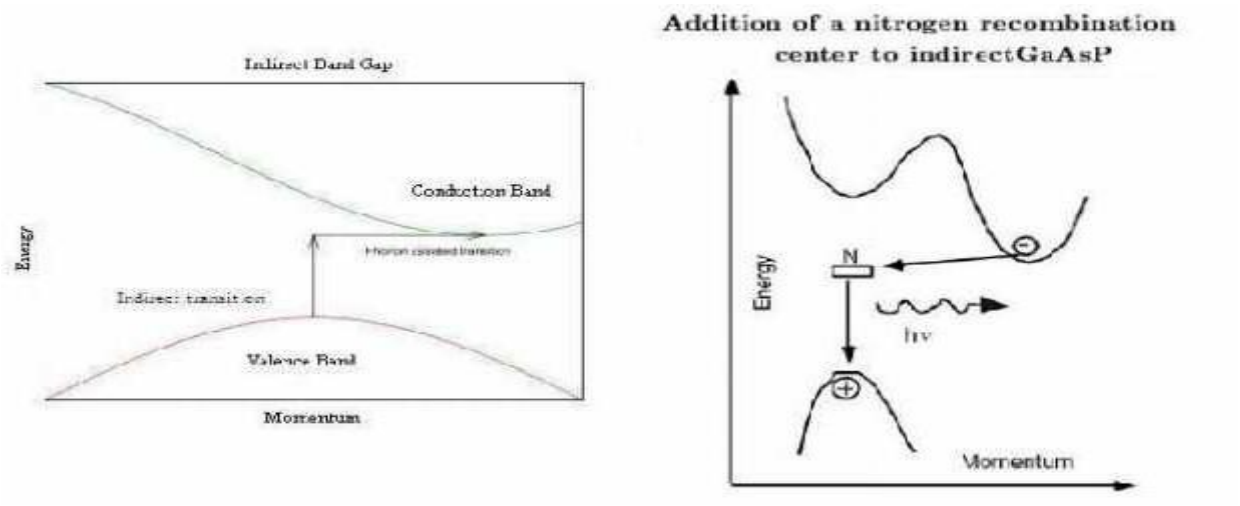


Figure 3: Indirect Bandgap and NonRadiative recombination

The wavelength of the light emitted, and hence the color, depends on the band gap energy of the materials forming the p-n junction.
The emitted photon energy is approximately equal to the band gap energy of the semiconductor. The following equation relates the wavelength and the energy band gap.

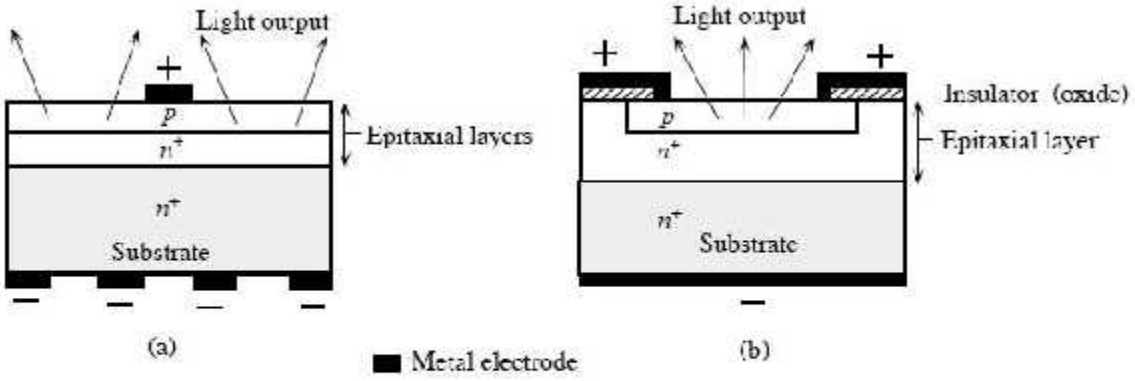
$$h\nu = Eg$$
$$hc/\lambda = Eg$$
$$\lambda = hc/Eg$$

Where h is Plank's constant, c is the speed of the light and Eg is the energy band gap Thus, a semiconductor with a 2 eV band-gap emits light at about 620 nm, in the red. A 3 eV band-gap material would emit at 414 nm, in the violet.

LED Materials:
An important class of commercial LEDs that cover the visible spectrum are the III-V. ternary alloys based on alloying GaAs and GaP which are denoted by GaAs1-yPy. InGaAlP is an example of a quarternary (four element) III-V alloy with a direct band gap. The LEDs realized using two differently doped semiconductors that are the same material is called a homojunction. When they are realized using different bandgap materials they are called a heterostructure device. A heterostructure LED is brighter than a homoJunction LED.
LED Structure:
The LED structure plays a crucial role in emitting light from the LED surface. The LEDs are structured to ensure most of the recombinations takes place on the surface by the following two ways.
• By increasing the doping concentration of the substrate, so that additional free minority charge carriers electrons move to the top, recombine and emit light at the surface.

• By increasing the diffusion length L = √ Dτ, where D is the diffusion coefficient and τ is the carrier life time. But when increased beyond a critical length there is a chance of re-absorption of the photons into the device.

The LED has to be structured so that the photons generated from the device are emitted without being reabsorbed. One solution is to make the p layer on the top thin, enough to create a depletion layer. Following picture shows the layered structure. There are different ways to structure the dome for efficient emitting.



A schematic illustration of typical planar surface emitting LED devices. (a) *p*-layer grown epitaxially on an *n*+ substrate. (b) First *n*+ is epitaxially grown and then *p* region is formed by dopant diffusion into the epitaxial layer.

LED structure

LEDs are usually built on an n-type substrate, with an electrode attached to the p-typelayer deposited on its surface. P-type substrates, while less common, occur as well. Manycommercial LEDs, especially GaN/InGaN, also use sapphire substrate.

LED efficiency:

A very important metric of an LED is the external quantum efficiency ηext. It quantifies the efficeincy of the conversion of electrical energy into emitted optical energy. It is defined as the light output divided by the electrical input power. It is also defined as the product of Internal radiative efficiency and Extraction efficiency.

ηext = Pout(optical) / IV

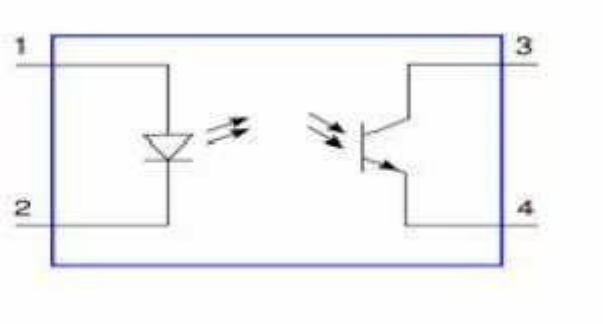
For indirect bandgap semiconductors ηext is generally less than 1%, where as for a direct band gap material it could be substantial.

ηint = rate of radiation recombination/ Total recombination

The internal efficiency is a function of the quality of the material and the structure and composition of the layer.

Applications: LED have a lot of applications. Following are few examples.

• Devices, medical applications, clothing, toys
• Remote Controls (TVs, VCRs)
• Lighting
• Indicators and signs
• Optoisolators and optocouplers
• Swimming pool lighting



Optocoupler schematic showing LED and phototransistor

Advantages of using LEDs:

• LEDs produce more light per watt than incandescent bulbs; this is useful inbattery powered or energy-saving devices.

• LEDs can emit light of an intended color without the use of color filters that traditional lighting methods require. This is more efficient and can lower initial costs.

• The solid package of the LED can be designed to focus its light. Incandescent and fluorescent sources often require an external reflector to collect light and direct itin a usable manner.

• When used in applications where dimming is required, LEDs do not change their color tint as the current passing through them is lowered, unlike incandescent lamps, which turn yellow.

• LEDs are ideal for use in applications that are subject to frequent on-off cycling, unlike fluorescent lamps that burn out more quickly when cycled frequently, or High Intensity Discharge (HID) lamps that require a long time before restarting.

• LEDs, being solid state components, are difficult to damage with external shock.Fluorescent and incandescent bulbs are easily broken if dropped on the ground.

• LEDs can have a relatively long useful life. A Philips LUXEON k2 LED has a life time of about 50,000 hours, whereas Fluorescent tubes typically are rated at about 30,000 hours, and incandescent light bulbs at 1,000–2,000 hours.

• LEDs mostly fail by dimming over time, rather than the abrupt burn-out of incandescent bulbs.

• LEDs light up very quickly. A typical red indicator LED will achieve full brightness in microseconds; Philips Lumileds technical datasheet DS23 for the Luxeon Star states "less than 100ns." LEDs used in communications devices can have even faster response times.

• LEDs can be very small and are easily populated onto printed circuit boards.

• LEDs do not contain mercury, unlike compact fluorescent lamps.

Disadvantages:

• LEDs are currently more expensive, price per lumen, on an initial capital cost basis, than more conventional lighting technologies. The additional expense partially stems from the relatively low lumen output and the drive circuitry and power supplies needed. However, when considering the total cost of ownership (including energy and maintenance costs), LEDs far surpass incandescent or halogen sources and begin to threaten the future existence of compact fluorescent lamps.

• LED performance largely depends on the ambient temperature of the operating environment. Over-driving the LED in high ambient temperatures may result in overheating of the LED package, eventually leading to device failure. Adequate heat-sinking is required to maintain long life .

• LEDs must be supplied with the correct current. This can involve series resistors or current-regulated power supplies.

• LEDs do not approximate a "point source" of light, so they cannot be used in applications needing a highly collimated beam. LEDs are not capable of providing divergence below a few degrees. This is contrasted with commercial ruby lasers with divergences of 0.2 degrees or less. However this can be corrected by using lenses and other optical devices.

Laser diodes:-

Laser diodes (also called .injection lasers.) are in effect anspecialised form of LED. Just like a LED, they.re a form of P-N junction diode with a thin depletion layer where electrons and holes collide to create light photons, when the diode is forward biased.

The difference is that in this case the .active. part of the depletion layer (i.e., where most of the current flows) is made quite narrow, to concentrate the carriers. The endsof this narrow active region are also highly polished, or coated with multiple very thin reflective layers to act as mirrors, so it forms a resonant optical cavity.

The forward current level is also increased, to the point where the current density reaches a critical level where carrier population inversion. occurs. This means there are more holes than electrons in the conduction band, and more electrons than holes in the valence band . or in other words, a very large excess population of electrons and holes which can potentially combine to release photons. And when this happens, the creation of new photons can be triggered not just by random collisions of electrons and holes, but lso by the influence of passing photons. Passing photons are then able to stimulate the production of more photons, without themselves being absorbed. So laser action is able to occur: Light Amplification by Stimulated Emission of Radiation. And the important thing to realise is that the photons that are triggered by other passing photons have the same wavelength, and arealso in phase with them. In other words, they end up .in sync. and forming continuous-wave coherent radiation.

Because of the resonant cavity, photons are thus able to travel back and forth from one end of the active region to the other, triggering the production of more and more photons in sync with themselves. So quite a lot of coherent light energy is generated.

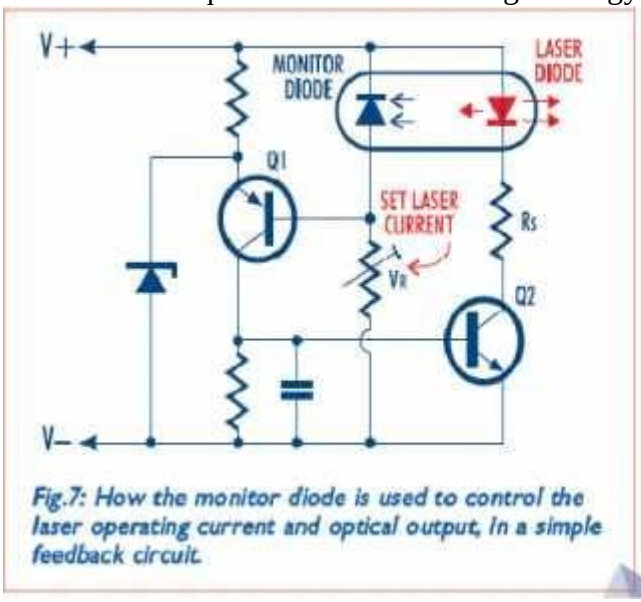


Fig.7: How the monitor diode is used to control the laser operating current and optical output, In a simple feedback circuit.

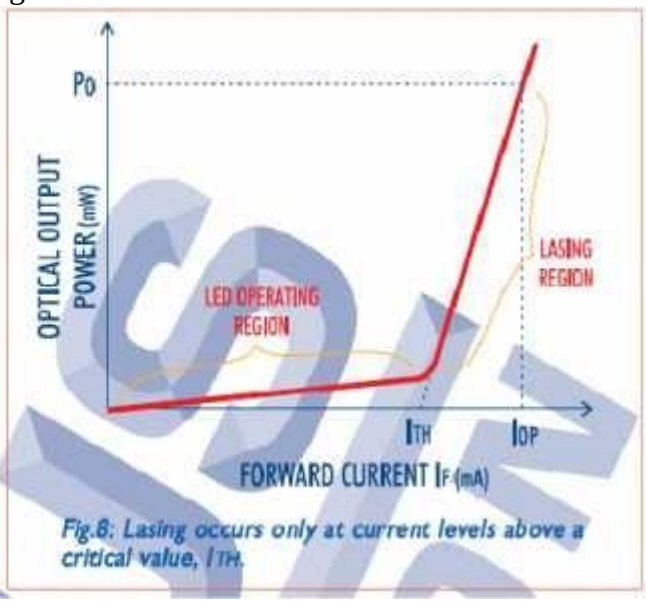


Fig.8: Lasing occurs only at current levels above a critical value, ITH.

And as the ends of the cavity are not totally reflective (typically about 90-95%), some of this coherent light can leave the laser chip . to form its output beam.

Because a laser.s light output is coherent, it is very low in noise and also more suitable for use as a .carrier. for data communications. The bandwidth also tends to be narrower and better defined

than LEDs, making them more suitable for optical systems where light beams need to be separated or manipulated on the basis of wavelength.
The very compact size of laser diodes makes them very suitable for use in equipment like CD, DVD and MiniDisc players and recorders. As their light is reasonably well collimated (although not as well as gas lasers) and easily focussed, they.re also used in optical levels, compact handheld laser pointers, barcode scanners etc. There are two main forms of laser diode: the horizontal type, which emits light from the polished ends of the chip, and the vertical or .surface emitting. type. They both operate in the way just described, differing mainly in terms of the way the active light generating region and resonant cavity are formed inside the chip. Because laser diodes have to be operated at such a high current density, and have a very low forward resistance when lasing action occurs, they are at risk of destroying themselves due to thermal runaway. Their operating light density can also rise to a level where the end mirrors can begin melting. As a result their electrical operation must be much more carefully controlled than a LED. This means that not only must a laser diode.s current be regulated by a .constant current. circuit rather than a simple series resistor, but optical negative feedback must generally be used as well . to ensure that the optical output is held to a constant safe level.
To make this optical feedback easier, most laser diodes have a silicon PIN photodiode built right into the package, arranged so that it automatically receives a fixed proportion of the laser.s output. The output of this monitor diode can then be used to control the current fed through the laser by the constant current circuit, for stable and reliable operation. Fig.6 shows a typical .horizontal. type laser chip mounted in its package, with the monitor photodiode mounted on the base flange below it so the diode receives the light output from the .rear. of the laser chip.
Fig.7 (page 3) shows a simple current regulator circuit used to operate a small laser diode, and you can see how the monitor photodiode is connected. The monitor diode is shunting the base forward bias for transistor Q1, which has its emitter voltage fixed by the zener diode. So as the laseroutput rises, the monitor diode current increases, reducing the conduction of Q1 and hence that of transistor Q2, which controls the laser current. As a result, the laser current is automatically stabilised to a level set by adjustable resistor VR.

Laser diode parameters

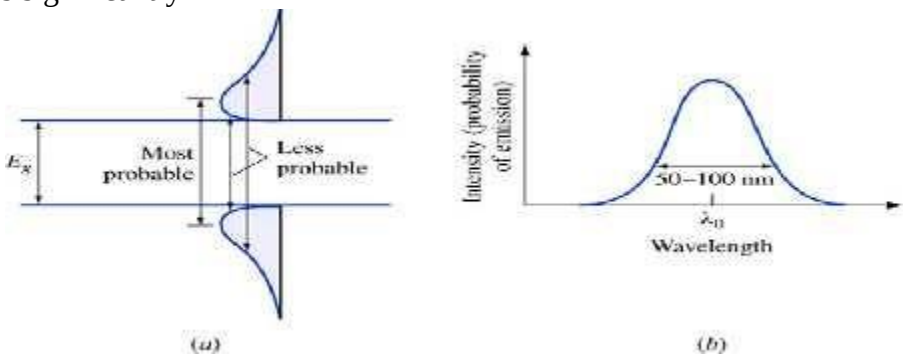
Perhaps the key parameter for a laser diode is the threshold current (ITH), which is the forward current level where lasing actually begins to occur. Below that current level the device delivers some light output, but it operates only as a LED rather than a laser. So the light it does produce in this mode is incoherent. Another important parameter is the rated light output (Po), which is the highest recommended light output level (in milliwatts) for reliable continuous operation. Not surprisingly there.s an operating current level (IOP) which corresponds to this rated light output (Fig.8). There.s also the corresponding current output from the feedback photodiode, known as the monitor current level (Im). Other parameters usually given for a laser diode are its peak lasing wavelength, using given in nanometres (nm); and its beam divergence angles (defined as the angle away from the beam axis before the light intensity drops to 50%), in the X and Y directions (parallel to, and normal to the chip plane).

Laser safety

Although most of the laser diodes used in electronic equipment have quite low optical output levels . typically less than 5mW (milliwatts) . their output is generally concentrated in a relatively narrow beam. This means that it is still capable of causing damage to a human or animal eye, and particularly to its light-sensitive retina.

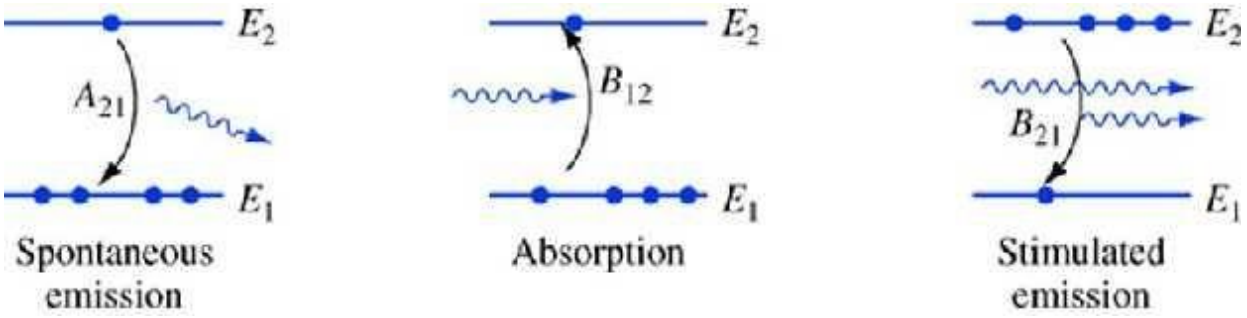
Infra-red (IR) lasers are especially capable of causing eye damage, because their light is not visible. This prevents the eye.s usual protective reflex mechanisms (iris contraction, eyelid closure) from operating. So always take special care when using devices like laser pointers, and especially when working on equipment which includes IR lasers, to make sure that the laser beam cannot enter either your own, or anyone else.s eyes. If you need to observe the output from a laser, either use protective filter goggles or use an IR-sensitive CCD type video camera. Remember that eye damage is often irreversible, especially when it.s damage to the retina.

•Light Emitting Diode
•Light is mostly monochromatic (narrow energy spread comparable to the distribution of electrons/hole populations in the band edges)
•Light is from spontaneous emission (random events in time and thus phase).
•Light diverges significantly



LASER
•Light is essentially single wavelength (highly monochromatic)
•Light is from "stimulated emission" (timed to be in phase with other photons
•Light has significantly lower divergence (Semiconductor versions have more than gas lasers though).

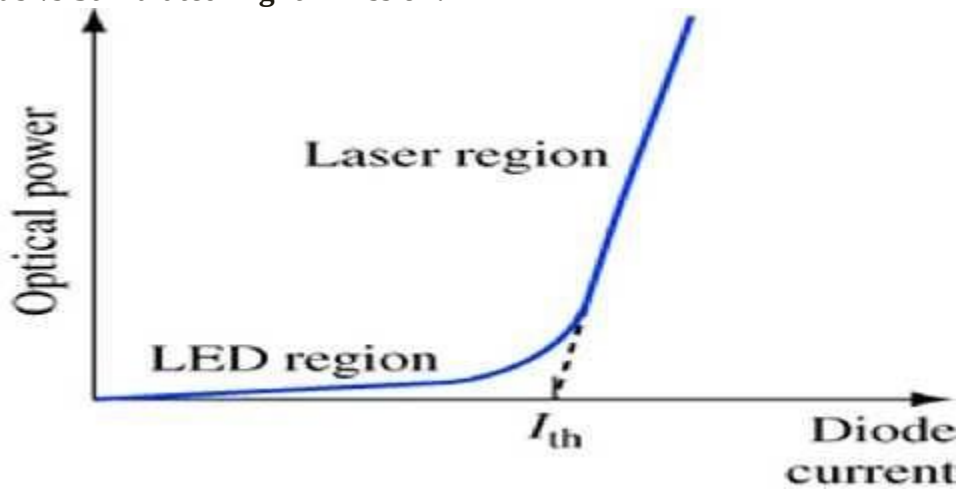
Spontaneous Light Emission



• We can add to our understanding of absorption and spontaneous radiation due to random recombination another form of radiation – Stimulated emission.
• Stimulated emission can occur when we have a "population inversion", i.e. when we have injected so many minority carriers that in some regions there are more "excited carriers" (electrons) than "ground state" carriers (holes).
• Given an incident photon of the band gap energy, a second photon will be "stimulated" by the first photon resulting in two photons with the same energy (wavelength) and phase.

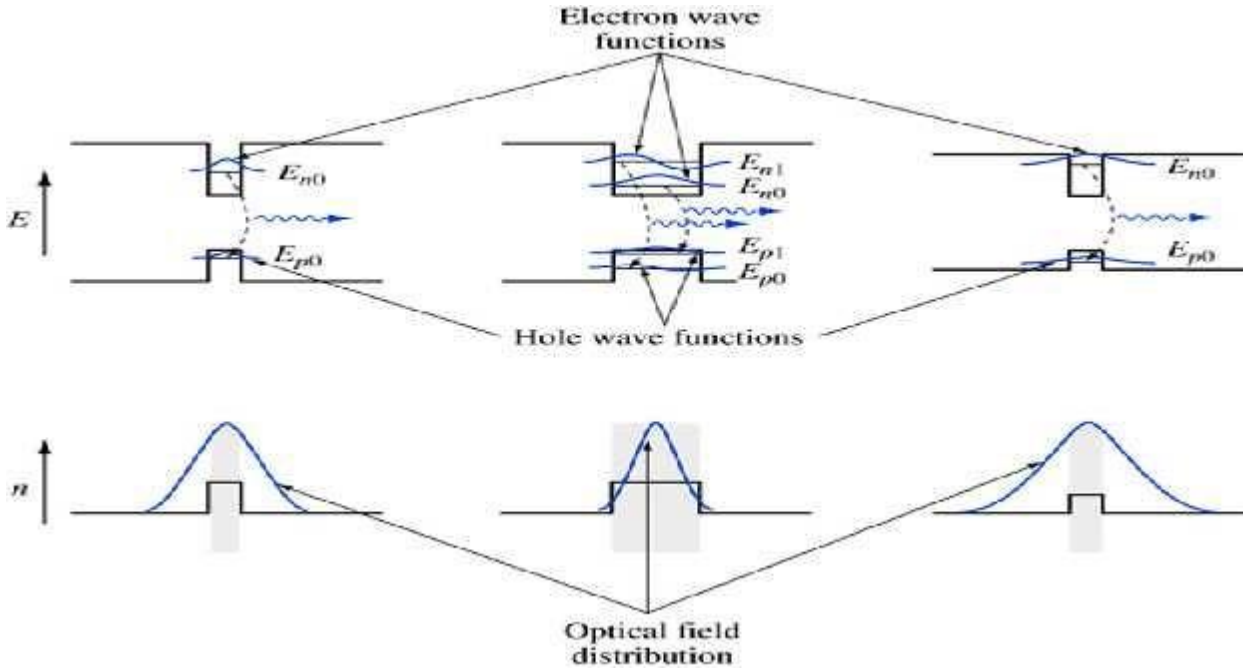
• **This phase coherence results in minimal divergence of the optical beam resulting in a directed light source.**

Spontaneous vs Stimulated Light Emission:



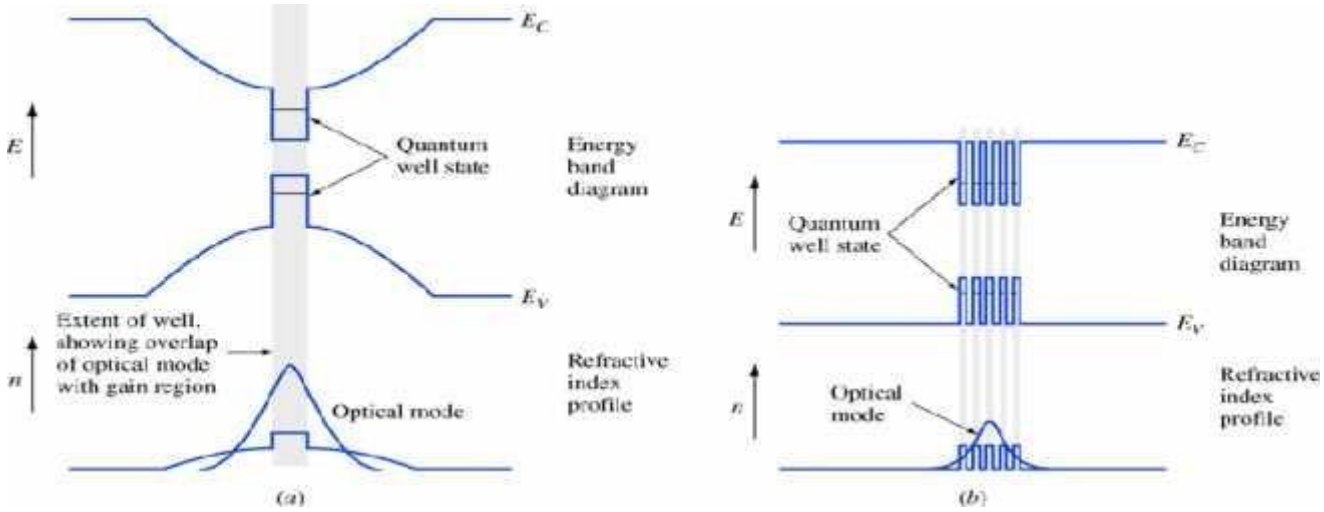
The power-current curve of a laser diode. Below threshold, the diode is an LED. Above threshold, the population is inverted and the light output increases rapidly.

LASER Wavelength Design:



Adjusting the depth and width of quantum wells to select the wavelength of emission is one form of band-gap engineering. The shaded areas indicate the width of the well to illustrate the degree of confinement of the mode.

Advanced LASER Wavelength Design:



(a) A GRINSCH structure helps funnel the carriers into the wells to improve the probability of recombination. Additionally, the graded refractive index helps confine the optical mode in the nearwell region. Requires very precise control over layers due to grading. Almost always implemented via MBE

(b) A multiple quantum well structure has improves carrier capture.

Sometimes the two are combined to give a "digitally graded" device where only two compositions are used but the well thicknesses are varied to implement an effective "index grade"

Photodetectors:-

These are **Opto-electric devices** i.e. to convert the optical signal back into electrical impulses.

The light detectors are commonly made up of semiconductor material.

When the light strikes the light detector a current is produced in the external circuit proportional to the intensity of the incident light.

Optical signal generally is **weakened** and distorted when it emerges from the end of the fiber,

the photodetector must meet following strict performance requirements.

A **high sensitivity** to the emission wavelength range of the received light signal.

A **minimum** addition of **noise** to the signal.

A **fast response** speed to handle the desired data rate.

Be **insensitive** to **temperature** variations.

Be **compatible** with the physical dimensions of the **fiber.**

Have a **Reasonable cost** compared to other system components.

Have a long **operating lifetime.**

Some important parameters while discussing photodetectors:

Quantum Efficiency

It is the ratio of primary electron-hole pairs created by incident photon to the photon incident on the diode material.

Detector Responsivity

This is the ratio of output current to input optical power. Hence this is the efficiency of the device.

Spectral Response Range

This is the range of wavelengths over which the device will operate.

Types of Light Detectors

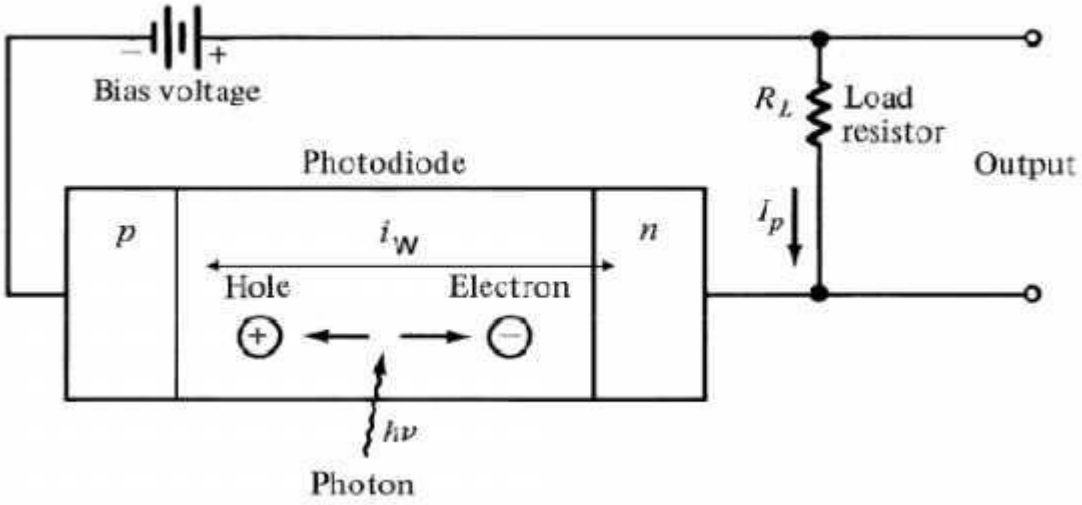
_ PIN Photodiode
_ Avalanche Photodiode

The Pin Photodetector:-

The **device structure** consists of **p and n** semiconductor regions separated by a very **lightly n-doped intrinsic (i) region.**

In **normal operation** a reverse-bias voltage is applied across the device so that **no free electrons or holes** exist in the **intrinsic region**.

Incident photon having energy **greater than or equal** to the **bandgap energy** of the semiconductor material, **give up itsenergy** and **excite an electron** from the valence band to the conduction band.



The high electric field present in the depletion region causes photogenerated carriers to separate and be collected across the reverse – biased junction. This gives rise to a current flow in an external circuit, known as **photocurrent**.

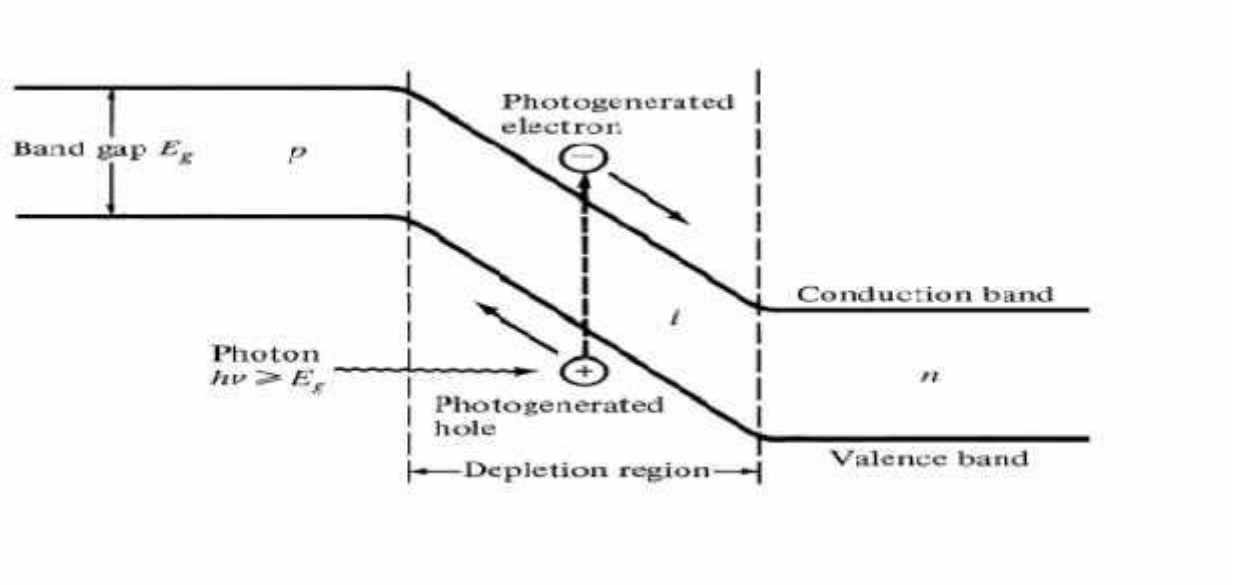
Photocarriers:

Incident photon, generates free (mobile) **electron-hole pairs in the intrinsic region**. These charge carriers are known as **photocarriers,** since they are generated by a photon.

Photocurrent:

The electric field across the device causes the **photocarriers to be swept out of the intrinsic region**, thereby giving rise to a **current flow in an external circuit**. This current flow is known as the **photocurrent.**

Energy-Band diagram for a *pin* photodiode:



An incident photon is able to boost an electron to the conduction band only if it has an energy that is greater than or equal to the bandgap energy

Thus, a particular semiconductor material can be used only over a limited wavelength range.

$$\lambda_c = \frac{hc}{E_g}$$

As the charge carriers flow through the material some of them recombine and disappear.

The charge carriers move a distance Ln or Lp for electrons and holes before recombining. This distance is known as diffusion length

The time it take to recombine is its life time _n or _p respectively.

$$L_n = (D_n \tau_n)^{1/2} \quad \text{and} \quad L_p = (D_p \tau_p)^{1/2}$$

Where Dn and Dp are the diffusion coefficients for electrons and holes respectively.

Photocurrent:-

As a photon flux penetrates through the semiconductor, it will be absorbed.

If *P*in is the optical power falling on the photo detector at x=0 and $P(x)$ is the power level at a distance x into the material then the incremental change be given as

$$dP(x) = -\alpha_s(\lambda)P(x)dx$$

where _s(_) is the photon absorption coefficient at a wavelength _. So that

$$P(x) = P_{in}\exp(-\alpha_s x)$$

Optical power absorbed, $P(x)$, in the depletion region can be written in terms of incident optical power, *Pin* :

$$P(x) = P_{in}(1 - e^{-\alpha_s(\lambda)x})$$

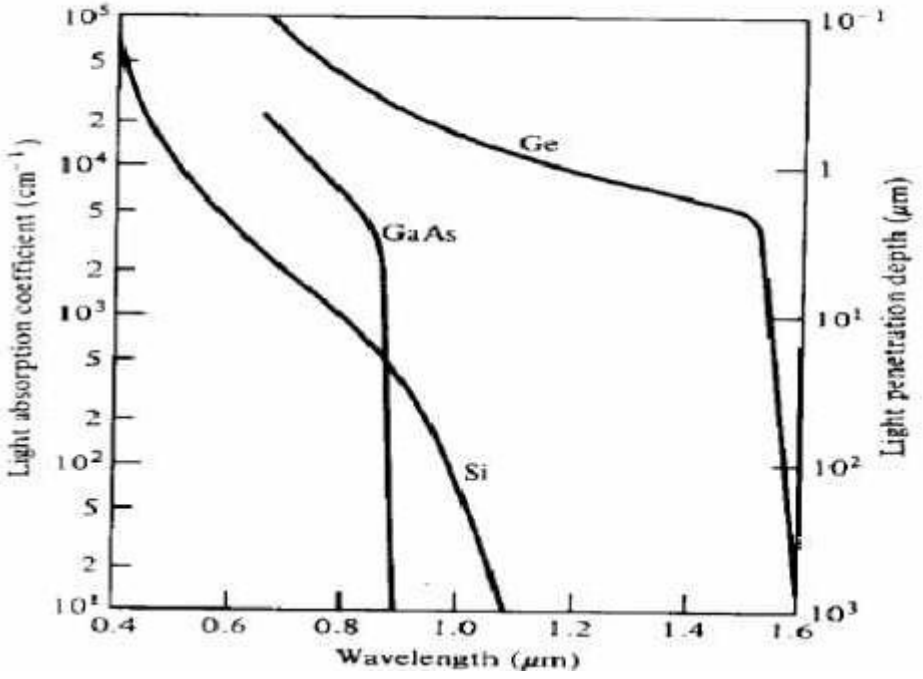
Absorption coefficient as (l) strongly depends on wavelength. The upper wavelength cutoff for any semiconductor can be

$$\lambda_c(\mu m) = \frac{1.24}{E_g(eV)}$$

Taking entrance face reflectivity into consideration, the absorbed power in the width of depletion region, *w*, becomes:

$$(1-R_f)P(w) = P_{in}(1-e^{-\alpha_s(\lambda)w})(1-R_f)$$

Optical Absorption Coefficient



The primary photocurrent resulting from absorption is:

$$I_p = \frac{q}{h\nu} P_{in}(1 - e^{-\alpha_s(\lambda)w})(1 - R_f)$$

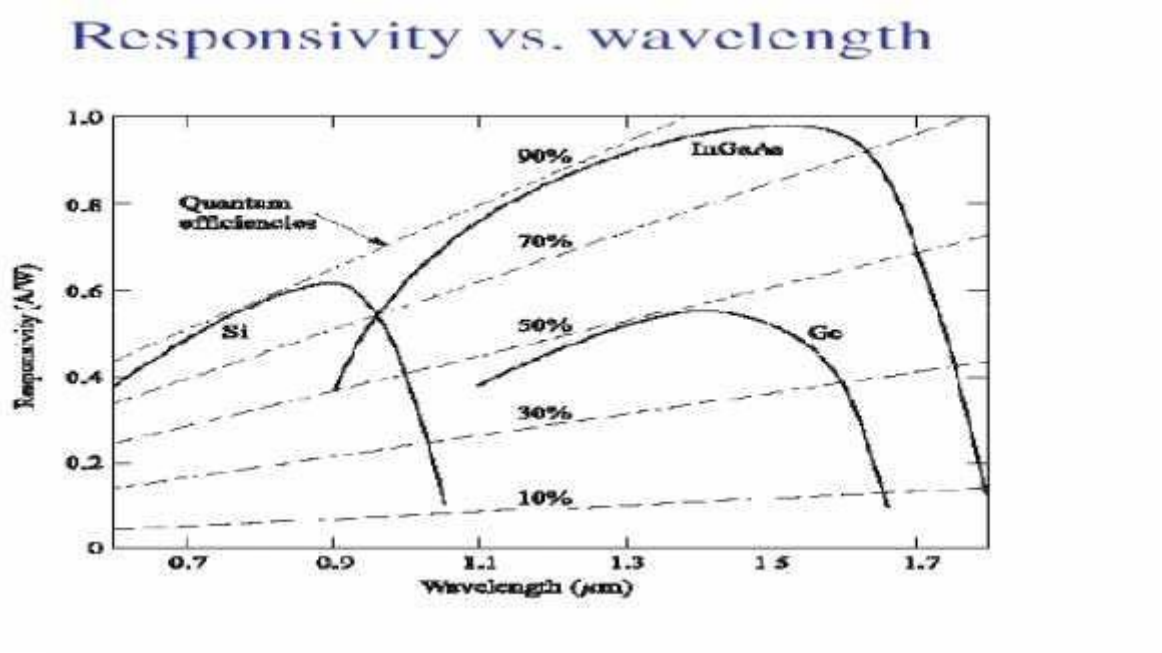
Quantum Efficiency:

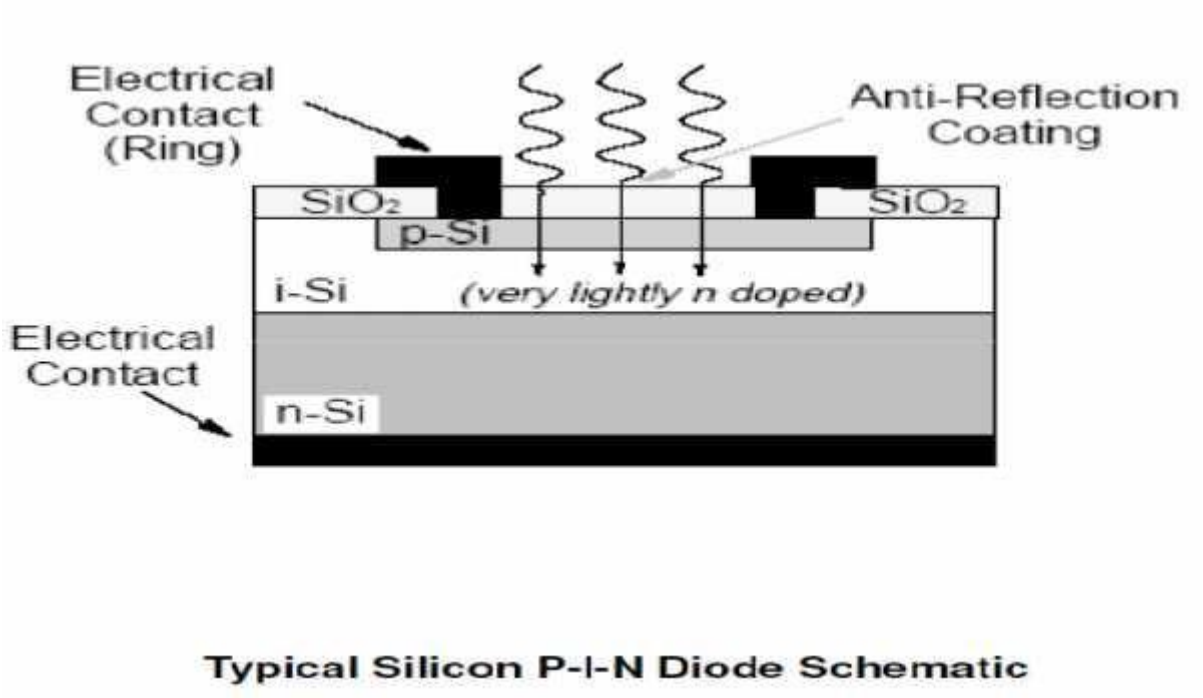
$$\eta = \frac{\#\text{ of electron - hole photogenerated pairs}}{\#\text{ of incident photons}}$$

$$\eta = \frac{I_p / q}{P_{in} / h\nu}$$

Responsivity:

$$\Re = \frac{I_p}{P_{in}} = \frac{\eta q}{h\nu} \quad [\text{A/W}]$$

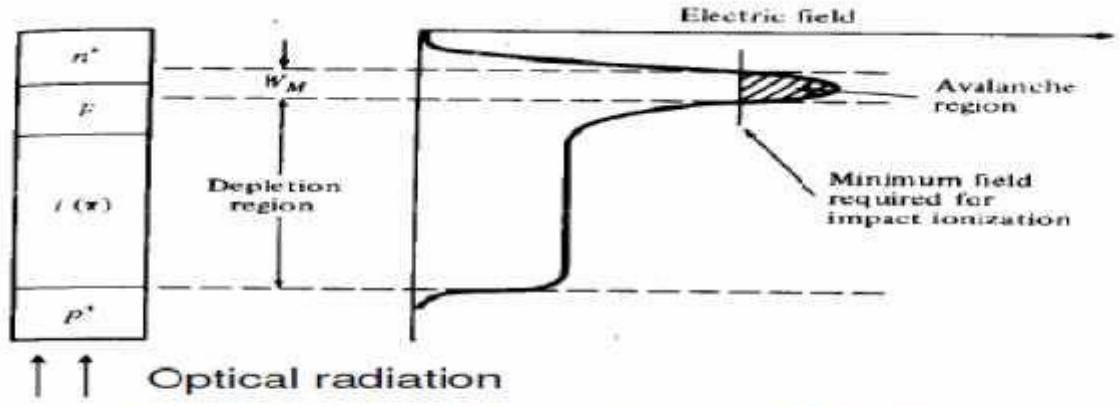




Typical Silicon P-I-N Diode Schematic

Avalanche Photodiode (APD):

APDs internally multiply the primary photocurrent before it enters to following circuitry. In order to carrier multiplication take place, the photogenerated carriers must traverse along a high field region. In this region, photogenerated electrons and holes gain enough energy toionize bound electrons in VB upon colliding with them. This multiplication is known as impact ionization. The newly created carriers in the presence of high electric field result in more ionization called avalanche effect.



Reach-Through APD structure (RAPD) showing the electric fields in depletion region and multiplication region.

Responsivity of APD:

The multiplication factor (current gain) *M* for all carriers generated in the photodiode is defined as:

$$M = \frac{I_M}{I_p}$$

where *IM* is the average value of the total multiplied output current & *Ip* is the primary photocurrent.

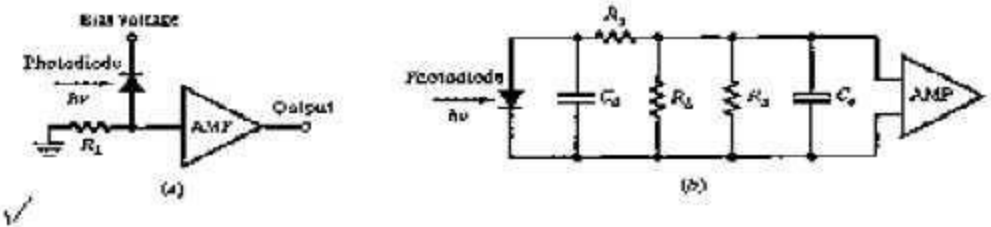
The responsivity of APD can be calculated by considering the current gain as:

$$\Re_{APD} = \frac{\eta q}{h\nu} M = \Re_0 M$$

Photodetector Noise & S/N:-

Detection of weak optical signal requires that the photodetector and its following amplification circuitry be optimized for a desired signal-to-noise ratio.

It is the noise current which determines the minimum optical power level that can be detected. This minimum detectable optical power defines the **sensitivity** of photodetector. That is the optical power that generates a photocurrent with the amplitude equal to that of the total noise current (*S/N=1*)



$$\frac{S}{N} = \frac{\text{signal power from photocurrent}}{\text{photodetector noise power} + \text{amplifier noise power}}$$

$$\frac{S}{N} = \frac{\langle i_P^2 \rangle M^2}{2q(I_P + I_D)BM^2F(M) + 2qI_LB + 4k_BTB/R_L}$$

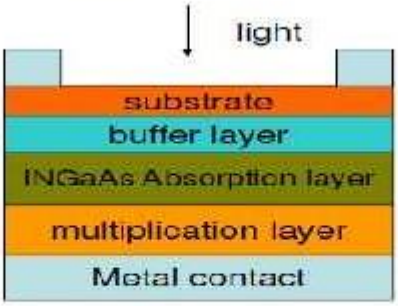
Since the noise figure $F(M)$ increases with M, there always exists an optimum value of M that maximizes the S/N. For sinusoidally modulated signal with $m=1$ and

$$F(M) \approx M^x$$

$$M_{\text{opt}}^{x+2} = \frac{2qI_L + 4k_BT/R_L}{xq(I_P + I_D)}$$

Structures for InGaAs APDs:-

Separate-absorption-and multiplication (SAM) APD



InGaAs APD superlattice structure (The multiplication region is composed of several layers of InAlGaAs quantum wells separated by InAlAs barrier layers.

Fiber Optic System Design:-

There are many factors that must be considered to ensure that enough light reaches the receiver. Without the right amount of light, the entire system will not operate properly.

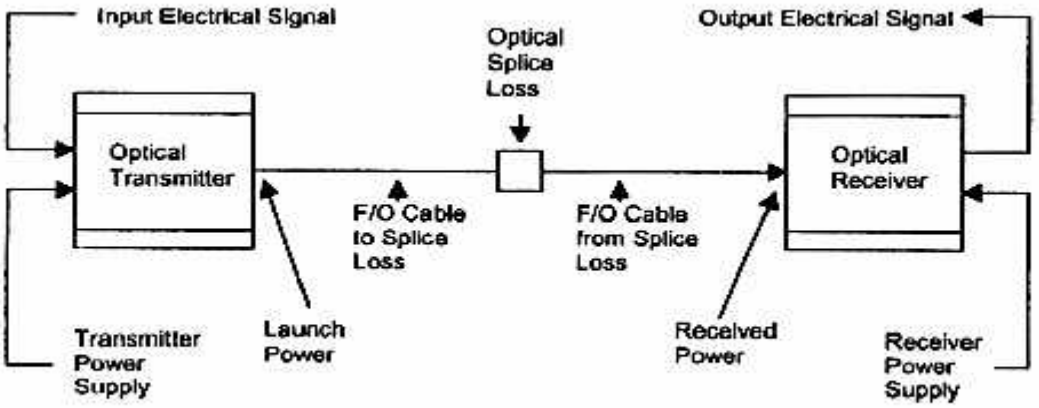


Figure 12, Important Parameters to Consider When Specifying F/O Systems

Fiber Optic System Design- Step-by-Step:-

Select the most appropriate optical transmitter and receiver combination based upon the signal to be transmitted

Determine the operating power available (AC, DC, etc.).

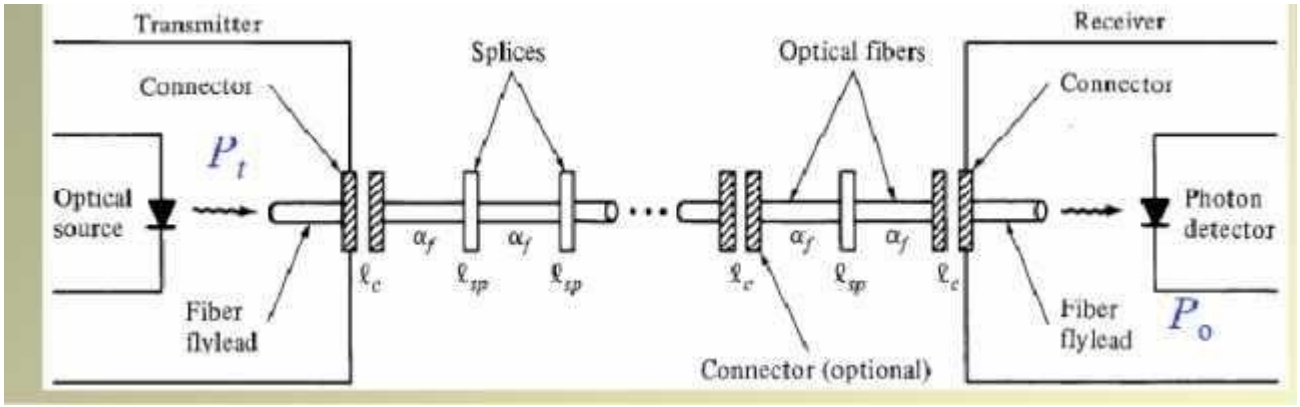
Determine the special modifications (if any) necessary (Impedances, bandwidths, connectors, fiber size, etc.).

Carry out system link power budget.

Carry out system rise time budget (I.e. bandwidth budget).

If it is discovered that the fiber bandwidth is inadequate for transmitting the required signal over the necessary distance, then either select a different transmitter/receiver (wavelength) combination, or consider the use of a lower loss premium fiber

Link Power Budget:-



Total loss $LT = \alpha f\, L + lc + lsp$

$Pt - Po = LT + SM$

Po = Receiver sensitivity (i.e. minimum power requirement)

SM = System margin (to ensure that small variation the system operating

parameters do not result in an unacceptable decrease in system performance)

Link Power Budget - Example 1:-

Parameters	Value	dB
▪ *Transmitter*		
▪ Average transmitted power	3 mW	4.8 dBm
▪ Fibre coupling losses		-3.7 dB
▪ *Channel*		
▪ Fibre loss		-15.7 dB
▪ Splitting losses		-10 dB
▪ Splice & Connector losses		-0.79 dB
▪ Fibre dispersion & nonlinearity		0 dB
▪ *Receiver*		
▪ Signal power at the receiver	All lossess	-26.79 dBm
▪ Receiver sensitivity		-31 dBm
System Margin (-20 dBm -(-30 dBm))		+4.1 dB

Link Power Budget - Example 2:-

Transmitter

Date rate – 500 Mb/s

- Source Laser @ 1300 nm
- Coupling power = 2 mW (3 dBm) into a 10 um fibre.

Channel

- Mono mode fibre of length 60 km and a loss of 0.3 dB/km
- Connector loss = 1 dB/connector
- Splicing every 5 km with a loss = 0.5 dB /splice

Receiver:

- PIN @ 1300 nm
- BER = 10^{-9}

System margin = ?

Link Power Budget - Example 2 *contd.*:-

Receiver sensitivity -29 dBm

$$P_t - P_o = L_T + SM$$

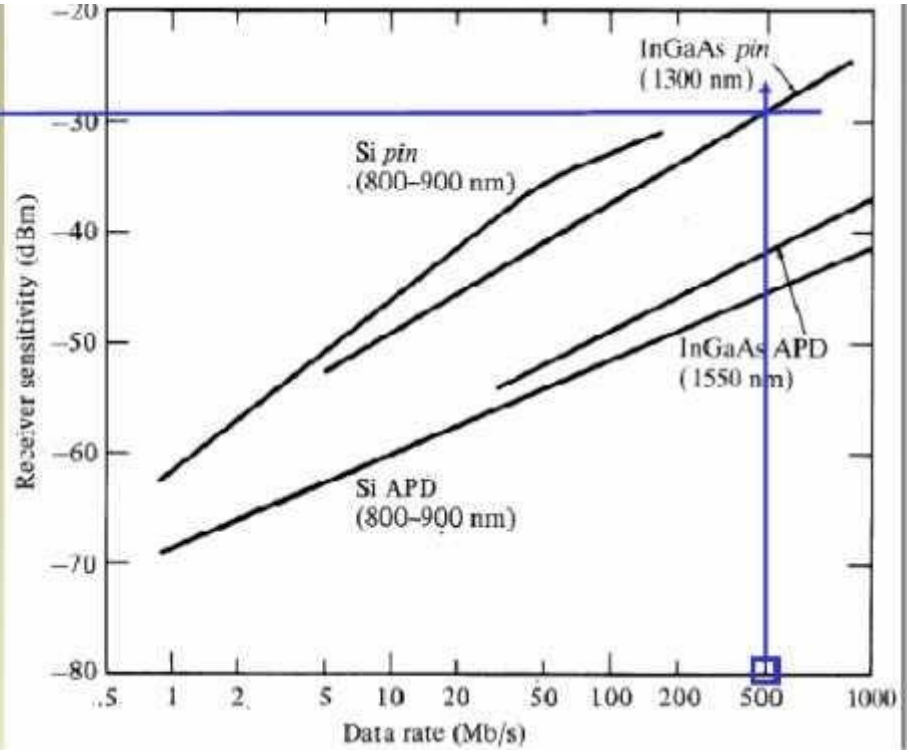
$L_T = 2(1\text{ dB}) + 0.3(60) + 0.5(11) - 25.5\text{ dB}$

thus

$3 - 29 = 25.5\text{ dB} - SM$

therefore

$SM = 5.5\text{ dB}$



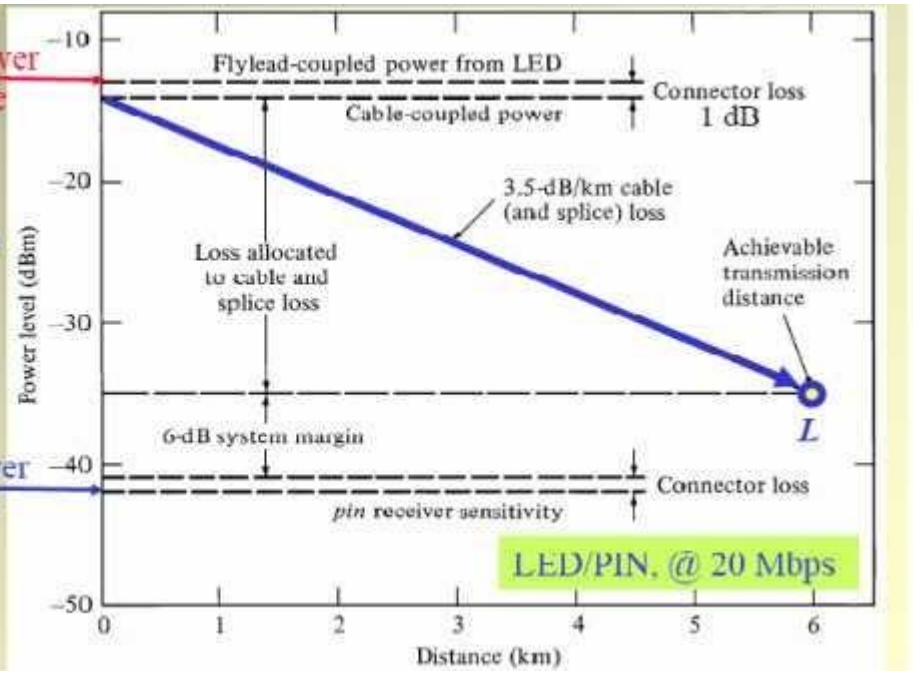
Link-Power Budget - Example 3:-

Launch power into fibre

Link power budget can be shown graphically in terms of receiver sensitivity Vs. the data rate

Launch power into fibre

G Keiser



Dispersion -equalisation penalty is given as:

$$D_L = 2\left(2\sigma B_1 \sqrt{2}\right)^4 \quad \textbf{(dB)}$$

Where B_1 is the bit rate, σ is the rms pulse width.

Therefore, the total channel loss is given as:

$$\textbf{Total loss } L_T = \alpha_f L + l_c + l_{sp} + D_L \quad \textbf{(dB)}$$

D_L is only significant in wideband multi-mode fibre systems

Rise Time Budget:-

The system design must also take into account the temporal response of the system components. The total loss LT (given in the power budget section) is determined in the absence of the any pulse broadening due to dispersion.

Finite bandwidth of the system (transmitter, channel, receiver) may results in pulse spreading (i.e. intersymbol interference), giving a **reduction in the receiver sencitivity**. I.e. worsening of BER or SNR

The additional loss penalty is known as **dispersion equalisation or ISI penalty.**

The total system rise time $t_{sys} = \left(\sum_{i-1}^{N} t_i^2\right)^{0.5}$

$$t_{sys} = \left(t^2{}_s + t^2{}_{inter} + t^2{}_{intra} + t^2{}_d\right)^{0.5}$$

Source, Fibre intermodal, Fibre intramodal, Detector

Note - 3 dB bandwidth of a simple low pass RC filter is given as:

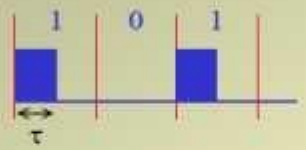
$$B = \frac{1}{2\pi RC}$$

With a step input voltage into the RC filter, the rise time of the output voltage is:

$$t_r = 2.2B = \frac{0.35}{B}$$

For a fibre optic link: $t_{sys} = t_r = \frac{0.35}{B}$

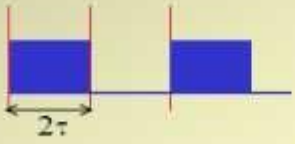
For RZ data format



Bit rate $R = B = 1/\tau$

$$B_{RZ} = \frac{0.35}{t_{sys}}$$

For NRZ data format



Bit rate $R = B = 1/2\tau$

$$B_{NRZ} = \frac{0.75}{t_{sys}}$$