NEAR EAST UNIVERSITY

Faculty of Engineering

Department of Electrical and Electronic Engineering

Noises In Fiber Optic Communication

Graduation Project
EE- 400

Student: Malik Taufiq-ur-Rehman (971375)

Supervisor: Prof. Dr. Fakhreddin Mamedov

Nicosia - 2001

TABLE OF CONTENTS

ACKNOWLEDGMENT

In this project several friends has contributed their time and expertise to review the chapters and lend good advice.

First, many thanks to Prof. Dr. Fakhreddin Mamedov, for understanding what I wanted to accomplish, having faith in the idea. And also to my friend Aneel and kashif, for pointing me in the right directions and taking out the mistakes in my project and also Khalid for using his computer.

ABSTRACT

Noise and distortion are important performance limiting factors in signal detection. They result in a smaller SNR or higher BER. In analog communications, the SNR should be maximized, and in digital communications, the BER should be minimized.

Two important characteristics of a noise are the PDF and PSD. They allow one to calculate the SNR and BER. In addition, an optimum filter can be designed to minimize the BER or maximize the SNR. Thermal noise is a white Gaussian noise due to random thermal radiation. Because of the central limit theorem and the flat spectrum of white noise, white Gaussian noise is often used to approximate other kids of noise. Shot noise in optical communications is caused by random EHP generations in a photodiode. The number of EHPs generated over a given time interval is a Poisson distribution. Shot noise defined as the photocurrent fluctuation is a filtered Poisson process. Its spectrum is often considered white for simplicity. Because shot noise is intrinsic to photocurrent generation, it places a fundamental performance limit called the quantum limit. When all other noise sources are ignored, the quantum limit is the minimum number of photons per bit required for a specified BER. At a BER of 10^{-9} and a 100 percent quantum efficiency, the quantum limit from incoherent detection is 10 photons per bit.

INTRODUCTION

Communication is an important part of our daily lives. It helps us to get closer to one another and exchange important information. An optical or lightwave communication system is a communication system that uses lightwaves as the carrier for transmission. This project focuses on the noises occurs from optical communications. In optical communications, noise can come from both transmitter and receiver. In addition to thermal noise, which occurs essentially every electronic circuit, there are phase noise, relative intensity noise (RIN), and mode partition noise (MPN) from the light source at the transmitter side, and shot noise and excess (avalanche gain) noise from the photodetector at the receiver side.

There are additional noises in advanced systems. For example, when optical amplifiers are used as overcome power loss, they add so-called amplified spontaneous emission (ASE) noise to the amplified. In wavelength-division multiplexing (WDM) and subcarrier multiplexing (SCM) systems in which multiple channels are transmitted through the same optical fiber, there can also be adjacent channel interference (ACI) or crosstalk, which is the interference from adjacent channels because of the power spectrum overlap. Because adjacent channels are statistically independent of the channel tuned to, they can be considered as another noise source.

Various noise and crosstalk sources discussed can be considered as waveform domain noise. That is, they are random distortion of the signal's waveform. More detailed analysis and equalization techniques for both noise and distortion will be discussed in chapter 6. under Incoherent detection.

In digital communications, there can also be time domain noise called jitter. Jitter is the timing error of the recovered bit clock with respect to the received data sequences. In digital communications, the recovered clock is used to sample the received signal for detection. As a result, a timing error will sample the received signal at a wrong timing and result in a large error detection probability. In general, jitter comes from imperfect bit time recovery.

CHAPTER 1
INTRODUCTION TO NOISES

1.1 Thermal Noise

Thermal noise, a white Gaussian noise, is one of the most common kinds of noise encountered in communication systems. Thermal noise is caused by radiation from random motion of electrons. Because it is a Gaussian noise, the PDF of thermal noise is Gaussian as given by Equation (1.1).

$$f_n(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/2\sigma^2} \qquad (1.1)$$

This Gaussian distribution comes from the fundamental central limit theorem, which states that if the number of noise contributors (such as the number of elections in a crystal) is large and they are statistically independent, the combined noise distribution is Gaussian. From thermodynamics, the PSD of thermal noise is given by

$$S_t(\omega)=\frac{h\omega}{2\pi}\left(\frac{1}{2}+\frac{1}{e^{h\omega/2\pi kT}-1}\right) \qquad (1.2)$$

where k is the Boltzmann constant (1.38×10^{-23} J/K) and T is the temperature in Kelvin. The first term in Equation (1.2) is from quantum mechanics. When $kT \geq h\omega/2\pi$, the power spectrum is almost a constant and equal to kT. From this approximation, thermal noise is a white noise with the following PSD:

$$S_t(\omega)=Kt \qquad (1.3)$$

The inverse Fourier transform gives the following autocorrelation for thermal noise:

$$R_T(\tau)=E[n_T(t)n_T(t+\tau)]=kT\delta(\tau) \qquad (1.4)$$

If the noise is filtered over a finite frequency band B, the filtered power spectrum will be zero outside the frequency band, and the average power is

$$\sigma^2=R_T(0)=\int_{frequencybands} kTdf=2kTB \qquad (1.5)$$

This calculation is illustrated in Figure 1.1.

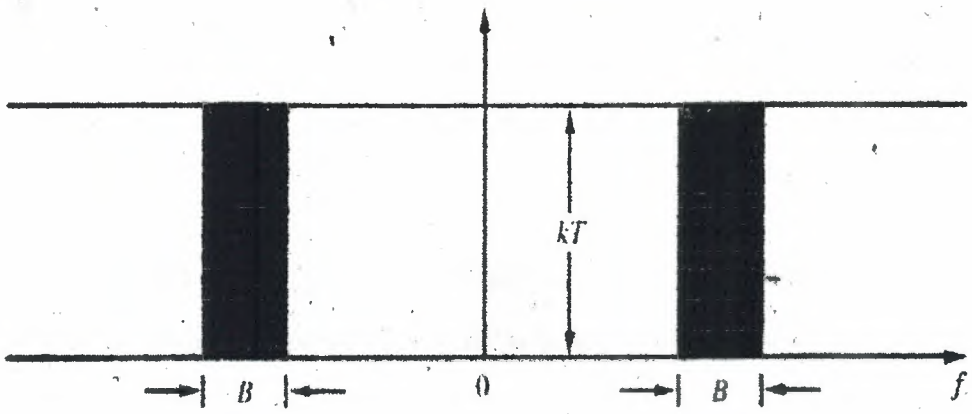

Figure 1.1. Power spectral density of thermal noise.

Thermal noise can be modeled as a voltage source of bandwidth *B by:*

$$\frac{v_{thermal}^2}{2R} = 2kTB \qquad (1.6)$$

In Equation (1.6), the factor of 2 in the denominator on the left-hand side is to account for the optimum power transfer efficiency. That is, 50 percent of the noise power from the equivalent voltage source contributes to the measurable noise power *2kTB.* The thermal current source can be similarly expressed as

$$i_{thermal}^2 = 4kTGB \qquad (1.7)$$

where G = 1/R is the conductance

If the thermal noise is included with the shot noise discussed earlier, the SNR at the photodiode output can be expressed as

$$\text{SNR} = \frac{(RP_{in})^2}{2qB(RP_{in} + I_d + 2V_T G)} \qquad (1.8)$$

where $V_T = == kT/q$ is the **thermal voltage** and G is the conductance of the load resistor Note that when the photocurrent I_{ph}, is large enough, thermal noise can be neglected. This motivates the use of APDs. However, there is an additional noise generated from the multiplication process.

Noise Equivalent Power An important parameter that is used to quantify the output noise power of a photodiode is called the noise equivalent power (NEP). It is defined as the required incident light power to have a zero dB SNR over a bandwidth of 1 Hz. Solving

Equation (1.9) for P_{in}, gives

$$\mathrm{NEP}_{pin} = \frac{1}{R}\hat{q}\left(1+\sqrt{1+2(I_d+V_T G)/\hat{q}}\right) \tag{1.9}$$

where $\hat{q} = q^*(1Hz)$ and the subscript *pin* indicates the use of a PIN diode in photo-detection. When the shot noise power due to *RPin* is negligible compared to $I_d + V_T G$ or when $2(I_d + V_T G) \geq \hat{q}$

$$\mathrm{NEP}_{pin} \approx \frac{1}{R}[2q(I_d + 2V_T G)]^{1/2} \tag{1.10}$$

Thus NEP is the noise power due to dark current and thermal noise.

1.2 Shot Noise

In practice, because of random EHP generation, the photocurrent has a random fluctuation from its average value. This random fluctuation is called shot noise and is the most fundamental noise in optical communications. This section gives a derivation of the PSD of a shot noise and explains its quantum limit as an ultimate detection performance limit in direct detection.

1.2.1 Power Spectral Density of Shot Noise

Shot noise $n_{shot}(t)$ as a function of time at the photodiode output is defined to be

$$n_{shot}(t) = i_{ph}(t) - I_{ph} \tag{1.11}$$

where $i_{ph}(t)$ is the photocurrent and I_{ph} is its average. The two-sided PSD of a shot noise is given by

$$S_{shot}(\omega) = q(I_{ph} + I_d)\left|H_{pin}(\omega)\right|^2 \approx q(I_{ph} + I_d) \tag{1.12}$$

where I_d is the dark current and $H_{pin}(\omega)$ is the Fourier transform of the impulse response of the PIN diode due to an EHR.Because $H_{pin}(\omega)$ is generally flat over a large frequency range, it can be dropped from equation (1.12). In otherwords, shot noise can be considered as a white noise over most relevant frequency ranges. If this is the case the shot noise power over a bandwidth *B* is

$$n_{shot}^2 = \int S_{shot}(\omega) * \frac{d\omega}{2\pi} \approx 2q(I_{ph} + I_d)B = 2q(RP_{in} + I_d)B \qquad (1.13)$$

1.2.2 Quantum Limit

As pointed out earlier, all noise sources except shot noise can theoretically be reduced to zero. Because the shot noise power from photo-detection is proportional to the incident light power or average photocurrent, however, as long as there is a light signal, there is shot noise. This section presents a derivation of the fundamental detection performance due to shot noise. At a specified BER, one must know what is the minimum number or photons per hit required. This minimum number is called the quantum limit.

The quantum limit due to shot noise can be derived from the following considerations. IF on-off keying is used to transmit binary bits, an optical pulse is transmitted for bit "1" and nothing (no pulse) for bit "O". At the receiver side, to detect whether a pulse is transmitted or not, one can count the number of incident photons **over** the bit interval T_0. When the number of photons counted is greater than a certain threshold, a pulse or "1" is detected; otherwise, "0" is detected. This photon counting process can be easily implemented by integrating the photocurrent generated for a duration T_0 and is called **integration-and-dump** in communications.

For an incident light signal of power P_{in}, the average number of EHPs generated over T_0 is

$$\bar{N} = \Lambda = \eta \frac{P_{in}}{hf} T_0 \qquad (1.14)$$

where η is the quantum efficiency of the photodiode. Because photocurrent generation is a Poisson process the actual number of EHPs generated over T_0 is a Poisson random variable, and the probability of having N EHPs counted over T_0 is given by

$$P[N] = \frac{\Lambda^N}{N!} e^{-\Lambda} \qquad (1.15)$$

Note that when $\Lambda = 0$ or $P_{in} = 0$, $P[0] = 1$. This means there is no possibility of having any EHPs generated. Therefore, to detect whether an optical pulse or bit "1" is transmitted, one can set die threshold at 0.5. That is, if N is greater than 0.5, one can be sure that an optical pulse is transmitted. On the oilier hand, if N counted is zero, it is determined that no

pulse is transmitted. Because P|N| can be zero even when P_{in} or A is nonzero, from Equation (1.15), the error detection probability is given by

$$P_E = e^{-\Lambda} p_1 \tag{1.16}$$

where p_1 is the prior probability of sending bit "1." At a given P_E value, the quantum limit N_q is the average number of EHPs per bit required to achieve the specified P_E. From equation (1.16), the quantum limit is given by

$$N_q = p_1 \Lambda = p_1 \ln\left(\frac{p_1}{p_E}\right) \tag{1.17}$$

When Λ is large and other noise in the system is considered, the threshold needs to be much larger. In this case, computation of the error detection probability becomes a series summation of the Poisson probability functions given by Equation (1.15). This is illustrated below.

If we use the central limit theorem and approximate the number of EHPs as a Gaussian distribution, then

$$P_E \approx 0.5 * \frac{1}{\sqrt{200\pi}} \int_{-\infty}^{39.5} e^{-(n-100)^2/200} \, dN = 5 * 10^{-10} \tag{1.18}$$

Therefore, Gaussian approximation in this case is a conservative estimation of the actual BER.

1.3 Effects of Noise and Distortion

To know the noise effects quantitatively, consider a basic point-to- point communication system in figure 1.2. Let the transmitted signal be s(t), the channel impulse response be h(t), and the channel noise be n(t). The received signal r(t) is thus given by

$$r(t) = s(t) \otimes h(t) + n(t). = q(t) + n(t) \tag{1.19}$$

where $\otimes$ denotes the convolution.

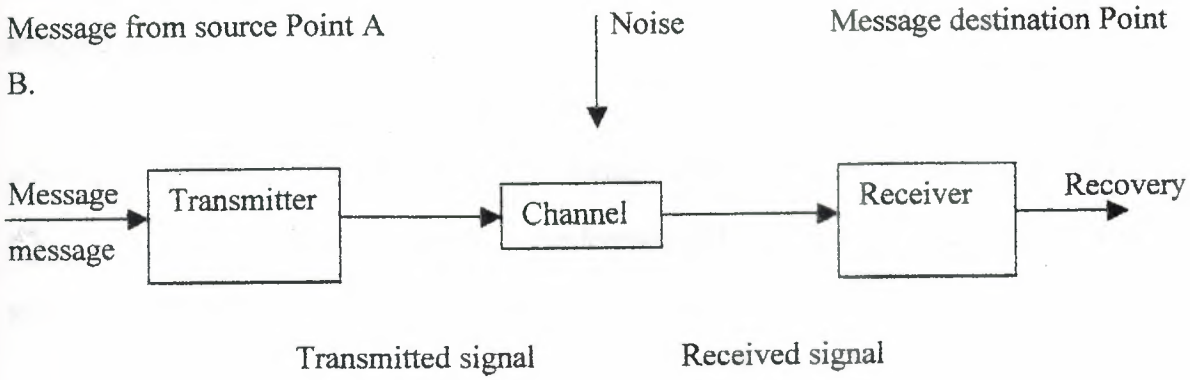

FIGURE 1.2. A point-to-point transmission link.

If the channel is ideal, it introduces only a certain delay and loss. Therefore, the impulse response of an ideal channel is given by

$$h(t) = a\,\delta(t-\tau) \tag{1.20}$$

where a is a constant factor representing transmission loss and τ is the propagation delay. Effect in Analog Communications In analog communications, the received signal quality can be characterized by the following ratio:

$$\gamma_Q = \frac{E\left[s(t)^2\right]}{E\left[\left|s(t)-r(t)\right|^2\right]} \tag{1.21}$$

where $E[x]$ denotes the expectation or average of signal x. Therefore, $E\left[s(t)^2\right]$ is the average signal power and $E\left[\left|s(t)-r(t)\right|^2\right]$ is the mean square error (MSE) with respect to the original signal s(t).

Effect in Digital Communications In digital communications, the consideration is a little bit different. Instead of minimizing the MSE, the objective is to recover the original bits transmitted with a minimal error detection probability. Consider a pulse amplitude modulated (PAM) signal transmitted over a channel. The received signal is given by

$$r(t) = \sum_k A_k\, p(t-kT_0) + n(t) \tag{1.22}$$

where A_k is the amplitude of the kth pulse, p(t) is the received pulse, and T_0 is the interval

between two consecutive pulses. To detect the transmitted amplitude A_k, the received signal is the first sampled at kT+τ for a certain τ within $(0, T_0)$. From equation (1.22), the sampled output is

$$r_k = r(kT+\tau) = \sum A_k p[(k-i)T+\tau] + n_k = A_k + ISI_k + n_k \quad (1.23)$$

where $p_i = p(iT+\tau)$ and $n_k = n(kT+\tau)$ .In digital communications, the distortion term ($\sum_{i \neq k} A_i p_{k-i}$) is called the intersymbol interference (ISI) because it is caused by adjacent symbols and pulses.

1.4 Noise Characterization

It is important to know the noise characteristics to evaluate the distortion and error detection probability. This section describes two primary noise characteristics: the **probability density function** (PDF) and the **power spectral density** (PSD).

1.4.1 Probability Density Function

The noise sample n_k considered earlier is a random variable. For continuous random variables, their PDFs are continuous functions; for discrete random variables, their PDFs are a summation of delta functions. When the PDF of a random variable is known, various statistics of the random variable can be computed.

Let fx(x) be the PDF of a continuous random variable *X*. By definition, the probability for $a < X < b$ is

$$Prob(a < X < b) = \int_a^b fx(z)dz$$

When the above integration is over (-∞, -x), the probability as a function of *x* is called the **probability distribution function** or **probability accumulation function.** That is,

$$F_X(x) = \int_{-\infty}^{x} f_X(z)dz \quad (1.24)$$

From this, *fx(x)* is the derivative of the probability accumulation function $F_X(x)$.

$$P(n_k < -A) = \int_{-\infty}^{-A} f_n(x)dx$$

Similarly,

$$P(n_k < A) = \int_{-\infty}^{-A} f_n(x)dx$$

Because of the importance of Gaussian noise, these two probabilities are commonly expressed in terms of the Q-function or the complementary error function erfc(x). The definition of the *Q*-function is

$$Q(x)=\frac{1}{\sqrt{2\pi}}\int_{x0}^{\infty} e^{-y^2/2}dy \tag{1.25}$$

Therefore, Q(0) == 0.5 and Q(∞) =*0.* The definition of the error function is

$$erf\,(x\}=\frac{2}{\sqrt{\pi}}\int_0^x e^{-y^2}\,dy=\frac{2}{\sqrt{2\pi\sigma^2}}\int_0^{x\sqrt{2\sigma^2}} e^{-y^2/2}dy \tag{1.26}$$

And the definition of the complementary error function is

$$erfc\{x)=\ 1\text{-}erf(x) \tag{1.27}$$

Therefore, *erf (∞) = erfc(0) =1* . From the definitions,

$$Q(x)=\frac{1}{2}erfc(x/\sqrt{2}) \tag{1.28}$$

As a result,

$$P(r_k > 0|A_k = -A) = P(r_k < 0|A_k = A)$$

$$= Q(A/\sigma)= \frac{1}{2}erfc(A/\sqrt{2\sigma^2}).$$

Because

$$SNR = \frac{E[s(t)^2]}{\sigma^2} = \frac{A^2}{\sigma^2}$$

$$P_E = P(r_k > 0|A_k = -A) = P(r_k < 0|A_k = A) \tag{1.29}$$

The following approximation *for Q(x)* makes calculation easier:

$$Q(x) \approx \frac{1}{\sqrt{2\pi x^2}}e^{-x^2/2} \text{ if } x \geq 1. \tag{1.30}$$

This equation shows that, the larger the SNR, the smaller the P_E. As a result, it is important to maximize the SNR to reduce the error probability.

In practice, a noise sample n_k can also be a discrete random variable. For example,

in optical communications, one can count the number of incident photons over a certain interval. In this case, n_k is the difference between the average number of photons and the actual number counted.

1.4.2 Power Spectral Density

Another important characteristic of noise is the power spectral density (PSD). Mathematically, it is defined as the Fourier transform of the autocorrelation function of the noise. Physically, it describes the frequency content of the noise power. In other words, for a given PSD $S_n(\omega)$ of noise n (t), the integration

$$\int_{\omega_1}^{\omega_2} S_n(\omega)\frac{d\omega}{2\pi}$$

gives the portion of the noise power within the frequency range from ω_1 to ω_2. If the integration is over the entire frequency range, it gives the average noise power. That is,

$$E\left[n(t)^2\right] = R_n(0) = \int_{-\infty}^{\infty} S_n(\omega)\frac{d\omega}{2\pi} \tag{1.31}$$

where $R_n(0)$ is the autocorrelation of n(t) at $\tau = 0$.

1.5 Mode Partition Noise

Mode partition noise (MPN) is caused by mode competition inside multimode FP laser cavity. As a result, even though the total power is constant, the power distribution over different modes is random. Because different modes have different propagation delays in fiber transmission, random power distribution results in random power variation at the receiving end. This power fluctuation due to mode competition is called MPN.

Because the power competition among all the longitudinal modes is not fully understood, an exact description of the PDF is not available. However, similar to RIN, it is well known that the noise power of MPN is proportional to the signal power. As a result, an error floor can be reached when MPN becomes dominant. This section represents basic properties of MPN and explains the error floor phenomenon.

Suppose a given laser diode has N longitudinal modes and each has a relative power a_i, i =1,....., N. By definition, the sum of these a_i's satisfies

$$\sum_{i=1}^{N} a_i = 1 \tag{1.32}$$

Because each a_i at a certain time is random variable, the average relative power for mode I is given by

$$\bar{a}_i = \mathrm{E}[a_i] = \int a_i * PDF(a_1 \ldots a_N) da_1 \ldots da_N .$$

If the waveform of mode i received is $f_i(t)$, the combined received signal is

$$\mathrm{r(t)} = \sum_{i,j} a_i f_i(t) \tag{1.33}$$

If the signal is sampled at time t_0, the variance of the sampled signal is

$$\sigma^2 = E[r(t_0)^2] - E[r(t_0)]^2 \tag{1.34}$$

From equations (1.33) and (1.34),

$$\sigma^2 = \sum_{i,j} f_i(t_0) f_j(t_0) \overline{(a_i a_j - a_i a_j)} \tag{1.35}$$

CHAPTER 2
OPTICAL WAVEGUIDES

2.1 Single-mode fibers

When a light-wave propagates inside the core of a fiber, it can have different EM field distributions over the fiber cross-section. Each field distribution that meets the Maxwell equations and the boundary condition at the core-cladding interface is called a **transverse mode.**

Several transverse modes are illustrated in Figure 2.1. As shown, they have different electric field distribution over the fiber cross-section. In general, different transverse modes propagate along the fiber at different speeds. These results are dispersion and are undesirable. Fibers that allow propagation of only one transverse mode are called **single-mode fibers** (SMF).

The key in fiber design to having single-mode propagation is to have a small core diameter. This can be understood from the dependence of the **cutoff wavelength** λ_c of the fiber on the core diameter. The cutoff wavelength is the wavelength above, which there can be only one single transverse mode. λ_c is expressed as

$$\lambda_c = \frac{2\pi a}{V}\left(n_1^2 - n_2^2\right)^{1/2} \tag{2.1}$$

where $V = 2.405$ for step-index fibers, a is the core radius, and n_1 and n_2 are the refractive indices of the core and cladding, respectively. This expression shows that fibers of a smaller core radius have a smaller cutoff wavelength.

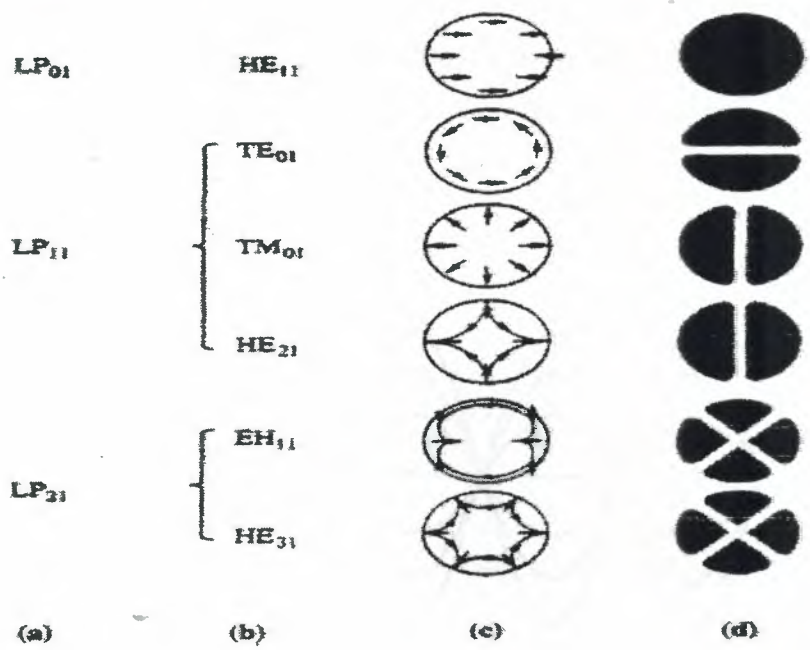

Figure 2.1. Some examples of low-order transverse modes of a step-index fiber. (a) Linear polarized (LP) mode designations, (b) exact mode designations, (c) electric field distribution, and (d) intensity distribution of the electric field component E_x

When the core diameter of a single-mode fiber is not much larger than the wavelength, there is a significant power portion or field penetration in the cladding. Therefore, it is necessary to define another parameter called **mode field diameter** (MFD). Intuitively, it is the' "width" of the transverse field. Specifically, it is the **root mean square** (RMS) width of the' field if the field distribution is Gaussian. When the field distribution is not Gaussian, the way to define the MFD is not unique. This MFD concept is useful when we want to determine the coupling or splicing loss of two fibers. In this case, it is the match of the MFD instead of the core diameter that is important to a smaller coupling or splicing loss.

We have already mentioned that when the fiber V parameter is less than 2.405 then only one mode (the HE_{11} mode, or, in the linearly polarized approximation, the LP_{01} mode) can propagate. Actually, strictly speaking two HE_{11} modes can be present with orthogonal polarizations, but for simplicity we will assume that are dealing with only one of these. In theory, the HE_{11} mode will propagate no matter how small the value of V. As V

decreases, however, the mode field will extend increasingly into the cladding if the field then becomes at all significant at the edge of the cladding, appreciable amounts of energy may be lost from the fiber, leading to the mode being highly attenuated.

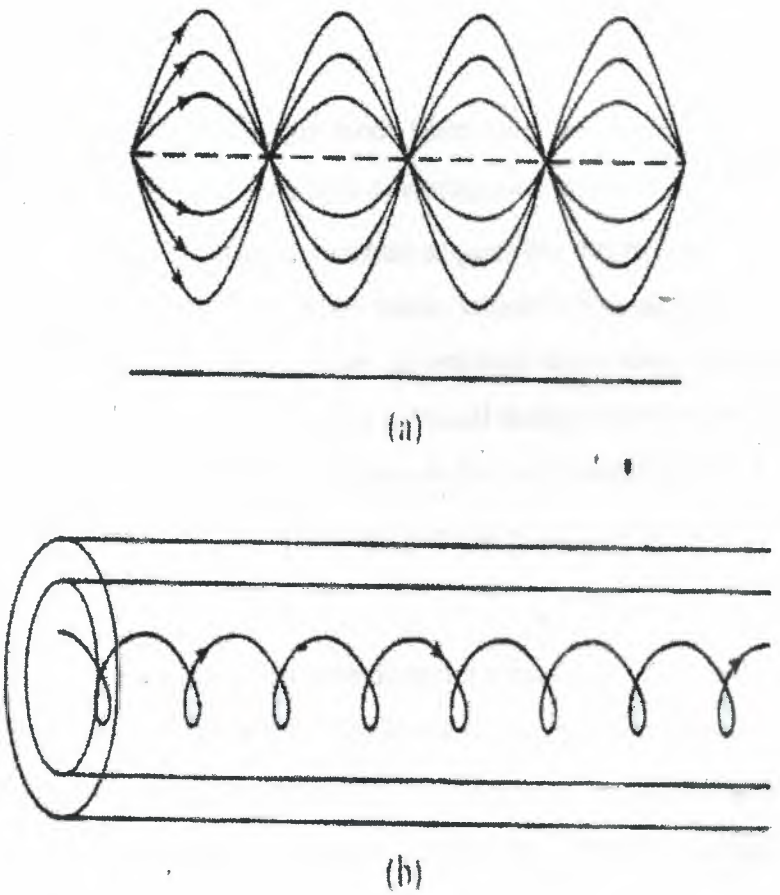

FIGURE 2.2. Ray paths in a graded index fiber for (a) Meridional rays and (b) helical rays which avoid the Center

In terms of the fiber core radius, a, for a single mode to propagate is

$$a \leq \frac{2.405\lambda_0}{2\pi\left(n_1^2 - n_2^2\right)^{1/2}}$$

or

$$a \leq \frac{2.405\lambda_0}{2\pi(NA)} \qquad (2.2)$$

This relationship implies dial single mode fibers will have cores that are only of the order of λ_0 (i.e. micro-meters) in radius. It is advantageous from a number of points of view, however, that they have as large a diameter as possible. From equation (2.2) we can sec that this may he done by reducing the NA value, which is by making the core and cladding refractive indices very close together. In practical terms single mode fibers are made with NA values of the order of 0.1, with a typical design criterion for a single mode fiber being $2 \leq V \leq 2.2$. When used with radiation in the wavelength region 1.3 μm to 1.6 μm, single mode fibers have core diameters that are typically between 5 μm and 10 μm.

It should be noted that for a given core diameter a particular fiber will only be single mode when the wavelength of radiation being used is greater than a critical value, λ_c, which is called the *cut-off wavelength* (since it represents the wavelength at which the mode above the lowest order mode cuts off). From equation (2) we have that

$$\lambda_c = 2\pi a(NA)/2.405 \qquad (2.3)$$

When a fiber is to be used as a single mode fiber, care must be taken to ensure that the wavelengths used never exceed the cut-off wavelength.

Although the mode field distribution in a single mode fiber is theoretically described by Bessel functions, it is convenient to represent the field irradiance distribution (with little loss inaccuracy) by the much simpler Gaussian function, that is

$$I(r)=I_0 \exp(-2r^2/\omega_0^2) \qquad (2.4)$$

where $2\omega_0$ is known as the *mode field diameter*. ω_0 thus represents the radial distance at which the mode irradiance has fallen to exp(-2) (i.e. 13.5%) of its peak value. As the V parameter of a fiber gets smaller we would expect that ω_0 would increase (i.e. the field will extend further into the cladding). A useful empirical relationship between ω_0 and V which is accurate to better than 1% if $1.2 < V < 3$.

$$\omega_0/a = 0.65 + 1.619V^{-3/2} + 2.879V^{-6} \qquad (2.4a)$$

2.2 Multimode fibers

Fibers that allow propagation of multiple transverse modes are called **multimode fibers** (MMF). Optical fiber sensors themselves can be divided into two main categories, namely 'intrinsic' and 'extrinsic' that are explained as follow:

2.2.1 Multimode extrinsic optical fiber sensors

Some of the simplest extrinsic fiber-based sensors are concerned with the measurement of movement or position. For example, when two fiber ends are moved out of alignment, the coupling loss depends on the displacement. A similar type of sensor uses a shutter moving between two fiber ends that are laterally displaced (Fig. 2.3a). Improvements in sensitivity are possible by placing a pair of gratings within the gap, one fixed, the other movable (Fig. 2.3b). Here, however, although the sensitivity has increased, the range has decreased, since the output will be periodic in the spacing of the grating. The range of movement possible for the single shutter sensor is obviously limited by the fiber core diameter. If a beam expander is employed between the fibers (Fig. 2.3c), then the range can be greatly expanded.

One of the first commercially available displacement sensors was the 'Fotonic sensor'. This uses a bundle of fibers, half of which are connected to a source of radiation, the other half to a detector (Fig. 2.4a). If the bundle is placed in close- proximity to a reflecting surface then light will be reflected back from the illuminating fibers into the detecting fibers. The amount detected will depend on the distance from the fiber ends to the surface.

surface. To analyze this dependence, we consider the somewhat simpler situation where there are

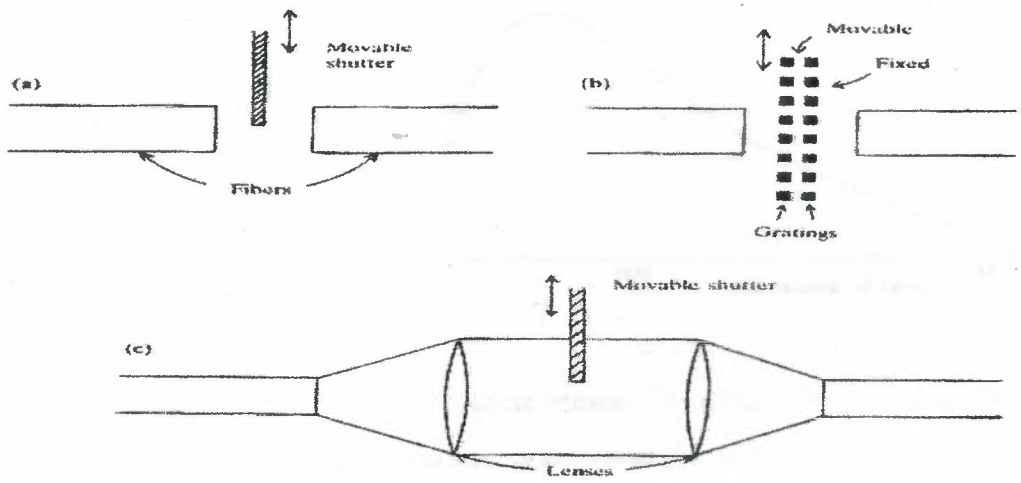

FIGURE 2.3 Simple displacement sensors. In (a) a movable shutter varies the light coupled between two longitudinally displaced fibers. In (b), the use of two gratings increases the sensitivity. In (c), a beam expansion system enables an increase in the range of measurable displacement to be increased.

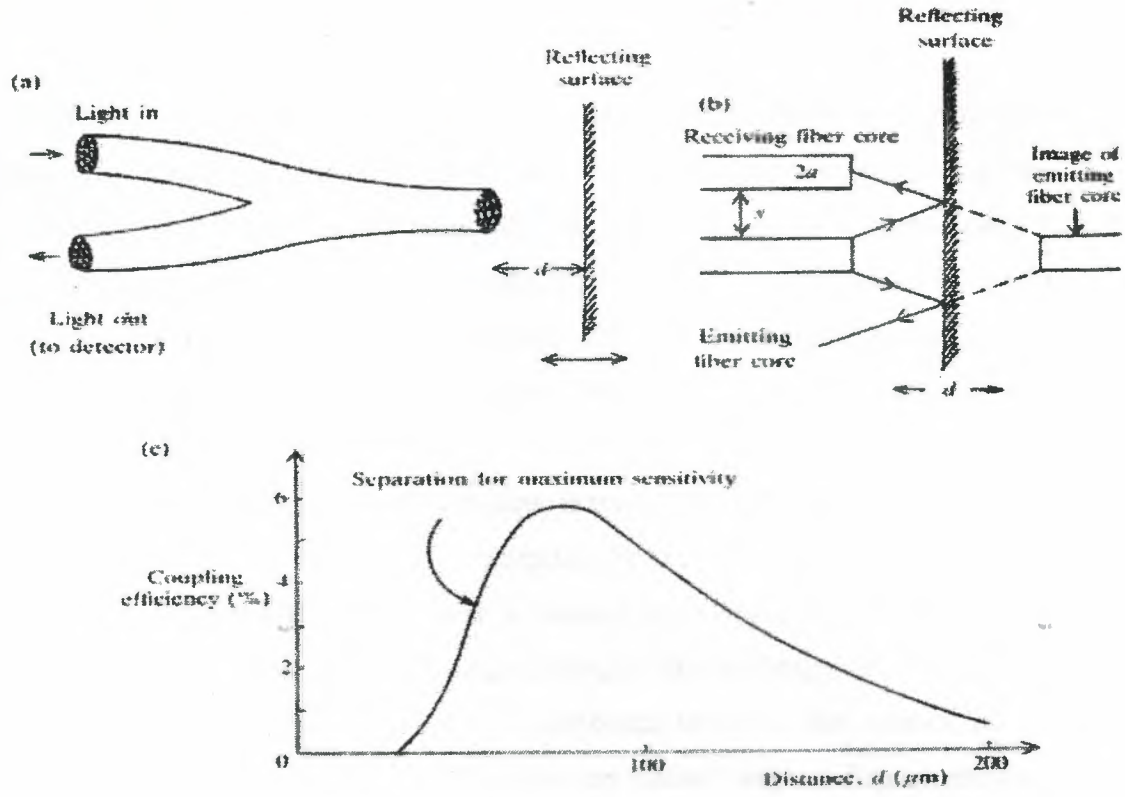

FIGURE 2.4 Illustrations of he Fotonic sensor. The general layout using fiber bundles is shown in (a). A two fiber version is shown in (b), which can be used to derive the form of the output-distance (c) relationship. The typcal result of such a calculation taking a=100 μm and NA = 0.4 is shown in (c).

just two fibers. If we regard the reflecting surface as a mirror, the problem then reduces to that of the coupling between two fibers that are displaced both laterally and longitudinally (Fig. 2.4b). The form of the relationship between displacement and light output may be determined by considering the overlap between the sensing fiber core area and the cross-section of the light cone emitted by the image of the emitting fiber. We can readily appreciate that at very small fiber-surface distances, no light will be coupled between the two fibers. Then, beyond a certain critical distance, there will be an increasing overlap between the above areas and the coupled radiation will increase rapidly. Once the detecting fiber area is completely filled, however, the output will fall with increasing distance. At large distances, an inverse square law will then be obtained.

In practice, when a fiber bundle is used instead of just two fibers the displacement-output characteristic will be somewhat different and will depend on how the emitting and receiving fibers are distributed (usually randomly), but the overall shape remains similar to that of Fig. 2.4(c). Because of the very non-linear nature of the curve, the sensor is not very suitable for the measurement of large displacements, although it is possible to increase the range by using a lens system. In fact, the sensor was developed originally for non-contact vibration analysis.

Any displacement measurement technique is readily adapted to the measurement of pressure, and those mentioned above are no exception. For example, the Fotonic sensor may be placed close to a reflecting diaphragm with a constant pressure maintained on the sensor side. Any change in external pressure will cause flexing of the diaphragm and a consequent change in the instrument's output. It should be remembered, however, that none of the instruments described above is linear except over a very limited range of displacements. Accurate calibration over the whole range is therefore required.

As well as displacement/pressure sensors, a number of extrinsic fiber temperature sensors have been proposed. For example, the band gap of semiconductors such as GaAs is temperature dependent (Fig. 2.5a) and a simple sensor can be made in which a piece of the semiconductor is placed in the gap between the ends of two fibers (Fig. 2.5b). Light with a wavelength corresponding approximately to the semiconductor band gap is sent down one of the fibers and the power emerging from the other is measured and can be related to the temperature.

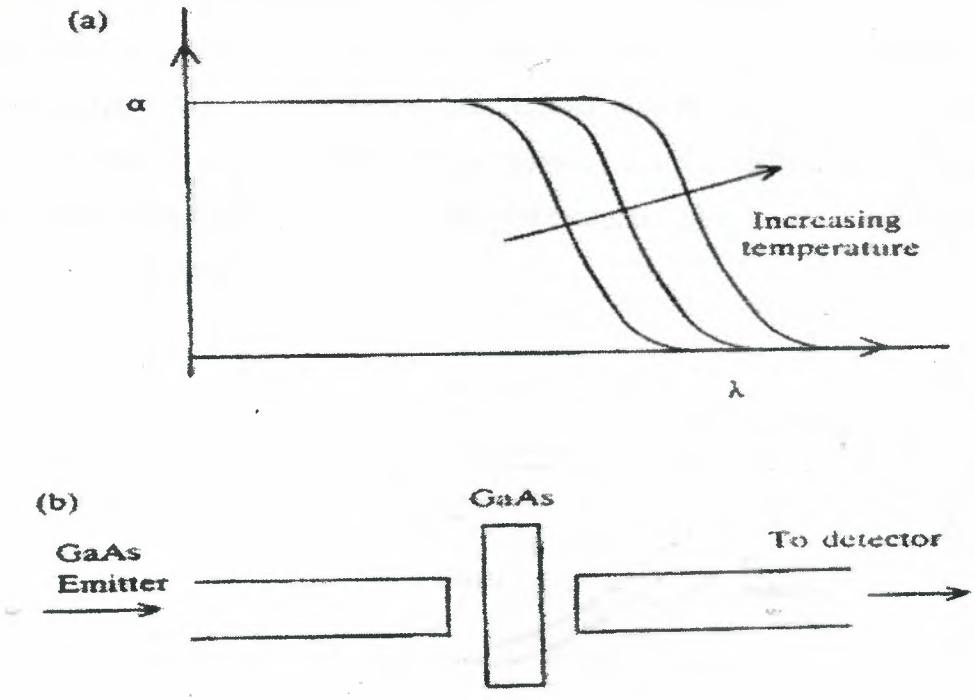

FIGURE 2.5 (a) Schematic variation of the absorption coefficient (α) of GaAs with both wavelength (λ) and temperature. (b) Temperature sensor utilizing in the transmission of GaAs with temperature

Another temperature sensor, the Fluoroptic sensor, is available commercially and is claimed to have a sensitivity of 0.1 °C over the range -50°C to 250°C. The instrument relics on the temperature variation of the fluorescence In europium-doped lanthanum oxysulfide (Eu: La_2O_2S). A small amount of this material is placed on the object whose temperature is to be measured, and the fluorescence is excited by illuminating it with ultraviolet light transmitted down a fairly large diameter (400 μm) plastic-coaled silica fiber. The source of radiation is a quartz-halogen lamp whose output has been filtered to remove any unwanted higher wavelengths. Another, similar, fiber picks up some of the emitted fluorescence and carries it back to the detector system (Fig. 2.6). In fact, the phosphor emits at more than one wavelength and it is the intensity ratio of two of the lines, which is measured. Because a ratio is measured, any fluctuations in the irradiance of the source arc not important.

At the detector end of the fiber, the radiation is split into two using a beam splitter.

the particular wavelength required. The ratio of the two signals then provides the temperature information, which is usually contained in a 'look-up' table. Because the ultraviolet output from a quartz-halogen lamp is small and the fiber absorption relatively large at short wavelengths, the output of the phosphor is quite small. Efficient detectors and low noise preamplifiers are required, and the maximum fiber length is restricted to about 15 m. Nevertheless, the

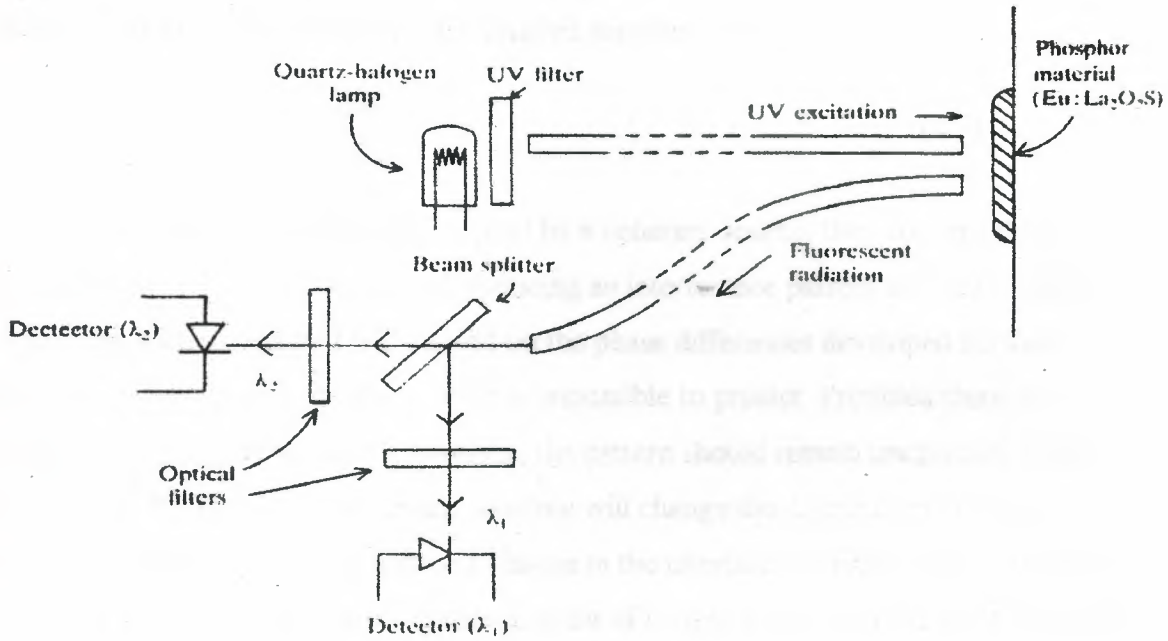

FIGURE 2.6. Schematic layout of the Fluorooptic temperature sensor. The fluorescent radiation generated in the phosphor is separated into its two main constituent wavelengths(λ_1 and λ_2) and the relative optical power of these wavelengths is determined by using a beam splitter followed by two optical band pass filters to isolate the two wavelengths.

device provides a performance superior to thermocouples, and allows point temperature measurements in semi remote hostile environments.

2.2.2 Multimode intrinsic optical fiber sensors

One of the ways in which we can influence the amount of radiation flowing down a

fiber is by means of micro bending loss, and this can therefore be made the basis of a displacement or pressure transducer. In a typical device, the fiber passes between a pair of ridged plates, which impart a periodic perturbation to the fiber. In fact, we have met such an arrangement before in the guise of a mode scrambler. If step index fiber is used, a particular periodic perturbation of wavelength Λ will only couple together a few modes. However, it may be shown (ref. 10.1) that, with graded index fiber where the profile parameter, a, is equal to 2, all modes are coupled together when

$$\Lambda_c = \frac{2\pi a}{\sqrt{2\Delta}} \qquad (2.5)$$

When the modes in a fiber are excited by a coherent source, they are capable of interfering with each other and thus of producing an interference pattern across the end of the fiber. The pattern obtained will depend on the phase differences developed between the modes as they travel along the fiber, which is impossible to predict. Provided there are no perturbations acting on the system, however, the pattern should remain unchanged. If the fiber is slightly flexed in any way, mode coupling will change the distribution of energy amongst the modes, and hence produce a change in the interference pattern across the fiber end. Of course, unless there is a significant amount of coupling into lossy modes, there will not be any great change in the total amount of energy emerging from the fiber. If, however, we consider only a small portion of the whole area of the fiber end, any change in the interference pattern as a whole is almost certain to produce quite significant changes in the emerging energy. Thus, if a detector is so placed as to intercept only a small portion of total light emerging from the fiber, its output should vary when there is any deformation of the fiber.

By its very nature, such a detector will be very non-linear, though in some circumstances this may not be a great disadvantage. For example, by laying the fiber just below ground level it may be possible to detect the presence of intruders, since their footsteps will cause deformation of the fiber. All that is required is for the output to trigger an alarm when the change in the signal exceeds some predetermined level.

The Bragg fiber grating structure can be used as a very useful sensing element. It

will be recalled that the grating will reflect radiation of wavelength λ_B, which satisfies the equation

$$\lambda_B = 2mn_1\Lambda \qquad (2.6)$$

where m== 1,2,3, etc., A is the periodicity of the grating and n_1 the refractive index of the core. The exact value of the product $n_1\Lambda$ will depend on both temperature and strain within the fiber. As far as temperature changes are concerned both the grating wavelength and the refractive index will be affected by temperature and we can write

$$\lambda_B = 2mn_1\Lambda\left(\frac{1}{n}\frac{dn_1}{dT} + \frac{1}{\Lambda}\frac{d\Lambda}{dt}\right)\Delta T$$

This can be written as

$$\Delta\lambda_B = \lambda_B(\beta + \alpha)\Delta T \qquad (2.7)$$

where $\beta = (dn_1/dT)/n_1$ and α is the linear expansion coefficient. Similarly the application of strain (ε) will affect both the grating spacing and the refractive index (via the photo elastic effect), and we may write

$$\Delta\lambda_B = \lambda_B(1 - p_e)\Delta E \qquad (2.8)$$

where p_e is an effective photo elastic coefficient given by

$$p_e = \frac{n_1^2}{2}[(1-\mu)P_{12} - \mu P_{11}] \qquad (2.9)$$

where P_{11} and P_{12} are Pockels coefficients and μ is Poisson's ratio. There are a number of ways in which the Bragg sensor can be 'interrogated' to obtain a measure of the reflection wavelength λ_B. For example, if radiation from a tunable laser is incident on the grating and its output wavelength scanned across the appropriate wavelength range then strong reflection will be obtained at λ_B. The magnitude of the back-reflected radiation is easily monitored using the set-up illustrated in Fig 2.7. Several such sensors may be employed at different positions along the fiber provided that the wavelength ranges associated with the sensors are mutually exclusive and also that the laser scanning range is sufficiently large. The wavelength of each sensor is determined by correlating the time at which a reflected

pulse is detected to the laser wavelength at that time. In another measurement technique radiation from a broadband source is sent down the fiber. Light at wavelength λ_B will be removed from the beam, leading to a 'notch' appearing in the transmitted spectrum and either the reflected or transmitted spectrum can be analyzed to obtain λ_B. However, the changes in λ_B are small and difficult to measure directly except with costly instruments such as the optical spectrum analyzer. A number of possible measurement schemes have been proposed most of which involve matching the wavelength λ_B to the resonant' wavelength of some other optical system such as a Fabry-Perot interferometer.

Bragg grating sensors offer a number of advantages over other types. For example, they offer a relatively high resolution of strain or temperature, the output is a linear function of the measured and they are insensitive to fluctuations in light intensity. In addition they are relatively easy to fabricate and do not compromise the structural integrity of the fiber.

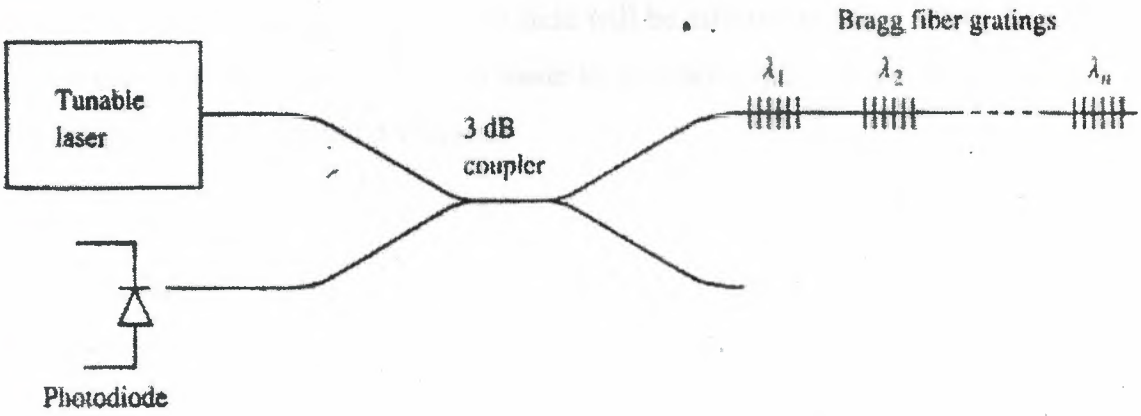

FIGURE 2.7. Illustration of a technique that can be used to interrogate an array of Bragg fiber diffraction gratings. The photodiode will only receive a signal when the output of the laser corresponds to one of the reflection wavelengths ($\lambda_1, \lambda_2 ..... , \lambda_n$).

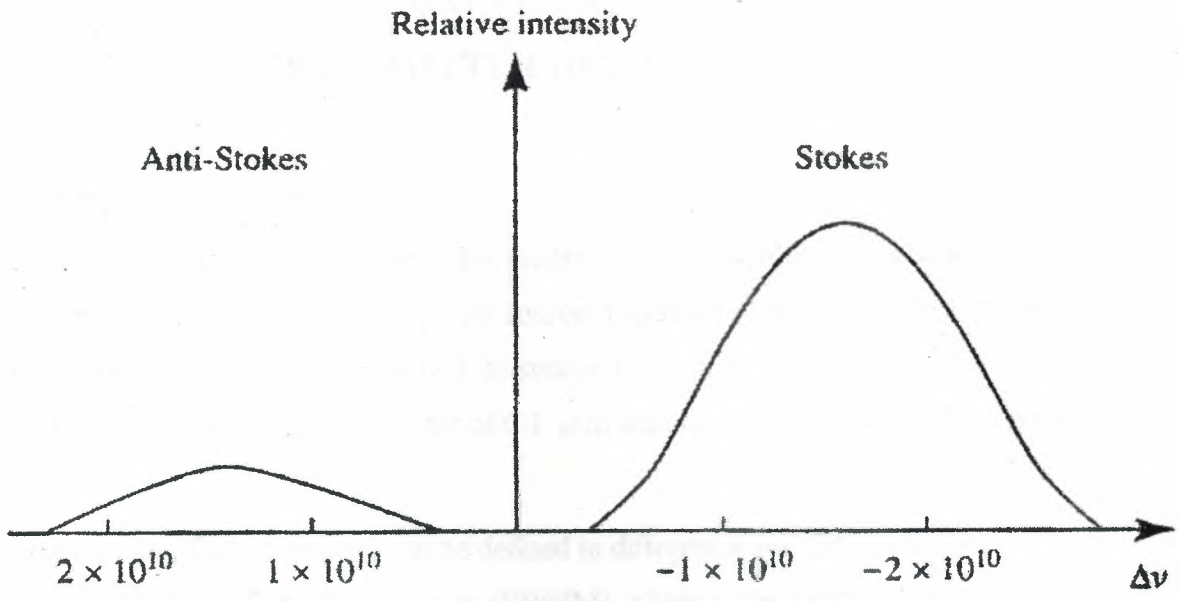

FIGURE 2.8 Raman scattering spectrum in silica; the scattered frequency differs from the incident frequency by an amount $\Delta\nu$.

changes in the surroundings, then the mode field will be affected to some extent. Sensors relying on this basic principle have been made to measure liquid refractive indices and various ionic concentrations and pH values.

CHAPTER 3
TRANSMITTER DEVICES

3.1 Light-Emitting Diodes

Light-emitting diodes are semiconductor diodes that emit **incoherent light** when they are biased by a forward voltage or current source. Incoherent light is an optical carrier with a rapidly drying random phase. Figure 1 illustrates a typical light spectrum of a GaAIAs LED. The line width is of the order of 0.1 μ m with the central wavelength around 0.87 μ m.

The line width of a light source can be defined in different ways. One common definition is called **full-width half-maximum** (FWHM), which is the width between two 50 percent points of the peak intensity. As a numerical example, the FWHM of the line width in figure 3.1 is approximately 0.03 μ m.

There exists a simple relationship between the line width and the spectrum width. Because

$$\lambda f = c$$

where c is the speed of light, by taking the total derivative, we have

$$f\partial\lambda + \lambda\partial f = 0$$

For a given line width $\Delta\lambda$, we thus have

$$\frac{|\Delta\lambda|}{\lambda} = \frac{|\Delta f|}{f}, |\Delta\lambda| = c\frac{|\Delta f|}{f^2}, |\Delta f| = c\frac{|\Delta\lambda|}{\lambda^2} \tag{3.1}$$

where Δf is the corresponding spectral width.

The spectrum width of LEDs depends on the material, temperature, doping level, and ing structure . For AlGaAs devices, the FWHM spectrum width of LEDs is about 2kT/h, where k is the Boltzmann constant and T is temperature in Kelvin. For InGaAs it is about 3 $k\,T/h$. As the doping level increases, the line width also increases.

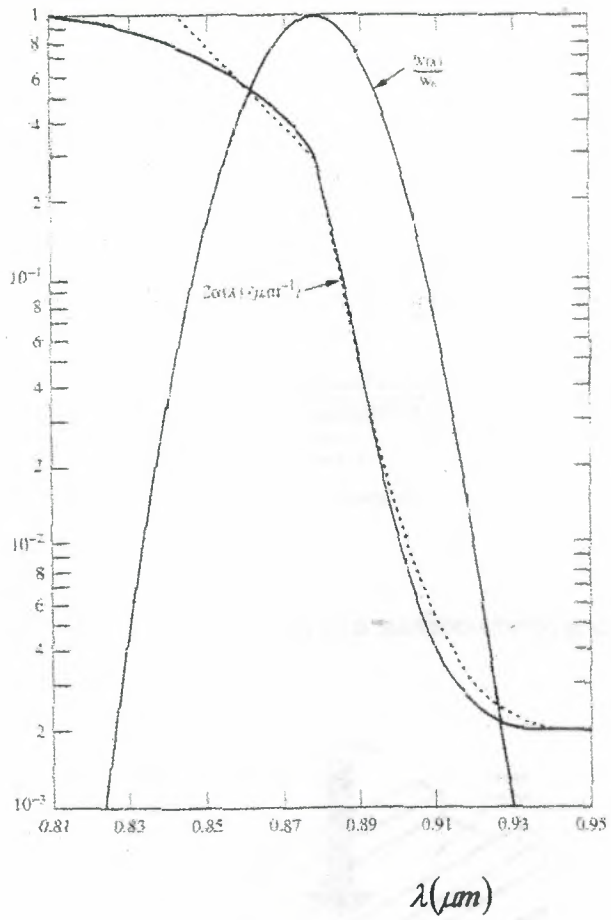

FIGURE 3.1 Line width of an LED

The spectrum width also depends on the light coupling structure of the LED. The light coupling structure couples photons out of the active layer. As illustrated in Figures 3.2 and 3.3, there are two different light coupling structures: **surface emitting** and **edge emitting.** The first type couples light vertically away from the layers and is called a surface emitting or Burrus LED. The second type couples light out in parallel to the layers and is called an edge-emitting LED.

Because of self-absorption along the length of the active layer, edge emitting LEDs have smaller line widths than those of surface-emitting diodes. In addition, because of the transverse wave guiding, the output light has an angle around 30° vertical to the active layer. On the other hand, because surface-emitting LEDs have a large coupling area, it is easier to interface them with fibers. Also, they can be better cooled because the heat sink is close to the active layer.

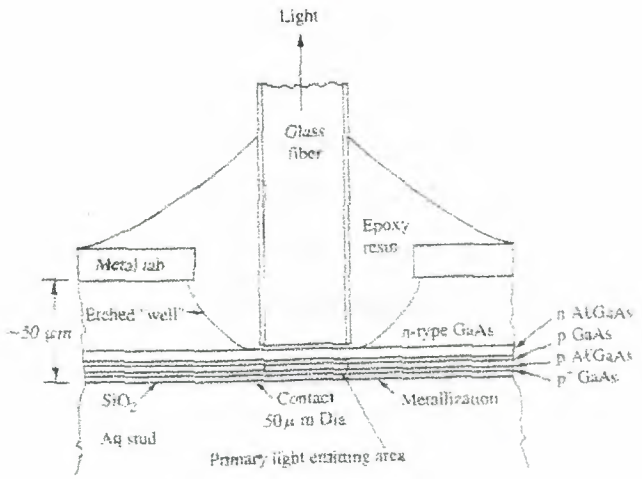

FIGURE 3.2. Illustration of a surface-emitting diode.

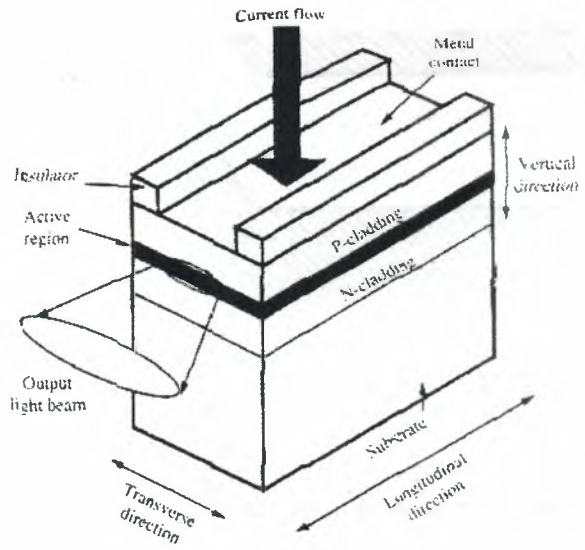

FIGURE 3.3. Illustration of an edge-emitting diode.

3.2 Semiconductor Lasers

Semiconductor lasers are not very different in principle from the light-emitting diodes. A p-n junction provides the active medium; thus, to obtain laser action we need only meet the other necessary requirements of population inversion and optical feedback. To obtain stimulated emission, there must be a region of the device where there arc many

excited electrons and vacant states (i.e. holes) present together. Forward biasing a junction formed from very heavily doped n and p materials achieves this. In such n-type material, the Fermi level lies within the conduction band. Similarly, for the p-type material the Fermi level lies in the valence band. The equilibrium and forward-biased energy band diagrams for a junction formed from such so-called degenerate materials are shown in Fig.4. When the junction is forward biased with a voltage that is nearly equal to the energy gap voltage E_g/e, electrons and holes are injected across the junction in sufficient numbers to create a population inversion in a narrow zone called the *active region* (Fig 3.5).

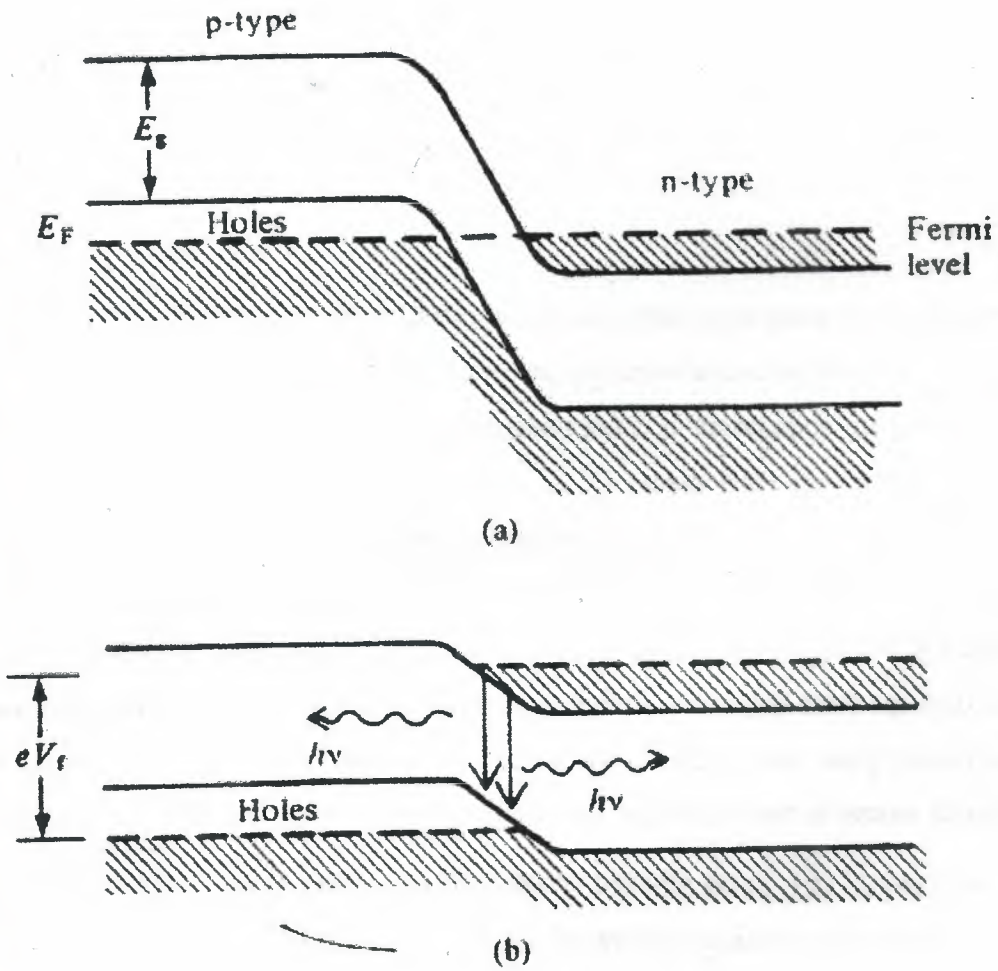

FIGURE 3.4. Heavily doped p-n junction: (a) in equilibrium and (b) with forward biased (the dashed lines represent the Fermi level in equilibrium (a) and with forward bias (b).

The thickness t of the active region can be approximated by the diffusion length L_e of the electrons injected into the p region, assuming that the doping level of the p region is less than that of the n region so that the junction current is carried substantially by electrons. For heavily doped GaAs at room temperature L_e, is 1 -3 μm.

In the case of those materials such as GaAs that have a direct band gap the electrons and holes have a high probability of recombining radiatively. The recombination radiation produced may interact with valence electrons and be absorbed, or interact with electrons in the conduction band thereby stimulating the production of further photons of the nine frequency ($v=E_g/h$). If the injected carrier concentration becomes large enough, the stimulated emission can exceed the absorption so that optical gain can be achieved in the active region. Laser oscillations occur, as usual, when the round trip gain exceeds the total losses over the same distance. In semiconductors, the principal, losses are due to scattering at optical in homogeneities in the semiconductor material and free carrier absorption; The latter results when electrons and holes absorb a photon move to higher energy states in conduction band or valence hand respectively. The carriers then return to lower energy states by non-radiative processes.

In the case of diode lasers, it is not necessary to use external mirrors to provide positive feedback. The high refractive index of the semiconductor material ensures that the reflectance at the material/air interface is sufficiently large even though it is only about 0.32.

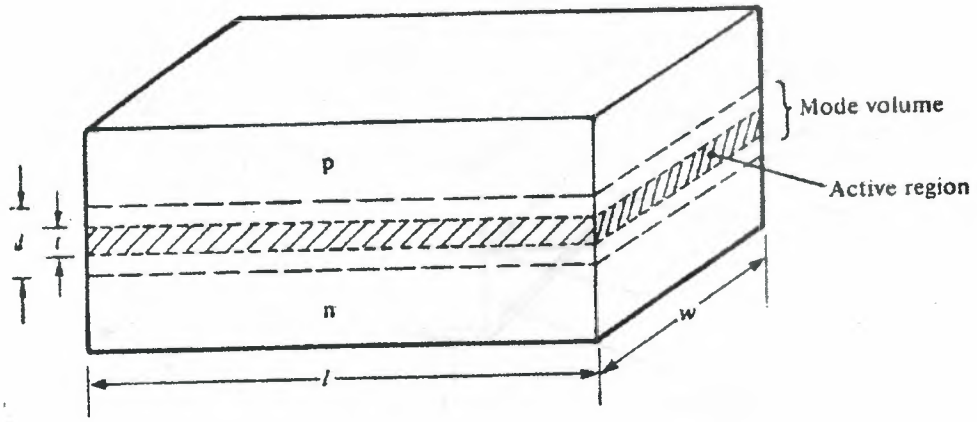

FIGURE 3.5. Diagram showing the active region and mode volume of a semi-conducting laser.

The diode is cleaved along natural crystal planes normal to the plane of the junction w that the end faces are parallel; no further treatment of the cleaved faces is usually necessary, although occasionally optical coatings are added for various purposes. For GaAs, the junction plane is (100) and the cleaved faces are (110) planes.

The radiation generated within the active region spreads out into the surrounding lossy GaAs, although there is, in fact, some confinement of the radiation within a region called the mode *volume* (Fig. 3.5). The additional carriers present in the active region increase a refractive index above that of the surrounding material, thereby forming a dielectric wave-guide. As the difference in refractive index between the centre waveguiding layer and the neighboring regions is only about 0.02, the waveguiding is very inefficient and the radiation extends some way beyond the active region, thereby forming the mode. The waveguiding achieved in simple homojunction laser diodes of the form shown Fig 3.6. only works just well enough to allow laser action to occur as a result of very vigorous pumping. Indeed homojunction lasers can usually only be operated in the pulsed side at room temperature because (lie threshold pumping current density required is so high, being typically of the order of 400 A mm^{-2}.

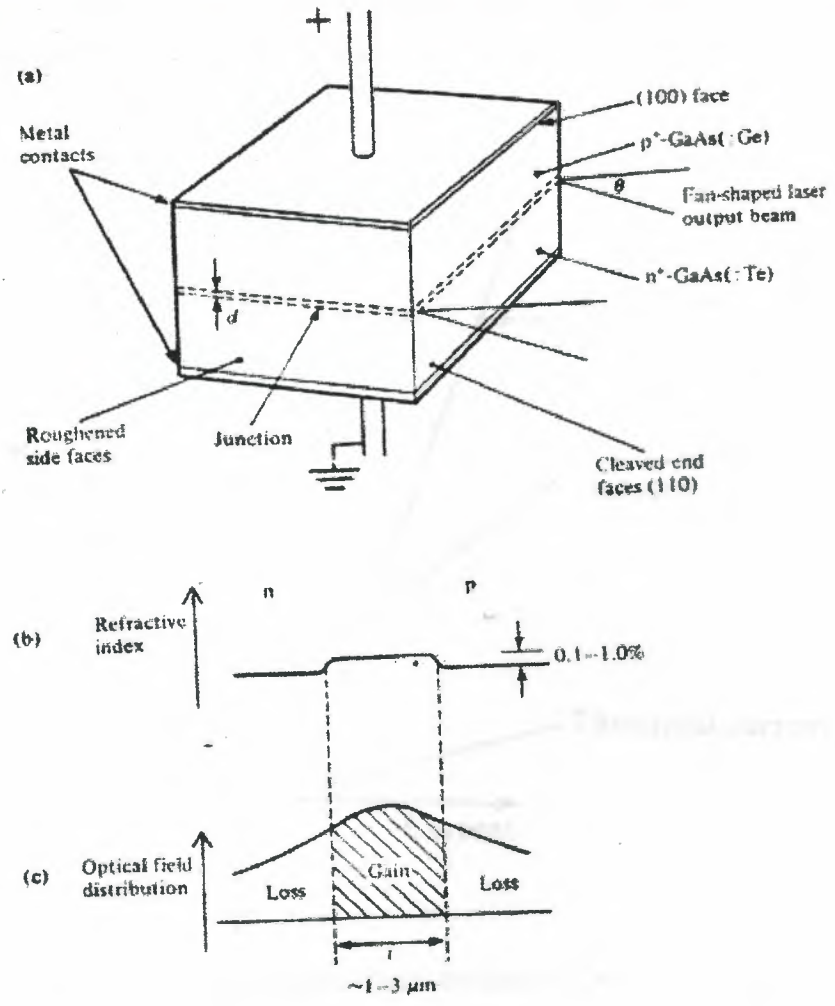

FIGURE 3.6. Schematic construction of GaAs homojunction semiconductor diode laser having side lengths 200-400 μm (a). The emission is confined to the junction region. The narrow thickness d of this region causes a large beam divergence. The very small change in refractive index in the junction region is shown in (b) and (c) shows the resulting poor confinement of the optical radiation to the gain region.

The onset of laser action at the threshold current density is detected by an abrupt increase in the radiance of the emitting region, as shown in Fig. 3.7, which is accompanied by a dramatic narrowing of the spectral width of emission. This is illustrated very clearly in Fig. 3.8, which is accompanied the mode structure below, and at threshold, where the energy has been channeled into relatively small number of modes. If the current is

increased substantially above threshold one mode usually predominates, with a further decrease in the spectral width of the emission.

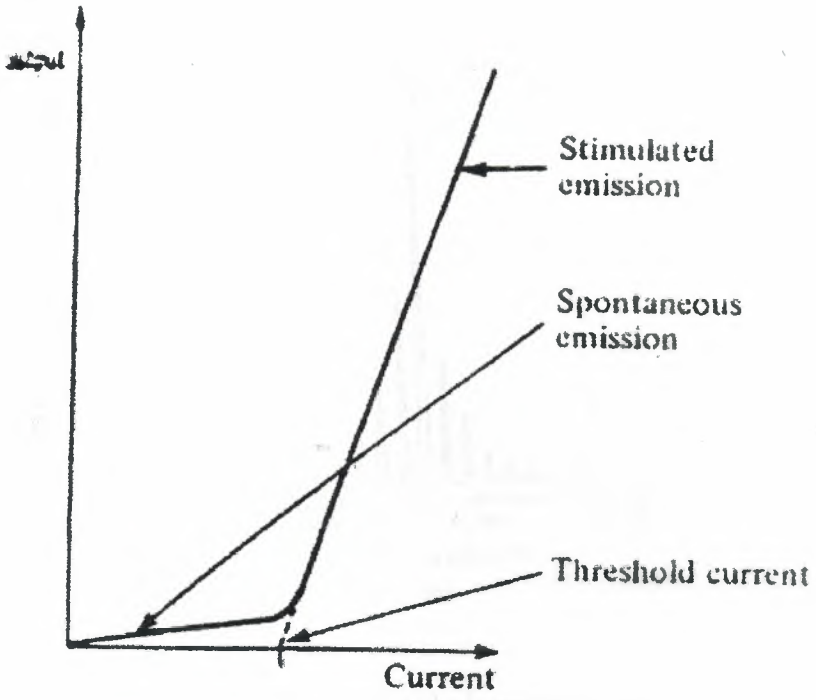

FIGURE 3.7. Light output-current characteristics of an ideal semiconductor laser.

3.2.1 Threshold current density for semiconductor lasers

An exact calculation of the threshold current for a semiconductor laser is complicated by the difficulty of defining what is meant by a population inversion between two *bands of* energy levels. To simplify the problem, however, and to gain some insight into the important factors, we use the idealized structure shown in Fig 3.5. We let the active volume, where population inversion is maintained, have thickness t and the mode volume, where the generated electromagnetic mode is confined, be of thickness *d (d> t)*. In other lasers, the mode volume is usually smaller than the volume within which population inversion is maintained.

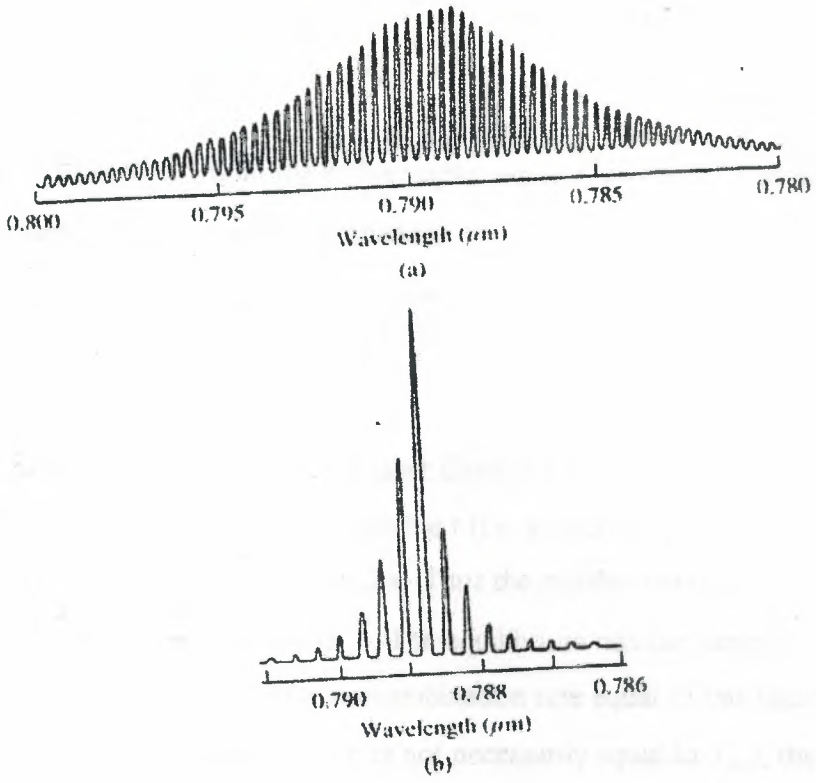

FIGURE 3.8. Emission spectrum of a GaAlAs laser diode both just below (a) and just above (b) threshold. Below threshold a large number of Fabry-Perot cavity resonance can be seen extending across a wide LED-type spectrum. Above threshold only a few modes close to the peak of the gain curve oscillate. For the particular laser shown here the threshold current was 37 mA while spectra (a) and (b) were taken with currents of 35 mA and 39 mA, respectively.

A consequence of the situation in semiconductor lasers is that the portions of the mode propagating outside the active region may be absorbed. Tills offsets to some extent the gain resulting from those parts of the mode propagating within the active region. We allow for this by assuming that the effective-population inversion within the mode volume (d*l*w) is given by reducing the actual population inversion in the active region by the factor *t/d.*

The threshold condition will thus be reached when

$$N_{th} = \left(N_2 - \frac{g_2}{g_1} N_1 \right)_{th} = \frac{d}{t} \left(\frac{8\pi v_0^2 k_{th} \tau_{21} \Delta v n^2}{c^2} \right)$$

We next assume that within the active region we can ignore N_1, that is there is a large number of holes in the valence band ;hence,

$$(N_2)_{th} = \frac{d}{t} \left(\frac{8\pi v_0^2 k_{th} \tau_{21} \Delta v n^2}{c^2} \right) \tag{3.2}$$

If the current density flowing through the laser diode is J A m^{-2}, then the number of electrons per second being injected into a volume t (i.e. a region of thickness t and of unit cross-sectional area) of the active region is *J/e.* Thus the number density of electrons being injected per second *is J/et* electrons s^{-1} m^{-3}. The equilibrium number density of electrons in the conduction band required to give a recombination rate equal to this injection rate is N_2 / τ_e, where τ_e is the electron lifetime (τ_e is not necessarily equal to τ_{21}), the spontaneous lifetime, since non-radiative recombination mechanisms are likely to be present).

The threshold current density is then given by

$$\frac{(J)_{th}}{et} = \frac{(N_2)_{th}}{\tau_e}$$

Substituting from equation. (2) we have

$$(J)_{th} = \frac{et}{\tau_e} \frac{d}{t} \left(\frac{8\pi v_0^2 k_{th} \tau_{21} \Delta v n^2}{c^2} \right)$$

3.2.2 Power output of semiconductor lasers

As the injection current increases above threshold, laser oscillations build up and the resulting stimulated emission reduces the population inversion until it is clamped at the threshold value. We can then express the power emitted by stimulated emission as

$$P = A[J - (J)_{th}] \frac{\eta_i h\nu}{e}$$

Part of this power is dissipated inside the laser cavity and the rest is coupled out via the end crystal faces. These two components are proportional to γ and $(1/2l)\ \ln(\ 1/R_1R_2)$ respectively. Hence we can write the output power as

$$P_0 = \frac{A[J-(J)_{th}]\eta_i h\nu}{e} \frac{[(1/2l)\ln(1/R_1R_2)]}{\gamma+(1/2l)\ln(1/R_1R_2)} \tag{3.3}$$

The external differential quantum efficiency η_{ex} is defined as the ratio of the increase in photon output rate resulting from an increase in the injection rate (i.e. carriers per second), that is

$$\eta_{ex} = \frac{d(P_0/h\nu)}{d\{(A/e)[J-(J)_{th}]\}}$$

From equation (3.3) we can write η_{ex} as

$$\eta_{ex} = \eta_i\left(\frac{\ln(1/R_1)}{\gamma l+\ln(1/R_1)}\right) \tag{3.4}$$

assuming that $R_1 = R_2$. Equation (3.4) enables us to determine the internal quantum efficiency from the experimentally measured dependence of η_{ex} on l; η_i in GaAs is usually in the range 0.7-1.0. Now if the forward bias voltage applied to the laser is V_f, then the power input is $V_f AJ$ and the efficiency of the laser in converting electrical input to laser output is

$$\eta = \frac{P_0}{V_f AJ} = \eta_i\left(\frac{J-(J)_{TH}}{J}\right)\left(\frac{h\nu}{eV_f}\right)\frac{\ln(1/R_1)}{\gamma l+\ln(1/R_1)} \tag{3.5}$$

$eV_f \approx l h\nu$ and therefore, well above threshold ($J \geq (J)_{th}$) where optimum coupling ensures that $(1/l)\ \ln(1/R_1) \geq \gamma$, η approaches η_i. As noted above, η_i is high (≈ 0.7) and thus semiconductor lasers have a very high power efficiency.

3.2.3 Heterojunction lasers

As we noted above, the threshold current density for homojunction lasers is very large owing to poor optical and carrier confinement Dramatic reductions in the threshold current density to values of the order of 10 A mm^{-2} at room temperature coupled with higher efficiency can he achieved using lasers containing heterojunctions. The properties of heterostructure lasers which permit a low threshold current density and CW operation at room temperature can be illustrated with the double heterostructure (DH) laser illustrated in Fig. 3.9. In this structure, a layer of GaAs, for example, is sandwiched between two layers of the primary compound

$Ga_{1-x}Al_xAs$ which has a wider energy gap than GaAs and also a lower refractive index. Both N-n-P and N-p-P structures show the same behaviour (where N and P represent the wider bandgap semiconductor, according to carrier type).

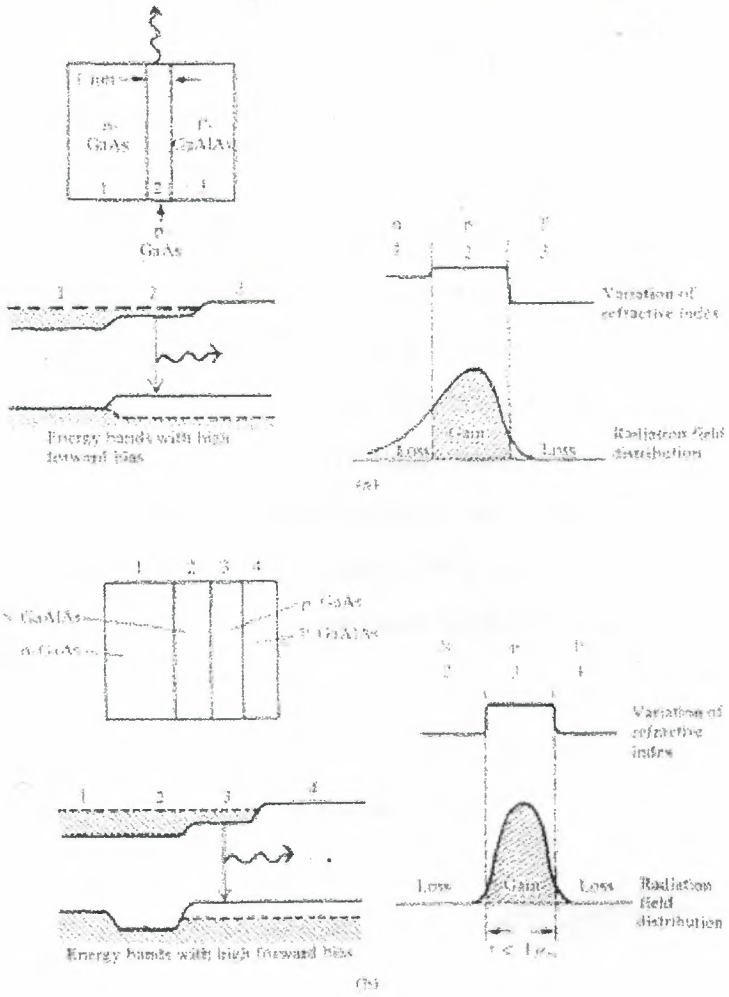

FIGURE 3.9 Diagram illustrating the action of single (a) and double (b) heterojunction structures in confining the carriers and radiation to the gain region (as before, in the diagrams of the energy bands, the dashed lines represent the Fermi levels after forward bias has been applied).

Figure 3.9(b) also shows that carrier and optical confinement may be achieved simultaneously. The bandgap differences form potential barriers in both the conduction and valence bands which prevent electrons and holes injected into the GaAs layer from diffusing away. The GaAs layer thus becomes the active region, and it can be made very narrow so that t is very small, typically about 0.2 μm. Similarly, the step change in refractive index provides a very much more efficient waveguide structure than was the case in homojunction lasers. The radiation is therefore confined mainly to the active region. In

addition, the fraction of the propagating mode which lies outside the active region is in a wider bandgap semiconductor and is therefore not absorbed, so that γ is much smaller than in homojunction lasers.

Further reductions in threshold current can be obtained by restricting the current along the junction plane into a narrow 'stripe' which may only be a few micrometers wide. Such stripe geometry lasers have been prepared in a variety of different ways; typical examples are shown in Fig.3.10. In Fig.3.10(a), the stripe has been defined by proton bombardment of the adjacent regions to form highly resistive material, whereas in Fig. 3.10(b) a mesa structure has been formed by etching; an oxide mask prevents shorting of the junction during metallization to form contacts. With stripe geometry structures, operating currents of less than 30 mA can produce output powers of about 10 mW.

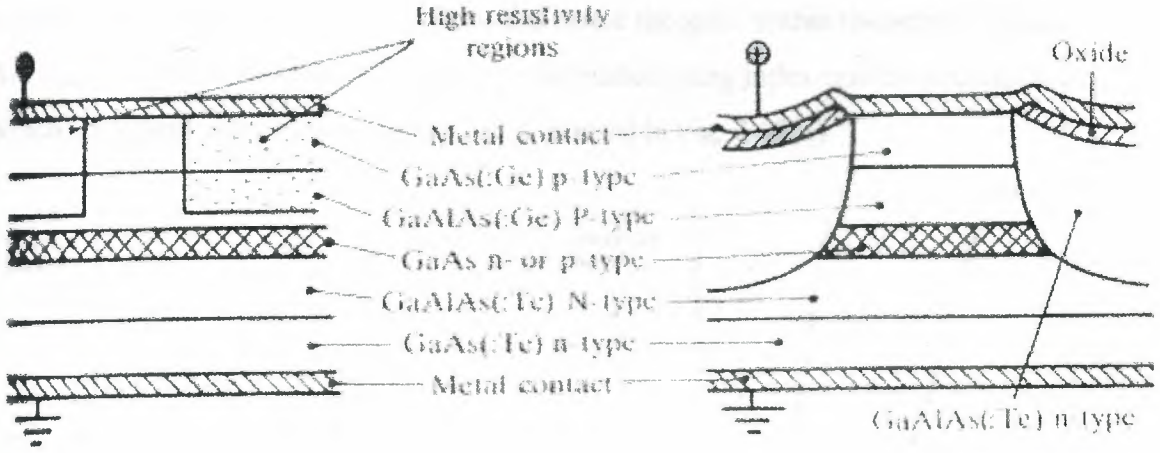

Figure 3.10 Schematic cross-section (end view) of two typical stripe geometry laser diodes: (a) the stripe is defined by proton bombardment of selected regions to form high resistivity material; (b) the stripe is formed by etching a mesa structure and then GaAlAs is grown into the previously etched outsides of the active region to form a 'buried stripe' structure.

Stripe geometry devices have further advantages including the facts that (a) the radiation is emitted from a small area which simplifies the coupling of the radiation into optical fibers and (b) the output is more stable than in other lasers. A close examination of typical light output-current characteristics reveals the presence of 'kinks' as shown in figure 3.11(a). These 'kinks' are associated with a sideways displacement of the radiating filaments within the active region (the radiation is usually produced from narrow filaments within Ac active region rather than uniformly from the whole active region). This lateral instability is caused by interaction between the optical and carrier distribution which arises because the refractive index profile, and hence the waveguiding characteristic, is determined, to a certain extent, by the carrier distribution within the active region. The use of very narrow stripe regions limits the possible movement of the radiating filament and eliminates the 'kinks' in the light output-current characteristics as shown in Fig. 3.11(b). The structures shown in Fig. 3.12 are referred to as *gain guiding* because the width of the gain region is determined by the restriction of the extent of the current flow, which of course creates the population inversion, and hence the gain, within the active region. Alternatively stripe geometry lasers can be fabricated using index-guided structures, in which an optical waveguide is created as illustrated in Fig. 3.12(a).

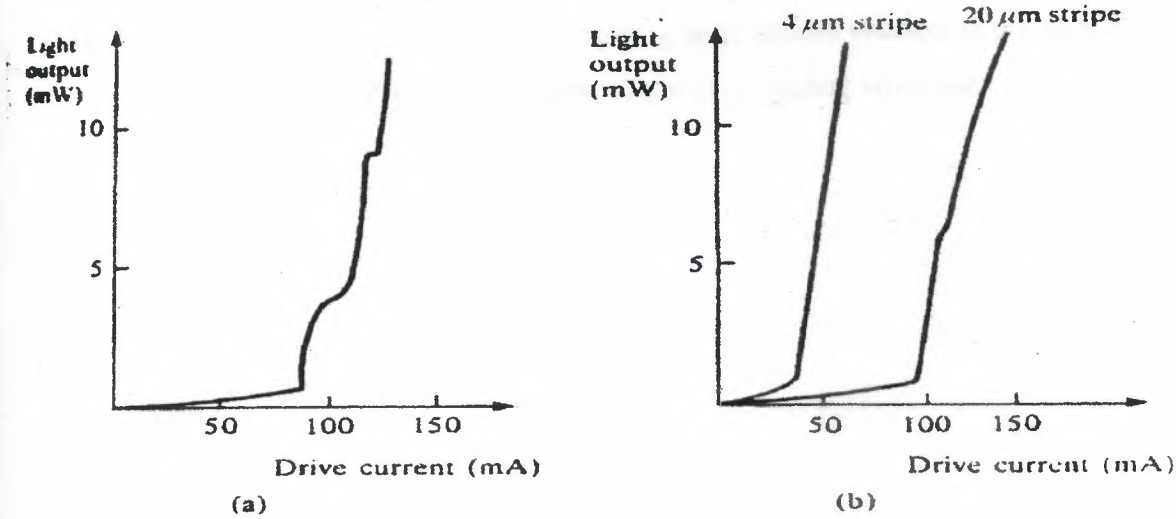

FIGURE 3.11 Light output-current characteristics of (a) a laser showing a lateral instability or 'kink' and (b) stripe lasers, in which the 'kinks' have disappeared.

structures in practice is quite complex; a relatively simple one is shown in Fig. 3.13 (a). One relatively straightforward alternative is to change the thickness of the semiconductor layer next to the waveguide (Fig. 12b) which creates an effective refractive index difference between the active region and those next to it in the same layer. A device based on this technique is shown in Fig. 13(b). Several others buried layer heterostructure devices.

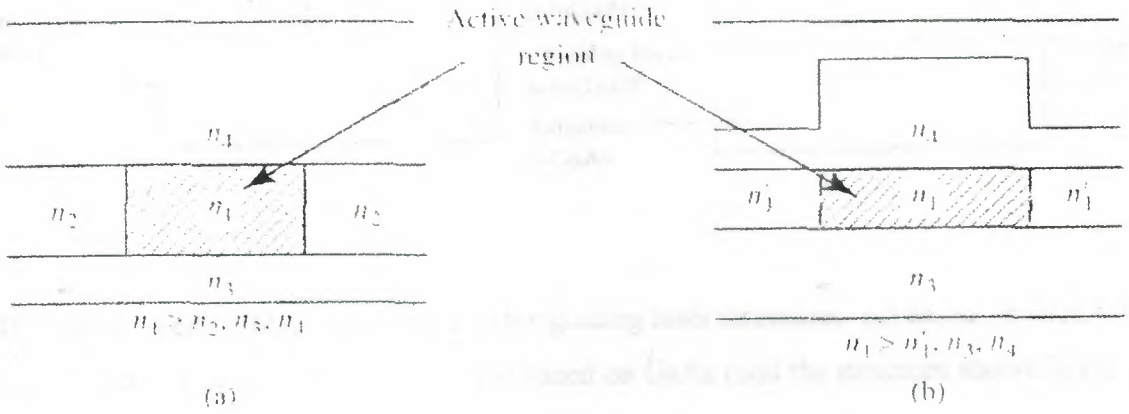

FIGURE 3.12 Schematic representation of (a) a buried heterostructure which acts as a waveguide (end view) and (b) a structure which behaves like a buried heterostructure; the varying thickness of the layer next to the guiding layer creates changes in the apparent refractive index, thereby achieving a waveguiding structure.

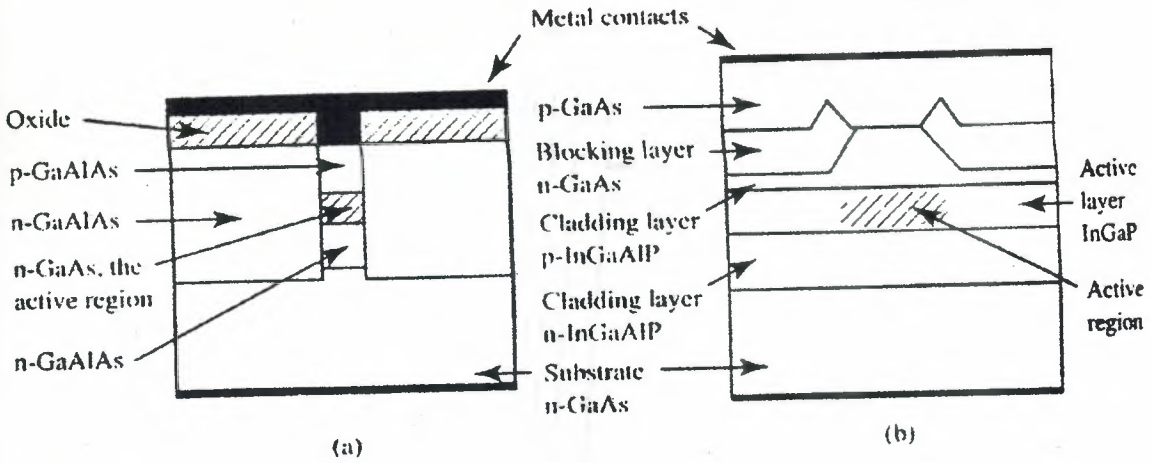

FIGURE 3.13 Buried heterostructure index guiding laser structures: (a) based on lnGaAsP (and the structure shown in fig 12a); (b) based on GaAs (and the structure shown in fig 12b).

In general gain-guided lasers are easier to fabricate than index-guided lasers, but their poorer optical confinement limits the beam quality, and makes stable, single mode operation difficult to achieve. On the other hand the fact that the beam spread is greater reduces the optical power density at the output face thereby reducing the risk of damage (see below)

These include the temperature dependence of the threshold current, output beam spread, degradation and the use of materials oilier than GaAlAs.

The threshold current density J_{th} increases with temperature in all types of semiconductor laser but, as many factors contribute to the temperature variation, no single expression is valid for all devices and temperature ranges. Above room temperature, which is usually the region of practical interest, it is found that the ratio of J_{th} at 70°C to J_{th} at 22°C for GaAlAs lasers is about 1.3-1.5 with the lowest temperature dependence occurring for an aluminium concentration such that the bandgap energy difference is 0.4 eV. Typical light output-current characteristics for a GaAlAs DH laser are shown in Fig. 3.14.

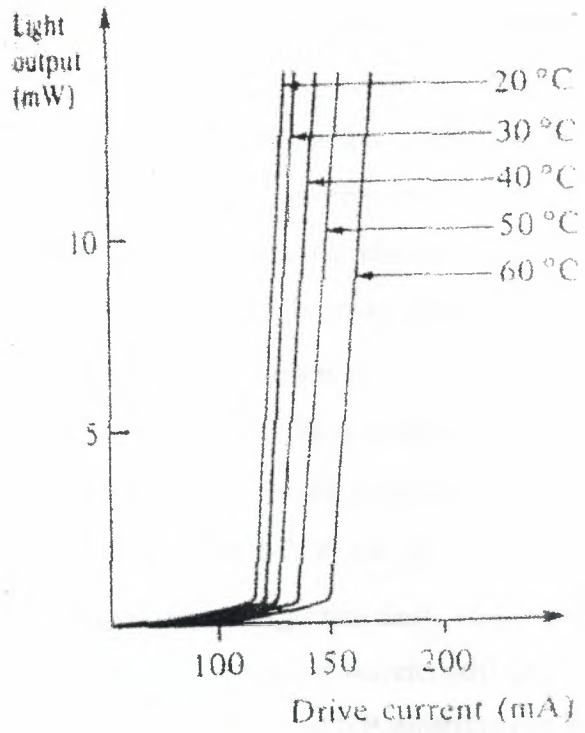

FIGURE 3.14. Light output-current characteristics of a 20 μm stripe laser as a function of temperature.

The angular spread of the output beam depends on the dimensions of the active region and the number of oscillating modes (which, in turn, depends on the dimensions of the active region, the refractive index and the pump power). For wide active regions, we find that the beam divergence both parallel to (θ_{11}) and perpendicular to ($\theta_{\perp}$) the plane of the junction is given approximately by simple diffraction theory. Thus, normal to the junction plane we have $\theta_{\perp}=1.22\lambda/d$. For DH lasers, where the active region is much narrower, $\theta_{\perp}$ is given approximately by $\theta_{\perp} \approx 1.1*10^{3}x(t/\lambda)$, where x is the mole fraction of aluminium. Thus for a DH laser with t= 0.1 μm, x=0.3 and λ=0.9 μm, we find $\theta_{\perp}$=37° (in good agreement with experimental observations).

Until recently, the system $Ga_{1-x}Al_{x}As$/GaAs was the most widely investigated and used for the production of DH lasers. There are many reasons for this, including the facts that (a) GaAs is a direct bandgap semiconductor which can easily be doped n- or p-type; (b)

the ternary compound $Ga_{1-x}Al_xAs$ can be grown over a wide range of compositions, and not only does it have a very close lattice match to GaAs ($\approx 0.1\%$) for all values of.x (thus there is low interfacial strain between adjacent layers and consequently very few strain-induced defects at which non-radiative recombination may occur), but it is also a direct bandgap semiconductor for x<0.45; and (c) the relative refractive indices and bandgaps of GaAs and $Ga_{1-x}Al_x$ provide for optical and carrier confinement.

In optical fiber communications, however, it is desirable to have a laser emitting at wavelengths in the region 1.1 to 1.6 μm where present optical fibers have minimum attenuation and dispersion. Wavelengths in this range can be obtained from lasers fabricated from quaternary compounds such as $Ga_xIn_{1-x}As_{1-y}P_y$ because of the wide range of bandgaps and lattice constants spanned by this alloy. Figure 3.15 shows the lattice constant variation with bandgap (and emission wavelength) for this alloy. By suitable choice of.x and y, exact lattice matching to an InP substrate can be achieved and strain-free heterojunction devices can be produced. The GaInAsP layers may be grown on InP substrates by liquid phase, vapour phase or molecular beam epitaxial methods. A typical DH stripe contact laser diode of GaInAsP/InP emitting at 1.1-1.3 pin is shown schematically in Fig. 3.16.

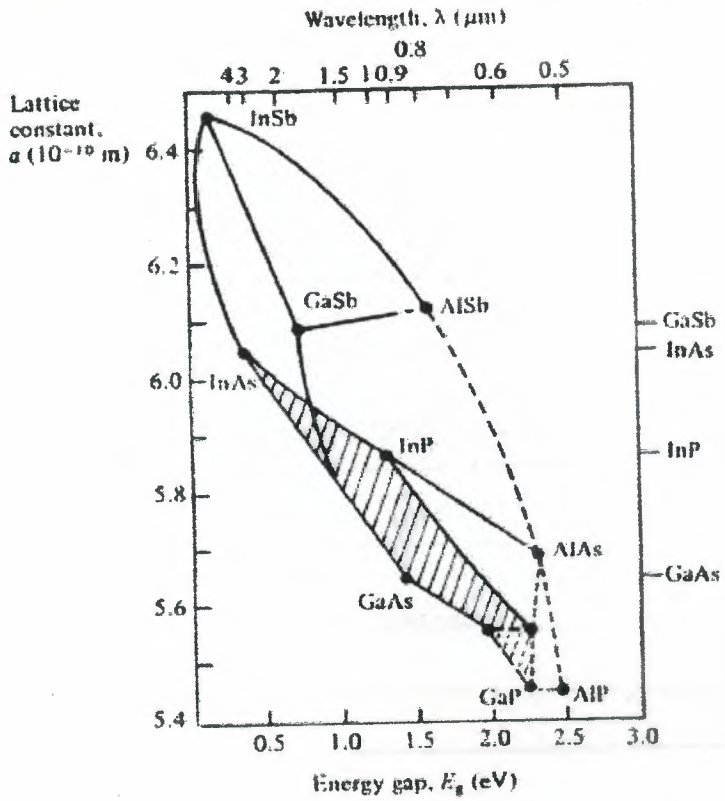

FIGURE 3.15 Lattice constant versus energy gap for various lll-V compounds.

The question of laser reliability is also important in relation to applications such as telecommunications. Laser life may be limited by 'catastrophic' or 'gradual' degradation. Catastrophic failure results from mechanical damage to the laser facets due to too great an optical flux density. The damage threshold is reduced by the presence of flaws on the facets: however, it may be increased by the application of half-wave coatings of materials such as Al_2O_3. While facet damage is more likely in lasers operating in the pulse mode, it can also occur in CW-operated lasers. This is so especially in the central portion of the active region of the lasers where the optical flux density is greatest. Uncoated lasers with stripes about 20 μm tend to fail catastrophically when the optical flux exceeds about 10^9Wm^{-2}.

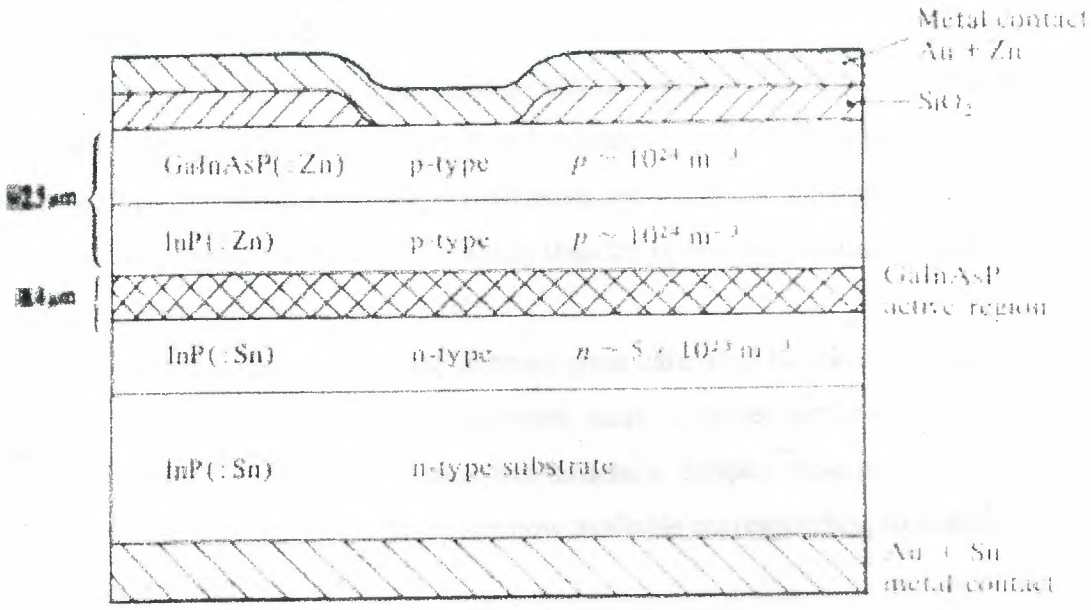

FIGURE 3.16 Schematic diagram of a double heterojunction stripe contact laser diode of the quaternary compound $Ga_x In_{1-x} As_{1-y} P_y$ on an InP substrate with (100) orientation.

Gradual degradation depends principally on the current density, but also on the duty cycle and fabrication process. It has been observed that as time elapses the threshold current density increases, 'dark' lines develop in the emission and then the CW output falls off drastically.

The development of dark lines is apparently related to the formation, in the vicinity of the active region, of so-called *dark-line defects,* which act as non-radiative recombination centres. Dark-line defects are attributed to defects such as dislocations, which may have a number of sources. These include (a) edge dislocations formed to relieve stress caused by interfacial lattice mismatch, (b) bonding of the laser to the heat sink and

Defects may be formed in the active region during device fabrication or penetrate into it during subsequent operation. Dark-line defects may grow owing to a process called dislocation (i.e. the movement of dislocations involving atomic transport to or away from the dislocation) and extend throughout the device structure.

Dislocation growth may be stimulated by carrier injection and recombination. GaAs lasers initially containing dislocations are found to degrade at a much higher rate than those that are initially dislocation free. Furthermore, devices with exposed edges that contain edge defects also degrade more rapidly than those in which recombination is restricted to internal regions of the crystal.

Thus, to produce lasers with long lifetimes great care must be taken with substrate selection and wafer processing and crystal growth must be carried out under, ultra clean conditions to fabricate a laser with a strain-free structure. Despite these problems, lasers with life times in excess of 40 000 hours are now available corresponding to continuous operation over a 5 year period.

3.2.4 Quantum Well Lasers

In very narrow semiconductor layers (i.e. the quantum wells) there is a very significant increase in the density of states near the bottom of the conduction band and the top of the valence band. The increased densities of states enable a population inversion to be obtained more easily and, as a consequence of this, and the very small active volume, the threshold currents in quantum well lasers are about a factor of 10 less than those in DH lasers. In addition, quantum well lasers have low temperature sensitivity and their output characteristics are 1 free from kinks. Such lasers are therefore increasingly replacing DH lasers as materials growth technology improves enabling the controlled fabrication of very thin structure in an increasingly wide range of semiconductors.

One of the problems with the single quantum well (SQW) structure described above that, because of the extreme narrowness of the active region, optical confinement is very poor. This causes higher losses and lends to negate the potential advantages of low threshold currents. One way of reducing these problems is to use the multiple quantum well (MQW) structure illustrated in Fig. 3.16(b), which because of its greater thickness gives better optical confinement and beam definition.

The single quantum well can be extended to coupled quantum wells, to form the MQW laser, Figs 13.6(a) and (b). In such devices very thin intervening GaAlAs harrier layers, for example, may couple several GaAs quantum wells. The overall active region is now thicker so that the carriers, which are not captured and therefore able to recombine in the first well, may be captured by the second or a subsequent well. Although MQW lasers have larger threshold currents than single quantum well lasers, where the threshold current may be as low as 1 mA or less, they can emit more optical power, and their structure results in better optical confinement.

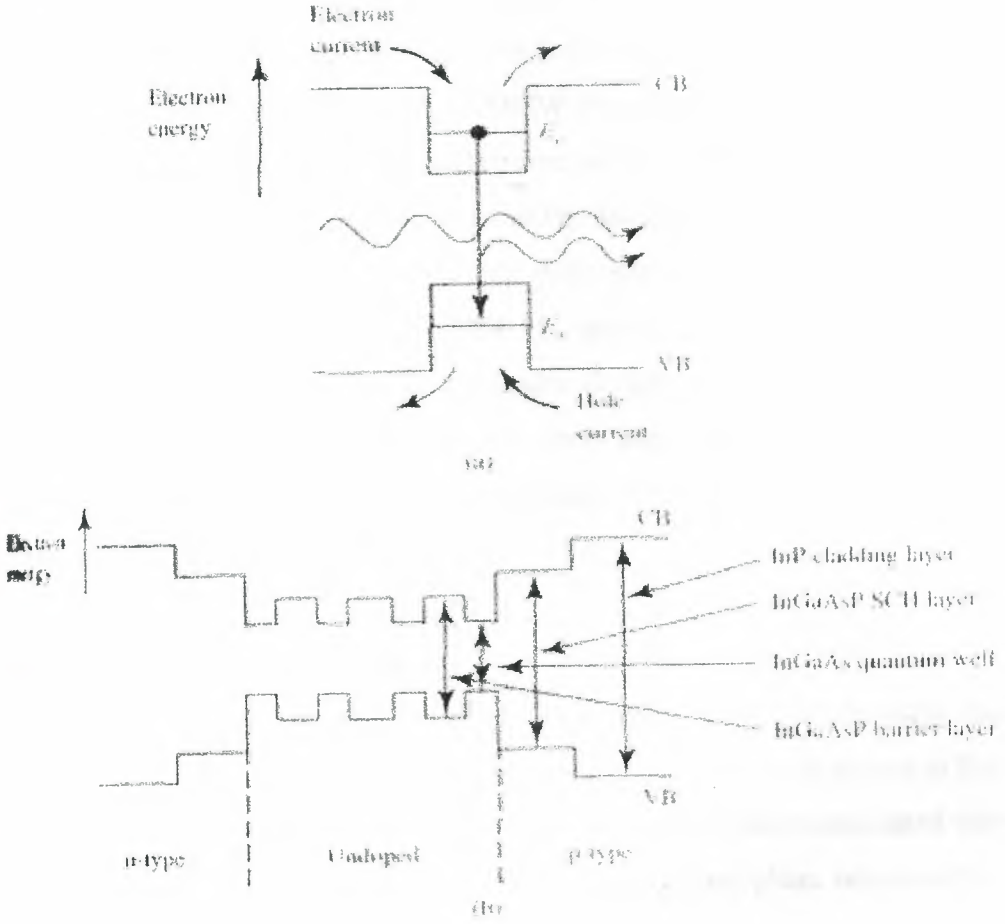

FIGURE 3.17. (a) Stimulated emission in a single quantum well. (b) The energy band diagram for a typical multiple quantum well(MQW) laser with separate confinement heterostructure (SCH) layers.

Further improvement in both optical and carrier confinement can be obtained by adding cladding layers and separate confinement heterostructure (SCH) layers as illustrated in Figure 3.17(b). The SCH layers are chosen to have a refractive index which is greater than that of the cladding layers, so that total internal reflection occurs at the boundary. The SCH layers also, together with the barrier layers, have an energy gap, E_g, between that of the cladding layers and quantum wells, so that the charge carriers are confined between the cladding regions - hence the SCH layers are so named because the carriers and photons are separately confined.The cladding layers are doped n- and p-type, while the MQW layers are undoped. Under forward bias the electrons and holes are injected from the cladding layers, diffuse across the SCH layers and enter the MQW structure where they recombine. The cavity mirrors are provided by the high reflectance of the device faces.

The lasing region of the active layer can be restricted to a narrow strip thereby in effect confining the carriers in two dimensions. Such structures are referred to as quantum wire microcavities, and are the basis of QWR-MC lasers. Further restriction, that is into three dimensions, gives rise to quantum dot lasers. Despite manufacturing difficulties quantum wire and quantum dot arrays are potentially important because, in addition to very low threshold currents, they have very high modulation bandwidths, narrow spectral linewidths and low temperature sensitivity.

3.2.5 Arrays - Vertical Cavity Lasers

The output power from semiconductor lasers may be increased by using one-dimensional arrays of single mode lasers on a bar of semiconductor as shown in Fig 3.18(a). Such arrays are called *phased* arrays since the electric fields associated with the individual elements interact with each other resulting in definite phase relationships between them.

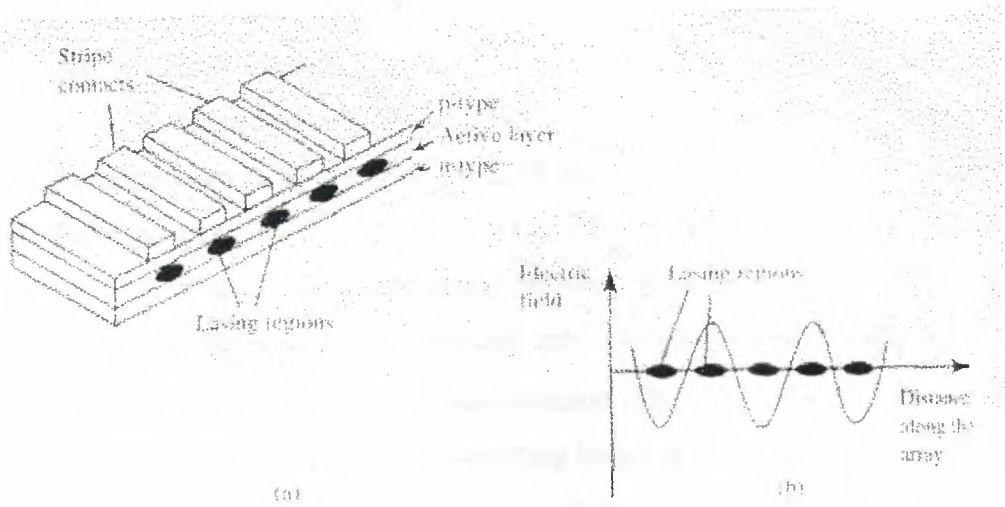

FIGURE 3.18 A Linear array of lasing formed within a single semiconductor bar: (a) shows the stripe contacts, which define the lasing regions; (b) shows the electric field distribution; the field is zero midway between the lasing elements where there is absorption rather than gain.

Frequently the phase difference between adjacent elements is arranged to be 180° (Fig. 3.18b), so that the resultant field midway between the active regions of adjacent elements is zero. These midway regions are more likely to exhibit absorption rather than gain so the overall losses are minimized. Unfortunately, this phase relationship results in a power distribution in the plane the active layer with an angular distribution comprising two lobes rather than a single one, In fact a single-lobed power output distribution can be achieved if the phase between adjacent elements is zero. The phase difference can be controlled by a number of techniques including variation of the lateral spacing between the elements in the array.

Linear arrays are available in widths up to 10 mm and can generate CW powers up to 20 W. Outputs of 10 kW or more can be achieved by stacking up to 200 linear array bars together It is important to realize that as the power output increases, so too do the cooling requirements; it is therefore vital to consider carefully how to remove excess heat to prevent the array from self-destructing. Very high power arrays, for example, require water-cooling

VERTICAL CAVITY LASERS

A structure, which particularly lends itself to the fabrication of laser arrays, is the *vertical cavity* surface-emitting laser (VCSEL). While in traditional, horizontal edge-emitting lasers the resonant cavity is in the plane of the active layer, in VCSELs (Fig. 3.19) it is perpendicular to this plane. The light resonates between mirrors on the top and bottom of the laser wafer so that the photons pass through only a very short length (typically $\leq 1\ \mu m$) of active medium, in which they can stimulate emission. Thus VCSELs have very much lower round trip gain than horizontal edge-emitting lasers, and consequently require highly reflecting mirrors (reflectance ≥ 0.9) to sustain oscillations. Clearly the reflectance of the semi-conductor facets at about 0.32 is insufficient and multilayer mirrors comprising several tens of alternate $\lambda/4$ coatings of AlAs and AlGaAs are often used as illustrated in Fig. 3.19.

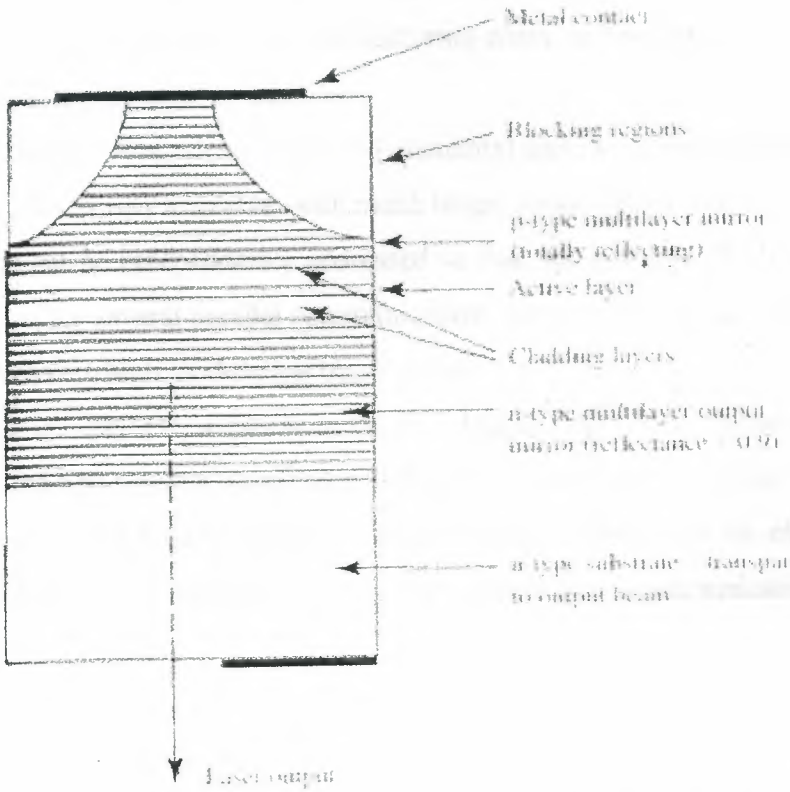

FIGURE 3.19. A Vertical cavity, surface-emitting laser (VCSEL).

The active layer comprises an SQW or MQW structure, which together with cladding and confinement layers forms an optical cavity, which is one wavelength thick. The active region is arranged to be at the peak of the standing wave formed between the mirrors.

All vertical cavity lasers emit from their surface rather than their edge (though surface-emitting lasers are available which do not have vertical cavities). The emission is typically from round or square areas, which are about 10 pm wide so that the output beams are highly symmetrical in contrast to those of edge-emitting devices. Divergence angles are only 7° to 10°, and by using microlenses integrated onto the device surface some 90% of the output may be coupled into optical fibers.

In addition to the symmetrical beam profile, low threshold currents and good tempera-lure stability of VCSELs, a major attraction of surface emission is the ability to fabricate monolithic one- and two-dimensional arrays of many elements. In practice it is possible to grow many thousands of VCSELs simultaneously on a 3 inch (75 mm) wafer and, equally importantly in relation to manufacturing costs, to test these and measure the optical and electrical properties in situ.

A range of one-dimensional (up to 64 elements) and two-dimensional (8*8) VCSEL arrays is now commercially available, with much larger arrays under development. Each laser in the array can be independently addressed so that, for example, the lasers in an array can act as sources for several parallel communication channels, particularly as vertical cavity lasers have very high modulation bandwidths.

VCSELs are currently available in the wavelength range 650-690 nm using GaAs/GaAlAs and 850-980 nm using InGaAs/GaAs semiconductor systems. Unfortunately efforts to fabricate VCSELs operating CW at room temperature in the wavelength range 1300-1550 nm, which is so important for long-range fiber optic communications, have not yet succeeded.

3.2.6 Short Wavelengths Lasers

Recently there has been increased demand for shorter wavelength semiconductor lasers for applications such as compact disc and optical storage, colour printing and semiconductor lithography. The shorter the wavelength the smaller is the area of a focused beam (- λ),

thereby allowing increased storage capacity, and similarly the narrower the features than can be created with optical lithography.

Recently red lasers based on AlGaInP have become available for use in barcode readers, while quantum well lasers with GaInP active layers have enabled wavelengths as short as 630 nm to be generated.

Despite the improved reliability of semiconductor lasers emitting in the red and yellow parts of the spectrum, reliable lasers emitting in the green and blue remain elusive. However, recent improvements in materials technology have enabled CW, room temperature operation to be demonstrated in so-called II-VI semiconductors such as ZnSe, ZnMgSSe and related compounds on GaAs substrates. Alternatively CW laser operation at a wavelength of 417 nm has been obtained from devices based on gallium nitride (GaN) which is rather difficult material to work with. These lasers contain an MQW structure of 26 quantum wells 2.5 nm thick $In_{0.2}Ga_{0.8}N$ separated by layers of $In_{0.05}Ga_{0.95}N$ barriers 5.0 nm thick giving a total thickness of some 200 nm. The threshold current densities and operating voltages are still rather high at about 10 kA cm^{-2} and 25 V respectively, but these are being steadily reduced as the technology develops.

The requirement of close lattice matching (i.e. $\approx 0.1\%$) for (lie components in a heterojunction structure made it difficult to cover some wavelength ranges. A recent development, which has helped in this respect, is the discovery that very thin layers (less than a few tens of nanometers) can accommodate a lattice mismatch of more than 1%. These layers are called strained lattice layers, and the technique was first used to enable the fabrication of InGaAs/GaAs lasers emitting at 980 nm. Strained layers are also used in quantum well structures to produce active layers, which need not be precisely matched to the surrounding layers. This technique was used to produce the lasers based on ZnSe, which emit in the green at a wavelength of 525 nm, and to enable GaN to be grown on mismatched substrates such as sapphire, which has the same crystal structure, or silicon nitride.

3.2.7 Superlumincscent Light-Emitting Diodes

We end this section with a discussion of a device, which, while not a laser, does depend on optical amplification, namely the super luminescent light-emitting diode (SLD).

SLDs have a structure, which is rather similar to that of conventional injection laser diodes and edge- emitting LEDs; indeed the SLD has optical properties, which are intermediate between these two devices. Both stripe geometry and burried heterostructure SLDs are available, emitting at a range of wavelengths. In contrast to laser diodes, however, the non-output end of the device is made optically lossy to minimize feedback and suppress laser oscillations. This can be achieved simply by roughening the cleaved surface of the device to scatter the light, or by adding an antireflection coating.

In operation the injection current is increased until stimulated emission, and hence amplification of spontaneous emission, just occurs. That is, operation is on the 'knee' of the laser diode output characteristic shown in Fig. 3.7. Although there are no oscillations the stimulated emission, within a single pass through the device, provides gain so that the device output increases rapidly with increase in current-this is termed *superradiance or superluminescence.* High optical output power can be obtained together with a narrowing of the spectral width, which also results from the stimulated emission.

These characteristics of the output from SLDs give a number of advantages over conventional LEDs in relation to their use in fiber optic communications. These include: higher power outputs (up to 60-100 mW), a more directional light beam, and a narrower spectral line width, all of which improve the source to fiber coupling. Moreover, the superradiant emission process within SLDs tends to increase their modulation bandwidth. In contrast to conventional LEDs, however, SLDs suffer from having a non-linear output characteristic and an increased temperature dependence of the output power. Compared with laser diodes they require substantially higher injection currents (by a factor of about three) to produce a similar power output.

CHAPTER 4
OPTICAL AMPLIFIERS

There are two primary types of optical amplifiers: **semiconductor amplifiers** and **fiber amplifiers**. A semiconductor amplifier is a laser diode operated below threshold. Therefore, it can amplify input signals but cannot generate a coherent light by itself. A fiber amplifier is a fiber section that has a positive medium gain. To achieve this, the fiber is doped with ions such as Er^{+3}. When external optical pumping excites carriers of the doped ions, they can be stimulated back to the ground state by the incident light. This results in stimulated emission and provides the positive optical gain. Among all optical fiber amplifiers, erbium-doped fiber amplifiers (EDFA) that amplify light at around 1.55 μm are the most mature. For amplification at around 1.3 μm, neodymium- and praseodymium-doped fiber amplifiers have also been recently developed. Compared to EDFA's, they are relatively immature but are promising.

4.1 Semiconductor Amplifiers

As mentioned earlier, semiconductor amplifiers are laser diodes that are biased below the threshold current. To provide amplification, the active layer of a semiconductor amplifier has a positive medium gain but not large enough for laser emission. This section describes and analyzes various semiconductor amplifier characteristics. In particular, the section quantifies the medium gain at a given current pumping, explains the gain saturation effect, characterizes the interchannel interference in multichannel amplification, derives the amplifier gain and bandwidth, and discusses two types of semiconductor amplifiers: Fabry-Perot (FP) and traveling wave (TW).

4.1.1 External Pumping And Rate Equation

Similar to laser diodes, a positive optical gain in semiconductor amplifiers comes from external current injection. From the rate equation of the carrier density which is given as

$$\frac{\partial N(t)}{\partial t} = R_p(t) - R_s(t) - \frac{N(t)}{\tau_r} = \frac{J(t)}{qd} - \Gamma v_g a[N(t) - N_{th}]N_{ph}(t) - \frac{N(t)}{\tau_r} \tag{4.1}$$

where

$$R_p(t) = \frac{J(t)}{qd} \tag{4.2}$$

is the external pumping rate from current injection,

$$R_s(t) = \Gamma v_g a[N(t) - N_{th}]\mathrm{N}_{ph} \tag{4.3}$$

is the net stimulated emission rate, and τ_r is the combined time constant due to spontaneous emission and various carrier recombination mechanisms. As w know that $\mathrm{R}_s = v_g g(N) N_{ph}$ and $\mathrm{g}(N)$ given by

$$\mathrm{g}(N) = \Gamma a(N - N_o) - \alpha_m = \Gamma a(N - N_0) \tag{4.4}$$

where α_m is the distributed medium loss, Γ is the confinement factor, v_g is the group velocity of the incident light, and N_{th}, is the threshold carrier density to have a positive gain

In the steady state, $\partial N/\partial t = 0$, and

Therefore,

$$\mathrm{g}(N) = \frac{J/qd - N_{th}/\tau_r}{v_g N_{ph} + 1/(\Gamma a \tau_r)} = \frac{g_0}{1 + N_{ph}/N_{ph,sat}} \tag{4.5}$$

where

$$N_{ph,sat} = \frac{1}{\Gamma a v_g \tau_r} \tag{4.6}$$

is called the **saturation photon density** and

$$\mathrm{g}_o = \Gamma a \tau_r \left\{ \frac{J}{qd} - \frac{N_{th}}{\tau_r} \right\} \tag{4.7}$$

is the medium gain at zero photon density.

From Equation (4.6), it is desirable to have a small $\Gamma a \tau_r$ product to have a large $\mathrm{N}_{ph,sat}$.

4.1.2 Amplifier gain, Pumping efficiency, And Bandwidth

When the medium gain *g(N) is* known, the light power *P(z)* as a function *of z* is determined by the following differential equation:

$$\frac{dP(z)}{dz} = g(N)P(z) \tag{4.8}$$

From Equation (4.5), N in turn is a function of N_{ph} or P(z). Therefore, *g(N)* is an implicit function of z. The one-trip amplifier gain within the amplifier is defined as

$$G_0 = \frac{P(L)}{P(0)} = e^{\int_0^L g(N)dz} \tag{4.9}$$

If gain saturation is negligible, *g(N) = go* is a constant and $Go = e^{g_0 L} = Go_{,nosat.}$

Amplifier Gain Considering Gain Saturation When gain saturation is considered, from Equations (4.5) and (4.8),

$$dP = g(z)P(z)dz = godz\frac{P(z)}{1+P(z)/P_{sat}}$$

where $P_{sat} = N_{ph,sat}(hf)(wd)v_g$ is the saturation optical power. With simple rearrangement,

$$g_0 dz = \left[\frac{1}{P(z)} + \frac{1}{P_{sat}}\right]dP.$$

integrating the above equation from $z = 0$ to $z = L$ gives

$$G_0 = 1 + \frac{P_{sat}}{P_{in}}\ln\left(\frac{G_{0,nosat}}{G_0}\right) \tag{4.10}$$

When $G_0 \geq 1$, Equation (10) can be rearranged as

$$10\log_{10}\left(\frac{G_{0,nosat}}{G_0}\right) \approx 4.34\frac{P_{in}}{P_{sat}} \tag{4.11}$$

This gain expression shows that the gain penalty $G_{0,nosat}/G_0$ in dB is linearly proportional to the actual amplifier gain G_0 and the power ratio P_{in}/P_{sat}. This gain penalty dependence is shown in Figure 4.1.

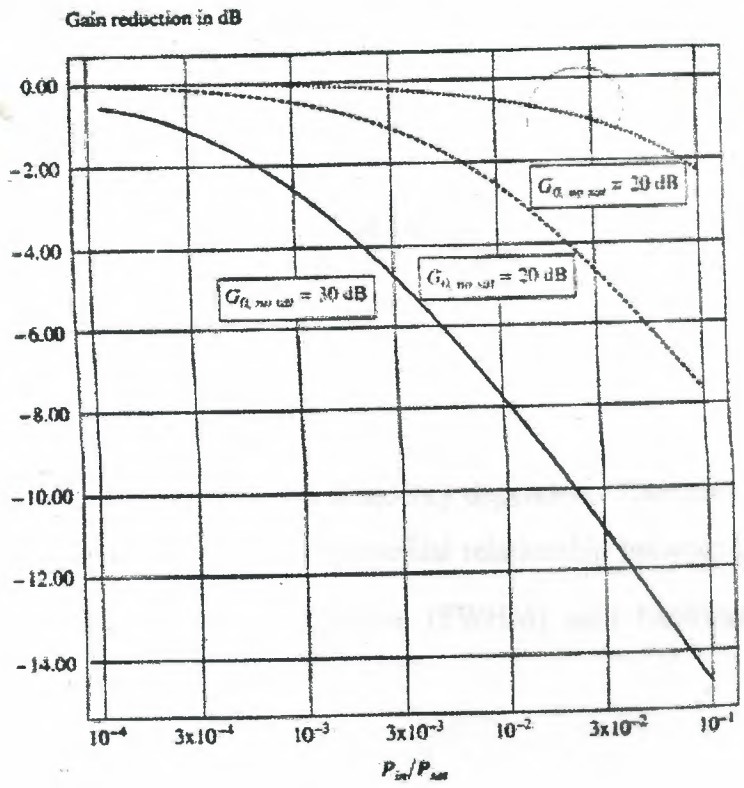

FIGURE 4.1. Gain reduction due to gain saturation of traveling wave semiconductor amplifiers.

Upper Bound of Amplifier Gain In general, to avoid laser emission, the amplifier gain G_0 cannot be arbitrarily large. Specifically, G_0 is bounded by

$$G_{rd} = G_0^2 R_L R_R \leq 1 \tag{4.12}$$

where G_{rd} is the round-trip gain and $r_L^2 = R_L$ and $r_R^2 = R_R$ are the reflectivities at the two cavity facets.

Pumping Efficiency From Equation (4.9), it appears that the amplifier gain can be increased by increasing the cavity length L. From Equation (4.7), however,

$$g_0 L = \Gamma a \tau_r \left\{ \frac{J}{qd} - \frac{N_{th}}{\tau_r} \right\} L = \Gamma a \tau_r \left\{ \frac{I}{qwd} - \frac{N_{th}}{\tau_r} L \right\}$$

Therefore, the gain in fact will decrease as L increases at a given injection current. 1 When

gain saturation is negligible, $G_0 = G_{0,nosat}$ in dB is

$$G_0 = G_{0,nosat}\,[dB] = 10\log_{10} G_0 = 10\log_{10} e^{g_0 L}$$

$$= 4.34 g_0 L \approx 4.34 \Gamma a \left[\frac{\tau_r}{qwd} 1 - N_{th} L \right] \qquad (4.13)$$

and the pumping efficiency η_i, in dB/A is

$$\eta_i = 4.34 \frac{\Gamma a \tau_r}{qwd} dB / A. \qquad (4.14)$$

Gain Bandwidth The gain constant a is frequency dependent. Therefore, the amplifier gain is also frequency dependent. From the exponential relationship between g and G_0 given by Equation (4.9), the full-width half-magnitude (FWHM) gain bandwidth of G_0 can be determined from the gain profile g(f).

4.1.3 Fabry-Perot Amplifiers

Because the two cavity facets of an amplifier can cause reflections, incident light can be bounced back and forth within the amplifier. Amplifiers that have strong internal reflections are called Fabry-Perot (FP) amplifiers. In this case, the one-trip amplifier gain G_0 is not the actual amplifier gain. Amplifiers that have negligible internal reflection or $R_L R_R \approx 0$ are called traveling-wave (TW) amplifiers. In general, FP amplifiers have poor time and frequency response. As a result, they are not attractive as compared to TW amplifiers.

To find the amplifier gain of FP amplifiers, assume a certain optical power distribution *P(z)* in the cavity. If *P(z)* is known, the medium gain *g(z)* can be obtained from Equation (4.8). If *g(z)* is known, *P(z)* can be expressed as

$$P(z) = \left(1 - r_L^2\right) \left| E^+(z) + E^-(z) \right|^2 \qquad (4.15)$$

Where

$$E^+(z) = \sqrt{P_{in}}\, e^{\int_0^z g(z')dz'/2} \sum_{m=0}^{\infty} \left(G_0 r_L r_R\right)^m e^{j(\beta z + m\theta_0)}$$

And

$$E^{-}(z) = \sqrt{P_{in}}\sqrt{G_0 r_R}\, e^{\int_z^L g(z')dz'/2}\, e^{j\theta_0/2} \sum_{m=0}^{\infty} (G_0 r_L r_R)^m e^{j[\beta(L-z)+m\theta_0]} \qquad (4.16)$$

represent the positive and negative traveling waves, respectively. Also in the above equations, θ_0 is the round-trip phase shift defined by and G_0 is the one-trip gain given by Equation (4.9). With some manipulation. Equation (4.15) can be simplified as

$$P(z)=P_{in}\left(1-r_L^2\right)\left|\frac{1}{1-G_0 r_L r_R e^{j\theta_0}}\right|^2 \left\{G(z)+\frac{G_0^2 r_R^2}{G(z)}+2r_R G_0 \cos(2\beta z-\theta_0)\right\} \qquad (4.17)$$

With

$$G(z) = e^{\int_0^z g(z')dz'} \qquad (4.18)$$

From Equation (4.6), note that *G(z)* is a function of *P(z)*. Therefore, Equations (4.17)

and (4.18) form a pair of integral equations for *P(z)* and can be solved numerically.

Once *P(z)* and *G(z)* are solved, from Equation (4.18), the net amplification gain of the FP amplifier is

$$G_{FP} = \left(1-r_L^2\right)\left(1-r_R^2\right)\frac{\left|E^{+}(L)\right|^2}{P_{in}} = \left(1-r_L^2\right)\left(1-r_R^2\right)\frac{G_0}{\left|1-G_0 r_R r_L e^{j2\beta L}\right|^2}. \qquad (4.19)$$

This equation shows strong frequency dependence when $G_0 r_R r_L$ is large. In Figure 4.2, the gain drops of G_{FP} and G_0 in dB with respect to $G_{0,nosat}$ at zero reflection and saturation are shown. The larger the gain $G_{0,nosat}$ and input power P_{in}, the larger the gain drop. In Figure 4.3, the gain is also strongly dependent on the round-trip phase when $G_{0,nosat}$ is large.

Because the round-trip phase $2\beta L$ is frequency dependent, this means the gain is strongly carrier frequency dependent.

To characterize the gain variation due to the round-trip phase dependence, the following maximum to minimum gain ratio is defined:

$$C=\left(\frac{1+G_0 r_L r_R}{1-G_0 r_L r_R}\right)^2 \qquad (4.20)$$

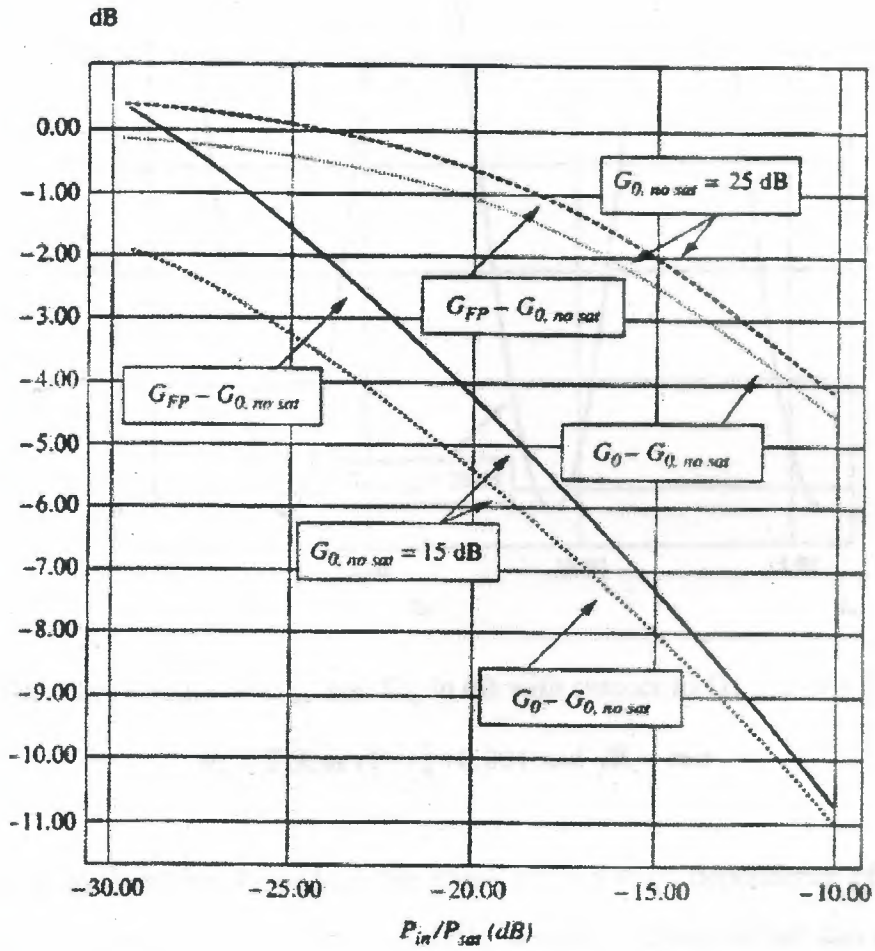

FIGURE 4.2. Changes of G_{FP} and G_0 in dB with respect to $G_{0,nosat}$ as a function of $G_{0,nosat}$ at $r_L^2 = r_R^2 = 0.001$ and $\beta L = m\pi$, where $G_{0,nosat}$ is the gain when $r_R = r_L = 0$ and other is no saturation effect or $P_{sat \to \infty}$

where the maximum gain is reached when $2\beta L = 2m\pi$ and the minimum gain is reached when $2\beta L = (2m+1)\pi$. Equation (19) shows that when $1 - G_0 r_L r_R$ is small, there is a large gain deviation.

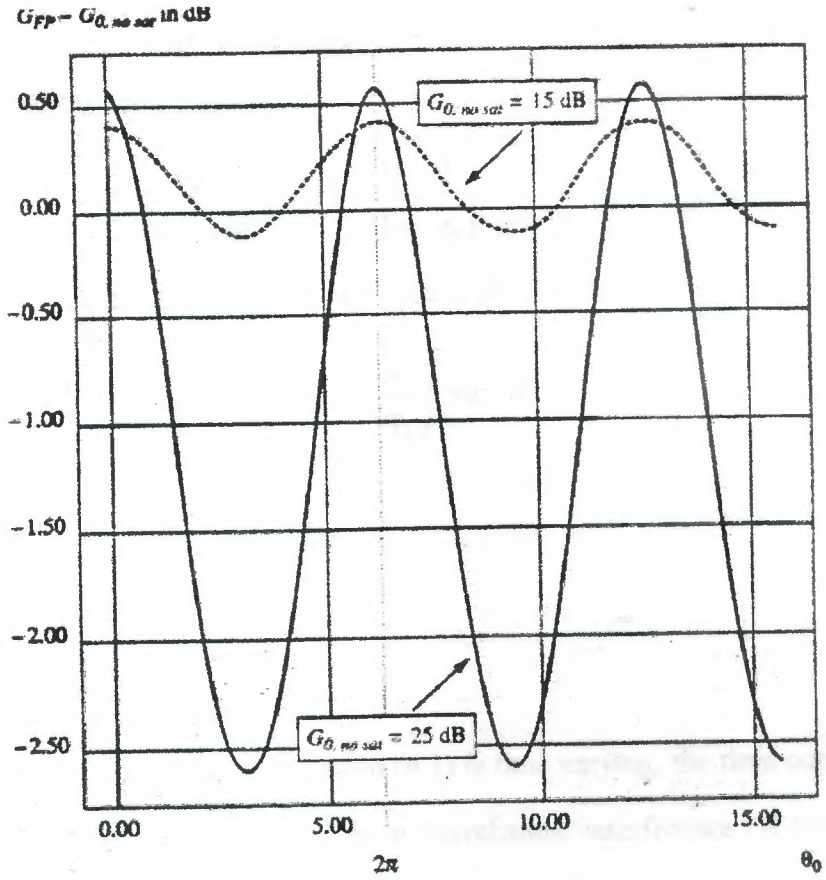

FIGURE 4.3. Changes of G_{FP} and G_0 in dB with respect to $G_{0,nosat}$ as a function of $\theta_0 = 2\beta L$ at $r_L^2 = r_R^2 = 0.001$ and $\beta L = m\pi$.

Reduction of Reflectivity To reduce the phase or frequency dependence of the amplifier gain requires a very small $r_R r_L$ value (≤ 0.001). Several techniques that can achieve a low reflectivity are depicted in Figure 4. As shown, the first technique uses antireflection coating to reduce reflection, and the second technique introduces a tilt of the cavity with respect to the cavity facets.

To see how zero reflection is achieved in the second technique, consider a transverse magnetic (TM) wave at an incident angle θ_1 and refracted angle θ_2. In this case, the reflection coefficient is

$$r = \frac{n_2 \cos(\theta_1) - n_1 \cos(\theta_2)}{n_2 \cos(\theta_1) + n_1 \cos(\theta_2)}.$$

From Snell's law,

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2).$$

Combining the above two equations,

$$r = \frac{n_2 \cos\theta_1 - n_1\sqrt{1-(n_1/n_2)^2 \sin^2\theta_1}}{n_2 \cos\theta_1 + n_1\sqrt{1-(n_1/n_2)^2 \sin^2\theta_1}}. \tag{4.21}$$

A zero-reflection condition can thus be achieved at $\theta_1 = \theta_p$ if

$$1 = \left(\frac{n_2}{n_1}\right)^2 \cos^2\theta_p + \left(\frac{n_1}{n_2}\right)^2 \sin^2\theta_p$$

or

$$\tan\theta_p = \frac{n_2}{n_1} \tag{4.22}$$

4.1.4 Interchannel Interference

When either R_S or R_p in Equation (4.1) is time varying, the time constant τ_r in the rate equation also plays an important role in interchannel interference (ICI) in multichannel amplification. Intuitively, the time constant determines how fast the carrier density *N(t)* follows the change of Rs or Rp. It is desirable to have a small τ_r to reduce the rise time and fall time in direct modulation. In optical amplification, however, a large is needed to reduce ICI.

To see the effect of τ_r on ICI, consider an amplitude change of one wavelength channel and see how it affects the change of the medium gain. Let ΔN_{ph} be the step change of one channel and let $N_{ph,0}$ be the average photon density of all channels. If the total number of wavelength channels is large, ΔN_{ph} is small compared to $N_{ph,0}$. As a result, the corresponding change of the carrier density $\Delta N(t)$ is also small compared to its steady state value *N*. In this case, Equation (4.1) can be made linear, producing the following small signal equation:

$$\frac{\partial}{\partial t}\Delta N = -\Gamma v_g a(N - N_{th})\Delta N_{ph} - \left[\Gamma v_g a N_{ph,0} + \frac{1}{\tau_r}\right]\Delta N.$$

With the initial condition $\Delta N(t) = 0$, it is easy to solve for $\Delta N(t)$:

$$\Delta N(t) = -\Gamma a v_g \tau' (N - N_{th}) \Delta N_{ph} (1 - e^{-t/\tau'}) \tag{4.23}$$

where

$$\frac{1}{\tau'} = \frac{1}{\tau_r}\left(1 + \frac{N_{ph,o}}{N_{pn,sat}}\right). \tag{4.24}$$

From Equation (4.24), the medium gain $g(N)$ has a change given by

$$\Delta g = \Gamma a \Delta N = g(N)\left[\Gamma a v_g \tau' \Delta N_{ph}\right]\left(1 - e^{-t/\tau}\right). \tag{4.25}$$

Using $N_{ph,sat}$ defined by Equation (6), one obtains

$$\Delta g = \Delta g_0 \left(1 - e^{-t/\tau'}\right) \tag{4.26}$$

where

$$\Delta g_0 = g(N)\frac{\tau'}{\tau_r}\frac{\Delta N_{ph}}{N_{ph,sat}} \tag{4.27}$$

Because the change ΔN_{ph} is random, Δg_0 is random. When the wavelength channels are asynchronous, the time t between the change and observation is also random. If the number of channels is large, a uniform distribution of t between 0 and T (the bit interval) can be assumed. As a result, the variance of Δg *is* given by

$$\mathrm{E}\left[\Delta g^2\right] = E\left[\Delta g_0^2\right]\frac{1}{T}\int_0^T \left(1 - e^{-t/\tau'}\right)^2 dt \tag{4.28}$$

Thus it is desirable to have a large τ' compared to T reduce the gain fluctuation.

4.2 ERBIUM-DOPED FIBER AMPLIFIERS

Although the use of rare-earth ions as a gain medium for optical fiber amplification was noted as early as 1964, erbium-doped fiber amplifiers (EDFAs) were not practical until low-loss doped fibers were made possible. Use of EDFAs in optical fiber communications is illustrated in Figure 4.4, where one small fiber section is doped with erbium ions, Er^{+3}, as the agency for stimulated emission. To excite the electrons of Er^{+3} to higher energy states, external optical pumping through a directional coupler is used.

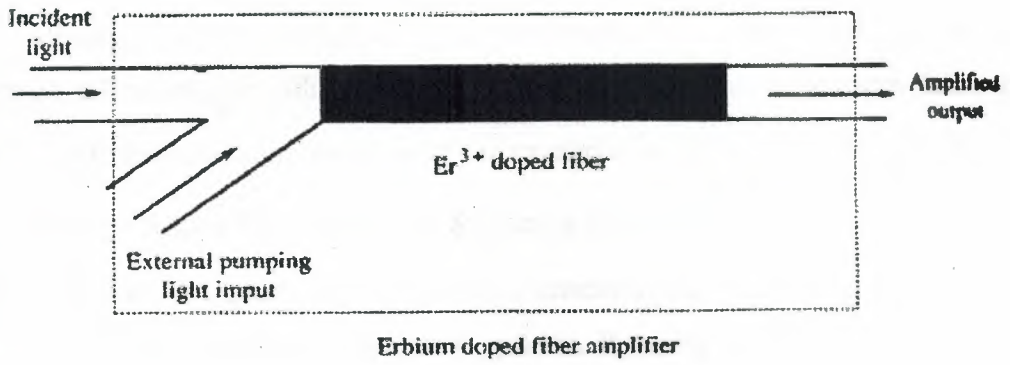

FIGURE 4. Erbium-doped fiber amplifier (EDFA).

4.2.1 OPTICAL PUMPING

To excite carriers to a higher energy level for stimulated emission, external pumping needs to be operated at a higher frequency than that of the amplified signal. The energy diagram of Er^{+3} is shown in Figure 4.5, where the ground level is labeled as ${}^4I_{15/2}$ and the metastable level (stimulated emission level) is ${}^4I_{13/2}$. The energy difference between these two levels gives an emission wavelength of 1530 nm.

In order to pump carriers from the ground level to the metastable level, a pumping source at wavelength 1450 nm, 980 nm, or 800 nm can be used. These will excite the carriers to ${}^4I_{13/2}$, ${}^4I_{11/2}$ **or** ${}^4I_{9/2}$, respectively. Excited carriers at ${}^4I_{11/2}$ **or** ${}^4I_{9/2}$ move down to the metastable level ${}^4I_{13/2}$ because of their short lifetime. At 1450 nm pumping, because of the Stark splitting effect that causes both the metastable level ${}^4I_{13/2}$ and the ground level ${}^4I_{15/2}$ to consist of several finer separated levels as shown in Figure 4.6, carriers are excited from the lower band of the ground level to the higher band of the metastable level. From thermal equilibrium or Boltzman distribution, **excited** carriers will quickly move down to the lower band of ${}^4I_{13/2}$.

The efficiency of external pumping is determined by the absorption **spectrum of** Er^{+3} ions. The absorption spectrum of a silicate glass is illustrated in Figure 4.7, where the absorption wavelengths correspond exactly to the energy level differences **from** level

shown in Figure 4.5.

Although external pumping at a wavelength lower than 700 nm has a higher absorption efficiency, the difficulty of finding good semiconductor sources limits pumping only 800, 980, and 1470 nm. Because of the excited-state absorption (ESA) from $^4I_{13/2}$ to $^2H_{11/2}$ shown in Figure 4.5, pumping at 800 nm is not good. Therefore, only 980 and 1470 nm pumping are used practically. In general, semiconductor sources at 1470 nm relatively more available and is used in early EDFA systems. Pumping at 980 nm, on the other hand, has a higher pumping efficiency (around 10 dB/mW) compared to 1470 nm pumping (around 6 dB/mW). In addition, 980 nm pumping has a lower pumping noise. As a result, as advanced lasers at 980 nm become more available, more systems are using 980 nm pumping.

Longitudinal Optical Pumping Different from current injection in semiconductor amplifiers, optical pumping in EDFAs is in the same direction as the incident light. As it lustrated in Figure 4.8, when the pumping direction is perpendicular to the propagation direction, as in semiconductor amplifiers, it is called **transverse** pumping. When the pumping direction is parallel to the incident light, it is called **longitudinal** pumping. In the latter case, the pumping rate is stronger at the input side of the amplifier. As power is transferred from the pumping light to the signal light, the pumping rate decreases along the light propagation direction.

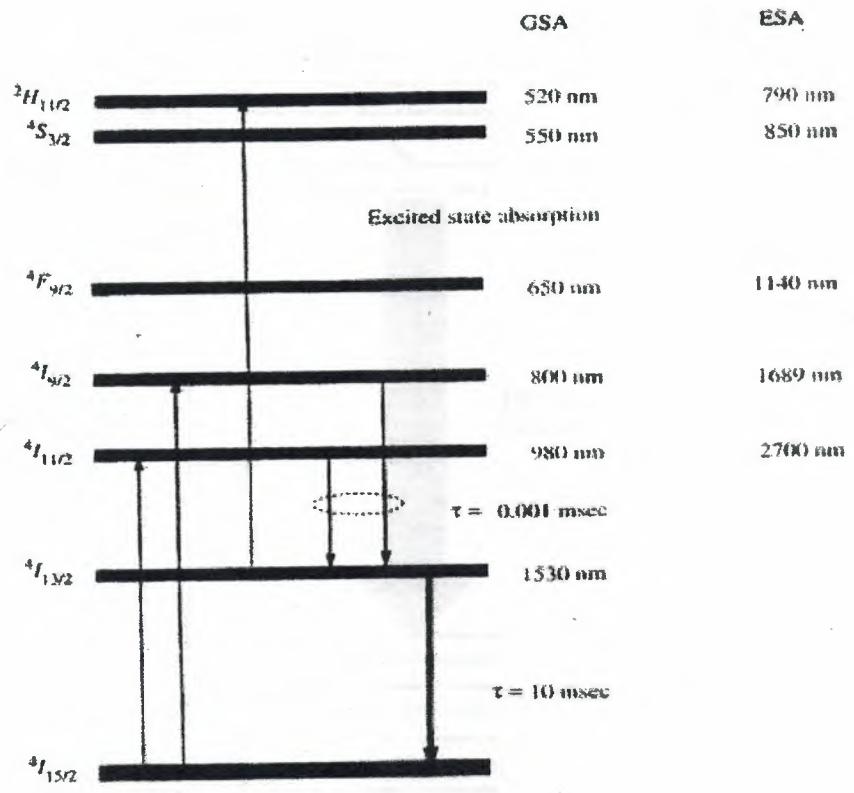

FIGURE 4.5. Energy level of Er^{3+}. For each level, the ground state absorption (GSA) wavelength is the light wavelength needed to excite carriers from the ground state to the given level, and the excited state absorption (ESA) wavelength is the light wavelength needed to excite carriers from the given level to the metastable state $^4 I_{13/2}$

Absorption and Emission Cross Sections To quantify the absorption efficiency in external pumping, a parameter called the absorption cross section σ_a is used. By definition, if the pumping power is P_p and the ground state population is N_1, the pumping rate is W_p / N_1, where

$$W_p = \frac{\sigma_a P_p}{h f_a A} \text{sec}^{-1} \tag{4.29}$$

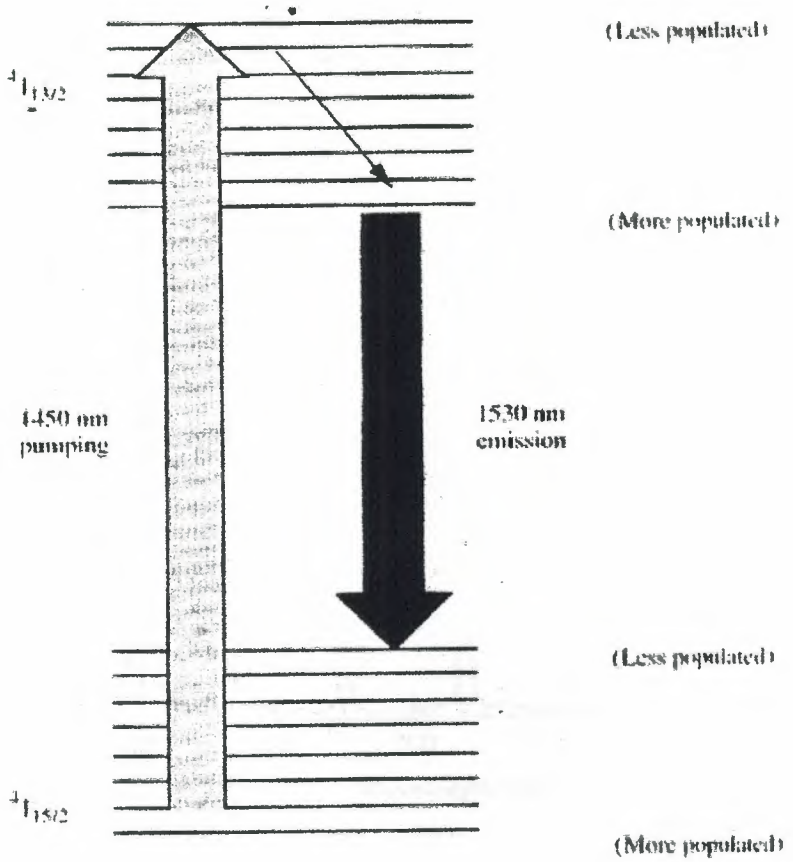

FIGURE 4.6. 1450 nm pumping. Because of the stark effect, there are splittings at the ground state and the metastable state.

hf_p is the photon energy of external pumping at Frequency f_p, and A is the core area of the EDFA. From Equation (28), a large absorption cross section produces a high pumping efficiency. Absorption cross sections at 800 nm, 980 nm, and 1450 nm are shown in Figures 4.9-4.11, respectively.

Because of longitudinal pumping, P_p, given in Equation (4.29) is also spatially dependent. At a given absorption cross section, the amount of power decrease over a short distance dz is

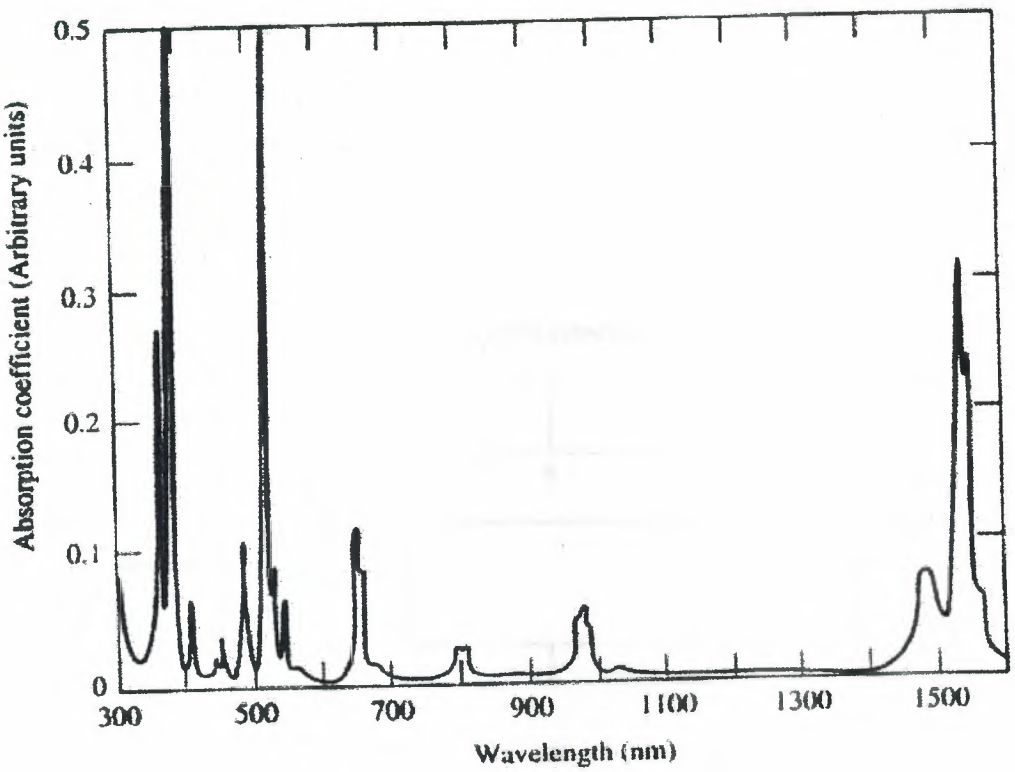

FIGURE 4.7. Absorption spectrum of Er^{+3}

$$dP_p(z) = -\sigma_a P_p(z) N_1 dz .$$

When $\sigma_a N_1$ is constant along the fiber, $P_P(z)$ has an exponential decay. In practice, N_1 increases with z because of a lower pumping power. As a result, $P_P(z)$ drops even faster.

In addition to the absorption cross section that determines the pumping rate, there is an **emission cross section** that determines the medium gain. Specifically, if σ_e is the emission cross section, the medium gain is given by

$$g = \sigma_e (N_2 - N_1) \tag{4.30}$$

where N_2 and N_1 are the carrier densities at the metastable and ground states, respectively. From Equation (4.29), the stimulated emission rate is

$$R_s = v_g g N_{ph} = W_s (N_2 - N_1) \tag{4.31}$$

where $P_{in} = v_g N_{ph} A$ is the incident light power, hf_s is the photon energy of the input signal, and

$$W_s = \frac{\sigma_e P_{in}}{hf_s A} \sec^{-1}. \tag{4.32}$$

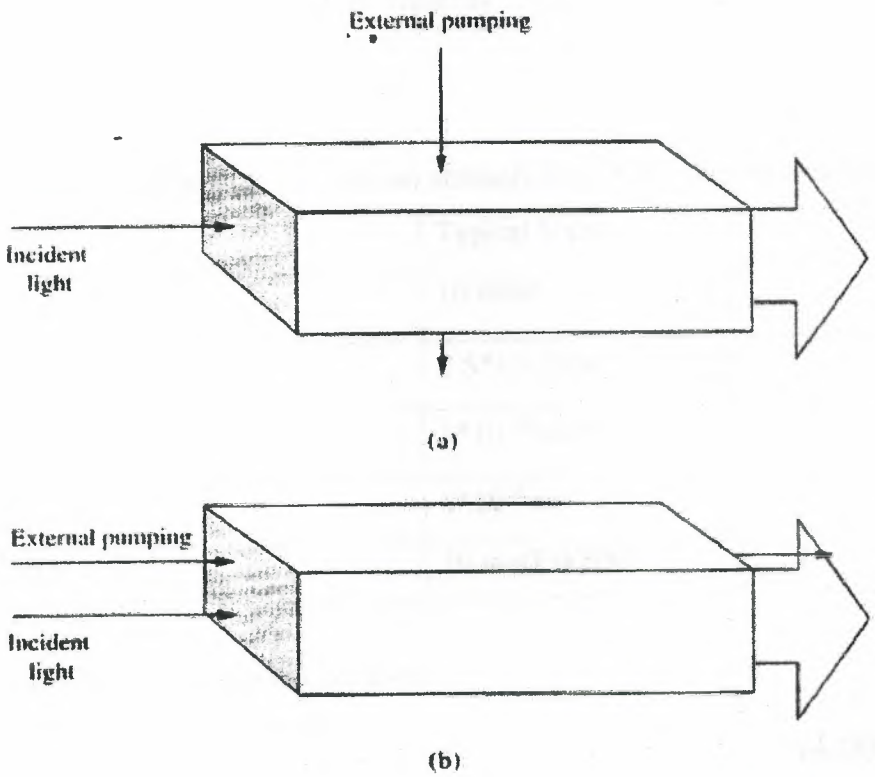

FIGURE 8. Two types of pumping: (a) transverse and (b) longitudinal.

4.2.2 Rate Equations And Amplifier Gain

Because the metastable energy level ${}^4I_{13/2}$ has a much longer lifetime than its upper levels, the energy diagram of Er^{+3} can be approximated as a two-level system, where ${}^4I_{15/2}$ is the ground level and ${}^4I_{13/2}$ is the upper level. The carrier rate equation of an EDFA thus be written as:

$$\frac{\partial N_2}{\partial t} = W_p N_1 - W_s (N_2 - N_1) - \frac{N_2}{\tau_{sp}} = -\frac{\partial N_1}{\partial t} \qquad (4.33)$$

The first term, $W_P N_1$ on the right-hand side is the pumping rate from the lower state to the upper state; the second term, $W_s(N_2 - N_1)$, is the net stimulated emission rate: and the third term, N_2/τ_{sp}, is the spontaneous recombination rate from the upper state to the lower state. The time constant τ_{sp} of EDFA is typically 10 msec Typical values of the above parameters are given in Table 1.

Table 1. Typical EDFA parameters, which can strongly depend on the materials doped.

Parameter	Typical Value
τ_{sp}	10 msec
σ_a	$2.5*10^{-21} cm^2$ @980nm
σ_e	$5*10^{-21} cm^2$ @1540nm
$N = N_1 + N_2$	$8*10^{18} cm^{-3}$
Emission bandwidth	30 nm(FWHM)

In the steady state, the rate equation (32) gives

$$N_2 - N_1 = \frac{W_p - 1/\tau_{sp}}{W_p + 2W_s + 1/\tau_{sp}} N_1 \qquad (4.34)$$

Where $N_t = N_1 + N_2$ is the total carrier density. When the pumping rate is high or $W_P \geq W_s$ and $W_p \geq 1/\tau_{sp}$, $N_2 - N_1 \approx N_1$. In this case, the medium gain is approximately

$$g = \sigma_e (N_2 - N_1) \approx \sigma_e N_1 = g^* \qquad (4.35)$$

where g^* is the upper limit of the medium gain constant.

As mentioned earlier, because of the longitudinal pumping, the pumping rate W_p is spatially dependent. Because

$$\frac{dP_P}{dz} = -\sigma_a N_1 P_P \qquad (4.36)$$

and

$$\frac{dP_{in}}{dz} = gP_{in} \approx \sigma_e (N_2 - N_1) P_{in} \tag{4.37}$$

from equations (4.34) and (4.35)

$$\frac{dP_P}{dz} = -(\sigma_a N_1) P_P \frac{W_S + 1/\tau_{sp}}{W_P + 2W + 1/\tau_{sp}} \tag{4.38}$$

and

$$\frac{dP_{in}}{dz} = \frac{g_0}{1 + W_S / W_{sat}} P_{in} \tag{4.39}$$

where

$$g_0 = g^* \frac{W_P - 1/\tau_{sp}}{W_P + 1/\tau_{sp}}$$

is the medium gain at zero incident signal and

$$W_{sat} = \frac{1}{2}\left(W_P + 1/\tau_{sp}\right) \tag{4.40}$$

is the saturation rate. Because W_{sat} becomes smaller as W_p gets smaller along the light propagation direction, gain saturation effect is stronger at the output end of the amplifier. When $W_p \leq 1/\tau_{sp}$ or $N_2 \leq N_1$, the gain can even be negative.

CHAPTER 5

RECEIVING DEVICES

5.1 Photodiodes

There are two main types of photodiodes: PINs and APDs. The structure of a typical PIN diode is shown in Figure 5.1, where photons are coupled to the left-hand side of the diode and pass through an intrinsic region. A photon with sufficient energy (hf) can excite an electron-hole pair. If the pair is in the presence of a large electric field, the electron and hole will be separated and move quickly in opposite directions, resulting in a photocurrent. If the pair is in the presence of a small or zero electric field, they move slowly and may even recombine and generate heat. Therefore, a strong electric field in the depletion region is essential.

Because one absorbed photon generates one EHP in PINs, the photocurrent is a linear function of the input optical power P_{in}:

$$I_{ph} = \eta \frac{q}{hf} P_{in} = \left(\eta \frac{\lambda}{1.24} \right) P_{in} = \Re P_{in} \qquad (5.1)$$

where η is the quantum efficiency discussed earlier and A is the wavelength in μm.

Figure 5.2. Shows the I-V characteristics at different input power levels. At zero input power, the reverse bias current is called the dark current. The total current is thus

$$\mathbf{I}_{tot} = I_d + I_{ph} = I_d + \Re P_{in} \qquad (5.2)$$

For APDs, because of the current gain from EHP multiplications, the generated photocurrent is

$$\mathrm{I}_{ph} = \mathrm{M}_{apd}\, \Re P_{in} \qquad (5.3)$$

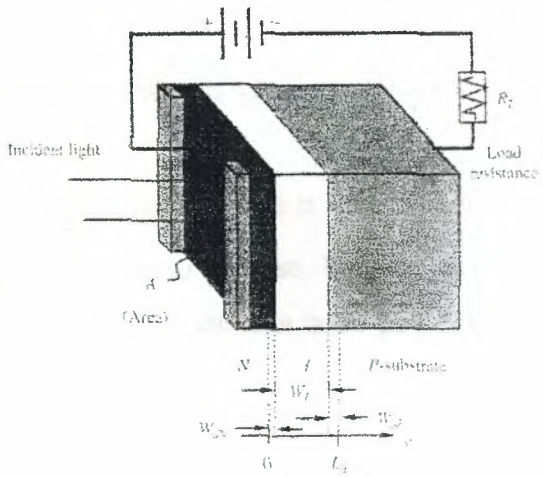

FIGURE 5.1. A PIN diode.

Where M_{apd} is the multiplication gain of the APD. For PINs, the same equation holds, with $M_{apd} = 1$.

Unlike LEDs and LDs, photodiodes are generally operated at reverse bias for detection in optical communications. There are several reasons for this reverse bias operation:

1. Photodiodes have a large resistance at reverse bias. This allows a large bias or load resistance for high impedance detection. A large input resistance can minimize the input current noise.
2. The electric field in the absorption region is large with reverse bias. As a result, carriers generated from photon absorption move quickly to the external circuit. This implies fast response.
3. The width of the depletion region is large at reverse bias. This results in a small junction capacitance and, consequently, a small RC time constant. This also means a fast response.

The reverse-bias detection is also called **photoconductive** (PC) **detection.** This is in contrast to another operation mode called **photovoltaic** (PV) **detection,** where the bias voltage is zero. The advantage of PV detection is its zero dark current at zero bias voltage. However, because it has a small depletion width and electric field, the response speed is low. Therefore. PV detection is mainly used in instrumentation and not appropriate for

high-speed detection. In high-speed optical communications, PIN and APD diodes are all operated in the PC mode.

The dark current I_d from the reverse bias is undesirable because it adds not only to the total current output but also to the total noise. It contributes to the so-called shot noise at the photodetection output. In general, the shot noise power is proportional

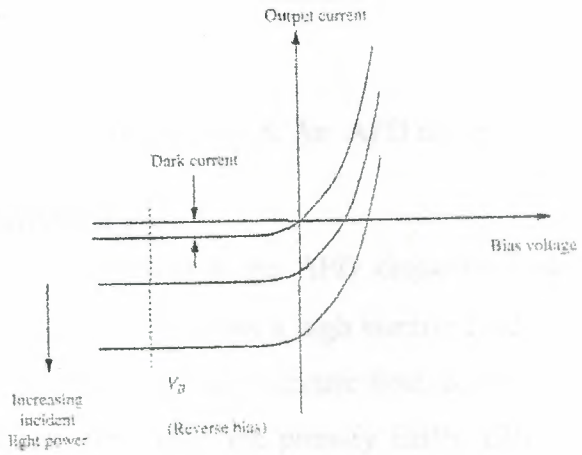

FIGURE 5.2. I-V characteristics of a reverse-bias PIN.

to the total current output. Therefore, it is important for a photodiode to have as small a dark current as possible.

5.2 Avalanche Photodiodes

A typical APD diode is illustrated in figure 5.3. In addition to N, I, and Players in PIN diodes, it has a high doping P^+ layer between the N and I layers. As a result, it has a high electric field that accelerates electrons and holes with high momenta, which in turn excites more electron-hole pairs.

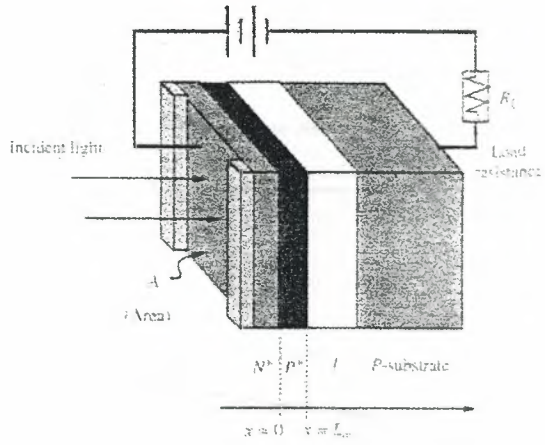

FIGURE 5.3. An APD diode.

5.2.1 Electric Field Distribution

The electric field distribution in an APD diode is illustrated in Figure 5.4. The region between the N^+ and P^+ layers has a high electric field where carrier multiplication takes place. The I-type layer has a smaller electric field, so there is no multiplication. EHPs generated in the I-type layer are called the primary EHPs. EHPs generated by the primary EHPs in the multiplication region are called the secondary EHPs.

The geometry and doping levels of an APD must be chosen carefully **to produce a** fast device that operates at a reasonable reverse-bias voltage. As illustrated in Figure 5.4 a, when the doping level of the multiplication region P^+ is too heavy, a high E_{max} is required to deplete the intrinsic region. Because a high E_{max} can cause device breakdown, this should be avoided. However, when E_{max} is not large enough to deplete the intrinsic region, the drift velocity is small. This implies a slow device.

On the other hand, as illustrated in Figure 5.4b, if the doping is too light in the multiplication region P^+, the intrinsic region has a high electric field. This implies a large reverse bias, which is undesirable from practical power supply considerations.

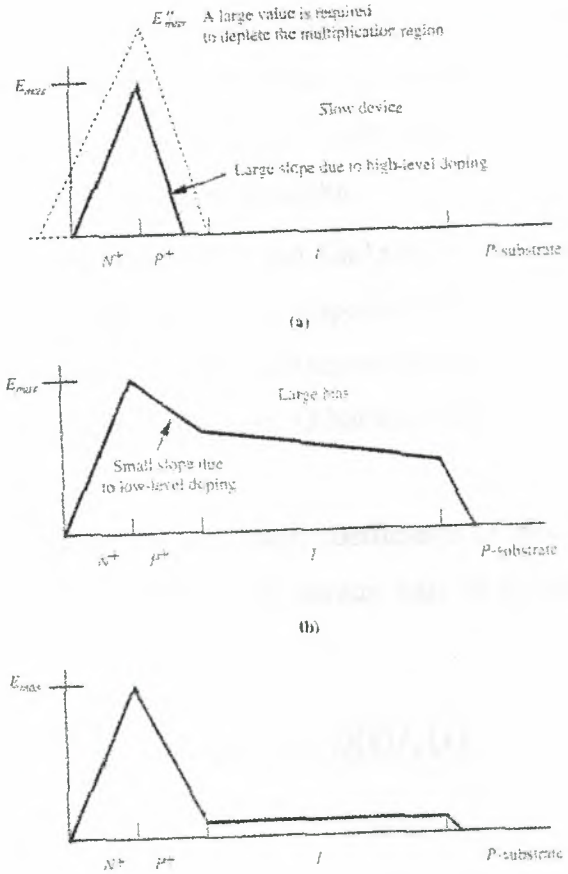

FIGURE 5.4. The electric field distribution at different doping levels of the multiplication region P^+: (a) doping level is too high and requires a large E_{max}, (b) doping level is too low and requires a large reverse bias, and (c) doping level is proper.

5.2.2 Current Multiplication

The avalanche multiplication process is illustrated in Figure 5.5. As shown, primary electrons come to the multiplication region and initiate the multiplication process. As illustrated, a primary electron can excite several secondary EHPs on its way to the anode. At a high electric field or bias, secondary EHPs generated can pick up large momenta and generate more secondary EHPs. If the reverse bias is larger than a certain threshold, this multiplication process can last forever. This is called the **avalanche breakdown** and the corresponding threshold voltage is called the **breakdown voltage.**

The capability for electrons and holes to excite EHPs is characterized by the ionization coefficients α and β, respectively, which represent the multiplication ratio per unit length (in the unit of 1/m). Typical ionization coefficients as a function of the electric field are shown in Figure 5.6. From the figure, note that

1. Electron ionization coefficients of Si and GaAsSb are higher than hole ionization coefficients. For Ge, GaAs, and InGaAs, the opposite is true.
2. Ge has the largest electron and hole ionization coefficients. On the other hand, Si has the smallest hole ionization coefficient. Otherwise, the ionization coefficients are approximately the same.

From the definitions of the ionization coefficients of electrons and holes, the transport equations of electrons and holes in the steady state in the multiplication region can be written as

$$-\frac{dJ_n(x)}{dx} = \alpha(x)J_N(x) + \beta(x)J_P(x) \qquad \textbf{(5.4)}$$

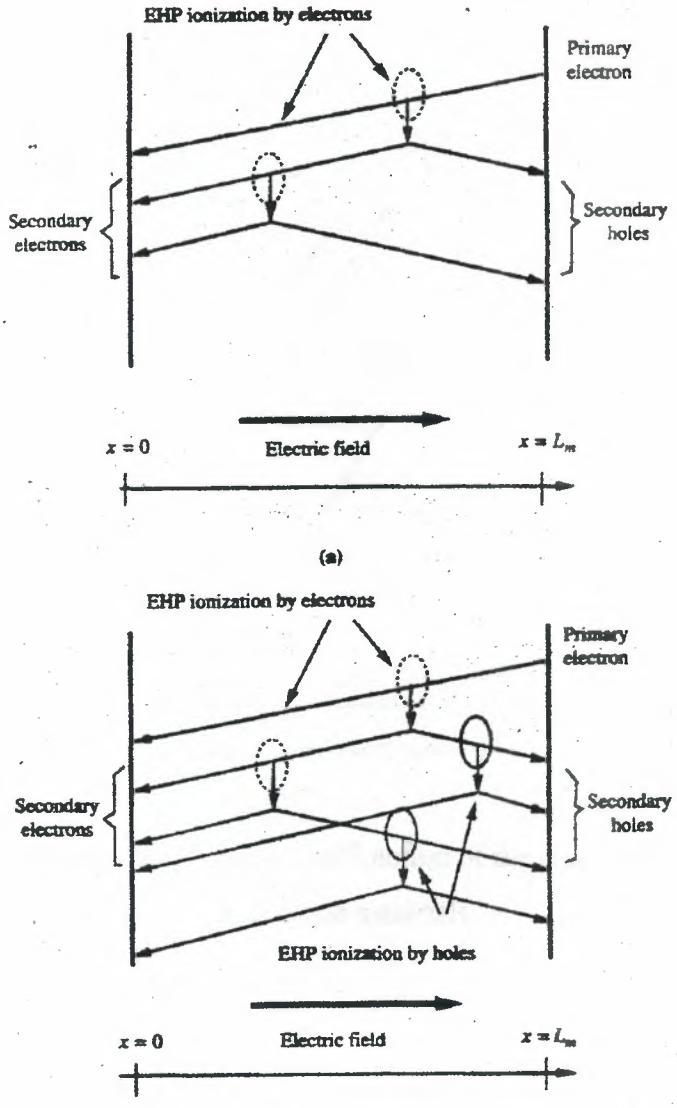

FIGURE 5.5 The multiplication process: (a) k=0; (b) k=1.

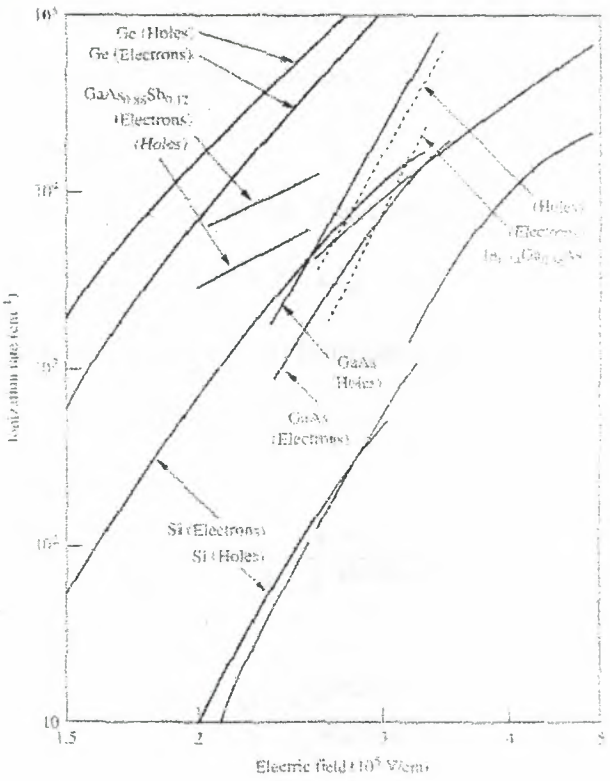

FIGURE 5.6 Ionization coefficients as a function of the electric field of important photodiode materials.

And

$$\frac{dJ_P(x)}{dx} = \alpha(x)J_n(x) + \beta(x)J_P(x) \tag{5.5}$$

where J_n and J_P are the current densities of electrons and holes in the multiplication region.

From Equations (5.4) and (5), note that the total current density ($J = J_n + J_P$) *is a* constant. Therefore, 5.5

$$-\frac{dJ_n}{dx} = (\alpha - \beta)J_n + \beta J \tag{5.6}$$

and

$$-\frac{dJ_P}{dx} = (\alpha - \beta)J_P - \alpha J \tag{5.7}$$

Because J is a constant, the above equations are first-order linear differential equations, and the solution of $J_n(x)$ to Equation (5.6) can be given by:

$$J_n(x) = J_1 e^{-(\alpha-\beta)x} + J_2 \tag{5.8}$$

Where J_1 and J_2 are constants to be solved. The corresponding J_p from Equation (5.6) is

$$J_p(x) = -J_1 e^{-(\alpha-\beta)x} + (J - J_2).$$

Substituting $J_n(x)$ given by Equation (5.8) in Equation (5.6) gives

$$(\alpha - \beta)J_2 + \beta J = 0$$

The total current density J is thus

$$J = J_n(x) + J_p(x) = \frac{k-1}{k - e^{-\alpha(1-k)L_m}} J_{PIN} \tag{5.9}$$

and the multiplication gain is

$$M_{apd,0} = \frac{J}{J_{PIN}} = \frac{(1-k)e^{\alpha L_m(1-k)}}{1 - ke^{\alpha L_m(1-k)}} \tag{5.10}$$

In the above derivation, electrons are assumed to be the primary carriers that initiate the secondary EHP generations. For a different type of APD for which holes are the carriers that initiate the multiplication process, symmetry gives the following gain expression:

$$M_{apd,h} \equiv \frac{J_p(0)}{J_p(L_m)} = \frac{(1-1/k)e^{\beta L_M(1-1/k)}}{1-(1/k)e^{\beta L_m(1-1/k)}}. \tag{5.11}$$

The multiplication gain as a function of α at different values of k is shown in Figure 5.7. When α becomes large enough, the gain becomes infinite. The voltage **to** achieve this infinite gain is called the **breakdown voltage.** Some typical values of the breakdown voltage are shown in Table 1. In general, Si has the largest breakdown voltage, and Ge the smallest. As a result, Si can sustain a high electric field, which allows it to have a high current gain. Although Ge has the lowest breakdown voltage, as Figure 6 shows, it has high ionization coefficients. Therefore, a good current gain still can be achieved.

An empirical equation to express the multiplication gain as a function of reverse bias voltage is

$$M_{apd,0} = \frac{1}{1 - \left(\frac{V}{V_B}\right)^{n_B}} \tag{5.12}$$

Where n_B depends on the material and doping profile. Typically, it can vary from 2 to 10.

TABLE 1. Typical APD characteristics

Devices	Ionization Ratio	Gain	Breakdown Voltage
Si	0.1	100-1000	200 V
Ge	2.0	50-500	5-50 V
InGaAs	1.5	20-50	50-100 V

5.2.3 Frequency Response

Because of the multiplication process, APDs have a slower frequency response compared to PIN diodes. As illustrated in Figure 5.5, carriers in the multiplication region can continue to generate secondary carriers at a large multiplication gain. As a result, the photocurrent impulse response has a long duration, and the frequency response is slow.

Similar to the simplified model used in the PIN diode frequency analysis, the frequency response of an APD can be analyzed by adding time dependent terms to Equations (5.5) and (5.6). As a result, the transport equations can be written as:

$$\frac{1}{v_n}\frac{\partial J_n(x,t)}{\partial t} - \frac{\partial J_n(x,t)}{\partial x} = \alpha(x)J_n(x,t) + \beta(x)J_P(x,t) \qquad (5.13)$$

and

$$\frac{1}{v_p}\frac{\partial J_p(x,t)}{\partial t} - \frac{\partial J_p(x,t)}{\partial x} = \alpha(x)J_n(x,t) + \beta(x)J_P(x,t) \qquad (5.14)$$

If the incident light is assumed to have the same form given by Equation (5.3), $J_n(x,t)$ and $J_p(x,t)$ can be expressed as

$$J_n(x,t) = J_{n0}(x) + J_{n1}(x)e^{j\omega_m t} \qquad (5.15)$$

$$J_p(x,t) = J_{p0}(x) + J_{p1}(x)e^{j\omega_m t} \qquad (5.16)$$

Where $J_{n0}(x)$ and $J_{p0}(x)$ are the dc components corresponding to P_0, and $J_{n1}(x)$ and $J_{p1}(x)$ are the ac components corresponding to $P_0 k_m e^{j\omega_m t}$. To derive the frequency response, only the ac components of the above transport equations will be used.

With these assumptions, in the multiplication region where $0 < x < L_m$, Equation (5.13) **and** (5.14) reduce to

$$\frac{\partial J_{n1}(x)}{\partial x} + aJ_{n1}(x) + \beta J_{p1}(x) = 0 \tag{5.17}$$

and

$$\frac{\partial J_{p1}(x)}{\partial x} - bJ_{p1}(x) - \alpha J_{n1}(x) = 0 \tag{5.18}$$

where

$$a = \alpha - \frac{j\omega_m}{v_n} \tag{5.19}$$

and

$$b = \beta - \frac{j\omega_m}{v_p} \tag{5.20}$$

In intrinsic region, the same equations for PIN diodes can be used. From the time dependence factor $e^{j\omega_m t}$,

$$\frac{\partial J_{n1}(x)}{\partial x} - \frac{j\omega_m}{v_n} J_{n1}(x) = 0 \tag{5.21}$$

and

$$\frac{\partial J_{p1}(x)}{\partial x} + \frac{j\omega_m}{v_n} J_{p1}(x) = 0 \tag{5.22}$$

for $L_m < x < L_m + L_d$, where L_d is the depletion width of the intrinsic region. The boundary conditions are

$$J_{n1}(x = L_m) = k_m J_0 \tag{5.23}$$

And

$$J_{p1}(x = 0) = 0 \tag{5.24}$$

The average current output is denned as

$$J_{ac} = \frac{1}{L_m + L_d} \int_0^{L_m + L_d} [J_{nl}(x) + J_{pl}(x)] dx. \qquad (5.25)$$

The multiplication gain M_{apd} is defined to be the ratio of the average current density over $0<x<L_m$. That is,

$$M_{apd} = \frac{(r_2 + a - \alpha) r_2 (e^{r_1 L_m} - 1) - (r_1 + a - \alpha) r_1 (e^{r_2 L_m} - 1)}{L_m r_1 r_2 [(r_2 + a) e^{r_1 L_M} - (r_1 + a) e^{r_2 L_m}]} \qquad (5.26)$$

At dc or $\omega_m = 0$, this gives the same result given by Equation (5.11). Figures 5.7 and 5.8 show the ac gain M_{apd} as a function of the normalized frequency $\omega_m \tau_t$ at different dc gains. In these results, it is assumed that $v_n = v_p$. Using Equation (5.25), one can find the 3 dB frequency at which

$$\left| \frac{J_{ac}}{k_m J_0} \right|^2 = 0.5.$$

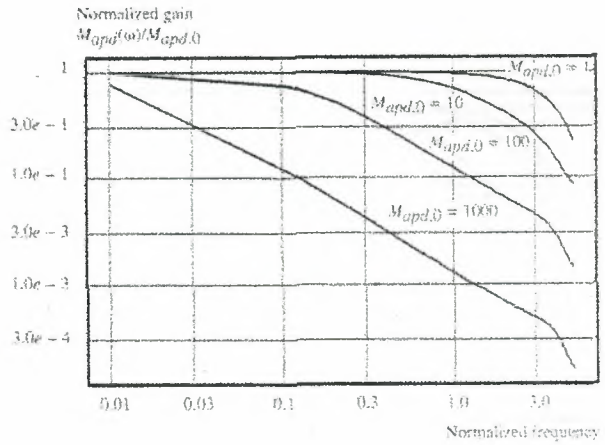

FIGURE 5.7.Frequency response at different values of $M_{apd,0}$; $k = 0.1$

Figure 5.9 shows the normalized 3 dB bandwidth $(\omega_B \tau_1)$ as a function of the dc multiplication gain M_{apd} at different k's and at $L_d = 0$, where $\tau_t = L_m / v_n$. In other words, there is no intrinsic region. In this case, the response bandwidth is entirely determined by the multiplication region. A line $k M_{apd} = 1$ is superimposed in the figure. When $k M_{apd,0}$=1 (above the line), the normalized 3 dB bandwidth stays in a small range around 1.0. When $k M_{apd,0} > 1$ (below the line), on the other hand, the 3 dB bandwidth is inversely proportional to $M_{apd,0}$. In other words, the following empirical formula holds when $k M_{apd,0} > 1$:

$$(\omega_{3dB\tau_t})(kM_{apd,0}) = N$$

Where N is between 3 (when k = 1.0) and 0.5 (when k = 0.001). In other words,

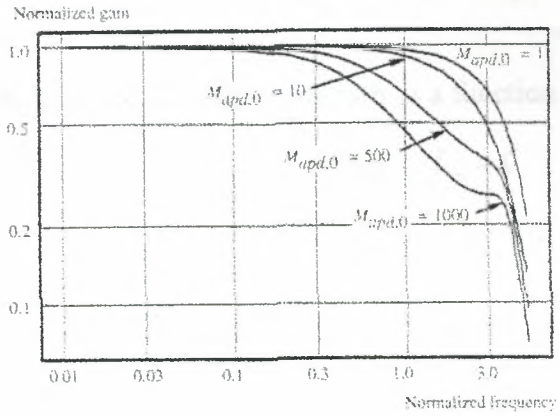

FIGURE 5.8. Frequency response at different values of $M_{apd,0}$: $k = 0.001$.

$$\omega_{3dB} \approx \frac{N}{\tau_t} \tag{5.27}$$

at $L_d = 0$. This shows that the larger the multiplication gains at dc, the smaller the 3 dB bandwidth when $k M_{apd,0} > 1$.

When $L_d = 0$, the time delay .due to the intrinsic region should be included, which results in a smaller 3 dB bandwidth. The 3 dB bandwidths as a function of the dc multiplication gain at various k's and L_d / L_m's is given in Figure 5.10.

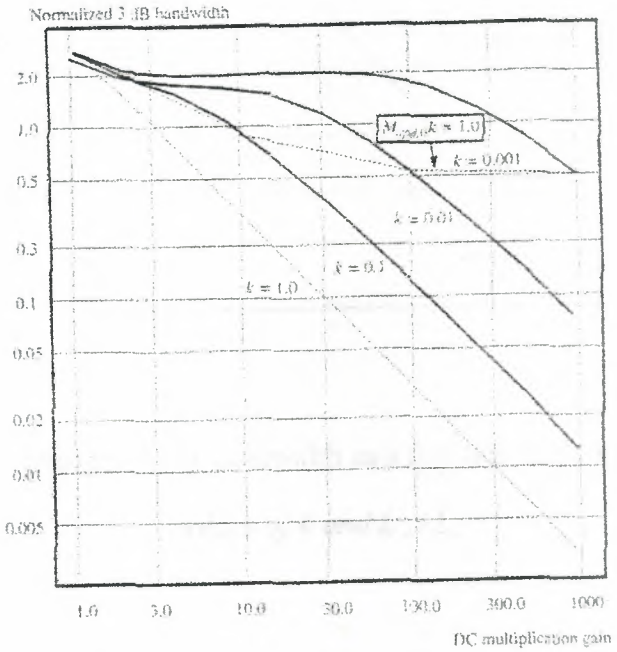

FIGURE 5.9. Normalized 3 dB bandwidth as a function of dc gain $M_{apd,0}$ *at various values of* $k; L_d / L_m = 0$

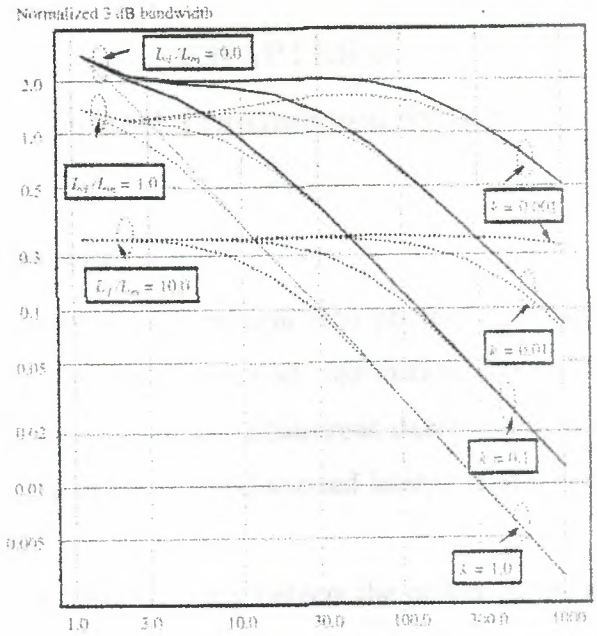

FIGURE 5.10. Normalized 3 dB bandwidth as a function of dc gain $M_{apd,0}$ *at various values of k and* L_d / L_m

CHAPTER 6

Optical Transmission Systems

6.1 Incoherent Detection

Photodetection converts incident light into photocurrent, which is proportional to the power of the incident light and carries no information about the phase of the incident light. As a result, this detection is called **incoherent detection** or **direct detection.** This is in contrast to **coherent detection,** to be discussed later, which detects both the power and phase of the incident light.

Because incoherent detection only detects the power of the incident light, it is used primarily for intensity or amplitude modulated transmission. When phase or frequency modulation is used, coherent detection is necessary. As will be explained later, coherent detection can also amplify the power of the incident light, which can thus improve detection performance and help to approach the quantum limit.

We will focus on incoherent detection for both analog and digital communications. In analog communications, both fiber dispersion and attenuation can be important. To minimize the effect of fiber dispersion, most existing analog transmission systems are based on 1.3 μm transmission. To minimize noise and to achieve a high signal-to-noise ratio (SNR) or carrier-to-noise ratio (CNR), laser RIN noise should be carefully controlled. In addition to dispersion and noise, nonlinear distortion can also be important. Nonlinear distortion in optical communications is primarily caused by the nonlinear characteristics of laser diodes near the threshold current. In a system that is not power limited or dispersion limited, nonlinear distortion can become the ultimate performance limit.

In digital communications, fiber dispersion and various noise sources are important degradation factors. Fiber dispersion can cause intersymbol interference (ISI), which can also be aggravated by inappropriate equalizer design and nonzero turn-on and turn-off delays of light sources and detectors. This chapter evaluates digital detection performance in terms of the bit error rate (BER) and discusses in detail various design considerations.

6.1.1 Analog Signal Detection

A receiver block diagram for analog communications is shown in Figure 1. In addition to the photocurrent generated, noise from the front-end amplifier such as thermal noise and transistor junction noise is added. Because there is a dc bias in analog communications, ac-coupling is used to reject the dc component. Furthermore, to compensate for any channel distortion and to maximize the SNR, an equalizer is commonly used before the final signal output.

In analog communications, the amplitude-modulated signal at the output of the photodiode can be expressed as

$$i_{ph}(t) = I_0[1 + k_m m(t)] \tag{6.1}$$

where I_0 is the dc current, *m(t) is* the message signal, and k_m is the amplitude modulation index. From this expression, the SNR is given by

$$\text{SNR} = \frac{k_m^2 I_0^2 m(t)^2}{\sigma_{n,out}^2} \tag{6.2}$$

Where $\sigma_{n,out}^2$ is the total noise power. Specifically, for a given receiver equalizer $H(\omega)$,

$$\sigma_{n,out}^2 = \int \left[q(I_0 M_{apd} + I_d M_{apd}^2) F_{apd} + \frac{RIN}{2} I_0^2 + S_a \right] |H(\omega)^2| \frac{d\omega}{2\pi} \tag{6.3}$$

where I_d is the dark current, M_{apd} is the current gain of the photodiode (equal to unity if a PIN diode is used), F_{apd} is the excess noise factor of the photodiode. RIN is the relative intensity noise factor, and S_a is the PSD of the equivalent front-end amplifier input noise. In practice, MPN is not important in analog communications and is not included.

An equivalent photocurrent circuit is shown in Figure 6.2, where Z_{in} is the input impedance of the front-end amplifier, and v_a and I_a are the equivalent input voltage and current noise sources, respectively. Therefore, the PSD of the total equivalent input noise source of the front-end amplifier is

$$S_a(\omega) = S_i(\omega) + \frac{S_v(\omega)}{|Z_{in}(\omega)|^2} \tag{6.4}$$

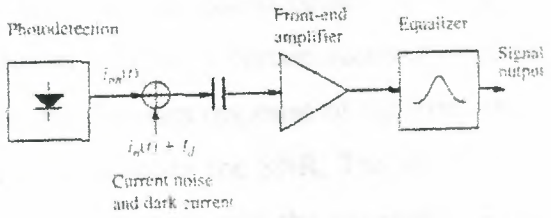

FIGURE 6.1 Block diagram of an analog receiver

A common choice for $H(\omega)$ is a low-pass filter at a cut-off frequency B equal to or greater than the signal's bandwidth. In this case, the output SNR is

$$\text{SNR} = \frac{k_m^2 I_0^2 m(t)^2}{\left[2qI_0 M_{apd} F_{apd} + RINI N_0^2 + (\Re M_{apd} NEP)^2\right] B} \tag{6.5}$$

Where

$$(\Re M_{apd} NEP)^2 B = 2qI_d M_{apd}^2 F_{apd} B + \int_{-2\pi B}^{2\pi B} S_a \frac{d\omega}{2\pi} \tag{6.6}$$

is the signal independent noise power, and NEP is the noise equivalent power. When $m(t)$ is a cosine carrier, or $m(t) = \cos(\omega_m(t)$, the above SNR reduces to the CNR given by

$$\text{CNR} = \frac{k_m^2 I_0^2}{2\sigma_{n,out}^2} \tag{6.7}$$

In addition to noise and bandwidth considerations, nonlinearity is another problem in analog communications. In optical communications, for example, nonlinear distortion can come from laser clipping at the threshold and saturation at a high current bias. Two importtant parameters used in community antenna TV (CATV) to characterize nonlinear distortion are composite second order (CSO) distortion and composite triple beats (CTB).

6.1.2 Binary Digital Signal Detection

A basic block diagram for digital signal detection is shown in Figure 6.2. As illustrated, consists of a photodetector, a front-end amplifier, an equalizer, a sheer (i.e. threshold detector), and a bit timing recovery circuit.

The photodetector converts incident light into photocurrent. The front-end amplifier amplifies the photocurrent with minimal added noise. The equalizer is used in combination with the front-end amplifier to achieve a certain receiver transfer function. For example, can be used to compensate the low-pass response of the front-end amplifier, and it can also be designed to reduce ISI and maximize the SNR. The slicer performs threshold detection. In the case of binary transmission, it detects the equalized output as either high (greater than the threshold) or low (smaller than the threshold). To regenerate the original bit stream, the bit timing recovery circuit recovers the origin transmitter clock from the received signal.

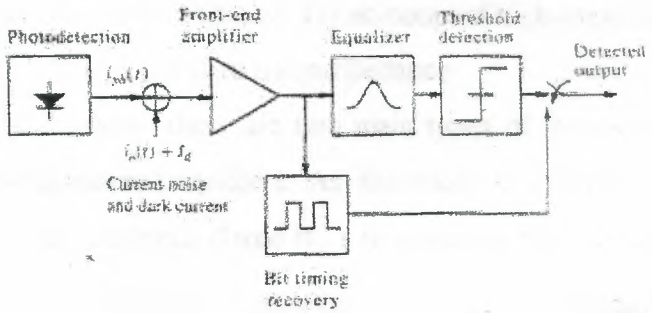

FIGURE 6.2 Block diagram of a typical digital receiver.

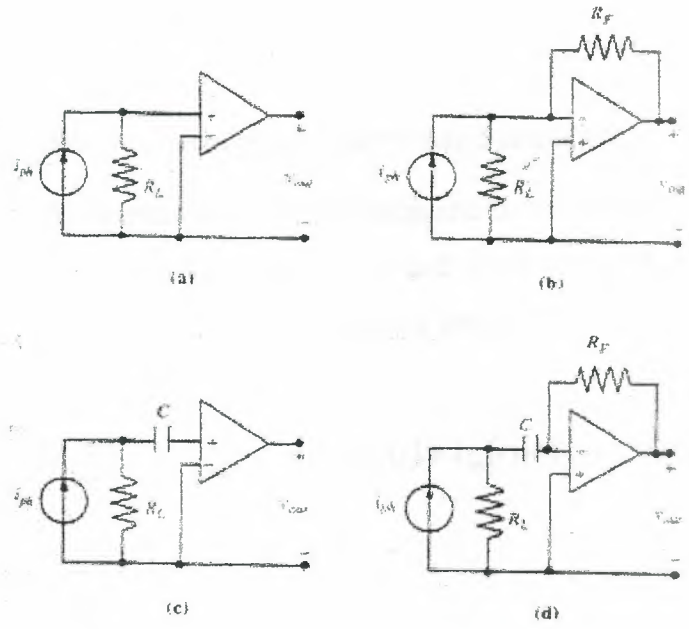

FIGURE 6.3 Illustration of different types of digital receivers: (a) dc-coupled high-impedance, (b) dc-coupled transimpedance, (c) ac-coupled high-impedance, and (d) ac-coupled transimpedance

In optical communications, there are two main types of front-end amplifiers: **high-impedance** and **transimpedance** amplifiers. As illustrated in Figure 6.3, high-impedance amplifiers have a high input resistance (large R_L) to minimize the thermal noise, and trans-impedance amplifiers have a feedback resistance (R_F) to accommodate a large dynamic range of input signals. The design and performance of these amplifiers will be discussed later in this chapter.

Between the photodetector and the front-end amplifier, there are two types of signal coupling: **dc coupling** and **ac coupling.** The classification is determined by whether there is a capacitor or equivalent on the signal path between the photodetector and front-end amplifier. As mentioned earlier, the purpose of ac coupling is to reject the undesirable dc component of the photocurrent output. For example, in digital communications, the nonzero dark current is an undesirable term. For high-impedance amplifiers, the dark current results in a high voltage input to the front-end amplifier, which limits the dynamic range of the signal. When ac coupling is used, however, the dc component of the signal is filtered out. If the signal is not **dc balanced** or has some **dc wander** (i.e. the local time

average of the signal is time varying), ac coupling can cause ISI. This ISI due to ac coupling will be explained later in the chapter.

6.1.3 Signal, Intersymbol Intereference, And Noise Formulation

To evaluate the transmission performance and to understand various design issues, it is useful to first formulate the signal, ISI, and noise at various stages of the digital receiver shown in Figure 6.3. To start, consider a binary digital signal at the photodetector output given by'

$$i_{tot}(t) = \sum A_k p(t - kT_0) + I_d + i_n(t) = i_{ph}(t) + I_d + i_n(t) \tag{6.8}$$

where

$$i_{ph}(t) = \sum A_k p(t - kT_0) \tag{6.9}$$

is the photocurrent due to a pulse-amplitude modulated (PAM) signal, I_d is the dark current of the photodiode, and $i_n(t)$ is the noise current. In binary transmission using on-off keying (OOK), A_k equals either a high value A_H for bit "1" or a low value A_L for bit '0'. The ratio

$$\epsilon = \frac{A_L}{A_H} \tag{6.10}$$

is called the **extinction ratio.** It is desirable to have $\epsilon = 0$ to allow for a larger noise margin. However, because of imperfect bias conditions in practice, it can be slightly greater than 0.

Signal If the front-end amplifier and equalizer have a combined transfer function $H(\omega)$ as illustrated in Figure 6.2, the output of the equalizer is

$$y_{out}(t) = i_{tot}(t) \otimes h(t)$$

where $\otimes$ denotes convolution and $h(t)$ is the impulse response corresponding to the transfer function $H(\omega)$. The signal component of the output signal is thus

$$y_s(t) = i_{ph}(t) \otimes h(t) = \sum_k A_k p_{out}(t - kT_0) \tag{6.11}$$

Intersymbol Interference To detect the transmitted amplitude A_k, y_{out} at the equalizer

output is sampled at the bit rate and compared with a threshold. As mentioned earlier, this is called **threshold detection.** From Equation (6.11), the sampled output at $kT_0+\tau(0 \le \tau \le T_0)$ is

$$y_{out,k} = y_{out}(kT_0 + \tau) = A_k p_{out}[0] + ISI_k + y_{n,k} \tag{6.12}$$

where the constant dark current term has been dropped for its irrelevance. In Equation (6.12),

$$ISI_k = \sum_{k' \neq k} A_{k'} p_{out}[k - k'] \tag{6.13}$$

is the ISI term with $p_{out}[k] = p_{out}(kT_0 + \tau)$, and $y_{n,k}$ is the noise term given by

$$y_{n,k} = y_{n,out}(kT_0 + \tau).$$

The characteristics of output noise $y_{n,out}(t)$ will be discussed shortly.

Equation (6.12) shows that ISI and noise are two primary sources that cause $y_{out,k}$ to deviate from $A_k p_{out}$ and result in error detection. Specifically, when $A_k = A_H$ and $y_{out,k} \le y_{th}$ or when $A_k = A_L$ and $y_{out,k} \ge y_{th}$ there is error detection, where y_{th} is the threshold used in the threshold detection. From this observation, the error detection probability is

$$P_E = p_0 P(y_{out,k} \ge y_{th} \mid A_K = A_L) + p_1 P(y_{out,k} \le y_{th} \mid A_K = A_H) \tag{6.14}$$

where p_0 and p_1 are a priori probabilities for bits "0" and "1"

The threshold y_{th} considered above can be optimized to minimize the BER or P_E. This will be explained in the next section. In practical implementation, the threshold can be directly derived from the dc average of $y_{out}(t)$ through low-pass filtering. For equally possible 1's and 0's, y_{th} is halfway between the high and low values of $y_{out}(t)$.

Noise The two-sided PSD at the photodiode output is

$$S_{n,ph}(\omega, t) = q[i_{ph}(t) M_{apd} + I_d M_{apd}^2] F_{apd} + \frac{1}{2} RIN i_{ph}^2 + S_{MPN}(\omega) \tag{6.15}$$

where the last term is due to mode portion noise. Because $i_{ph}(t)$ is not a constant but depends on $A_k's$, the noise power spectrum is signal dependent and time varying. Because both the MPN and RIN are proportional to i_{ph}^2, the subsequent discussion uses RIN to represent both noise sources for simplicity. The time-dependent noise PSD at the equalizer output can be expressed as:

$$S_{n,out}(\omega,t) = \int S_{n,ph}(\omega, t-t')H(\omega)h(t')e^{j\omega t'}dt' + S_a(\omega)|H(\omega)|^2. \qquad (6.16)$$

If $S_{n,ph}$ is not a function of time. Equation (6.16) reduces to the standard form:

$$S_{n,out} = [S_{n,ph}(\omega) + S_a(\omega)]|H(\omega)|^2. \qquad (6.17)$$

when $S_{n,ph}(\omega)$ is white or relatively independent of frequency (but time varying), the total noise power at the equalizer output is shown to be

$$\sigma_{n,out}^2(t) = S_{n,ph}(0,t) \otimes h(t)^2 + \int S_a(\omega)|H(\omega)|^2 \frac{d\omega}{2\pi}. \qquad (6.18)$$

6.1.4 Received Pulse Determination

Determining the received pulse *p(t)* in Equation (6.9) is important to the subsequent receiver filter design and consequently to the detection performance. When it is known, a proper receiver filter to minimize the BER can be chosen.

In general, the waveform of the received pulse *p(t)* depends on the light source, modulation, line coding, fiber dispersion, and photodetector. For example, an LED light source has a wide spectrum. As a result, fiber dispersion can significantly broaden the pulse. **On** the other hand, if a single-frequency laser diode and external modulation are used, there is no chirping effect and pulse broadening due to fiber dispersion is minimal. In this case, the received pulse *p(t)* is close to the original one that modulates the external modulator,

Light Pulse $p_s(t)$ at the Transmitter Output To derive the input pulse *p(t)*, start from the light source. From the step input response, the pulse output from directly modulating a laser diode can be expressed as

$$p_s(t) = 1 - \cos(\omega_r [t - t_d]) e^{-\alpha(t-t_d)} \qquad \text{if } t_0 \le t \le T_0 \qquad (6.19)$$

where t_d is the initial turn-on delay, ω_r is the relaxation oscillation frequency, α is the damping constant of the relaxation oscillation, and $\alpha_{off} \ge \alpha$ is the decay constant when the laser is turned off. A pulse given by Equation (6.19) is illustrated in Figure 6.4.

In Equation (6.19), the chirping effect is ignored for simplicity. When the turn-on delay t_d and the relaxation oscillation of the laser diode are neglected, $p_s(t)$ can be further approximated as

$$p_s(t) \approx p_m(t) \otimes h_{LD}(t) \qquad (6.20)$$

where $p_m(t)$ is the input pulse that drives the laser diode (a rectangle pulse for NRZ signaling), and $h_{LD}(t)$ is the impulse response of the laser diode. For simplicity, it can be modeled as a first-order low-pass filter with a cutoff frequency of $1/(2\pi TLD)$. Therefore, the impulse response can be expressed as

$$h_{LD}(t) = \frac{1}{\tau_{LD}} e^{-t/\tau_{LD}} \qquad \text{if } t \ge 0 \qquad (6.21)$$

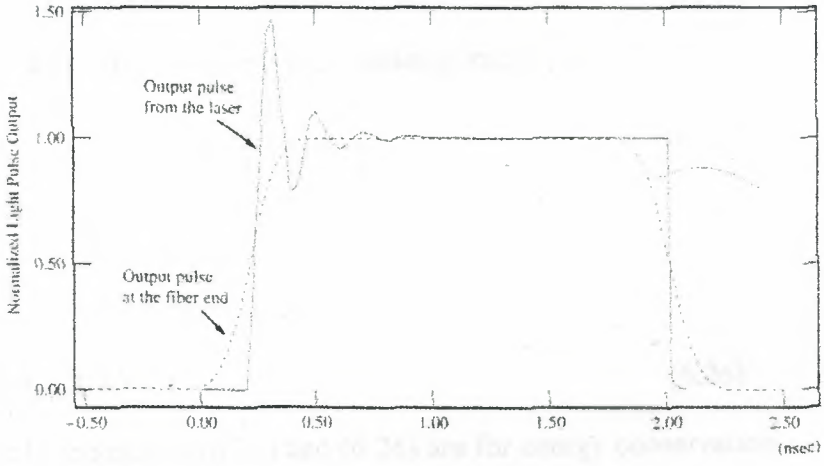

FIGURE 6.4 Pulse output from a laser diode and at the fiber end. The total fiber dispersion is $D_{intra} L\Delta\lambda = 0.2n\,\text{sec}$.

If an LED is used, there is no time delay or relaxation oscillation. Instead, one must consider its rise time and fall time. For simplicity, it can be similarly modeled with the following impulse response:

$$h_{LED}(t) = \frac{1}{\tau_{LED}} e^{-t/\tau_{LED}} \qquad \text{if } t \geq 0 \qquad (6.22)$$

where τ_{LED} is the time constant of the LED. Similar to Equation (20), the output pulse $p_s(t)$ is given by

$$p_s(t) = p_m(t) \otimes h_{LED}(t) \qquad (6.23)$$

Optical Channel The channel response of an optical fiber is determined by the fiber dispersion, fiber length, and source's spectrum. Consider a single-mode fiber of length L and intramodal dispersion D_{intra}, the propagation delay of a photon at wavelength λ is $\tau_g(\lambda) = \tau_{g0} + (\lambda - \lambda_0) D_{intra} L$, where τ_{g0} is the propagation delay at the reference wavelength λ_0 Therefore, if the light source has a normalized spectrum $g_s(\lambda - \lambda_0)$ so that

$$\int g_s(\lambda - \lambda_0) d\lambda = 1 \qquad (6.24)$$

the fiber channel can be modeled with the following impulse response:

$$h_{fiber}(t) = g_s\left(\frac{t - \tau_{g0}}{D_{intra} L}\right) \frac{1}{D_{intra} \Delta\lambda L}. \qquad (6.25)$$

The factor $1/(D_{intra} \Delta\lambda L)$ is introduced to have

$$\int h_{fiber}(t) dt = 1 \qquad (6.26)$$

Conditions given by Equations (6.24) and (6.26) are for energy conservation.

For LEDs or multimode laser diodes, the output light spectrum is commonly assumed to be Gaussian. If the linewidth is $\Delta\lambda$,

$$g_s(\lambda - \lambda_0) = \sqrt{\frac{2}{\pi \Delta\lambda^2}} e^{-2(\lambda - \lambda_0)^2/(\Delta\lambda)^2} \qquad (6.27)$$

Note that this expression meets the condition given by Equation (6.24). From this, the channel impulse response is

$$h_{fiber}(t) = \sqrt{\frac{2}{\pi}} \frac{1}{D_{int\,ra} L\Delta\lambda} e^{-2(\lambda-\lambda_0)^2/(D_{int\,ra}L\Delta\lambda)^2} \qquad (6.28)$$

For single-mode laser diodes, from the Lorentzian spectrum

$$g_s(\lambda-\lambda_0) = \frac{2}{\pi\Delta\lambda} \frac{1}{1+4(\lambda-\lambda_0)^2/\Delta\lambda^2} \qquad (6.29)$$

and

$$h_{fiber}(t) = \frac{2}{\pi D_{int\,ra} L\Delta\lambda} \frac{1}{1+4(t-\tau_{g0})^2/(D_{int\,ra}L\Delta\lambda)^2}. \qquad (6.30)$$

Received Pulse At the receiver end, the impulse response of the photodiode can be similarly modeled as a first-order low-pass filter. The impulse response is thus

$$h_{ph}(t) = \frac{1}{\tau_{ph}} e^{-t/\tau_{ph}} \qquad \text{if } t \geq 0 \qquad (6.31)$$

where τ_{ph}, is the time constant of the photodiode.

Given the output pulse $p_s(t)$, the channel impulse response $h_{fiber}(t)$, and the photodetector response $h_{ph}(t)$, the received pulse at the front-end amplifier input is given by

$$p(t) = p_s(t) \otimes h_{fiber}(t) \otimes h_{ph}(t). \qquad (6.32)$$

An output pulse according to Equations (19) and (27) at $D_{int\,ra} L\Delta\lambda$ = 0.2 nsec is illustrated in Figure 4, where $h_{ph}(t) = \delta(t)$ is assumed.

6.1.5 Receiver Equalizer Design

As can be seen from Equation (6.18), the choice of the total receiver transfer function $H(\omega)$ determines the noise power and the signal. If an improper $H(\omega)$ is used, there will be either excessive noise power or significant signal distortion (i.e., ISI). This section discusses receiver design and the trade-off between noise and ISI.

When ISI due to fiber dispersion is important, the matched filter is not necessarily

the best choice to minimize the BER. Instead, optimum detection involves matched filtering, sampling, and sequence estimation. In this optimum detection, the use of matched filtering and sampling at the bit rate generates a set of sufficient statistics (i.e., no information loss). All samples are then jointly detected to minimize the error detection probability.

In optical communications, this optimum detection is impractical for high-speed transmission. Furthermore, the detection technique is applicable only to Gaussian noise. When signal-dependent noise such as shot noise is important, the technique may not even be applicable. Therefore, depending on whether noise or ISI is stronger, two main types of filters are used in practice. When noise is a stronger factor, a low-pass filter or integration-and-dump is used. When ISI is a stronger factor, a raised-cosine filter is preferred. These two filters are discussed below.

Integration-and-Dump Filtering Integration-and-dump has been considered in the previous examples. An implementation is illustrated in Figure 6.5, where the generated photocurrent is integrated every bit interval. At the end of integration, the integrated value is sampled and threshold detected. From Equation (6.9), **the** integrated output at time kT_0+T_0 is

$$v_{out,k} = \frac{1}{C}\int_{kT_0}^{(k+1)T_0} i_{tot}(t)dt$$

$$= \frac{1}{C}\left[rect(t,T_0)\otimes i_{tot}(t)\right]_{t=kT_0+T_0} \qquad (6.33)$$

where *rect(t. T_0)* is the unit rectangle function from 0 to T_0 . The last convolution expression shows that integration-and-dump is equivalent to matched filtering if *p(t)* is an NRZ pulse. Furthermore, because the impulse response *rect(t, T_0)* has a transfer function of a cutoff frequency 1/ T_0, integration-and-dump is low-pass filtering of bandwidth *I/* T_0.

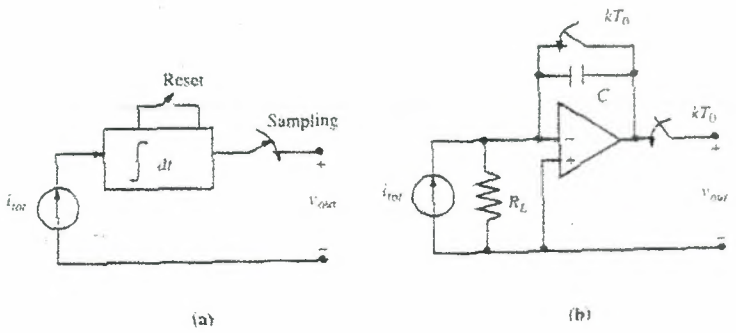

FIGURE 6.5 (a) Integration and dump detection and (b) implementation of integration and dump.

Raised-Cosine Filtering When the received pulse *p(t)* has a finite duration greater than one bit interval, the use of integration-and-dump results in ISI. When ISI is a stronger factor than noise, an equalizer must be used to reduce ISI. Although in general this can enhance the noise power at the same time. it is still good if the final BER is reduced.

To reduce ISI, one approach is to force ISI to zero. This kind of equalizer is called the **zero-forcing** equalizer. When a given output pulse $p_{out}(t)$ zero ISI is chosen, the zero-forcing equalizer has a transfer function (including that of the front-end amplifier) given by

$$H(\omega) = \frac{P_{out}(\omega)}{P(\omega)} \tag{6.34}$$

where P (ω) is the Fourier transform of p(t).

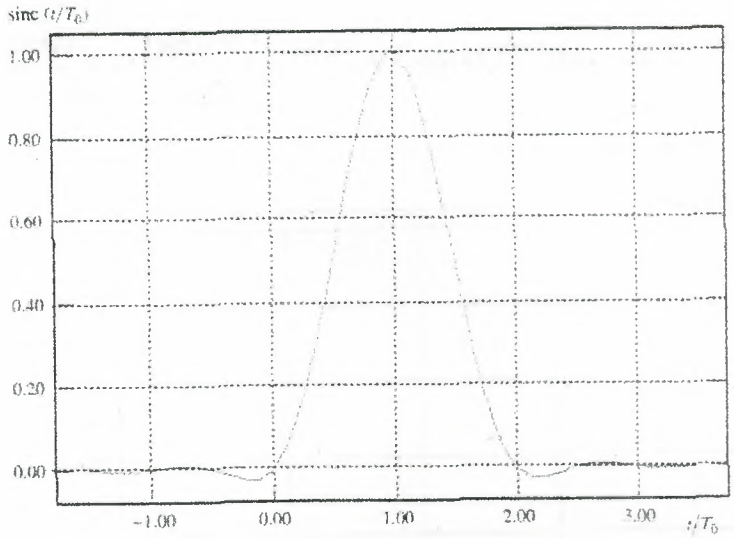

FIGURE 6.6. Raised cosine waveform of zero ISI at 100 percent excess bandwidth.

An important zero-forcing equalizer is called the **raised-cosine filter.** Its equalized output is given by

$$p_{out}(t) = \sin c(2t/T_0 - 2) + \frac{1}{2}\sin c(2t/T_0 - 3) + \frac{1}{2}\sin c(2t/T_0 - 1)$$

$$= \frac{1}{1 - 4[(t/T_0 - 1) - 1]^2}\sin c(2t/T_0 - 2) \qquad (6.35)$$

This pulse is illustrated in Figure 6. This definition says that

$$p_{out}(t)|_{kT_0} = \begin{cases} 1 & \text{if } k=1 \\ 0 & \end{cases} \qquad (6.36)$$

Therefore,

$$y_s(kT_0 + T_0) = \left[\sum_{k'} A_{k'} p_{out}(t - k'T_0)\right]_{kT_0+T_0} = A_k. \qquad (6.37)$$

Thus there is no ISI at the sample time A_k. The corresponding Fourier transform of $p_{out}(t)$ *is*

$$P_{out}(\omega)=\frac{T_0}{2}[1+\cos(\omega T_0/2)]e^{-j\omega T_0} \quad \text{if } |\omega T_0|\le 2\pi \tag{6.38}$$

The corresponding Fourier transform is

$$P_{in}(\omega)=2\frac{\sin(\omega T_0/2)}{\omega}\frac{\sin(\omega\delta)}{\omega\delta}e^{-j\omega T_0/2}=T_0\sin c(fT_0)\sin c(2\delta f)e^{-j\omega T_0/2} \tag{6.39}$$

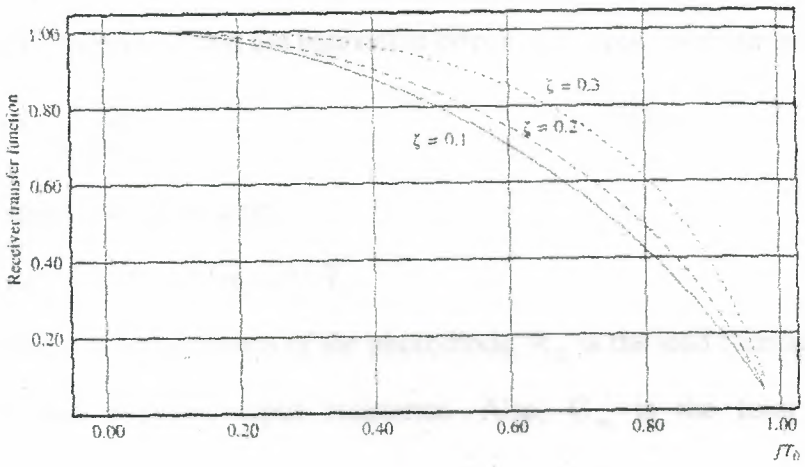

FIGURE 6.7 Frequency response of the raised-cosine filter; $\xi=\delta/T_0$.

Therefore, the equalizer has the following transfer function

$$\mathrm{H}(\omega)=\frac{P_{out}(\omega)}{P_{in}(\omega)}=\frac{1+\cos(\omega T_0/2)}{2\sin c(\omega T_0/2\pi)\sin c(\omega\delta/\pi)}e^{-j\omega T_0/2} \quad \text{if } |\omega T_0|\le 2\pi \tag{6.40}$$

This transfer function at different values of $\xi=\delta/T_0$ is shown in Figure 6.7.

6.1.6 Front-End Amplifiers

The objective of the front-end amplifier is to amplify the signal with minimal added noise. Since the photocurrent signal can be very weak at the front-end amplifier, the added amplifier noise is very critical to the subsequent detection.

As illustrated in Figure 6.3, two important types of front-end amplifiers are (1) high impedance and (2) transimpedance amplifiers. High-impedance amplifiers are optimized

from low noise consideration, which is important in long-distance point-to-point communications. Transimpedance amplifiers, on the other hand, are optimized from wide dynamic range consideration, which is **important to multiple access.**

6.1.6.1 High-Impedance Amplifier

An equivalent circuit for a high impedance amplifier is depicted in Figure 6.8, where the amplifier can be a bipolar junction transistor (BJT), a field effect transistor (FET), or an operational amplifier. From the equivalent circuit, the input impedance is

$$Z_{in}(\omega) = \frac{R_{in}}{1 + j\omega R_{in} C_{in}} \qquad (6.41)$$

where R_{in} *is* the total input resistance with

$$G_{in} = 1/R_{in} = 1/R_L + 1/R_d + 1/R_a .$$

In the equation, R_d is the output resistance of the photodiode, R_L is the load resistance, And R_a is the front-end amplifier input resistance. Also, C_{in} is the total input capacitance, which can be expressed as

$$C_{in} = C_d + C_s + C_a$$

Where C_d is the diode junction capacitance, C_s is the stray capacitance, and C_a is the front-end amplifier input capacitance.

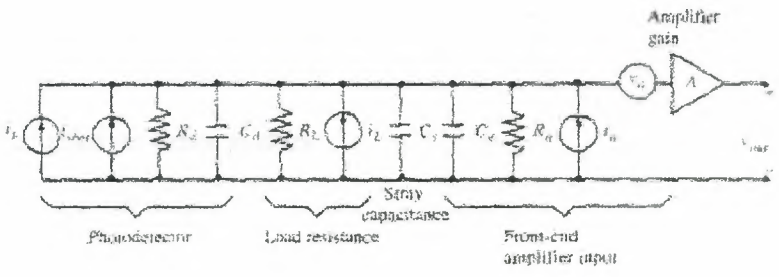

FIGURE 6.8. Equivalent circuit of high impedance front-end amplifier.

From the input resistance R_{in}, there is a current thermal noise with the PSD given by

$$S_{th} = 2kTG_{in} \tag{6.42}$$

From equations (4) and (18), the PSD of the receiver noise

$$S_a(\omega) = S_i(\omega) + S_v(\omega)\frac{1+(\omega R_{in}C_{in})^2}{R_{in}^2} = S_i(\omega) + \frac{S_v(\omega)}{R_{in}^2} + \omega^2 C_{in}^2 S_v(\omega) \tag{6.43}$$

Therefore, high-impedance amplifiers can have a minimal noise power by using a large R_{in} When R_{in} is large enough, the middle term in Equation (6.43) can be dropped. Specific current and voltage sources at the input of different types of front-end amplifiers are shown below.

BJT Amplifier For BIT devices, the PSD of the current noise source is qI_B, where I_B is the base current. Therefore, the total current noise source is

$$S_i = qI_B + 2kTG_{in} \approx qI_B \tag{6.44}$$

where the term $2kTG_{in}$ is dropped if G_{in} is small. Also, the voltage noise source for BIT is

$$S_v = \frac{2kT}{g_m} \tag{6.45}$$

Where g_m is the equivalent transconductance of the transistor equal to

$$g_m = \frac{qI_C}{kT} \tag{6.46}$$

The total PSD of receiver noise is thus

$$S_{a,BJT} \approx S_i + 2kT\frac{\omega^2 C_{in}^2}{g_m} \tag{6.47}$$

From the PSD derived, if the total receiver transfer function is $H(\omega)$, the output noise power due to the front-end amplifier is

$$\sigma_{a,BJT}^2 = \int S_a(\omega)|H(\omega)|^2 \frac{d\omega}{2\pi} = \int \left[S_i(\omega) + \omega^2 C_{in}^2 S_v\right] |H(\omega)|^2 \frac{d\omega}{2\pi} \tag{6.48}$$

which can be conveniently expressed as

$$\sigma_{a,BJT}^2 = qI_B B J_0 + 2kT \frac{(2\pi C_{in})^2}{g_m} B^3 J_2 \tag{6.49}$$

where

$$J_i = \int \left|\hat{H}(x)\right|^2 x^i dx \qquad \text{for i = 0,1,2} \tag{6.50}$$

is a normalized parameter with

$$\hat{H}(x) = H(2\pi Bx) \tag{6.51}$$

FET Amplifier For FET devices, the current noise source is

$$S_i = qI_G + 2kTG_{in} \approx 0 \tag{6.52}$$

where I_G *is* the gate current and is close to zero. Similar to the voltage noise source in BJT, the voltage noise source is

$$S_v = \frac{2kT\Gamma}{g_m} \tag{6.53}$$

where Γ is a material dependent parameter. A typical value for Γ is from 0.5 to 3.0. Therefore,

$$S_{a,FET} \approx \omega^2 C_{in}^2 S_v = 2kT\Gamma \frac{\omega^2 C_{in}^2}{g_m} \tag{6.54}$$

and the receiver noise output power is

$$\sigma_{a,BJT}^2 = \int \omega^2 C_{in}^2 S_v \left|H(\omega)^2\right| \frac{d\omega}{2\pi} = 2kT\Gamma \frac{(2\pi C_{in})^2}{g_m} B^3 J_2 \tag{6.55}$$

6.1.6.2 Transimpedance Amplifier

As illustrated in Figure 6.3, a transimpedance amplifier has a negative feedback resistance.

The equivalent circuit is shown in Figure 6.10. In this circuit, in addition to the input photocurrent and noise, as in the case of high-impedance amplifiers, there is thermal nose (i_F) from the feedback resistor R_F.

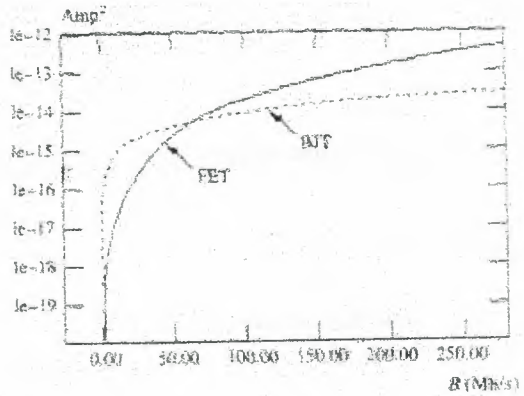

FIGURE 6.9 Noise power comparison between FET and BJT front-end amplifier.

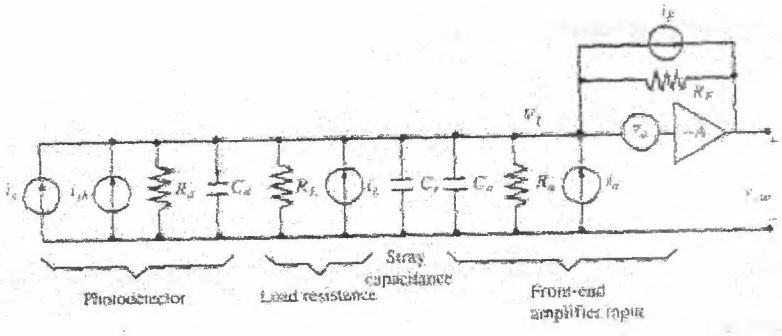

FIGURE 6.10. Equivalent circuit of transimpedance front-end amplifier.

The circuit analysis is more complicated than for high-impedance amplifiers. For simplicity, the amplifier gain A is assumed to be much greater than one and its input impedance is assumed to be infinity. Under these approximations, the signal output V_{out} satisfies the following equation:

$$i_{ph} = \frac{v_1}{Z_{in}} + \frac{v_1 - v_{out}}{R_F}$$

where

$$v_{out} = -Av_1 \tag{6.56}$$

This gives

$$v_{out} = -\frac{i_{ph}R_F}{1+1/A+R_F/AZ_{in}} \approx -i_{ph}R_F \tag{6.57}$$

Thus the output voltage is controlled by the feedback resistance and not by the amplifier Gain. Furthermore, because the pole of Equation (6.57) is approximately $A/(R_F C_{in})$, at a large A, it is much higher than the signal's bandwidth. Therefore, the transimpedance amplifier can be considered as an all-pass filter with the transimpedance gain equal to $-R_F$.

PSD of Parallel Current Noise Sources The output noise power spectrum is the superposition of the output due to each noise source. Because shot noise, RIN noise, thermal noise, and current noise of the front-end amplifier are all parallel to the signal photocurrent, their total output spectrum is R_F^2, times the sum of the individual spectra. That is,

$$S_{i,out} = R_F^2\left(qF_{apd}M_{apd}i_{apd} + \frac{RIN}{2}i_{ph}^2 + \frac{2kT}{R_{in}} + S_i\right) \tag{6.58}$$

PSD of Feedback Resistor For noise output due to thermal current noise i_F,

$$\frac{v_1}{Z_{in}} + i_F + \frac{v_1 - v_{F,out}}{R_F} = 0 \tag{6.59}$$

Using Equation (6.56),

$$v_{F,out} \approx i_F R_F$$

Therefore, the output noise power spectrum due to if is

$$S_{F,out} = 2kTR_F = R_F^2\frac{2kT}{R_F}$$

PSD of Voltage Noise Source For noise output due to the voltage source v_a of the front-end amplifier,

$$-A(v_a + v_1) = v_{a,out}$$

and

$$\frac{v_1 - v_{a,out}}{R_F} + \frac{v_1}{Z_{in}} = 0$$

These gives

$$v_{a,out} \approx -R_F v_a \left(\frac{1}{R_F} + \frac{1}{Z_{in}} \right)$$

If

$$\frac{1}{R_P} = \frac{1}{R_F} + \frac{1}{R_{in}} \tag{6.60}$$

then

$$S_{v,out} = R_F^2 S_v \left(\frac{1}{R_F^2} + \omega^2 C_{in}^2 \right) \tag{6.61}$$

Total PSD Adding all the noise terms, the total noise output spectrum is

$$S_{n,out} = R_F^2 \left(qF_{apd} M_{apd}^2 i_{ph} + \frac{RIN}{2} i_{ph}^2 + \frac{2kT}{R_P} + S_i \right) + R_F^2 S_v \left(\frac{1}{R_P^2} + \omega^2 C_{in}^2 \right) \tag{6.62}$$

Dividing the output spectrum by the factor R_F^2 gives the equivalent input noise spectrum:

$$S_{n,in,trans} = qF_{apd} M_{apd}^2 i_{ph} + \frac{RIN}{2} i_{ph}^2 + \frac{2kT}{R_P} + S_i + S_v \left(\frac{1}{R_P^2} + \omega^2 C_{in}^2 \right) \tag{6.63}$$

Comparing this with that of the high-impedance amplifier gives

$$S_{n,in,trans} = S_{n,in,high} + \frac{2kT}{R_F} + \frac{S_V}{R_P^2} \tag{6.64}$$

And

$$\sigma_{a,trans}^2 = \sigma_{a,high}^2 + \frac{2kT}{R_F} BJ_0 + \frac{2kT}{g_m R_P^2} BJ_0 \tag{6.65}$$

Therefore, the transimpedance amplifier has two extra noise power terms due to (1) feedback resistance noise and (2) front-end amplifier voltage noise.

6.1.6.3 Allowable Dynamic Range

For a given front-end amplifier design, there is a window of the received signal power within which satisfactory performance can be achieved. The lower limit of this window is determined by the receiver sensitivity. The higher limit of the window is determined by the receiver amplifier gain saturation discussed below.

Consider a high-impedance or transimpedance amplifier. Let

$$A_H G = v_{out}$$

where G is the transimpedance gain equal to either $R_{in}A$ (high-impedance amplifiers) or $R_F A_H$ (transimpedance amplifiers), and V_{out}, is the voltage at the amplifier output. If the output V_{out} cannot be higher than V_D due to either bias or other circuit constraints, the photocurrent high level A_H needs to be lower than V_D/G for linear response. In other words, if $A_H \geq V_D/G$, the output v_{out}, stays at the same value V_D. Although this distortion is fine in digital communications, the output noise power can continue to increase after signal saturation. At a high photocurrent level, the noise power can be dominated by the RIN noise. In this case,

$$\sigma_H + \sigma_L \approx \sigma_H = G\sqrt{RIN * A_H^2 B}$$

6.2 COHERENT DETECTION

In previous section we discussed about incoherent detection, where only the intensity of the incident light is detected. Although incoherent detection is simple in implementation, it cannot detect the phase and frequency of the received signal. In other words, it can detect only amplitude modulated (AM) signals. When phase modulation (PM) or frequency modulation (FM) is desirable, such as when intensity noise is strong, coherent detection becomes a better choice. This is familiar from radio communications, where FM is much better than AM in transmission quality.

Coherent detection is also important in applications such as wavelength division multiplexing (WDM), where multiple channels are transmitted at the same time. As discussed before, coherent detection is one important technique used to tune in or select one particular frequency channel. Although passive tunable filters **can** be used to avoid

coherent detection, a larger channel separation is necessary because of limited filter resolution.

In the history of lightwave technology development, a more important reason behind the active coherent detection research work is its ability to amplify the received signal optically for a better signal-to-noise ratio (SNR). Practical incoherent detection receivers, however, still have a performance far worse than the quantum limit because of the excess noise. Coherent detection can avoid this problem and at the same time provide signal amplification. As a result, coherent detection can have a performance close to the quantum limit.

Optical amplifiers developed over the last few years provide another attractive alternative. For example, Erbium-doped fiber amplifiers (EDFAs) can be easily inserted into regular optical fibers for a power gain of 20-30 dB and at a pumping efficiency of 5-10dB/mW. One disadvantage is that the amplifier also introduces noise because of amplified spontaneous emission (ASE).

6.2.1 Basic Principles of Coherent Detection

Although coherent detection is relatively new in optical communications, it has been around in radio communications for a long time. In both radio and optical communications. the essence of coherent detection is to generate a product term of the received signal and a local carrier. As a result, the received passband signal can be demodulated or shifted back to baseband.

As an example, consider a passband signal m(r) cos (ω_{inc} t)shown in Figure 6.11a. To recover the original baseband signal *m(t)*, the received signal is multiplied by a local oscillator $cos(\omega_{toc}t)$. If the local carrier is synchronized to the received signal *m(t)* cos(ω_{inc}) in frequency, i.e.($\omega_{loc} = \omega_{inc}$), the product term is

$$m(t)\cos(\omega_{inc})*cos(\omega_{too}t)=\frac{1}{2}m(t)+\frac{1}{2}m(t)\cos(2\omega_{inc}t).$$

Therefore, the baseband signal can be recovered using a low-pass filter.

The above scheme is called **homodyning** because $\omega_{loc} = \omega_{inc}$. In practice, the local

carrier frequency does not have to be equal to ω_{inc}. In this case, the coherent detection scheme is called **heterodyning** and demodulation is performed in two stages. As will be explained in detail in this chapter, there are various techniques that can be used for the second-stage demodulation.

6.2.1.1 Optical Mixing

Although the use of the multiplier to generate the product term is common in radio communications, it is not practical in optical communications. An alternative way is to mix the incident signal with a local optical carrier. As illustrated in Figure 11 *b*, if the two signals have the same polarizations, the magnitudes of their fields can be scalarly added. In this case, because the photocurrent output is proportional to the combined intensity,

$$I_{ph} = \Re\left[P_{inc} + P_{loc} + 2\sqrt{P_{inc}P_{loc}}\cos(\omega_{inc}t - \omega_{loc}t)\right] \qquad (6.66)$$

Where $\Re$, is the responsivity of the photodiode and P_{loc} is the local oscillator power. Among the three terms, P_{loc} is a constant term that can be simply filtered out by ac-coupling. The third term is the product term of interest. Because $P_{loc} \geq P_{inc}$, $\sqrt{P_{inc}P_{loc}}$ is much larger than P_{inc} - Therefore, the latter term can be dropped.

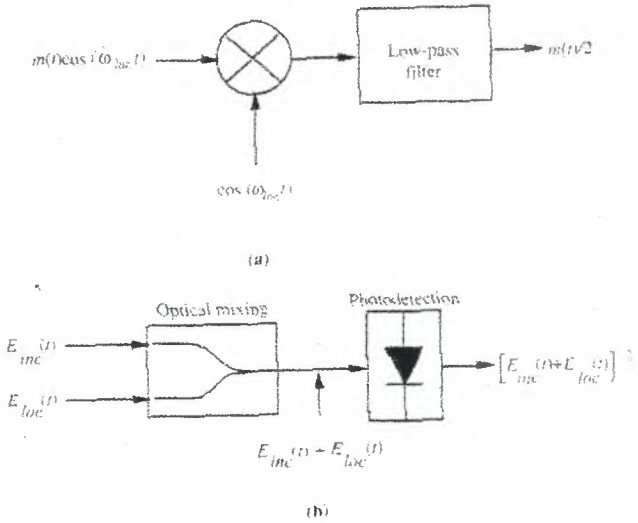

FIGURE 6.11 Coherent detection in (a) radio and (b) optical communication

6.2.1.2 Homodyne and Heterodyne Detection

A more detailed block diagram of coherent detection is shown in Figure 6.12. As mentioned earlier, there are two different types: homodyne and heterodyne detection. In the latter case, the two frequencies differ by a radio frequency called **intermediate frequency** (IF) and denoted by $\omega_{IF} = \omega_{inc} - \omega_{loc}$. Also, the photocurrent output is filtered by an IF or bandpass filter. In general, it is easier to implement heterodyne detection because of simpler carrier synchronization (see Section 6.2.3). However, the trade-off is a lower receiver sensitivity by a few dBs.

As shown in Figure 6.12, there are some common blocks in both homodyne and heterodyne detection. In addition to photodetection and a local oscillator, they both use a carrier recovery loop for local carrier synchronization, a device for polarization control, and a hybrid for optical mixing. The functions of these common blocks are described below.

Carrier Recovery In homodyne detection, the carrier recovery loop uses a photodetector output to drive the carrier loop. The photodetector output carries the phase difference information of the signal and the local oscillator. In heterodyne detection, on the other hand, the output of the IF filter is used to drive an **automatic frequency control** (AFC) device in the carrier loop. The APC generates an output that is proportional to the difference of the frequency of the IF filter output and the specified *uiif* value. This thus maintains the frequency difference between the local oscillator output and the received signal. Detailed discussion on carrier recovery is given in Section 6.2.3.

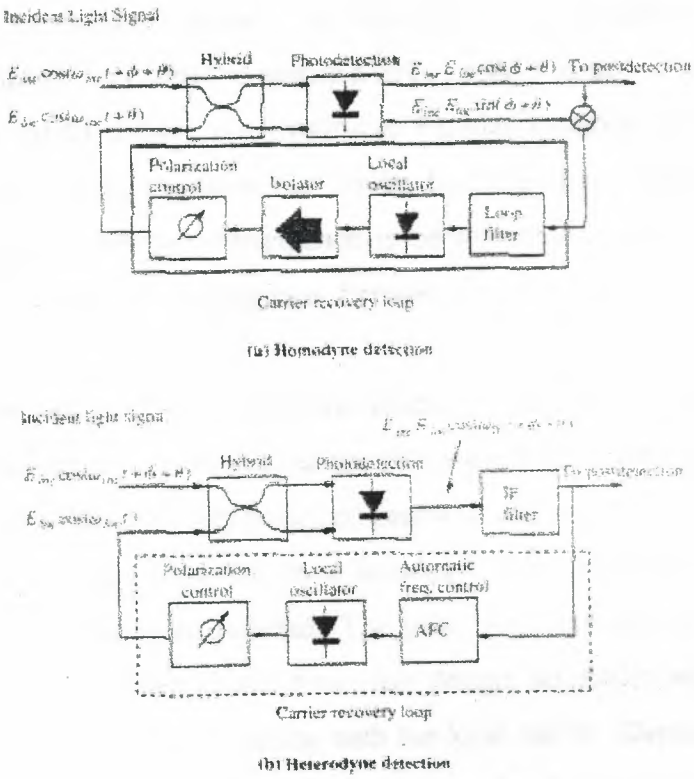

FIGURE 6.12 Block diagram of coherent detection: (a) homodyne detection and (b) heterodyne detection.

Polarization Control in Coherent Detection As mentioned earlier, the photocurrent given in Equation (6.1) assumes the two light signals have the same polarization. In general, this may not be the case. Let $E_{inc}(t) = S(t)x$ be the electric field of the incident light, where x is the unit vector in the direction of the polarization, and let $\mathrm{E}_{inc}(t) = S(t)\overset{\Lambda}{x}$ be the electric field of the local oscillator. When the two signals are mixed, the output photocurrent is proportional to

$$|E_{inc} + E_{loc}|^2 = |S(t)|^2 + |L(t)|^2 + 2\Re\{S(t)L(t)^*\}\overset{\Lambda}{x}.\overset{\Lambda'}{x}$$

where $\overset{\Lambda}{x}.\overset{\Lambda'}{x}$ is the inner product of the two unit vectors. As mentioned earlier, the cross term $2\Re\{S(t)L(t)^*\}\overset{\Lambda}{x}.\overset{\Lambda'}{x}$ carries the signal information for detection. To maximize this term, it is

desirable to maximize the inner product or align the two polarizations. Therefore, it is important to use polarization control to ensure a large product term.

To implement polarization control, polarization or Faraday rotators can be used. They are made of an isotropic media and have the similar birefringence property. Different from electro-optic modulators, whose birefringence is between two *linearly* polarized waves, polarization rotators have a birefringence between two opposite, *circularly* polarized waves.

An alternative approach to polarization control is the use of a polarization diversity receiver, as shown in Figure 6.13. In this design, two **polarizing beam splitters** (PBSs) are used to separate the two orthogonally polarized beams of the local laser output and the incident light. From the separations, the same polarized beams from the incident light and local laser output are mixed and detected. The two photocurrent signals from the two orthogonal polarizations are then added. From this design, no matter what the polarization of the incident light, there is always mixing with the local carrier. Depending on the power partition among the two orthogonal polarizations, it can be shown that the diversity design can maintain 70 percent of the peak photocurrent .

Hybrids The device that mixes two light signals is called a hybrid, which in general is a four-port device, whose two inputs and two outputs can be related by a 2 x 2 matrix:

$$\begin{bmatrix} E_{01} \\ E_{02} \end{bmatrix} = \begin{bmatrix} T_1 & X_1 \\ X_2 & T_2 \end{bmatrix} \begin{bmatrix} E_{inc} \\ E_{loc} \end{bmatrix} = \bar{H} \begin{bmatrix} E_{inc} \\ E_{loc} \end{bmatrix}. \qquad (6.67)$$

In coherent detection, there are two important types of hybrids that deserve further consideration. The first type is called the 180° hybrid, with the transfer matrix given by

$$\bar{H}_{180} = \frac{1}{\sqrt{2}} e^{j\theta} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \qquad (6.68)$$

Note that there is a 180° phase shift between T_1 *and* T_2, and the hybrid is lossless because $|T_1|^2 + |X_2|^2 = |T_2|^2 + |X_1|^2 = 1$.

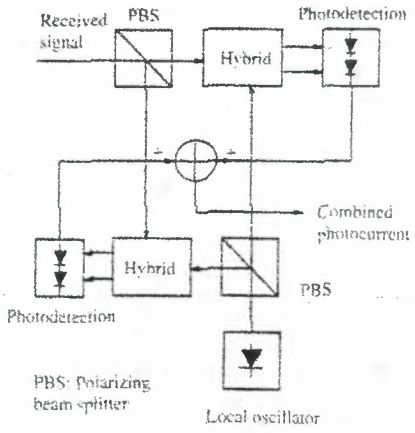

FIGURE 6.13 A polarization diversity receiver. A polarizing beam splitter separates two orthogonally polarized beams.

Another important kind of hybrid has a transfer matrix given by

$$\bar{H}_{90} = \frac{a}{\sqrt{2}} e^{j\theta} \begin{bmatrix} 1 & 1 \\ 1 & j \end{bmatrix} \tag{6.69}$$

where $.0 < a < 1$ is a certain loss factor from practical implementation. This hybrid is called the 90° hybrid because there is a 90° phase shift between T_1 and T_2. As Section 6.2.3 explains, 90° hybrids are needed for carrier recovery based on the Costas loop.

In practical 90" four-port hybrid design, the loss factor a cannot be greater than $1/\sqrt{2}$ because of the limitation of physics. This implies at least a 3 dB power loss and is undesirable.

Fortunately, when hybrids are used for signal detection and carrier recovery, a 90° six-port hybrid, illustrated in Figure 6.14. In this design, 50 percent of the signal power is used for signal detection, and the remaining 50 percent is used for carrier recovery.

6.2.2 Signal and Noise Formulations in Coherent Detection

After the two light signals are mixed by the hybrid, there are two main configurations used in photodetection: **single detection** and **balanced detection.** As illustrated in Figure 6.15, single detection uses only one photodiode. This is the same as in incoherent detection. In

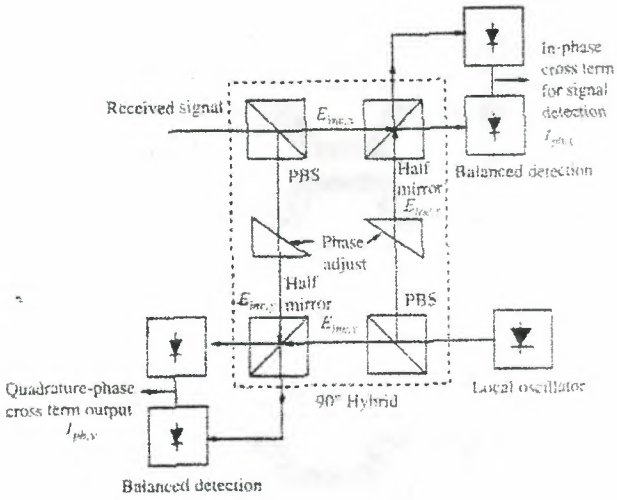

FIGURE 6.15 A six-port hybrid with two input ports and four output ports.

this case, one of the hybrid's outputs is not used and can be used for carrier recovery as discussed later. Balanced detection feeds the two outputs to two photodiodes whose current outputs are subtracted. As will be explained shortly, one major advantage of balanced detection is that it cancels the relative intensity noise (RIN) from the local oscillator.

Signal Formulations Using Balanced Detection Without loss of generality, consider the use of a 180° hybrid. The two outputs from the hybrid can thus be expressed as

$$\mathrm{E}_{01} = \frac{1}{\sqrt{2}}\left(E_{inc} + E_{loc}\right)$$

And

$$\mathrm{E}_{02} = \frac{1}{\sqrt{2}}\left(E_{inc} - E_{loc}\right).$$

After photodetection,

$$\mathrm{I}_{ph,1} = \frac{1}{2}\Re\left\{P_{inc} + P_{loc} + 2\sqrt{P_{inc}P_{loc}}\cos\left[\left(\omega_{inc} - \omega_{loc}\right)t + \phi(t)\right]\right\} \qquad (70)$$

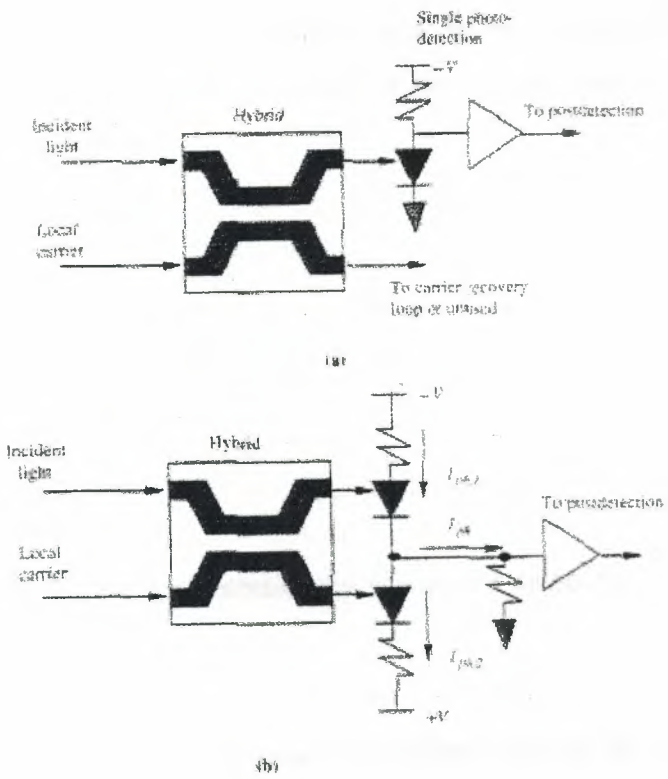

FIGURE 6.16 (a) Single detection versus (b) balanced detection.

$$I_{ph,2} = \frac{1}{2}\Re\left\{P_{inc} + P_{loc} - 2\sqrt{P_{inc}P_{loc}}\cos\left[(\omega_{inc} - \omega_{loc})t + \phi(t)\right]\right\} \tag{6.71}$$

where P_{inc} is the incident light power and P_{loc} is the local carrier power. In amplitude modulation, P_{inc} is modulated according to the transmitted data. Also, $\phi(t)$ is the phase of the carrier and can be used for phase modulation.

With balanced detection, the difference between the photocurrents is

$$I_{ph} = I_{ph,1} - I_{ph,2} = \left\{2\Re\sqrt{P_{inc}P_{loc}}\cos\left[(\omega_{inc} - \omega_{loc})t + \phi(t)\right]\right\} \tag{6.72}$$

This subtracted current has no dc terms and is twice that of the individual photodiode output. Therefore, use of single detection has a 3 dB (factor 1/2) power loss compared to bal-

anced detection.

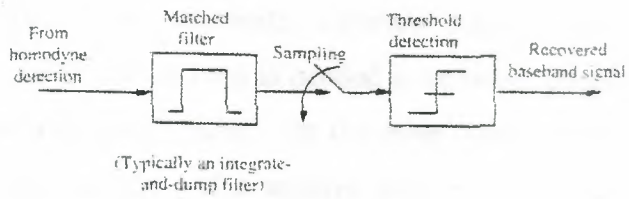

FIGURE 6.17. Postdetection for homodyne detection.

Based on the balanced detection, when homodyne detection is used or $\omega_s = \omega_{loc}$,

$$I_{ph,homo}(t) = \{2\Re\sqrt{P_{inc}P_{loc}}[\cos\phi(t)]\} \qquad (6.73)$$

Similarly, when heterodyne detection is used, or $\omega_{inc} - \omega_{loc} = \omega_{IF}$,

$$I_{ph,hetero}(t) = \{2\Re\sqrt{P_{inc}P_{loc}}\cos[\omega_{IF}t + \phi(t)]\} \qquad (6.74)$$

Signal Detection in Homodyne Detection In the case of homodyne detection, the photocurrent signal given by Equation (6.73) is a baseband signal and immediately ready for detection. Specifically, as shown in Figure 6.17, the photocurrent output from homodyne detection is first equalized by a matched filter and then followed by threshold detection. When the shot noise is approximated as Gaussian and there is no ISI, this matched filtering structure gives the optimum detection performance. When the input pulse is rectangular or a NRZ pulse, the matched filtering is equivalent to integrate-and-dump.

To convert the incident light signal directly to baseband, the carrier frequency of the local optical carrier needs to be synchronized by a carrier loop. As Section 6.2.3 explains, the loop has a feedback circuit that drives the local laser dio de according to the photocurrent until the two carriers have the same optical frequency and a small but fixed phase difference.

Signal Detection in Heterodyne Detection In the case of heterodyne detection, the photocurrent signal given by Equation (6.14) is still a passband signal and consequently needs to be demodulated again. Because detection the carrier loop for frequency synchronization can be relaxed and only needs to ensure that the frequency difference is within the IF band (a fixed phase relationship is unnecessary). To perform postdemodulation, there are two methods: coherent and incoherent postdetections. As shown in Figure 6.18, an IF carrier loop is needed in coherent postdetection to generate a carrier that is in phase with the IF signal. On the other hand, in incoherent postdetection, envelope detection, which consists of a squarer and low-pass filter, is used. Incoherent postdetection can be used to detect amplitude and frequency modulated signals.

Noise Formulation in Balanced Detection The current outputs given in Equations (6.70) and (6.71) contain only signal terms. In practice, there are additional noise terms that need to be added. In addition to receiver noise, two important noise terms are the shot noise from photodetection and the RIN from the local oscillator. Because the RIN power is proportional to the local optical power, which is much larger than the received signal power, the RIN can greatly affect detection performance. When balanced detection is used, the same RIN occurs at the two photodiode outputs. Therefore, by subtracting the two current outputs from balanced detection, the RIN can be cancelled.

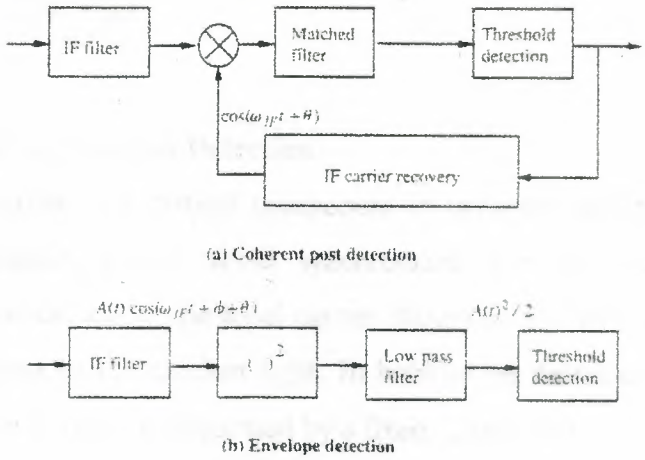

FIGURE 6.17 Use of (a) coherent detection and (b) envelope detection in postdetection for heterodyne detection.

After the RIN is cancelled, the only noise term to consider is the shot noise because of the high local optical power. The two-sided power spectral density (PSD) of noise at each photodiode output is

$$S_{n,i}(\omega) = \frac{1}{2} q\Re P_{loc}$$

Where i is either 1 or 2. When the two current outputs are subtracted in balanced detection, the total noise power is

$$S_n(\omega) = S_{n,1}(\omega) + S_{n,2}(\omega) = q\Re P_{loc} B . \qquad (6.75)$$

As discussed in chapter 1, shot noise can be assumed to be Gaussian when the noise power is large. If an integrate-and-dump filter is used in Figure 16 as the matched filter for homodyne detection, the noise power at the threshold detector input is

$$\sigma^2_{n,\text{hom}\,o} = q\Re P_{loc} T . \qquad (6.76)$$

When heterodyning is used, an additional IF demodulation is needed. In the case of coherent IF demodulation, as in Figure 6.15 *a*, an IF carrier $\cos(\omega_{IF} t)$ is used to multiply the combined photocurrent. The corresponding noise power after integrate-and-dump is scaled down by a factor of 2. That is,

$$\sigma^2_{n,hetero} = \frac{1}{2} q\Re P_{loc} T \tag{6.77}$$

6.2.3 Carrier Recovery in Coherent Detection

As mentioned earlier, one critical component in coherent detection is the carrier recovery loop that generates a local carrier synchronized with the incident light signal. Specifically, in homodyne detection, the local carrier should be synchronized in both phase and frequency with respect to the incident light. In heterodyne detection, the local carrier should be synchronized in frequency (separated by a fixed $_{IF}$ amount).

Compared to RF carrier recovery, the primary difficulty of optical carrier recovery comes from the need for a similar implementation in the optical domain. For example, optical sources in general have much larger phase noise than their RF counterparts. Therefore, a He-Ne laser instead of a semiconductor diode laser is needed in homodyne detection.

Although there can be many different implementations in RF and optical communications, a carrier recovery loop in general has three components. As illustrated in Figure 6.18, they are (1) phase detector, (2) loop filter, and (3) voltage controlled oscillator (VCO). The VCO generates the local carrier, whose frequency and phase are determined by the voltage (or current) input to the oscillator. The loop filter is generally a low-pass filter. It is used to determine the time response for frequency locking and tracking. The phase detector is used to compare the phases of the received carrier and local carrier. In practice, most carrier recovery loops differ only in the phase detection implementation.

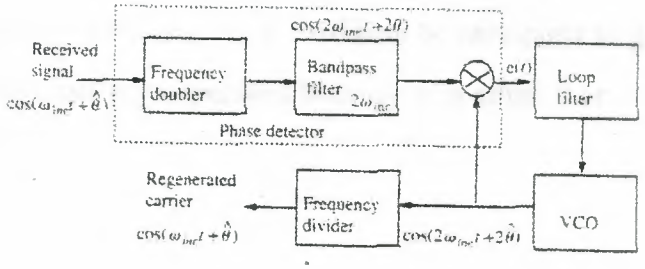

FIGURE 6.18 Use of frequency doubler and divider for binary PSK carrier recovery.

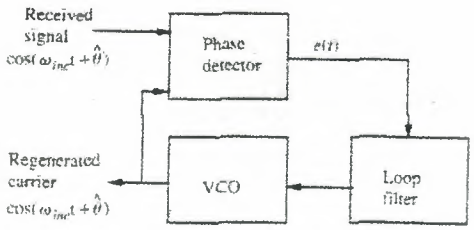

FIGURE 6.19 A block diagram of a carrier recovery phase-locked loop.

The following sections explain phase detection techniques in optical carrier recovery and describe their operation in the steady state. Analysis of a carrier recovery loop for carrier acquisition is beyond the scope of this book.

Homodyne PSK Carrier Recovery An implementation of the Costas loop for optical PSK homodyne detection is shown in Figure 6.20. To get the 90° phase shift from the regenerated carrier, a six-port 90° hybrid (shown in Figure 6.15) is used, where the two PBSs split each of the two inputs into two beams with orthogonal polarizations. In general, the input is assumed or made to be linearly polarized, and the beam splitter is set at 45° with respect to the input polarization. As a result, the two split outputs have equal power. Two half mirrors are then used to mix the output from the PBSs of the same polarization. With proper phase adjustment, a 45° phase shift can be introduced from each phase adjuster. After photodetection and balanced detection, the two outputs are proportional to $\cos(\phi + \Delta\theta)$ and $\sin(\phi + \Delta\theta)$.

The cosine term, $\cos(\phi + \Delta\theta)$, can be used for subsequent detection. At the same time, the two terms $\cos(\phi + \Delta\theta)$ and $\sin(\phi + \Delta\theta)$ can be multiplied to give $2\sin(2\phi + 2\Delta\theta)$. This product term is information independent because ϕ is either 0 or π.

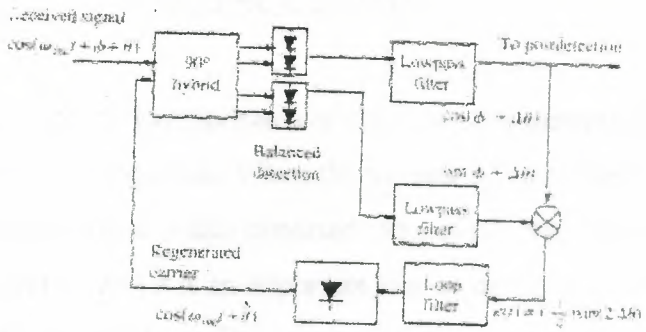

FIGURE 6.20. Costas loop implementations for homodyne PSK carrier recovery using a 90^0 hybrid.

Heterodyne Carrier Recovery In heterodyne detection, the requirement in carrier recovery is much relaxed. For example, it is unnecessary to lock the receiver carrier in phase. Instead, it is necessary only to ensure that the frequency difference of the two carriers be close to the IF frequency. The phase difference is either unimportant in envelope detection or can be taken care of by coherent postdetection.

CONCLUSION

I have concluded that the various kinds of noise are also generated, transmitted, and added to the final detected photocurrent. When the transmission channel is not ideal, the waveform of the transmitted signal is also distorted. As a result, the transmitted signal cannot be perfectly recovered, and it is an important task to minimize the effects of noise and distortion at the receiver end. In analog communications, this means maximizing the signal-to-noise ratio (SNR); in digital communications, this means minimizing the bit error rate (BER).

I have also concluded that unlike thermal noise, most noise sources in optical communications are signal dependent. That is, when the signal level increases, the noise level also increases. For example, shot noise is linearly proportional to the photocurrent generated. Relative intensity noise and mode partition noise power are even worse, being proportional to the photocurrent squared.

In addition to noise and cross talk, there can be signal distortion because of a nonideal channel. In optical communications, distortion can come from fiber dispersion and device nonlinearity. Depending on the signal transmission, channel distortion results in different effects. In analog communications, signal distortion results in intersymbol interference (ISI), which in turn causes an exclusively high bit error rate. Because of noise and ISI, the detection and amplitude of pulse are not necessarily the same. A digital receiver thus needs to 'guess' what amplitude is transmitted from the received detection.

In this project I have seen that the noise from the fiber channel is negligible. On other hand, there are multiple noise sources from the both the transmitter (light source) and the optical receiver. In addition to noise, the received signal can also be corrupted by distortion from a nonideal channel. Therefore, the challenge of the receiver design is to recover the transmitted signal from the corrupted form.

REFERENCES

1. John M. Senior 'Optical Fiber Communication Principles And Practice', Prentice-Hall Inc., 1992
2. Wim Van Etten, Jan Vam der Plaats'Fundamental of Optical Fiber Communication' Prentice-Hall Inc., 1991
3. Max Ming, Kang Liu, 'Principles And Applications of Optical Communication' IRWIN, 1996
4. John Wilson, John Hawkes, 'Optoelectronics', Prentice-Hall Inc., 1998
5. www.shehzada.com
6. www.msn.com
7. www.google.com
8. www.yahoo.com