1 WAVE NATURE OF LIGHT

1.1 LIGHT WAVES IN A HOMOGENEOUS MEDIUM

A. Plane Electromagnetic Wave

The wave nature of light, quite aside from its photonic behavior, is well recognized by such phenomena as interference and diffraction. We can treat light as an electromagnetic (EM) wave with time-varying electric and magnetic fields, E_x and B_y, respectively, which are propagating through space in such a way that they are always perpendicular to each other and the direction of propagation z as illustrated in Figure 1.1. The simplest traveling wave is a sinusoidal wave that, for propagation along z, has the general mathematical form

Traveling wave along z

$$E_x = E_o \cos(\omega t - kz + \phi_o) \tag{1.1.1}$$

in which E_x is the electric field at position z at time t, k is the **propagation constant**[2] given by $2\pi/\lambda$, where λ is the wavelength, ω is the angular frequency, E_o is the amplitude of the wave, and ϕ_o is a phase constant, which accounts for the fact that at $t = 0$ and $z = 0$; E_x may or may not necessarily be zero depending on the choice of origin. The argument $(\omega t - kz + \phi_o)$ is called the **phase** of the wave and denoted by ϕ. Equation (1.1.1) describes a **monochromatic plane wave** of infinite extent traveling in the positive z direction as depicted in Figure 1.2. In any plane perpendicular to the direction of propagation (along z), the phase of the wave, according to Eq. (1.1.1), is constant, which means that the field in this plane is also constant. A surface over which the phase of a wave is constant at a given instant is referred to as a **wavefront**. A wavefront of a plane wave is obviously an infinite plane perpendicular to the direction of propagation as shown in Figure 1.2.

We know from electromagnetism that time-varying magnetic fields result in time-varying electric fields (Faraday's law) and vice versa. A time-varying electric field would set up a time-varying magnetic field with the same

[2] Some authors also call k the *wave number*. However, in spectroscopy, the wave number implies $1/\lambda$, reciprocal wavelength. To avoid any confusion, propagation constant would be preferred for k.

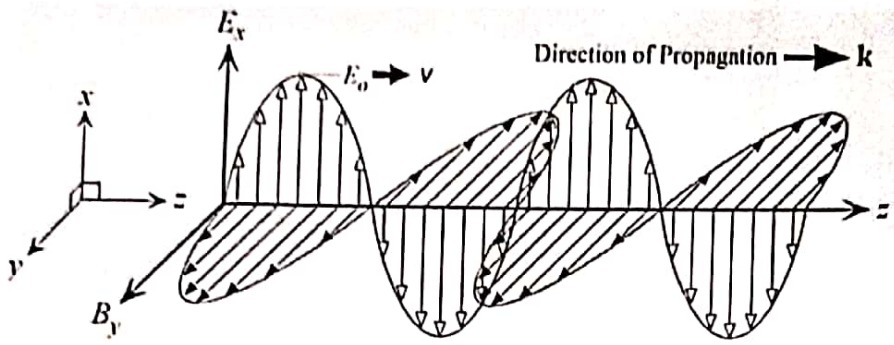


FIGURE 1.1 An electromagnetic wave in a homogenous and isotropic medium is a traveling wave that has time-varying electric and magnetic fields which are perpendicular to each other and the direction of propagation z. This is a snapshot at a given time of a particular harmonic or a sinusoidal EM wave. At a time δt later, a point on the wave, such as the maximum field, would have moved a distance $v\delta t$ in the z-direction.

frequency. According to electromagnetic principles,[3] a traveling electric field E_x as represented by Eq. (1.1.1) would always be accompanied by a traveling magnetic field B_y with the same wave frequency and propagation constant (ω and k) but the directions of the two fields would be orthogonal as in Figure 1.1. Thus, there is a similar traveling wave equation for the magnetic field component B_y. We generally describe the interaction of a light wave with a non-conducting matter (conductivity, $\sigma = 0$) through the electric field component E_x rather than B_y because it is the electric field that displaces the electrons in molecules or ions in the crystal and thereby gives rise to the polarization of matter. However, the two fields are linked, as in Figure 1.1, and there is an intimate relationship between them. The **optical field** refers to the electric field E_x.

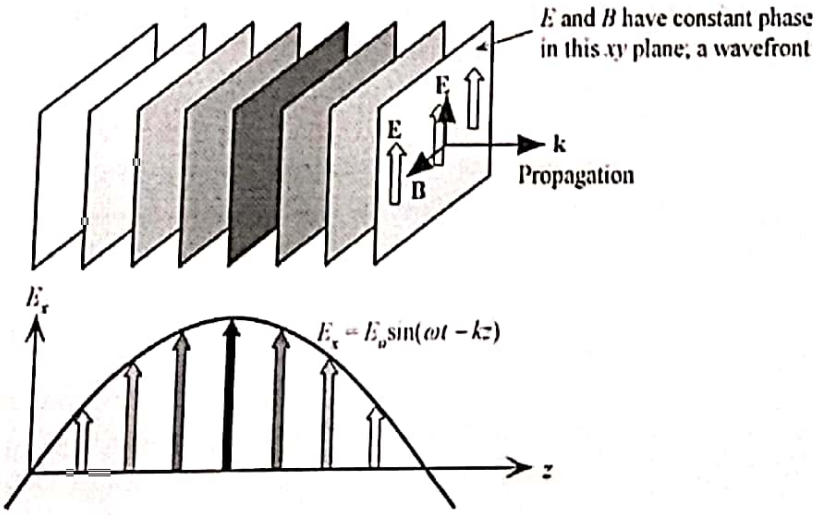


FIGURE 1.2 A plane EM wave traveling along z has the same E_x (or B_y) at any point in a given xy plane. All electric field vectors in a given xy plane are therefore in phase. The xy planes are of infinite extent in the x and y directions.

[3]Maxwell's equations formulate electromagnetic phenomena and provide relationships between the electric and magnetic fields and their space and time derivatives. We need only to use a few selected results from Maxwell's equation without delving into their derivations. The magnetic field B is also called the magnetic induction or magnetic flux density. The magnetic field *intensity* H and magnetic field B in a non-magnetic material are related by $B = \mu_o H$ in which μ_o is the absolute permeability of the medium.

We can also represent a traveling wave using the exponential notation since $\cos\phi = \text{Re}\ \exp(j\phi)$ in which Re refers to the real part. We then need to take the real part of any complex result at the end of calculations. Thus, we can write Eq. (1.1.1) as

$$E_x(z, t) = \text{Re}\ E_o \exp(j\phi_o) \exp j(\omega t - kz)$$

or

$$E_x(z, t) = \text{Re}\ E_c \exp j(\omega t - kz) \tag{1.1.2}$$

Traveling wave along z

in which $E_c = E_o \exp(j\phi_o)$ is a complex number that represents the amplitude of the wave and includes the constant phase information ϕ_o. Note that in Eq. (1.1.2), $\exp j(\omega t - kz)$ represents $e^{j(\omega t - kz)}$.

We indicate the direction of propagation with a vector **k**, called the **wave vector** (or **propagation vector**), whose magnitude is the **propagation constant**, $k = 2\pi/\lambda$. It is clear that **k** is perpendicular to constant phase planes as indicated in Figure 1.2. Consider an electromagnetic wave that is propagating along some arbitrary direction **k**, as indicated in Figure 1.3. The electric field $E(\mathbf{r}, t)$ at an arbitrary point **r** is given by

$$E(\mathbf{r}, t) = E_o \cos(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_o) \tag{1.1.3}$$

Traveling wave in 3D with wave vector k

because the dot product $\mathbf{k}\cdot\mathbf{r}$ is along the direction of propagation similar to kz as indicated in Figure 1.3. The latter can be shown by drawing a plane that has the point **r** and is perpendicular to **k** as illustrated in Figure 1.3. The dot product is the product of **k** and the projection of **r** onto **k**, which is $\mathbf{r}'$ in Figure 1.3, so that $\mathbf{k}\cdot\mathbf{r} = kr'$. Indeed, if propagation is along z, $\mathbf{k}\cdot\mathbf{r}$ becomes kz. In general, if **k** has components k_x, k_y, and k_z along x, y, and z, then from the definition of the dot product, $\mathbf{k}\cdot\mathbf{r} = k_x x + k_y y + k_z z$.

The relationship between time and space for a given phase, ϕ for example, that corresponds to a maximum field, according to Eq. (1.1.1), is described by

$$\phi = \omega t - kz + \phi_o = \text{constant}$$

During a time interval δt, this constant phase (and hence the maximum field) moves a distance δz. The phase velocity of this wave is therefore $\delta z/\delta t$. Thus the **phase velocity** v is

$$v = \frac{dz}{dt} = \frac{\omega}{k} = \upsilon\lambda \tag{1.1.4}$$

Phase velocity

in which υ is the frequency ($\omega = 2\pi\upsilon$) of the EM wave. For an EM wave propagating in free space v is the speed of light in vacuum or c.

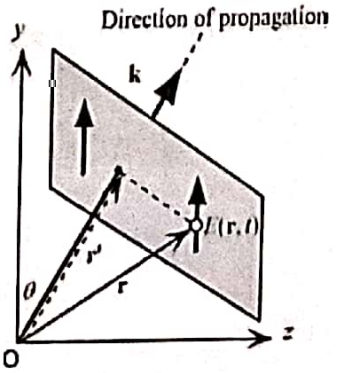


FIGURE 1.3 A traveling plane EM wave along a direction **k**.

We are often interested in the phase difference $\Delta\phi$ at a given time between two points on a wave (Figure 1.1) that are separated by a certain distance. If the wave is traveling along z with a wave vector k, as in Eq. (1.1.1), then the phase difference between two points separated by Δz is simply $k\Delta z$ since ωt is the same for each point. If this phase difference is 0 or multiples of 2π then the two points are in phase. Thus phase difference $\Delta\phi$ can be expressed as $k\Delta z$ or $2\pi\Delta z/\lambda$.

B. Maxwell's Wave Equation and Diverging Waves

Consider the plane EM wave in Figure 1.2. All constant phase surfaces are xy planes that are perpendicular to the z-direction. A cut of a plane wave parallel to the z-axis is shown in Figure 1.4 (a) in which the parallel dashed lines at right angles to the z-direction are wavefronts. We normally show wavefronts that are separated by a phase of 2π or a whole wavelength λ as in the figure. The vector that is normal to a wavefront surface at a point such as P represents the direction of wave propagation (**k**) at that point P. Clearly, the propagation vectors everywhere are all parallel and the plane wave propagates without the wave diverging; *the plane wave has no divergence.* The amplitude of the planar wave E_o does not depend on the distance from a reference point, and it is the same at all points on a given plane perpendicular to **k** (*i.e.*, independent of x and y). Moreover, as these planes extend to infinity there is infinite energy in the plane wave. A plane wave such as the one in Figure 1.4 (a) is an idealization that is useful in analyzing many wave phenomena. In reality, however, the electric field in a plane at right angles to **k** does not extend to infinity since the light beam would have a finite cross-sectional area and finite power. We would need an infinitely large EM source with infinite power to generate a perfect plane wave!

In practice there are many types of possible EM waves. These waves must obey a special wave equation that describes the time and space dependence of the electric field. In an isotropic and linear dielectric medium, the relative permittivity (ε_r) is the same in all directions and is independent of the electric field. The field E in such a medium obeys **Maxwell's EM wave equation**

Maxwell's wave equation

$$\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} + \frac{\partial^2 E}{\partial z^2} - \varepsilon_o \varepsilon_r \mu_o \frac{\partial^2 E}{\partial t^2} = 0 \qquad (1.1.5)$$

in which μ_o is the absolute permeability, ε_o is the absolute permittivity, and ε_r is the relative permittivity of the medium. Equation (1.1.5) assumes an isotropic medium (as discussed later) and

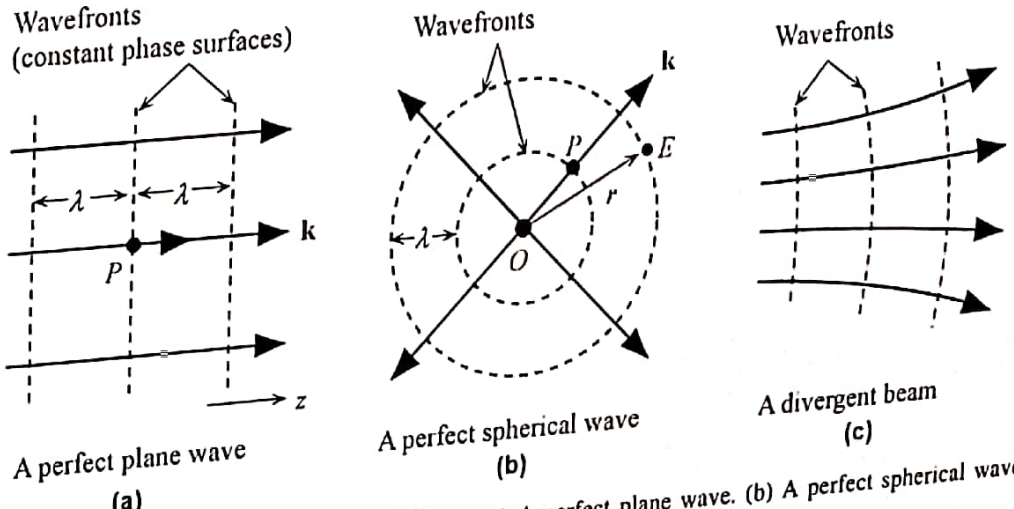


FIGURE 1.4 Examples of possible EM waves. (a) A perfect plane wave. (b) A perfect spherical wave. (c) A divergent beam.

that the conductivity of the medium is zero. To find the time and space dependence of the field, we must solve Eq. (1.1.5) in conjunction with the initial and boundary conditions. We can easily show that the plane wave in Eq. (1.1.1) satisfies Eq. (1.1.5). There are many possible waves that satisfy Eq. (1.1.5) that can therefore exist in nature.

A **spherical wave** is described by a traveling field that emerges from a point EM source and whose amplitude decays with distance r from the source. At any point r from the source, the field is given by

$$E = \frac{A}{r}\cos(\omega t - kr) \tag{1.1.6}$$

Spherical wave

in which A is a constant. We can substitute Eq. (1.1.6) into Eq. (1.1.5) to show that Eq. (1.1.6) is indeed a solution of Maxwell's equation (transformation from Cartesian to spherical coordinates would help). A cut of a spherical wave is illustrated in Figure 1.4 (b) where it can be seen that *wavefronts are spheres* centered at the point source O. The direction of propagation $\mathbf{k}$ at any point such as P is determined by the normal to the wavefront at that point. Clearly $\mathbf{k}$ vectors diverge out and, as the wave propagates, the constant phase surfaces become larger. **Optical divergence** refers to the angular separation of wave vectors on a given wavefront. The spherical wave has 360° of divergence in all planes through the point source. It is apparent that plane and spherical waves represent two extremes of wave propagation behavior from perfectly parallel to fully diverging wave vectors. They are produced by two extreme sizes of EM wave source: an infinitely large source for the plane wave and a point source for the spherical wave. In reality, an EM source is neither of infinite extent nor in point form, but would have a finite size and finite power. Figure 1.4 (c) shows a more practical example in which a light beam exhibits some inevitable divergence while propagating; the wavefronts are slowly bent away thereby spreading the wave. Light rays of geometric optics are drawn to be normal to constant phase surfaces (wavefronts). Light rays therefore follow the wave vector directions. Rays in Figure 1.4 (c) slowly diverge away from each other. The reason for favoring plane waves in many optical explanations is that, at a distance far away from a source, over a *small spatial region*, the wavefronts will appear to be plane even if they are actually spherical. Figure 1.4 (a) may be a small part of a huge spherical wave.

Many light beams, such as the output from a laser, can be described by assuming that they are **Gaussian beams**. Figure 1.5 illustrates a Gaussian beam traveling along the z-axis. The beam still has an $\exp j(\omega t - kz)$ dependence to describe propagation characteristics but the amplitude varies spatially away from the beam axis and also *along* the beam axis. Such a beam has similarities to that in Figure 1.4 (c); it slowly diverges[4] and is the result of radiation from a source of finite extent. The light intensity (*i.e.*, the radiation energy flow per unit area per unit time) distribution across the beam cross-section anywhere along z is Gaussian as shown in Figures 1.5 (b) and (c). The **beam diameter** $2w$ at any point z is defined in such a way that the cross-sectional area πw^2 at that point contains 86% of the beam power. Thus, the beam diameter $2w$ increases as the beam travels along z. The Gaussian beam shown in Figure 1.5 (a) starts from O with a finite width $2w_o$ where the wavefronts are parallel and then the beam slowly diverges as the wavefronts curve out during propagation along z. The finite width $2w_o$ where the wavefronts are parallel is called the **waist** of the beam; w_o is the **waist radius**

[4]The divergence is due to the self-diffraction of the beam—the beam is diffraction limited. Diffraction is covered later in this chapter. Further, the intensity of light will be defined quantitatively in Section 1.4. For the present discussion it represents the radiation energy flow per unit time per unit area.

For stable He-Ne single mode lasers, it is very close to unity. We can apply the above Gaussian equations to a real laser beam by replacing w_o with w_{or}/M^2. For example, the width at a distance z becomes

Beam width at distance z

$$2w = 2w_{or}\left[1 + \left(\frac{z\lambda M^2}{\pi w_{or}}\right)^2\right]^{1/2} \tag{1.1.11}$$

Far away from the Rayleigh range, $2w_r = M^2(2w)$, where $2w$ is the ideal Gaussian beam width at the same location. Suppose that we put a Gaussian beam with a waist w_o onto the real beam and adjust w_o to be the same as w_{or}, *i.e.*, $w_{or} = w_o$. Then, from Eq. (1.1.10) the divergence of the real beam is greater inasmuch as $\theta_r = M^2\theta$, which is shown in Figure 1.6 (b).

EXAMPLE 1.1.1 A diverging laser beam

Consider a He-Ne laser beam at 633 nm with a spot size of 1 mm. Assuming a Gaussian beam, what is the divergence of the beam? What are the Rayleigh range and the beam width at 25 m?

Solution

Using Eq. (1.1.7), we find

$$2\theta = \frac{4\lambda}{\pi(2w_o)} = \frac{4(633 \times 10^{-9}\ \text{m})}{\pi(1 \times 10^{-3}\ \text{m})} = 8.06 \times 10^{-4}\ \text{rad} = 0.046°$$

The Rayleigh range is

$$z_o = \frac{\pi w_o^2}{\lambda} = \frac{\pi\ (1 \times 10^{-3}\ \text{m})/2\ ^2}{(633 \times 10^{-9}\ \text{m})} = 1.24\ \text{m}$$

The beam width at a distance of 25 m is

$$2w = 2w_o\ 1 + (z/z_o)^{2\ 1/2} = (1 \times 10^{-3}\ \text{m})\ 1 + \text{[redacted]}^{\ 2\ 1/2}$$

$$= 0.0202\ \text{m} \quad \text{or} \quad 20\ \text{mm}.$$

1.2 REFRACTIVE INDEX AND DISPERSION

When an EM wave is traveling in a dielectric medium, the oscillating electric field polarizes the molecules of the medium at the frequency of the wave. Indeed, the EM wave propagation can be considered to be the propagation of this polarization in the medium. The field and the induced molecular dipoles become coupled. The relative permittivity ε_r measures the ease with which the medium becomes polarized and hence it indicates the extent of interaction between the field and the induced dipoles. To find the nature of propagation of an EM wave in a dielectric medium, and hence the phase velocity, we have to solve Maxwell's equations in a dielectric medium. If we assume the medium is insulating, nonmagnetic, and also **isotropic**, that is the relative permittivity is independent of the direction of propagation of the EM wave and the optical field, then the solution becomes quite simple and leads to Maxwell's wave equation stated in Eq. (1.1.5). We can continue to represent the EM wave in a similar fashion to its propagation in vacuum but

we need to assign to it a new phase velocity v, and a new wavelength, both of which depend on ε_r. In a dielectric medium of relative permittivity ε_r, the phase velocity v is given by

$$v = \frac{1}{\sqrt{\varepsilon_r \varepsilon_o \mu_o}} \tag{1.2.1}$$

Phase velocity in a medium

It is important to use the relative permittivity at the frequency of operation in Eq. (1.2.1) since ε_r depends on the frequency. Typical frequencies that are involved in optoelectronic devices are in the infrared (including far infrared), visible, and UV, and we generically refer to these frequencies as *optical frequencies*; they cover a somewhat arbitrary range from roughly 10^{12} Hz to 10^{16} Hz.

For an EM wave traveling in free space, $\varepsilon_r = 1$ and $v_{\text{vacuum}} = 1/(\varepsilon_o \mu_o)^{1/2} = c = 3 \times 10^8 \text{m s}^{-1}$, the velocity of light in vacuum. The ratio of the speed of light in free space to its speed in a medium is called the **refractive index** n of the medium, that is,

$$n = \frac{c}{v} = \sqrt{\varepsilon_r} \tag{1.2.2}$$

Definition of refractive index

If k is the propagation constant ($k = 2\pi/\lambda$) and λ is the wavelength, both in free space, then in the medium[7] $k_{\text{medium}} = nk$ and $\lambda_{\text{medium}} = \lambda/n$. Equation (1.2.2) is in agreement with our intuition that light propagates more slowly in a denser medium that has a higher refractive index. We should note that the frequency υ (or ω) remains the same.[8]

The refractive index of a medium is not necessarily the same in all directions. In noncrystalline materials such as glasses and liquids, the material structure is the same in all directions and n does not depend on the direction. The refractive index is then **isotropic**. In crystals, however, the atomic arrangements and interatomic bonding are different along different directions. Crystals, in general, have nonisotropic, or **anisotropic**, properties. Depending on the crystal structure, the relative permittivity ε_r is different along different crystal directions. This means that, in general, the refractive index n seen by a propagating electromagnetic wave in a crystal will depend on the value of ε_r along the direction of the oscillating electric field (*i.e.*, along the direction of polarization). For example, suppose that the wave in Figure 1.1 is traveling along the z-direction in a particular crystal with its electric field oscillating along the x-direction. If the relative permittivity along this x-direction is ε_{rx}, then $n_x = (\varepsilon_{rx})^{1/2}$. The wave therefore propagates with a phase velocity that is c/n_x. The variation of n with direction of propagation and the direction of the electric field depends on the particular crystal structure. With the exception of cubic crystals (such as diamond), all crystals exhibit a degree of optical anisotropy that leads to a number of important applications as discussed in Chapter 6. Typically, noncrystalline solids, such as glasses and liquids, and cubic crystals are **optically isotropic**; they possess only one refractive index for all directions.

Relative permittivity ε_r or the dielectric constant of materials, in general, depends on the frequency of the electromagnetic wave. The relationship $n = (\varepsilon_r)^{1/2}$ between the refractive index n and ε_r must be applied at the same frequency for both n and ε_r. The relative permittivity for many materials can be vastly different at high and low frequencies because different polarization

[7] On occasions, we will need to use k_o and λ_o for the free-space propagation constant and wavelength (as in the next section) and use k and λ for those values inside the medium. In each case, these quantities will be clearly defined to avoid confusion.

[8] We are accustomed to describing light in terms of its wavelength and often quote wavelengths in nm (or Å) in the visible and IR regions. However, there would be certain advantages to using the frequency instead of the wavelength, for example, terahertz (THz) instead of nm, one of which is that the frequency υ does not change in the medium. (See Roger A. Lewis, *Am. J. Phys*, *79*, 341, 2011.)

TABLE 1.1 Low-frequency (LF) relative permittivity ε_r(LF) and refractive index n

Material	ε_r(LF)	$[\varepsilon_r(\text{LF})]^{1/2}$	n (at λ)	Comment
Si	11.9	3.44	3.45 (at 2.15 μm)	Electronic bond polarization up to optical frequencies
Diamond	5.7	2.39	2.41 (at 590 nm)	Electronic bond polarization up to UV light
GaAs	13.1	3.62	3.30 (at 5 μm)	Ionic polarization contributes to ε_r(LF)
SiO_2	3.84	2.00	1.46 (at 600 nm)	Ionic polarization contributes to ε_r(LF)
Water	80	8.9	1.33 (at 600 nm)	Dipolar polarization contributes to ε_r(LF), which is large

mechanisms operate at these frequencies.[9] At low frequencies all polarization mechanisms present can contribute to ε_r, whereas at optical frequencies only the electronic polarization can respond to the oscillating field. Table 1.1 lists the relative permittivity ε_r(LF) at low frequencies (*e.g.*, 60 Hz or 1 kHz as would be measured, for example, using a capacitance bridge in the laboratory) for various materials. It then compares $\varepsilon_r(\text{LF})^{1/2}$ with n.

For silicon and diamond there is an excellent agreement between $\varepsilon_r(\text{LF})^{1/2}$ and n. Both are covalent solids in which electronic polarization (electronic bond polarization) is the only polarization mechanism at low and high frequencies. Electronic polarization involves the displacement of light electrons with respect to positive ions of the crystal. This process can readily respond to the field oscillations up to optical or even ultraviolet frequencies.

For GaAs and SiO_2 $\varepsilon_r(\text{LF})^{1/2}$ is larger than n because at low frequencies both of these solids possess a degree of ionic polarization. The bonding is not totally covalent and there is a degree of ionic bonding that contributes to polarization at frequencies below far-infrared wavelengths.

In the case of water, the ε_r(LF) is dominated by orientational or dipolar polarization, which is far too sluggish to respond to high-frequency oscillations of the field at optical frequencies.

It is instructive to consider what factors affect n. The relative permittivity depends on the polarizability α per molecule (or atom) in the solid. (α is defined as the induced electric dipole moment per unit applied field.) The simplest and approximate expression for the relative permittivity is

$$\varepsilon_r \approx 1 + \frac{N\alpha}{\varepsilon_o}$$

in which N is the number of molecules per unit volume. Both the atomic concentration, or density, and polarizability therefore increase n. For example, glasses of given type but with greater density tend to have higher n.

The frequency or wavelength dependence of ε_r and hence n is called the **dispersion relation**, or simply dispersion. There are various theoretical and empirical models that describe the n vs. λ behavior. The Cauchy dispersion equation in its simplest form is given by[10]

Cauchy short form dispersion equation

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \qquad \textbf{(1.2.3)}$$

[9] Chapters 7 and 9 in *Principles of Electronic Materials and Devices*, 3rd Edition, S. O. Kasap (McGraw-Hill, 2006) provides a semiquantitative description of the frequency dependence of ε_r and hence the wavelength dependence of n.

[10] Dispersion relations like the one in Eq. (1.2.3) are always in terms of the free-space wavelength λ. (It does not make sense to give them in terms of the actual wavelength in the medium.)

TABLE 1.2 Sellmeier and Cauchy coefficients

Sellmeier	A_1	A_2	A_3	λ_1 (μm)	λ_2 (μm)	λ_3 (μm)
SiO_2 (fused silica)	0.696749	0.408218	0.890815	0.0690660	0.115662	9.900559
86.5%SiO_2-13.5%GeO_2	0.711040	0.451885	0.704048	0.0642700	0.129408	9.425478
GeO_2	0.80686642	0.71815848	0.85416831	0.068972606	0.15396605	11.841931
Sapphire	1.023798	1.058264	5.280792	0.0614482	0.110700	17.92656
Diamond	0.3306	4.3356	–	0.1750	0.1060	–

Cauchy	Range of $h\nu$ (eV)	n_{-2} (eV^2)	n_0	n_2 (eV^{-2})	n_4 (eV^{-4})
Diamond	0.05–5.47	-1.07×10^{-5}	2.378	8.01×10^{-3}	1.04×10^{-4}
Silicon	0.002–1.08	-2.04×10^{-8}	3.4189	8.15×10^{-2}	1.25×10^{-2}
Germanium	0.002–0.75	-1.0×10^{-8}	4.003	2.2×10^{-1}	1.4×10^{-1}

Source: Sellmeier coefficients combined from various sources. Cauchy coefficients from D. Y. Smith *et al.*, *J. Phys. CM*, *13*, 3883, 2001.

where A, B, and C are material-specific constants. A more general Cauchy dispersion relation is of the form

Cauchy dispersion equation in photon energy

$$n = n_{-2}(h\nu)^{-2} + n_0 + n_2(h\nu)^2 + n_4(h\nu)^4 \tag{1.2.4}$$

where $h\nu$ is the photon energy, and n_0, n_{-2}, n_2, and n_4 are constants; values for diamond, Si, and Ge are listed in Table 1.2. The general Cauchy equation is usually applicable over a wide photon energy range.

Another useful dispersion relation that has been widely used, especially in optical fibers, is the Sellmeier equation given by

Sellmeier equation

$$n^2 = 1 + \frac{A_1\lambda^2}{\lambda^2 - \lambda_1^2} + \frac{A_2\lambda^2}{\lambda^2 - \lambda_2^2} + \frac{A_3\lambda^2}{\lambda^2 - \lambda_3^2} \tag{1.2.5}$$

where A_1, A_2, A_3 and λ_1, λ_2, λ_3 are constants, called **Sellmeier coefficients.**[11] Equation (1.2.5) turns out to be quite a useful semi-empirical expression for calculating n at various wavelengths if the Sellmeier coefficients are known. Higher terms involving A_4 and higher A coefficients can generally be neglected in representing n vs. λ behavior over typical wavelengths of interest. For example, for diamond, we only need the A_1 and A_2 terms. The Sellmeier coefficients are listed in various optical data handbooks.

EXAMPLE 1.2.1 Sellmeier equation and diamond

Using the Sellmeier coefficients for diamond in Table 1.2, calculate its refractive index at 610 nm (red light) and compare with the experimental quoted value of 2.415 to three decimal places.

[11] This is also known as the Sellmeier-Herzberger formula.

Solution

The Sellmeier dispersion relation for diamond is

$$n^2 = 1 + \frac{0.3306\lambda^2}{\lambda^2 - 175\,\text{nm}^2} + \frac{4.3356\lambda^2}{\lambda^2 - 106\,\text{nm}^2}$$

$$n^2 = 1 + \frac{0.3306(610\,\text{nm})^2}{(610\,\text{nm})^2 - (175\,\text{nm})^2} + \frac{4.3356(610\,\text{nm})^2}{(610\,\text{nm})^2 - (106\,\text{nm})^2} = 5.8308$$

So that

$$n = 2.4147$$

which is 2.415 to three decimal places and matches the experimental value.

EXAMPLE 1.2.2 Cauchy equation and diamond

Using the Cauchy coefficients for diamond in Table 1.2, calculate the refractive index at 610 nm.

Solution

At $\lambda = 610$ nm, the photon energy is

$$h\nu = \frac{hc}{\lambda} = \frac{(6.626 \times 10^{-34}\,\text{Js})(2.998 \times 10^{8}\,\text{ms}^{-1})}{(610 \times 10^{-9}\,\text{m})} \times \frac{1}{1.602 \times 10^{-19}\,\text{JeV}^{-1}} = 2.0325\,\text{eV}$$

Using the Cauchy dispersion relation for diamond with coefficients from Table 1.2,

$$\begin{aligned} n &= n_{-2}(h\nu)^{-2} + n_0 + n_2(h\nu)^2 + n_4(h\nu)^4 \\ &= (-1.07 \times 10^{-5})(2.0325)^{-2} + 2.378 + (8.01 \times 10^{-3})(2.0325)^2 \\ &\quad + (1.04 \times 10^{-4})(2.0325)^4 \\ &= 2.4140 \end{aligned}$$

which is slightly different than the value calculated in Example 1.2.1; one reason for the discrepancy is due to the Cauchy coefficients quoted in Table 1.2 being applicable over a wider wavelength range at the expense of some accuracy. Although both dispersion relations have four parameters, A_1, A_2, λ_1, λ_2 for Sellmeier and n_{-2}, n_0, n_2, n_4 for Cauchy, the functional forms are different.

1.3 GROUP VELOCITY AND GROUP INDEX

Since there are no perfect monochromatic waves in practice, we have to consider the way in which a group of waves differing slightly in wavelength will travel along the z-direction. Figure 1.7 shows how two perfectly harmonic waves of slight different frequencies $\omega - \delta\omega$ and $\omega + \delta\omega$ interfere to generate a periodic **wave packet** that contains an oscillating field at the mean frequency ω that is amplitude modulated by a slowly varying field of frequency $\delta\omega$. We are interested in the velocity of this wave packet. The two sinusoidal waves of frequencies $\omega - \delta\omega$ and $\omega + \delta\omega$ will propagate with propagation constants $k - \delta k$ and $k + \delta k$ respectively inside the material so that their sum will be

$$E_x(z, t) = E_o \cos\,(\omega - \delta\omega)t - (k - \delta k)z \; + E_o \cos\,(k + \delta k)z$$

The Rayleigh range z_o was calculated previously as $z_o = \pi w_o^2/\lambda = 1.24\,\text{m}$ in Example 1.1.1. At $z = 25\,\text{m}$, the axial irradiance is

$$I_{axis} = (1.273 \times 10^4\,\text{W m}^{-2})\frac{(1.24\,\text{m})^2}{(25\,\text{m})^2} = 31.3\,\text{W m}^{-2} = 3.13\,\text{mW cm}^2$$

1.5 SNELL'S LAW AND TOTAL INTERNAL REFLECTION (TIR)

Willebrord Snellius (Willebrord Snel van Royen, 1580–1626) was a Dutch astronomer and a mathematician, who was a professor at the University of Leiden. He discovered his law of refraction in 1621 which was published by Réne Descartes in France 1637; it is not known whether Descartes knew of Snell's law or formulated it independently. *(Courtesy of AIP Emilio Segre Visual Archives, Brittle Books Collection.)*

René Descartes (1596–1650) was a French philosopher who was also involved with mathematics and sciences. He has been called the "Father of Modern Philosophy." Descartes was responsible for the development of Cartesian coordinates and analytical geometry. He also made significant contributions to optics, including reflection and refraction. *(Courtesy of Georgios Kollidas/Shutterstock.com.)*

We consider a traveling plane EM wave in a medium (1) of refractive index n_1 propagating toward a medium (2) with a refractive index n_2. Constant phase fronts are joined with broken lines and the wave vector $\mathbf{k}_i$ is perpendicular to the wavefronts as shown in Figure 1.11. When the wave reaches the plane boundary between the two media, a transmitted wave in medium 2 and a reflected wave in medium 1 appear. The transmitted wave is called the **refracted light.** The angles θ_i, θ_t, θ_r define the directions of the incident, transmitted, and reflected waves, respectively, with respect to the normal to the boundary plane as shown in Figure 1.11. The wave vectors of the reflected and transmitted waves are denoted as $\mathbf{k}_r$ and $\mathbf{k}_t$. Since both the incident and reflected waves are in the same medium, the magnitudes of $\mathbf{k}_r$ and $\mathbf{k}_i$ are the same, $k_r = k_i$.

Simple arguments based on constructive interference can be used to show that there can be only one reflected wave that occurs at an angle equal to the incidence angle. The two waves along A_i and B_i are in phase. When these waves are reflected to become waves A_r and B_r then they must still be in phase, otherwise they will interfere destructively and destroy each other. The only way the two waves can stay in phase is if $\theta_r = \theta_i$. All other angles lead to the waves A_r and B_r being out of phase and interfering destructively.

The refracted waves A_t and B_t are propagating in a medium of refracted index $n_2(<n_1)$ that is different than n_1. Hence the waves A_t and B_t have different velocities than A_i and B_i. We consider what happens to a wavefront such as AB, corresponding perhaps to the maximum field, as it propagates from medium 1 to 2. We recall that the points A and B on this front are always in phase. During the time it takes for the phase B on wave B_i to reach B', phase A on wave A_t has

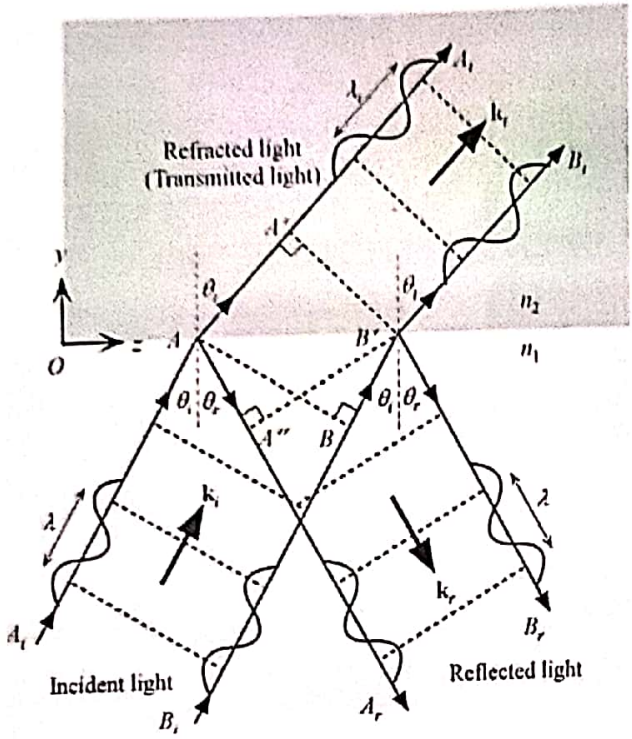


FIGURE 1.11 A light wave traveling in a medium with a greater refractive index ($n_1 > n_2$) suffers reflection and refraction at the boundary. (Notice that λ_t is slightly longer than λ.)

progressed to A'. The wavefront AB thus becomes the front $A'B'$ in medium 2. Unless the two waves at A' and B' still have the same phase, there will be no transmitted wave. A' and B' points on the front are in phase only for one particular transmitted angle, θ_t.

If it takes time t for the phase at B on wave B_i to reach B', then $BB' = v_1 t = ct/n_1$. During this time t, the phase A has progressed to A' where $AA' = v_2 t = ct/n_2$. A' and B' belong to the same front just like A and B so that AB is perpendicular to $\mathbf{k}_i$ in medium 1 and $A'B'$ is perpendicular to $\mathbf{k}_t$ in medium 2. From geometrical considerations, $AB' = BB'/\sin\theta_i$ and $AB' = AA'/\sin\theta_t$ so that

$$AB' = \frac{v_1 t}{\sin\theta_i} = \frac{v_2 t}{\sin\theta_t}$$

or

$$\frac{\sin\theta_i}{\sin\theta_t} = \frac{v_1}{v_2} = \frac{n_2}{n_1} \qquad \textbf{(1.5.1)}$$ *Snell's Law*

This is **Snell's law,**[16] which relates the angles of incidence and refraction to the refractive indices of the media.

If we consider the reflected wave, the wavefront AB becomes $A''B'$ in the reflected wave. In time t, phase B moves to B' and A moves to A''. Since they must still be in phase to constitute the reflected wave, BB' must be equal to AA''. Suppose it takes time t for the wavefront B to move to B' (or A to A''). Then, since $BB' = AA'' = v_1 t$, from geometrical considerations,

$$AB' = \frac{v_1 t}{\sin\theta_i} = \frac{v_1 t}{\sin\theta_r}$$

so that $\theta_i = \theta_r$. Angles of incidence and reflection are the same.

[16] Snell's law is known as Descartes's law in France as he was the first to publish it in his "Discourse on Method" in 1637.

When $n_1 > n_2$ then obviously the transmitted angle is greater than the incidence angle as apparent in Figure 1.11. When the refraction angle θ_t reaches 90°, the incidence angle is called the **critical angle** θ_c, which is given by

Total internal reflection (TIR)

$$\sin\theta_c = \frac{n_2}{n_1} \tag{1.5.2}$$

When the incidence angle θ_i exceeds θ_c then there is no transmitted wave but only a reflected wave. The latter phenomenon is called **total internal reflection (TIR)**. The effect of increasing the incidence angle is shown in Figure 1.12. It is the TIR phenomenon that leads to the propagation of waves in a dielectric medium surrounded by a medium of smaller refractive index as shown in Chapter 2. Although Snell's law for $\theta_i > \theta_c$ shows that $\sin\theta_t > 1$ and hence θ_t is an "imaginary" angle of refraction, there is however a wave called the evanescent wave, whose amplitude decays exponentially with distance into the second medium as discussed below. The wave exists only in the interface region from which the reflected wave emerges (not outside).

Snell's law can also be viewed as the **k**-vector of light parallel to the interface being continuous through the interface, that is, having the same value on both sides of the interface. In medium n_1, $\mathbf{k}_i$ parallel to the interface is $k_i \sin\theta_i$ or $kn_1 \sin\theta_i$, where $k_i = kn_1$, and k is the magnitude of the wave vector in free space. In medium n_2, $\mathbf{k}_t$ parallel to the interface is $k_t \sin\theta_t$ or $kn_2 \sin\theta_t$. If **k**'s component tangential to the interface remains constant, $kn_1 \sin\theta_i = kn_2 \sin\theta_t$, then we obtain Snell's law in Eq. (1.5.1). Put differently, Snell's law is equivalent to

Snell's Law

$$n\sin\theta = \text{constant through an interface between different media} \tag{1.5.3}$$

Snell's law of refraction and TIR play a very important role in many optoelectronic and photonic devices. A prism is a transparent optical component that can deflect a light beam as illustrated in Figure 1.13. There are two basic types of prism. In a **refracting prism**, the light deflection is caused by refractions whereas in a **reflecting prism** it is caused by one or more TIRs. (Some prisms such as composite prisms need both refraction and TIR to achieve their desired deflection.) The deflection δ depends not only on the incidence angle of the light beam on the prism, the prism material (n), and geometry, but also on the wavelength and the polarization state of the incident light. The reason is that the refractive index n of the prism material normally depends on the wavelength, and further, for certain materials (*e.g.*, quartz, calcite), it depends on the polarization state (direction of the electric field) of light as well.

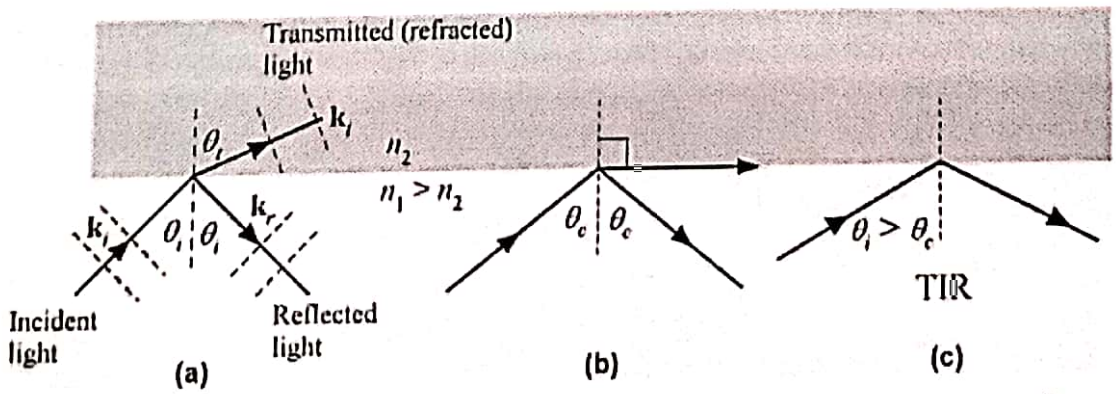


FIGURE 1.12 Light wave travelling in a more dense medium strikes a less dense medium. Depending on the incidence angle with respect to θ_c, which is determined by the ratio of the refractive indices, the wave may be transmitted (refracted) or reflected. (a) $\theta_i < \theta_c$ (b) $\theta_i = \theta_c$ (c) $\theta_i > \theta_c$ and total internal reflection. (Wavefronts are only indicated in (a).)

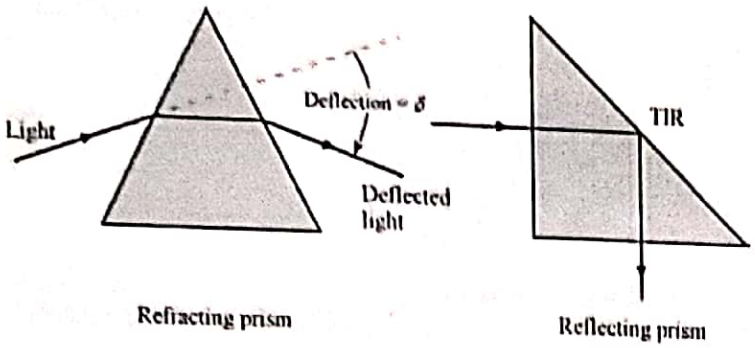


FIGURE 1.13 Basic types of prism: refracting and reflecting prisms.

EXAMPLE 1.5.1 Beam displacement

Lateral displacement of light, or **beam displacement**, occurs when a beam of light passes obliquely through a plate of transparent material, such as a glass plate. When a light beam is incident on a plate of transparent material of refractive index n, it emerges from the other side traveling parallel to the incident light but displaced from it by a distance d, called *lateral displacement*, as illustrated in Figure 1.14. Find the displacement d in terms of the incidence angle the plate thickness. What is d for a glass of $n = 1.600$, $d = 10$ mm if the incidence angle is 45°?

Solution

The displacement $d = BC = AB\sin(\theta_i - \theta_t)$. Further, $L/AB = \cos\theta_t$ so that combining these two equation we find

$$d = L\left[\frac{\sin(\theta_i - \theta_t)}{\cos\theta_t}\right]$$

We can expand $\sin(\theta_i - \theta_t) = \sin\theta_i\cos\theta_t - \cos\theta_i\sin\theta_t$, use $\cos\theta_t = \sqrt{1 - \sin^2\theta_t}$ and then apply Snell's law $n\sin\theta_t = n_o\sin\theta_i$ at the top surface to find

$$\frac{d}{L} = \sin\theta_i\left[1 - \frac{\cos\theta_i}{\sqrt{(n/n_o)^2 - \sin^2\theta_i}}\right]$$

which is maximum with $d = L$ when $\theta_i \simeq 90°$, glazing incidence. Substituting $n = 1.600, n_o = 1, \theta_i = 45°$, and $L = 10$ mm, we find, $d = 3.587$ mm. If the refractive index increases by 1%, $n = 1.616$, then $d = 3.630$ and the change in d is 0.043 mm or 43 μm, which can be measured electronically by using, for example, CCD or CMOS photodiode arrays.

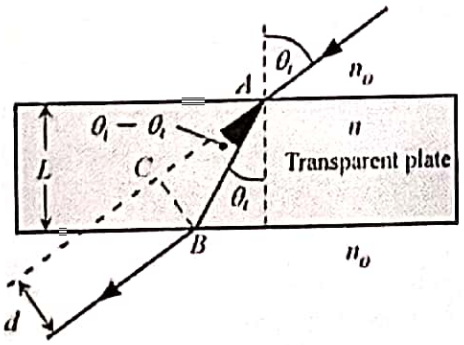


FIGURE 1.14 Lateral displacement of light passing obliquely through a transparent plate.

Solution

The frequency and wavelength are related through $\upsilon = c/\lambda$, so that differentiating the latter we can find the frequency width $\Delta\upsilon$ from the wavelength width $\Delta\lambda$

$$\frac{\Delta\upsilon}{\Delta\lambda} \approx \left|\frac{d\upsilon}{d\lambda}\right| = \left|-\frac{c}{\lambda^2}\right|$$

so that

$$\Delta\upsilon = \Delta\lambda(c/\lambda^2) = (22 \times 10^{-9}\,\text{m})(3 \times 10^{8}\,\text{m s}^{-1})/(650 \times 10^{-9}\,\text{m})^2$$

$$= 1.562 \times 10^{13}\,\text{Hz}$$

Thus, the coherence time is

$$\Delta t \approx 1/\Delta\upsilon = 1/(1.562 \times 10^{13}\,\text{Hz}) = 6.40 \times 10^{-14}\,\text{s} \quad \text{or} \quad 64.0\,\text{fs}$$

The coherence length is

$$l_c = c\Delta t = 1.9 \times 10^{-5}\,\text{m} \quad \text{or} \quad 19\ \text{microns}$$

The above very short coherence length explains why LEDs are not used in interferometry.

1.10 SUPERPOSITION AND INTERFERENCE OF WAVES

Optical interference involves the superposition of two or more electromagnetic waves in which the electric field vectors are added; the fields add vectorially. The waves are assumed to be nearly monochromic, and have to have the same frequency. Two waves can only interfere if they exhibit *mutual temporal coherence* as in Figure 1.29 (a) at a point in space where they interact. Indeed, interference phenomena can be used to infer on the mutual coherence of the waves. When two waves with the same frequency with fields $\mathbf{E}_1$ and $\mathbf{E}_2$ interfere, they generate a resultant field $\mathbf{E}$ that corresponds to the superposition of **individual fields**, that is, $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$. Consider two linearly polarized plane waves that originate from O_1 and O_2, as schematically shown in Figure 1.30, so that the field oscillations at some arbitrary point of interest P is given by

$$\mathbf{E}_1 = \mathbf{E}_{o1}\sin(\omega t - kr_1 - \phi_1) \quad \text{and} \quad \mathbf{E}_2 = \mathbf{E}_{o2}\sin(\omega t - kr_2 - \phi_2) \qquad \textbf{(1.10.1)}$$

where r_1 and r_2 are the distances from O_1 and O_2 to P. These waves have the same ω and k. Due to the process that generates the waves, there is a constant phase difference between them given

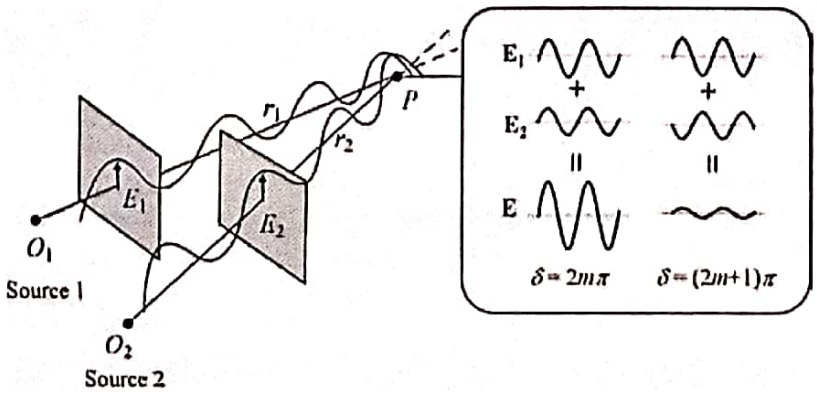


FIGURE 1.30 Interference of two mutually coherent waves of the same frequency originating from sources O_1 and O_2. We examine the resultant at P. The resultant field $\mathbf{E}$ depends on the phase angle δ which depends on the optical path difference $k(r_2 - r_1)$.

by $\phi_2 - \phi_1$. The resultant field at P will be the sum of these two waves, that is, $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$. Its irradiance depends on the time average of $\mathbf{E} \cdot \mathbf{E}$, that is, $\overline{\mathbf{E} \cdot \mathbf{E}}$, so that

$$\overline{\mathbf{E} \cdot \mathbf{E}} = \overline{(\mathbf{E}_1 + \mathbf{E}_2) \cdot (\mathbf{E}_1 + \mathbf{E}_2)} = \overline{\mathbf{E}_1^2} + \overline{\mathbf{E}_2^2} + 2\overline{\mathbf{E}_1 \cdot \mathbf{E}_2}$$

It is clear that the interference effect is in the $2\overline{\mathbf{E}_1 \cdot \mathbf{E}_2}$ term. We can simply the above equation a little further by assuming $\mathbf{E}_{o1}$ and $\mathbf{E}_{o2}$ are parallel with magnitudes E_{o1} and E_{o2}. Further, irradiance of the interfering waves are $I_1 = \frac{1}{2}c\varepsilon_o E_{o1}^2$ and $I_2 = \frac{1}{2}c\varepsilon_o E_{o2}^2$ so that the resultant irradiance is given by the sum of individual irradiances, I_1 and I_2, and has an additional third term I_{21}, that is,

Interference

$$I = I_1 + I_2 + 2(I_1 I_2)^{1/2}\cos\delta \tag{1.10.2}$$

Interference for mutually coherent beams

where the last term is usually written as $2(I_1 I_2)^{1/2}\cos\delta = I_{21}$, and δ is a phase difference given by

$$\delta = k(r_2 - r_1) + (\phi_2 - \phi_1) \tag{1.10.3}$$

Since we are using nearly monochromatic waves, $(\phi_2 - \phi_1)$ is constant, and the interference therefore depends on the term $k(r_2 - r_1)$, which represents the phase difference between the two waves as a result of the *optical path difference* between the waves. As we move point P, $k(r_2 - r_1)$ will change because the optical path difference between the two waves will change; and the interference will therefore also change.

Suppose $(\phi_2 - \phi_1) = 0$, the two waves are emitted from a spatially coherent source. Then, if the path difference $k(r_2 - r_1)$ is 0, 2π or a multiple of 2π, that is, $2m\pi$, $m = 0, \pm 1, \pm 2 \ldots$, then the interference intensity I will be maximum; such interference is defined as **constructive interference**. If the path difference $k(r_2 - r_1)$ is π or 3π or an odd multiple of π, $(2m + 1)\pi$, then the waves will be 180° out of phase, and the interference intensity will be minimum; such interference is defined as **destructive interference**; both constructive and destructive intensity are shown in Figure 1.30. The maximum and minimum irradiances are given by

Constructive and destructive interference

$$I_{\max} = I_1 + I_2 + 2(I_1 I_2)^{1/2} \quad \text{and} \quad I_{\min} = I_1 + I_2 - 2(I_1 I_2)^{1/2} \tag{1.10.4}$$

If the interfering beams have equal irradiances, then $I_{\max} = 4I_1$ and $I_{\min} = 0$.

It is important to emphasize that we have considered the interference of two nearly monochromatic waves that exhibited mutual temporal and spatial coherence. If the waves do not have any mutual coherence, that is, they are incoherent, then we cannot simply superimpose the electric fields as we did above, and the detector at P, which averages the measurement over its response time, will register an irradiance that is simply the sum of individual irradiances,

Irradiance of two superimposed incoherent beams

$$I = I_1 + I_2 \tag{1.10.5}$$

One of the most described interference experiments is **Young's two slit experiment** that generates an interference fringe. In the modern version of this, a coherent beam of light, as available from a laser, is incident on two parallel slits S_1 and S_2. There is a screen far away from the slits on which the waves emanating from the slits interfere, as shown in Figure 1.31. The result is an interference pattern that is composed of light and dark regions, corresponding to $I_{\max}$ and $I_{\min}$. Since S_1 and S_2 are excited by the same wavefront, they emit coherent waves, and we can take $\phi_2 - \phi_1 = 0$. Consider a point P at a distance y on the screen. The phase difference δ at P is then $k(r_2 - r_1)$. As we move P along y, $\delta = k(r_2 - r_1)$ changes and the irradiance on the screen goes through minima and maxima with distance y, following

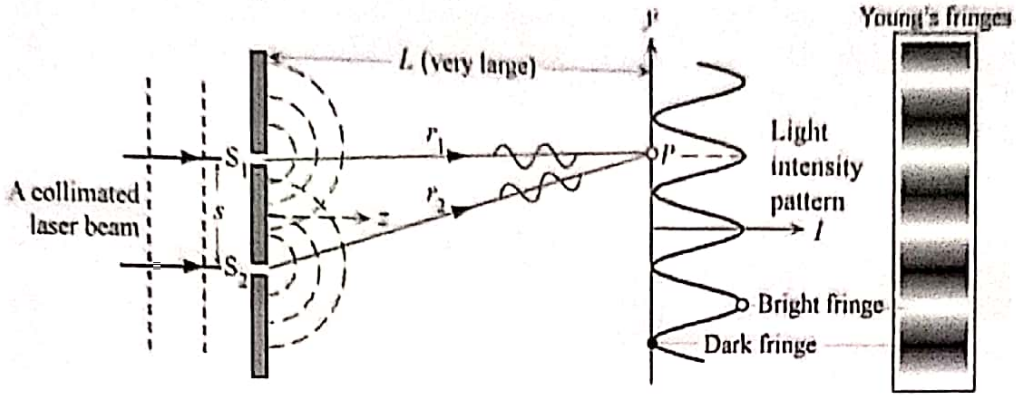


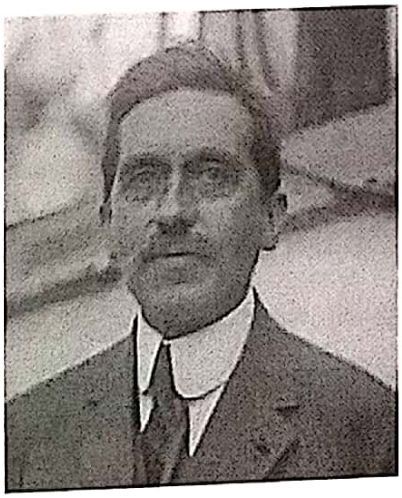
FIGURE 1.31 Young's two slit experiment. Slits S_1 and S_2, separated by s are illuminated at the same time by coherent (nearly monochromatic) collimated laser beam. The irradiance at the screen shows bright and dark fringes due to the interference of waves emanating from the two slits. The screen is assumed to be far away at a distance L from the slits ($L >> s$).

Eq. (1.10.2), generating dark and light bands or fringes. The maxima occur when $\delta = 2m\pi$, and minima when $\delta = (2m + 1)\pi$, where $m = 0, \pm 1, \pm 2$. It is not difficult to show that, if the screen is far away, $(r_2 - r_1) \simeq (s/L)y$ in which L is the distance from the slits to the screen, and s is the separation of the slits, so that δ is proportional to y. Assuming equal irradiances are emitted from S_1 and S_2, $I_1 + I_2 = I_o$, the brightness on the screen changes with y periodically as

$$I = I_o \ 1 + \cos \ (s/L)ky \qquad \textbf{(1.10.6)}$$ Interference

The resulting periodic interference pattern on the screen, as shown in Figure 1.31, is often called **Young's interference fringes**. In a better treatment one needs to consider not only the coherence length of the waves from S_1 and S_2, but also the diffraction that takes place at each slit due to the finite width of the slit. (See diffraction in Section 1.12.)

1.11 MULTIPLE INTERFERENCE AND OPTICAL RESONATORS



Charles Fabry (1867–1945), left, and Alfred Perot (1863–1925), right, were the first French physicists to construct an optical cavity for interferometry. *(Perot:* The Astrophysical Journal, *Vol. 64, November 1926, p. 208, courtesy of the American Astronomical Society. Fabry: Courtesy of Library of Congress Prints and Photographs Division, Washington, DC 20540, USA.)*

Solution

The wavelength of radiation inside the cavity is λ/n, where n is the refractive index of the medium, which is 3.6, and λ is the free-space wavelength. We need to use Eq. (1.11.8), so that the mode number for the wavelength 1310 nm is

$$m = \frac{2nL}{\lambda} = \frac{2(3.6)(250 \times 10^{-6})}{(1310 \times 10^{-9})} = 1374.05$$

which must be an integer (1374) so that the actual mode wavelength is

$$\lambda_m = \frac{2nL}{m} = \frac{2(3.6)(250 \times 10^{-6})}{(1374)} = 1310.04 \text{ nm}$$

Thus, for all practical purposes the mode wavelength is 1310 nm.

The mode frequency is $\upsilon_m = c/\lambda_m$ so that the separation of the modes, from Eq. (1.11.9), is

$$\Delta\upsilon_m = \upsilon_f = \frac{c}{2nL} = \frac{(3 \times 10^8)}{2(3.6)(250 \times 10^{-6})} = 1.67 \times 10^{11} \text{ Hz} \quad \text{or} \quad 167 \text{ GHz}$$

The finesse is

$$F = \frac{\pi R^{1/2}}{1 - R} = \frac{\pi 0.90^{1/2}}{1 - 0.90} = 29.8$$

and the spectral width of each mode is

$$\delta\upsilon_m = \frac{\upsilon_f}{F} = \frac{1.67 \times 10^{11}}{29.8} = 5.59 \times 10^9 \text{ Hz} \quad \text{or} \quad 5.59 \text{ GHz}$$

The mode spectral width $\delta\upsilon_m$ will correspond to a certain spectral wavelength width $\delta\lambda_m$. The mode wavelength $\lambda_m = 1310$ nm corresponds to a mode frequency $\upsilon_m = c/\lambda_m = 2.29 \times 10^{14}$ Hz. Since $\lambda_m = c/\upsilon_m$, we can differentiate this expression to relate small changes in λ_m and υ_m,

$$\delta\lambda_m = \delta\left(\frac{c}{\upsilon_m}\right) = \left|-\frac{c}{\upsilon_m^2}\right|\delta\upsilon_m = \frac{(3 \times 10^8)}{(2.29 \times 10^{14})^2}(5.59 \times 10^9) = 3.2 \times 10^{-11} \text{ m} \quad \text{or} \quad 0.032 \text{ nm}$$

The Q-factor is

$$Q = mF = (1374)(29.8) = 4.1 \times 10^4$$

An optical cavity like this is used in so-called Fabry–Perot semiconductor laser diodes, and will be discussed further in Chapter 4.

1.12 DIFFRACTION PRINCIPLES

A. Fraunhofer Diffraction

An important property of waves is that they exhibit diffraction effects; for example, sound waves are able to bend (deflect around) corners and a light beam can similarly "bend" around an obstruction (though the bending may be very small). Figure 1.34 shows an example of a collimated light beam passing through a circular aperture (a circular opening in an opaque screen). The passing beam is found to be divergent and to exhibit an intensity pattern that has bright and dark rings, called *Airy rings*.[32] The passing beam is said to be diffracted and its light intensity pattern is called

[32] Sir George Airy (1801–1892), Astronomer Royal of Britain from 1835 to1881.

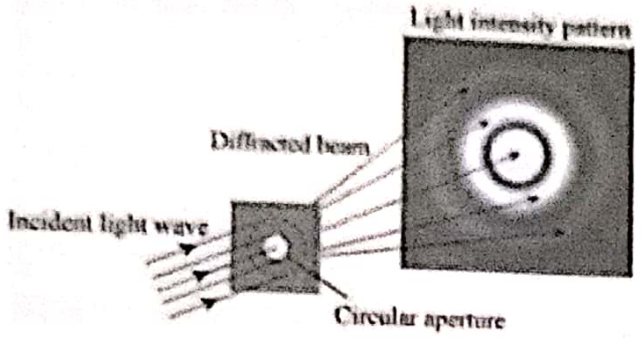


FIGURE 1.34 A collimated light beam incident on a small circular aperture becomes diffracted and its light intensity pattern after passing through the aperture is a diffraction pattern with circular bright rings (called Airy rings). If the screen is far away from the aperture, this would be a Fraunhofer diffraction pattern.

a **diffraction pattern**. Clearly, the light pattern of the diffracted beam does *not* correspond to the geometric shadow of the circular aperture. Diffraction phenomena are generally classified into two categories. In **Fraunhofer diffraction,**[33] the incident light beam is a plane wave (a collimated light beam) and the observation or detection of the light intensity pattern (by placing a photographic screen, *etc.*) is done far away from the aperture so that the waves received also look like plane waves. Inserting a lens between the aperture and the photographic screen enables the screen to be closer to the aperture. In **Fresnel diffraction,** the incident light beam and the received light waves are not plane waves but have significant wavefront curvatures. Typically, the light source and the photographic screen are both close to the aperture so that the wavefronts are curved. Fraunhofer diffraction is by far the most important; and it is mathematically easier to treat.

Diffraction can be understood in terms of the interference of multiple waves emanating from the aperture in the obstruction.[34] We will consider a plane wave incident on a one-dimensional slit of length a. According to the Huygens-Fresnel principle,[35] *every unobstructed point of a wavefront, at a given instant in time, serves as a source of spherical secondary waves (with the same frequency as that of the primary wave). The amplitude of the optical field at any point beyond is the superposition of all these wavelets (considering their amplitudes and relative phases).* Figures 1.35 (a) and (b) illustrates this point pictorially showing that, when the plane wave reaches the aperture, points in the aperture become sources of coherent spherical secondary waves. These spherical waves interfere to constitute the new wavefront (the new wavefront is the envelope of the wavefronts of these secondary waves). These spherical waves can interfere constructively not just in the forward direction as in (a) but also in other appropriate directions, as in (b), giving rise to the observed bright and dark patterns (rings for a circular aperture) on the observation screen.

We can divide the unobstructed width a of the aperture into a very large number N of coherent "point sources" each of extent $\delta y = a/N$ (obviously δy is sufficiently small to be nearly a point), as in Figure 1.36 (a). Since the aperture a is illuminated uniformly by the plane wave, the strength (amplitude) of each point source would be proportional to $a/N = \delta y$. Each would be a source of spherical waves. In the forward direction ($\theta = 0$), they would all be in phase and constitute a forward wave, along the z-direction. But they can also be in phase at some angle θ to the z-direction and hence give rise to a diffracted wave along this direction. We will evaluate the intensity of the received wave at a point on the screen as the *sum of all*

[33] Joseph von Fraunhofer (1787–1826) was a German physicist who also observed the various dark lines in Sun's spectrum due to hydrogen absorption.

[34] "No one has been able to define the difference between interference and diffraction satisfactorily" [R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics* (Addison-Wesley, 1963)].

[35] Eugene Hecht, *Optics*, 4th Edition (Pearson Education, 2002), Ch. 10, p. 444.

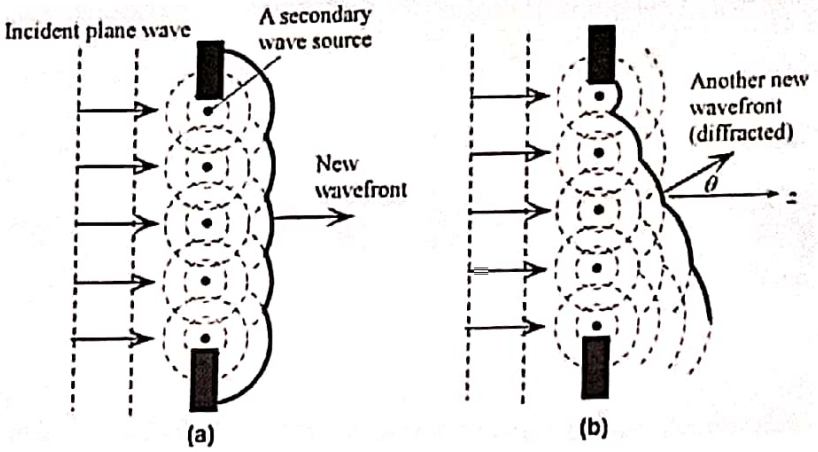


FIGURE 1.35 (a) Huygens-Fresnel principle states that each point in the aperture becomes a source of secondary waves (spherical waves). The spherical wavefronts are separated by λ. The new wavefront is the envelope of all these spherical wavefronts. (b) Another possible wavefront occurs at an angle θ to the z-direction, which is a diffracted wave.

waves arriving from all point sources in the aperture. The screen is far away from the aperture so the waves arrive almost parallel at the screen (alternatively a lens can be used to focus the diffracted parallel rays to form the diffraction pattern).

Consider an arbitrary direction θ, and consider the phase of the emitted wave (Y) from an arbitrary point source at y with respect to the wave (A) emitted from source at $y = 0$ as shown in Figure 1.36 (a). If k is the propagation constant, $k = 2\pi/\lambda$, the wave Y is out of phase with respect to A by $ky\sin\theta$. Thus the wave emitted from the point source at y has a field δE,

$$\delta E \propto (\delta y)\exp(-jky\sin\theta) \tag{1.12.1}$$

All of these waves from point sources from $y = 0$ to $y = a$ interfere at the screen at a point P that makes an angle θ at the slit, and the resultant field at the screen at this point P is

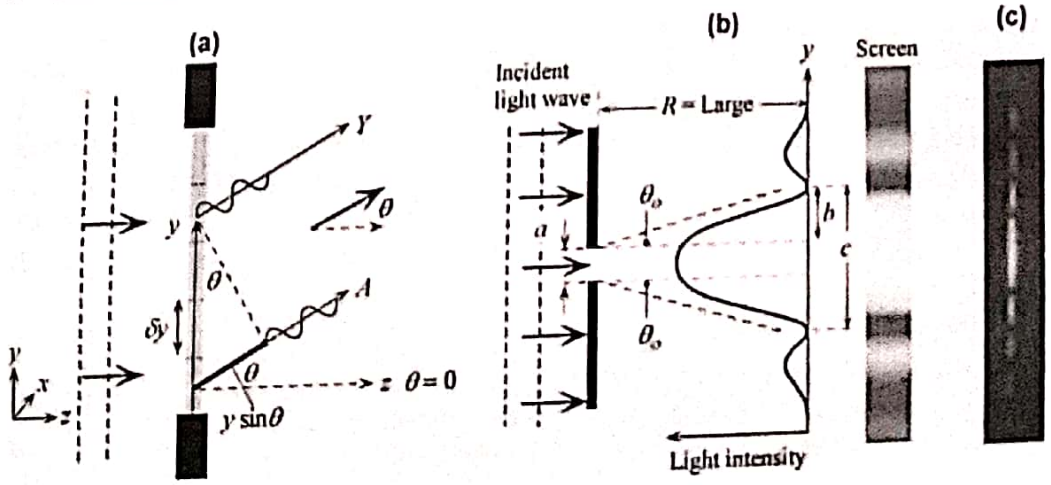


FIGURE 1.36 (a) The aperture has a finite width a along y, but it is very long along x so that it is a one-dimensional slit. The aperture is divided into N number of point sources each occupying δy with amplitude proportional to δy since the slit is excited by a plane electromagnetic wave. (b) The intensity distribution in the received light at the screen far away from the aperture: *the diffraction pattern.* Note that the slit is very long along x so that there is no diffraction along this dimension. The incident wave illuminates the whole slit. (c) Typical diffraction pattern using a laser pointer on a single slit. The difference from the pattern in (b) is due to the finite size of the laser pointer beam along x that is smaller than the length of the slit.

their sum. Because the screen is far away, a point on the screen is at the same distance from anywhere in the aperture. This means that all the spherical waves from the aperture experience the same phase change and decrease in amplitude in reaching the screen. This simply scales δE at the screen by an amount that is the same for all waves coming from the aperture. Thus, the resultant field $E(\theta)$ at point P at the screen is

$$E(\theta) = C\int_{y=0}^{y=a} \delta y \exp(-jky\sin\theta) \tag{1.12.2}$$

in which C is a constant. Integrating Eq. (1.12.2) we get

$$E(\theta) = \frac{Ce^{-j\frac{1}{2}ka\sin\theta}\, a\sin\ \frac{1}{2}ka\sin\theta}{\frac{1}{2}ka\sin\theta}$$

The light intensity I at a point at P at the screen is proportional to $|E_\theta|^2$, and thus

$$I(\theta) = \left[\frac{C'a\sin\ \frac{1}{2}ka\sin\theta}{\frac{1}{2}ka\sin\theta}\right]^2 = I(0)\,\mathrm{sinc}^2(\beta);\quad \beta = \tfrac{1}{2}(ka\sin\theta) \tag{1.12.3}$$

Single slit diffraction equation

in which C' is a constant and β is a convenient new variable representing θ, and sinc ("sink") is a function that is defined by $\mathrm{sinc}\,(\beta) = \sin(\beta)/(\beta)$.

If we were to plot Eq. (1.12.3) as a function of θ at the screen we would see the intensity (diffraction) pattern schematically depicted in Figure 1.36 (b). First, observe that the pattern has bright and dark regions, corresponding to constructive and destructive interference of waves emanating from the aperture. Second, the center bright region is wider than the aperture width a, which mean that the transmitted beam must be *diverging*. The zero intensity occurs when, from Eq. (1.12.3),

$$\sin\theta = \frac{m\lambda}{a};\quad m = \pm 1, \pm 2, \ldots \tag{1.12.4}$$

Zero intensity points

The angle θ_o for the first zero, corresponding to $m = \pm 1$, is given by $\theta_o = \pm\lambda/a$, where we assumed that the divergence is small (usually the case) so that $\sin\theta_o \approx \theta_o$. Thus, the divergence $\Delta\theta$, the angular spread, of the diffracted beam is given by

$$\Delta\theta = 2\theta_o \approx \frac{2\lambda}{a} \tag{1.12.5}$$

Divergence from single slit of width a

A light wave at a wavelength 1300 nm, diffracted by a slit of width $a = 100$ μm (about the thickness of this page), has a divergence $\Delta\theta$ of about 1.5°. From Figure 1.36 (b), it is apparent that, using geometry, we can easily calculate the width c of the central bright region of the intensity pattern, given θ_o from Eq. (1.12.5) and the distance R of the screen from the aperture.

The diffraction patterns from two-dimensional apertures such as rectangular and circular apertures are more complicated to calculate but they use the same principle based on the multiple interference of waves emitted from all point sources in the aperture. The diffraction pattern of a rectangular aperture is shown in Figure 1.37. It involves the multiplication of two individual single slit (sinc) functions, one slit of width a along the horizontal axis, and the other of width b along the vertical axis. (Why is the diffraction pattern wider along the horizontal axis?)

The diffraction pattern from a circular aperture, known as **Airy rings**, was shown in Figure 1.34, and can be roughly visualized by rotating the intensity pattern in Figure 1.36 (b)

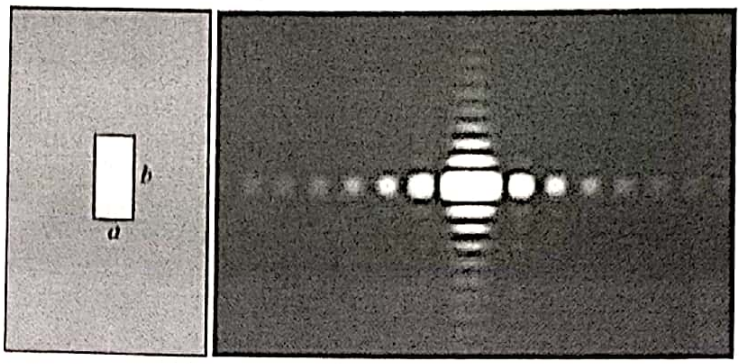


FIGURE 1.37 The rectangular aperture of dimensions $a \times b$ on the left gives the diffraction pattern on the right (b is twice a).

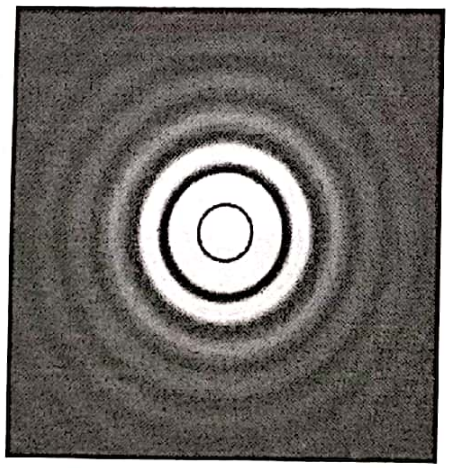
about the z-axis. We can, as we did for the single slit, sum all waves emanating from every point in the circular aperture, taking into account their relative phases when they arrive at the screen to obtain the actual intensity pattern at the screen. The result is that the diffraction pattern is a Bessel function of the first kind,[36] and not a simply rotated sinc function. The central white spot is called the **Airy disk**; its radius corresponds to the radius of the first dark ring. We can still use Figures 1.36 (a) and (b) to imagine how diffraction occurs from a circular aperture by taking this as a cut through the aperture so that a is now the diameter of the aperture, denoted as D. The angular position θ_o of the first dark ring, as defined as in Figure 1.36 (b), is determined by the diameter D of the aperture and the wavelength λ, and is given by

Angular radius of Airy disk

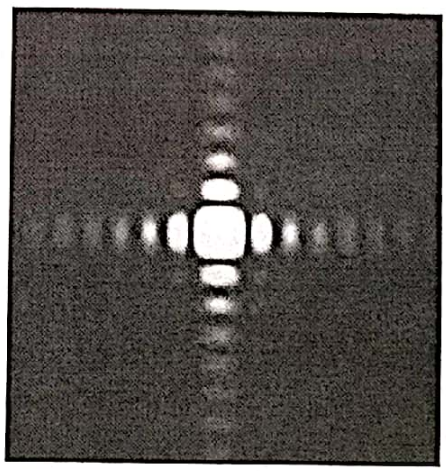
$$\sin\theta_o = 1.22\frac{\lambda}{D} \tag{1.12.6}$$

The divergence angle from the aperture center to the Airy disk circumference is $2\theta_o$. If R is the distance of the screen from the aperture, then the radius of the Airy disk, approximately b, can be calculated from the geometry in Figure 1.36 (b), which gives $b/R = \tan\theta_o \approx \theta_o$. If a lens is used to focus the diffracted light waves onto a screen, then $R = f$, focal length of the lens.

It is worth commenting on the Gaussian beam at this point. Suppose we now examine a Gaussian beam with a waist $2w_o$ that is the same as the aperture size D. The far field



Diffraction pattern far away from a circular aperture.



Diffraction pattern far away from a square aperture.

[36] Bessel functions are special mathematical functions, which can be looked up in mathematics handbooks. They are used in various engineering problems.

half-divergence angle θ of this Gaussian beam would be $\theta = (2/\pi)(\lambda/D)$ or $0.637(\lambda/D)$ as in Eq. (1.1.7). The Gaussian beam has a smaller divergence than the diffracted beam from a circular aperture. The difference is due the fact that each point in the circular aperture emits with the same intensity because the aperture is illuminated by a plane wave. If we were to change the emission intensity within the aperture to follow a Gaussian distribution, we would see a Gaussian beam as the "diffracted beam." The Gaussian beam is a self-diffracted beam and has the smallest divergence for a given beam diameter.

EXAMPLE 1.12.1 Resolving power of imaging systems

Consider what happens when two neighboring point light sources are examined through an imaging system with an aperture of diameter D (this may even be a lens). The two sources have an angular separation of $\Delta\theta$ at the aperture. The aperture produces a diffraction pattern of the sources S_1 and S_2, as shown in Figure 1.38. As the points get closer, their angular separation becomes narrower and the diffraction patterns overlap more. According to the **Rayleigh criterion**, the two spots are just resolvable when the principal maximum of one diffraction pattern coincides with the minimum of the other, which is given by the condition

$$\sin(\Delta\theta_{\min}) = 1.22\frac{\lambda}{D} \tag{1.12.7}$$

Angular limit of resolution

The human eye has a pupil diameter of about 2 mm. What would be the minimum angular separation of two points under a green light of 550 nm and their minimum separation if the two objects are 30 cm from the eye? The image will be a diffraction pattern in the eye, and is a result of waves in this medium. If the refractive index $n \approx 1.33$ (water) in the eye, then Eq. (1.12.7) is

$$\sin(\Delta\theta_{\min}) = 1.22\frac{\lambda}{nD} = 1.22\frac{(550 \times 10^{-9}\ \text{m})}{(1.33)(2 \times 10^{-3}\ \text{m})}$$

giving

$$\Delta\theta_{\min} = 0.0145°$$

Their minimum separation s would be

$$s = 2L\tan(\Delta\theta_{\min}/2) = 2(300\ \text{mm})\tan(0.0145°/2) = 0.076\ \text{mm} = 76\ \mu\text{m}$$

which is about the thickness of a human hair (or this page).

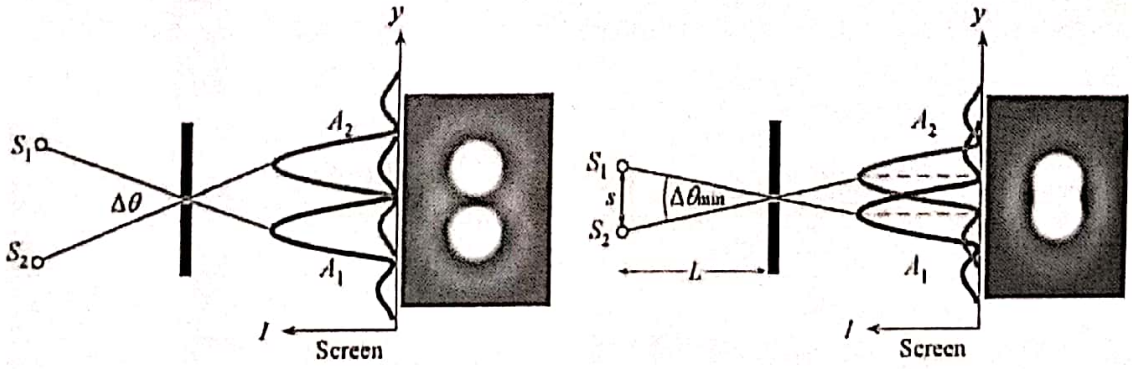


FIGURE 1.38 Resolution of imaging systems is limited by diffraction effects. As points S_1 and S_2 get closer, eventually the Airy patterns overlap so much that the resolution is lost. The Rayleigh criterion allows the minimum angular separation of two of the point sources to be determined. (Schematic illustration inasmuch as the side lobes are actually much smaller than the center peak.)

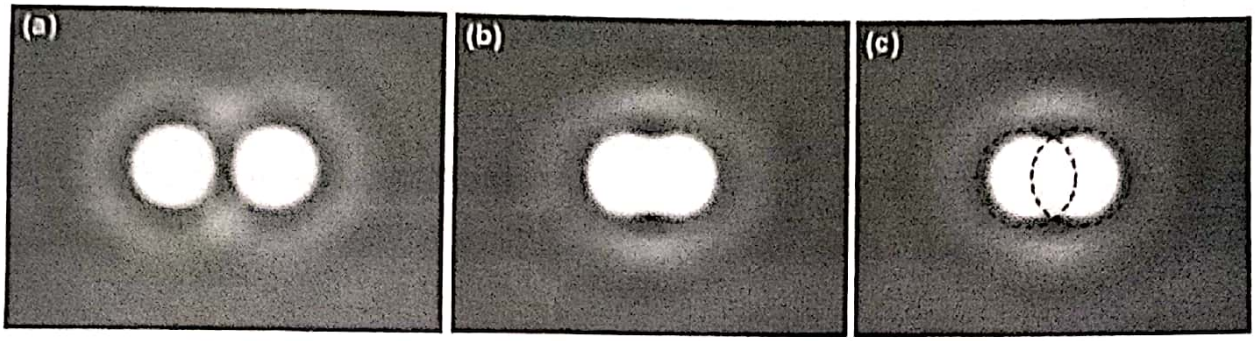


Image of two-point sources captured through a small circular aperture. (a) The two points are fully resolved since the diffraction patterns of the two sources are sufficiently separated. (b) The two images are near the Rayleigh limit of resolution. (c) The first dark ring of one image passes through the center of the bright Airy disk of the other. (Approximate.)

B. Diffraction Grating

A **diffraction grating** in its simplest form is an optical device that has a periodic series of slits in an opaque screen as shown in Figure 1.39 (a). An incident beam of light is diffracted in certain well-defined directions that depend on the wavelength λ and the grating properties. Figure 1.39 (b) shows a typical intensity pattern in the diffracted beam for a finite number of slits. There are "strong beams of diffracted light" along certain directions (θ) and these are labeled according to their occurrence: zero-order (center), first-order, either side of the zero-order, and so on. If there are an infinite number of slits then the diffracted beams have the same intensity. In reality, any periodic variation in the refractive index would serve as a diffraction grating and we will discuss other types later. As in Fraunhofer diffraction we will assume that the observation screen is far away, or that a lens is used to focus the diffracted parallel rays on to the screen (the lens in the observer's eye does it naturally).

We will assume that the incident beam is a plane wave so that the slits become coherent (synchronous) sources. Suppose that the width a of each slit is much smaller than the separation

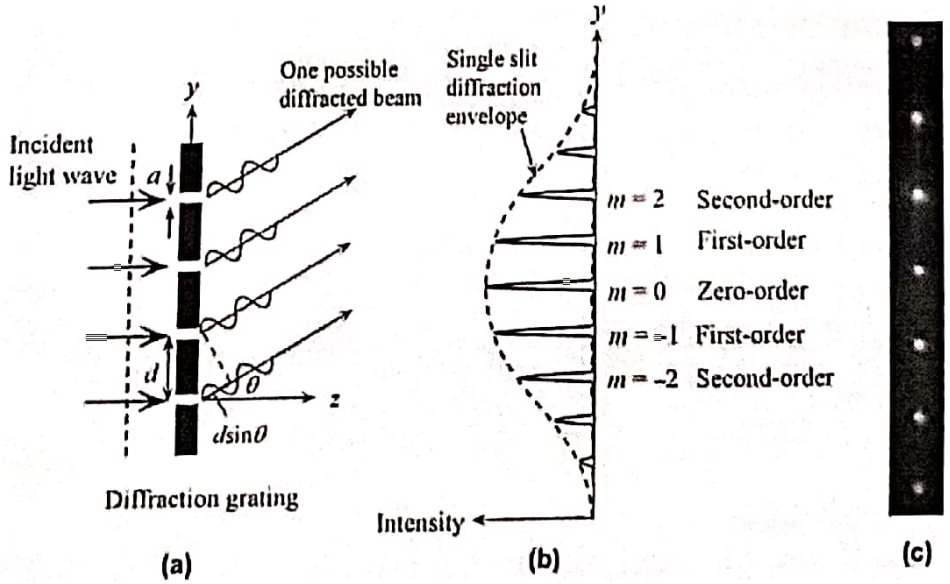


FIGURE 1.39 (a) A diffraction grating with N slits in an opaque screen. Slit periodicity is d and slit width is a; $a \ll d$. (b) The diffracted light pattern. There are distinct, that is diffracted, beams in certain directions (schematic). (c) Diffraction pattern obtained by shining a beam from a red laser pointer onto a diffraction grating. The finite size of the laser beam results in the dot pattern. (The wavelength was 670 nm, red, and the grating has 200 lines per inch.)

d of the slits as shown in Figure 1.39 (a). Waves emanating at an angle θ from two neighboring slits are out of phase by an amount that corresponds to an optical path difference $d\sin\theta$. Obviously, all such waves from pairs of slits will interfere constructively when this is a multiple of the whole wavelength

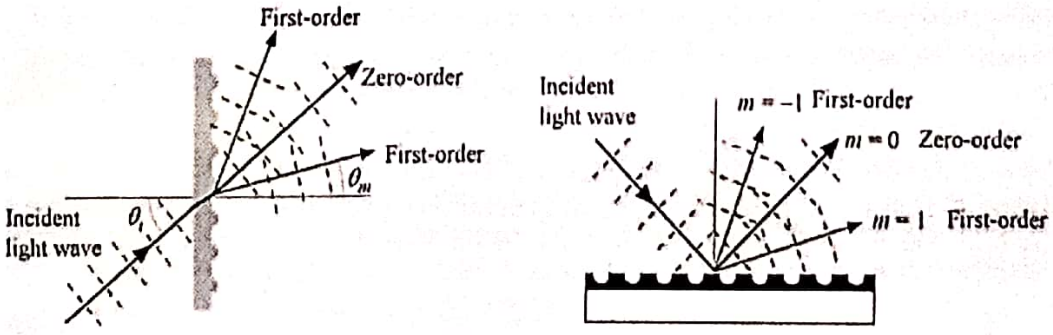
$$d\sin\theta = m\lambda; \quad m = 0, \pm 1, \pm 2, \ldots \tag{1.12.8}$$

Grating equation

which is the well-know **grating equation**, also known as the **Bragg**[37] **diffraction condition**. The value of m defines the diffraction order; $m = 0$ being zero-order, $m = \pm 1$ being first-order, *etc.* If the grating in Figure 1.39 (a) is in a medium of refractive index n, that is, the incident and diffracted beams are all in the same medium of index n, then we should use λ/n for the wavelength in Eq. (1.12.8), where λ is the free-space wavelength, that is, $d\sin\theta = m\lambda/n$.

The problem of determining the actual intensity of the diffracted beam is more complicated as it involves summing all such waves at the observer and, at the same time, including the diffraction effect of each individual narrow slit. With a smaller than d as in the Figure 1.39 (a), the amplitude of the diffracted beam is modulated by the diffraction amplitude of a single slit since the latter is spread substantially, as illustrated in Figure 1.39 (b). It is apparent that the diffraction grating provides a means of deflecting an incoming light by an amount that depends on its wavelength—the reason for their use in *spectroscopy*.

The diffraction grating in Figure 1.40 (a) is a **transmission grating**. The incident and diffracted beams are on opposite sides of the grating. Typically, parallel thin grooves on a glass plate would serve as a transmission grating as in Figure 1.40 (a). A **reflection grating** has the incident beam and the diffracted beams on the same side of the device as in Figure 1.40 (b). The surface of the device has a periodic reflecting structure, easily formed by etching parallel grooves in a metal film, *etc.* The reflecting unetched surfaces serve as synchronous secondary sources that interfere along certain directions to give diffracted beams of zero-order, first-order, *etc.* Among transmission gratings, it is customary to distinguish between *amplitude gratings* in which the transmission amplitude is modulated, and so-called *phase gratings* where only the refractive index is modulated, without any losses.



(a) Transmission grating (b) Reflection grating

FIGURE 1.40 (a) Ruled periodic parallel scratches on a glass serve as a *transmission grating*. (The glass plate is assumed to be very thin.) (b) A *reflection grating*. An incident light beam results in various "diffracted" beams. The zero-order diffracted beam is the normal reflected beam with an angle of reflection equal to the angle of incidence.

[37] William Lawrence Bragg (1890–1971), Australian-born British physicist, won the Nobel Prize with his father, William Henry Bragg, for his "famous equation" when he was only 25 years old.

When the incident beam is not normal to the diffraction grating, then Eq. (1.12.8) must be modified. If θ_i is the angle of incidence with respect to the normal to the grating, then the diffraction angle θ_m for the m-th mode is given by

Grating equation

$$d(\sin\theta_m - \sin\theta_i) = m\lambda; \quad m = 0, \pm 1, \pm 2, \ldots \tag{1.12.9}$$

The same equation can be used for transmission and reflection gratings provided that we define the angles θ_i and θ_m as positive on either side of the normal as in Figure 1.40 (b).[38] Consider a grating with N slits. The slit width is a (very narrow), and d is the periodicity as before. The detector is at a distance L, far away from the grating. While the periodicity in the slits gives rise to the diffracted beams, the diffraction at each narrow slit defines the envelope of the diffracted intensities as shown in Figure 1.39 (b). If the incident plane wave is normal to the grating, the intensity distribution along y at the screen is given by

Grating diffraction pattern

$$I(y) = I_o\left[\frac{\sin\frac{1}{2}k_y a}{\frac{1}{2}k_y a}\right]^2\left[\frac{\sin\frac{1}{2}Nk_y d}{N\sin\frac{1}{2}k_y d}\right]^2 \tag{1.12.10}$$

where k_y is the scattering wave vector defined by $k_y = (2\pi/\lambda)(y/L) = (2\pi/\lambda)\sin\theta$, and I_o is the maximum intensity along $\theta = 0$. The second term represents the oscillations in the intensity due to interference from different slits. The first term is the envelope of the diffraction pattern, and is the diffraction pattern of a single slit.

The **resolvance** or the **resolving power** R of a diffraction grating is its ability to be able to separate out adjacent wavelengths. If $\lambda_2 - \lambda_1 = \Delta\lambda$ is the minimum wavelength separation that can be measured, as determined by the Rayleigh criterion (the maximum of the intensity distribution at λ_1 is at the first minimum of the intensity distribution at λ_2), and λ is the average wavelength $(1/2)(\lambda_1 + \lambda_2)$ in $\Delta\lambda$, then the **resolving power** is defined by

Resolving power

$$R = \lambda/\Delta\lambda \tag{1.12.11}$$

The separation $\Delta\lambda$ is also called the spectral resolution. If N grooves on a grating are illuminated and the order of diffraction is m, the theoretical resolving power is given simply by $R = mN$. The resolving power is also called the **chromatic resolving power** since it refers to the separation of wavelengths.

Diffraction gratings are widely used in spectroscopic applications because of their ability to provide light deflection that depends on the wavelength. In such applications, the undiffracted light that corresponds to the zero-order beam (Figure 1.40) is clearly not desirable because it wastes a portion of the incoming light intensity. Is it possible to shift this energy to a higher order? Robert William Wood (1910) was able to do so by ruling grooves on glass with a controlled shape as in Figure 1.41 (a) where the surface is angled periodically with a spatial period d. The diffraction condition in Eq. (1.12.9) applies with respect to the normal to the grating plane, whereas the first-order reflection corresponds to reflection from the flat surface, which is at an angle γ. Thus it is possible to "blaze" one of the higher orders (usually $m = 1$) by appropriately choosing γ. Most modern diffraction gratings are of this type. If the angle of incidence is θ_i with respect to the grating normal, then specular reflection occurs at an angle $(\gamma + \theta_i)$ with respect to the face normal and $(\gamma + \theta_i) + \gamma$ with respect to the grating normal. This reflection at $(\gamma + \theta_i) + \gamma$ should occur at diffraction angle θ_m so that

Blazing angle

$$2\gamma = \theta_m - \theta_i \tag{1.12.12}$$

[38] Some books use $d(\sin\theta_m + \sin\theta_i) = m\lambda$ for a transmission grating but the angles become positive on the incidence side and negative on the transmitted side with respect to the normal. It is a matter of sign convention.

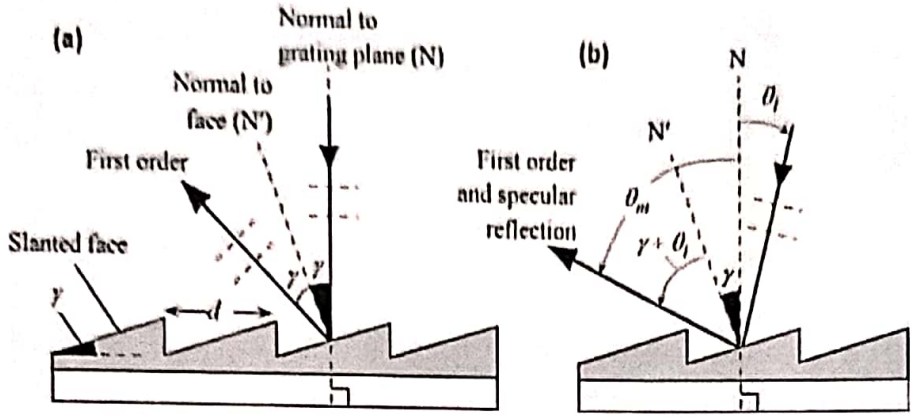
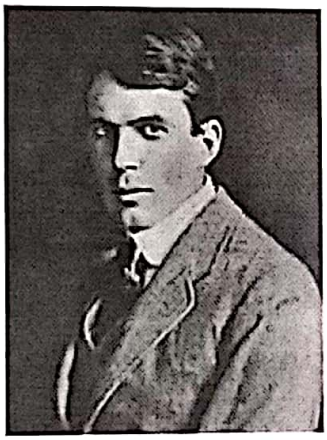


FIGURE 1.41 (a) A blazed grating. Triangular grooves have been cut into the surface with a periodicity d. The side of a triangular groove makes an angle γ to the plane of the diffraction angle. For normal incidence, the angle of diffraction must be 2γ to place the specular reflection on the diffracted beam. (b) When the incident beam is not normal, the specular reflection will coincide with the diffracted beam when $(\gamma + \theta_i) + \gamma = \theta_m$.



William Lawrence Bragg (1890–1971), Australian-born British physicist, won the Nobel Prize with his father, William Henry Bragg, for his "famous equation" when he was only 25 years old. *(SSPL via Getty Images.)*

"The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them."

EXAMPLE 1.12.2 A reflection grating

Consider a reflection grating with a period d that is 10 μm as in Figure 1.42 (a). Find the diffracted beams if a collimated light wave of wavelength 1550 nm is incident on the grating at an angle of 45° to its normal. What should be the blazing angle γ if we were to use a blazed grating with the same periodicity? What happens to the diffracted beams if the periodicity is reduced to 2 μm?

Solution

If we put $m = 0$ in Eq. (1.12.9) we would find the zero-order diffraction, which is at an angle 45°, as expected, and shown in Figure 1.42 (a). The general Bragg diffraction condition is

$$d(\sin\theta_m - \sin\theta_i) = m\lambda$$

so that

$$(10\ \mu\text{m})(\sin\theta_m - \sin(45°) = (+1)(1.55\ \mu\text{m})$$

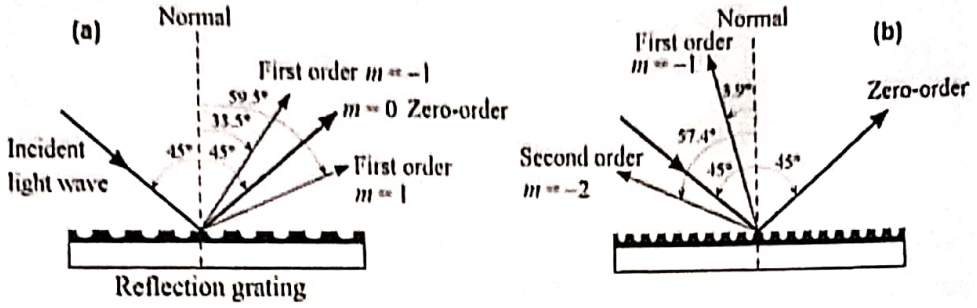


FIGURE 1.42 A light beam is incident at an angle 45° to the normal on a reflection grating. (a) The grating periodicity is 10 μm. (b) The periodicity is 2 μm.

and

$$(10\ \mu\text{m})(\sin\theta_m - \sin(45°) = (-1)(1.55\ \mu\text{m})$$

Solving these two equations, we find $\theta_m = 59.6°$ for $m = 1$, and $\theta_m = 33.5°$ for $m = -1$.

Consider Figure 1.41 (b) in which the secular reflection from the grooved surface coincides with the m_{th} order diffraction when $2\gamma = \theta_m - \theta_i$. Thus

$$\gamma = (1/2)(\theta_m - \theta_i) = (1/2)(59.6° - 45°) = 7.3°$$

Suppose that we reduce d to 2 μm. Recalculating the above we find $\theta_m = -3.9°$ for $m = -1$ and imaginary for $m = +1$. Further, for $m = -2$, there is a second-order diffraction beam at $-57.4°$. Both are shown in Figure 1.42 (b). It is left as an exercise to show that if we increase the angle of incidence, for example, $\theta_i = 85°$ on the first grating, the diffraction angle for $m = -1$ increases from 33.5° to 57.3° and the other diffraction peak ($m = 1$) disappears.

Additional Topics

1.13 INTERFEROMETERS

An **interferometer** is an optical instrument that uses the wave-interference phenomena to produce interference fringes (*e.g.*, dark and bright bands or rings) which can be used to measure the wavelength of light, surface flatness, or small distances. A nearly monochromatic light wave is split into two coherent waves traveling two different paths, and then the two waves are brought together and made to interfere on a screen (or a detector array); the result is an interference pattern as in the Young's fringes in Figure 1.31. In some interferometers, the intensity of the resultant interference is measured at one location, and this intensity is monitored due to changes in one of the optical path lengths. Even a small change in the optical path, distance × refractive index, nL, can cause a measurable shift in the diffraction pattern or a displacement in the fringes, which can be used to infer on nL. There are many types of interferometers.

The **Fabry–Perot interferometer** is a Fabry–Perot cavity–based interferometer that produces an interference ring pattern (bright and dark rings) when illuminated from a broad monochromatic light source as illustrated in Figure 1.43. The resonator consists of two parallel flat glass plates facing each other and separated by an adjustable spacer. A piezoelectric transducer can provide small changes in the spacing between the plates. The inside surface of the glass plates are coated to enhance reflections within the cavity. (Dielectric mirrors can also be used on the inner surfaces.)

so that

$$E_c - E_{Fp} = -k_B T \ln n_i^2/(N_a N_c) \tag{3.9.2}$$

Thus, subtracting Eq. (3.9.2) from (3.9.1) gives

Built-in potential V_o

$$eV_o = E_{Fn} - E_{Fp} = k_B T \ln (N_a N_d)/n_i^2 \tag{3.9.3}$$

3.10 HETEROJUNCTIONS

The *pn* junction with the band diagram in Figure 3.25 is a junction within the same crystal (Si), and hence the bandgap does not change along the device; it represents a **homojunction**. A **heterojunction** is a junction between two different semiconductor crystals with different bandgaps E_{g1} and E_{g2}, for example, between GaAs and Al_xGa_{1-x} As ternary alloys. Inasmuch as the bandgaps are now different, their alignment becomes important. If the bandgap difference is $\Delta E_g = E_{g2} - E_{g1}$, then this difference is taken up by a difference ΔE_c ($= E_{c2} - E_{c1}$) in the CB edges, and ΔE_v ($= E_{v1} - E_{v2}$) in the VB edges. The energy discontinuities in ΔE_c and ΔE_v are called **band offsets** and play an important role in heterojunction devices.

The terms *heterojunction* and *heterostructure* are frequently used interchangeably, though a heterostructure usually has more than one heterojunction. There may or may not be a change in the doping across the heterojucntion. The doping in the wider bandgap semiconductor is usually denoted with a capital letter N or P, and that in the narrower bandgap semiconductor with lower case n or p. There are two cases of particular importance, called Type I and Type II heterojunctions. In a Type I **straddled bandgap alignment heterojunction**, as illustrated in Figure 3.27 (a), the smaller bandgap material offers the lowest energy for both electrons and holes as in GaAs/Al_xGa_{1-x}As heterostructures in which GaAs has a smaller bandgap than Al_xGa_{1-x}As. Tpe I is the most common heterostructure in optoelectronic devices, for example, Ga_xIn_{1-x}As/InP, GaAs/Ga_xIn_{1-x}P. In a Type II **staggered lineup heterojunction**, as illustrated

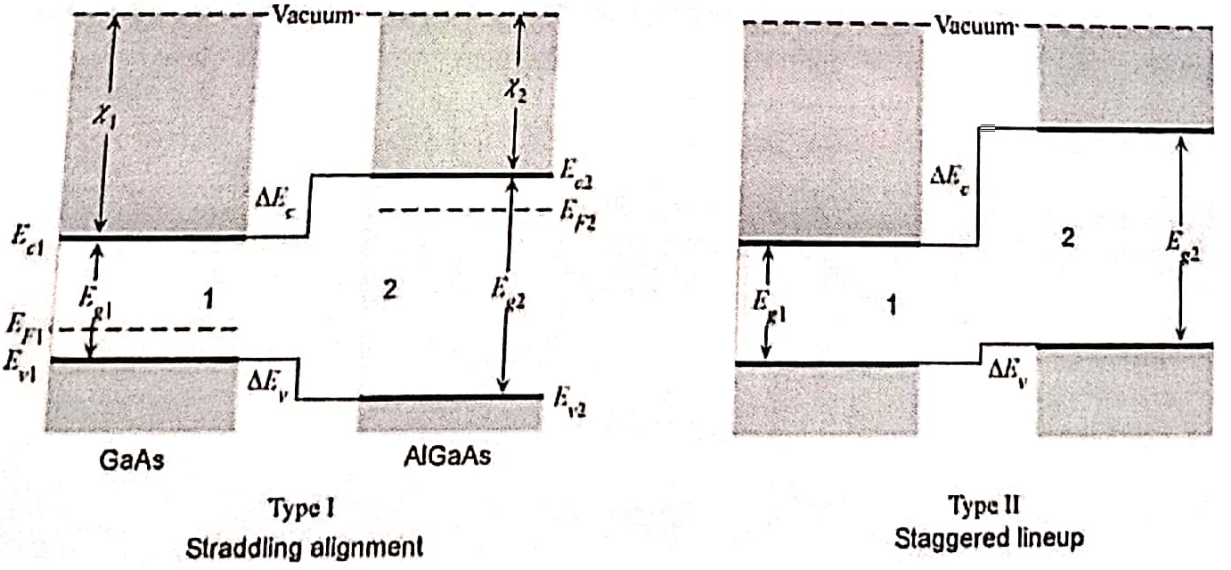


FIGURE 3.27 Two types of heterojunction and the definitions of band offsets Type I and Type II between two semiconductor crystals 1 and 2. Crystal 1 has a narrower bandgap E_{g1} than E_{g2} for crystal 2. Note that the semiconductors are not in contact so that the Fermi level in each is different. In the Type I example, crystal 1 (GaAs) is *p*-type and crystal 2 (AlGaAs) is *N*-type. (Note that the subscripts 1 and 2 refer to semiconductors on the left and right respectively.)

in Figure 3.27 (b), minimum energies for holes and electrons are in the different materials. $Ga_xIn_{1-x}As/GaAs_ySb_{1-y}$ heterojunctions over wide compositions follow the Type II behavior.

The band edge profiles for a given heterostructure are determined by the doping levels, carrier transport, and recombination around the junction, as they are for a simple *pn*-homojunction. The open-circuit heterojunction band diagrams are straightforward based on the principle that the Fermi level must be uniform once a contact is made, and we know how ΔE_g is shared between ΔE_c and ΔE_v. Figure 3.28 (a) shows the energy band diagram of an *Np* heterojunction between *n*-type AlGaAs and *p*-type GaAs. First, notice that E_F is uniform through the device as an equilibrium requirement. Far away from the junction on the *N*-side, we have an *n*-type wide bandgap AlGaAs with E_F close to E_c. Far away from the junction on the right, we have a *p*-type narrower bandgap GaAs with E_F close to E_v. In the depletion regions, around the junction, E_c and E_v must bend because there is an internal field, as we know from the ordinary *pn* junction. E_{c1} and E_{v1} of AlGaAs bend upwards and E_{c2} and E_{v2} of GaAs bend downwards as in the normal *pn* junction. We need to join E_{v1} to E_{v2} but also need to account for ΔE_v, which is easily done by putting ΔE_v between E_{v1} and E_{v2} at the junction as shown in Figure 3.28 (a). Similarly, we need to join E_{c1} to E_{c2} but also need to account for ΔE_c. In this case, we can only join E_{c1} and E_{c2} by having a narrow "spike" whose height must be ΔE_c at the junction as shown in Figure 3.28 (a).

One obvious conclusion is that the potential barrier for hole injection, $(E_{v2} - E_{v1})$, from *p* to *N* is greater than the barrier $(E_{c2} - E_{c1})$ for electron injection from *N* to *p*. Under forward bias, the current will be dominated by the injection of electrons from the *N*-side into the *p*-side. ΔE_v has increased the potential barrier against holes and ΔE_c has decreased the potential barrier against electrons.

The energy band diagram for a *pP*-type heterojunction is shown in Figure 3.28 (b). The basic principle for drawing the diagram is the same as before. In this case, there is a small spike

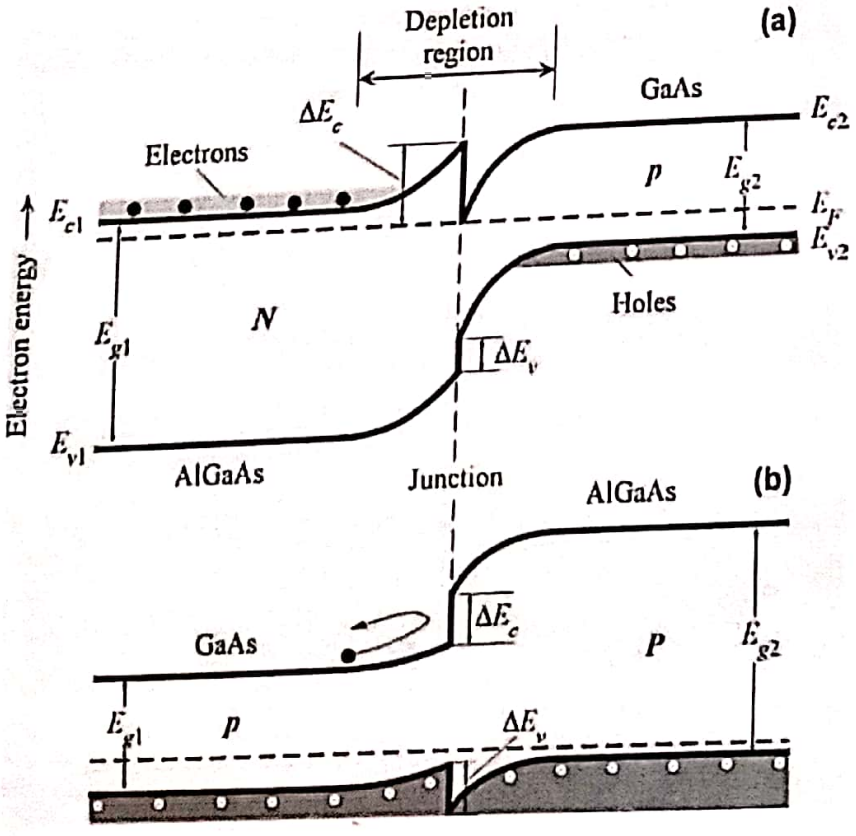


FIGURE 3.28 (a) *nP* and (b) *pP* heterojunctions and their energy band diagrams (schematic only to illustrate general features). Under open circuit and equilibrium conditions, the Fermi level E_F must be uniform, that is, continuous throughout the device. If E_F is close to the conduction band edge, E_c, it results in an *n*-type, and if it is close to the valence band edge, E_v, it results in a *p*-type semiconductor. There is a discontinuity ΔE_c as in E_c, and ΔE_v in E_v, right at the junction.

in E_v at the junction. ΔE_c increases the potential barrier from E_{c1} to E_{c2}. An electron in the CB in the p-side cannot simply overcome this barrier and enter the P-side. Holes in the VB of p- and P-sides can easily cross through the spike by tunneling. Both the Np and pP heterojunctions are used extensively in LED and semiconductor laser diode heterostructures.

3.11 LIGHT-EMITTING DIODES: PRINCIPLES

A. Homojunction LEDs

A light-emitting diode is essentially a pn junction diode typically made from a direct bandgap semiconductor, for example, GaAs, in which the electron–hole pair recombination results in the emission of a photon. The emitted photon energy is therefore approximately equal to the bandgap energy, $h\upsilon \approx E_g$. Figure 3.29 (a) shows the energy band diagram of an unbiased pn^+ junction device in which the n-side is more heavily doped than the p-side. The band diagram is drawn to keep the Fermi level, E_{Fp} and E_{Fn} on the p- and n-sides, uniform through the device which is a requirement of equilibrium with no applied bias. The depletion region in a pn^+ device extends mainly into the p-side. There is a potential energy (PE) barrier eV_o from E_c on the n-side to E_c on the p-side, that is, $\Delta E_c = eV_o$, where V_o is the *built-in voltage*. The higher concentration of conduction (free) electrons in the n-side encourages the diffusion of these electrons from the n- to the p-side. This net electron diffusion, however, is prevented by the electron PE barrier eV_o.

As soon as a forward bias V is applied, this voltage drops almost entirely across the depletion region since this is the most resistive part of the device. Consequently, the built-in potential V_o is reduced to $V_o - V$, which then allows the electrons from the n^+ side to diffuse, or become injected, into the p-side as illustrated in Figure 3.29 (b). The hole injection component from p into the n^+ side is much smaller than the electron injection component from the n^+ to p-side. The recombination of injected electrons in the depletion region as well as in the neutral p-side results in the *spontaneous emission* of photons. Recombination primarily occurs within the depletion region and within a volume extending over the diffusion length L_e of the electrons in the p-side. (Electron injection is preferred over hole injection in GaAs LEDs because electrons have a higher mobility and hence a larger diffusion coefficient.) This recombination zone is

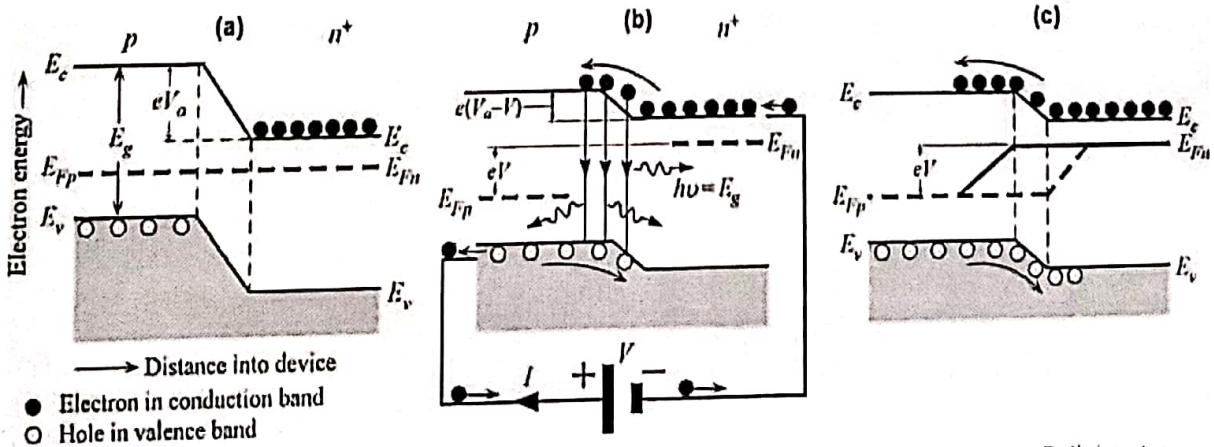


FIGURE 3.29 (a) The energy band diagram of a pn^+ (heavily n-type doped) junction without any bias. Built-in potential V_o prevents electrons from diffusing from n^+ to p-side. (b) The applied bias potential V reduces V_o and thereby allows electrons to diffuse, be injected, into the p-side. Recombination around the junction and within the diffusion length of the electrons in the p-side leads to spontaneous photon emission. (c) Quasi-Fermi levels E_{Fp} and E_{Fn} for holes and electrons across a forward-biased pn-junction.

laser diode. Chapter 18 is devoted to semiconductor detectors.

16.1 SEMICONDUCTORS

As discussed in Sec. 13.1C, a semiconductor is a crystalline or amorphous solid whose electrical conductivity is typically intermediate between that of a metal and that of an insulator. Its conductivity can be significantly altered by modifying the temperature or doping concentration of the material, or by illuminating it with light. The band structure of semiconductors, and the ability to form junctions and heterostructures, offer unique properties. Quantum-confined semiconductor structures further extend the range of available properties. Electronic semiconductor devices are principally fabricated from silicon (Si), while optoelectronic semiconductor devices often make use of ternary or quaternary semiconductor compounds such as InGaAsP and AlInGaN (see Sec. 16.1B).

A. Energy Bands and Charge Carriers

Energy Bands in Semiconductors

The atoms comprising solid-state materials have sufficiently strong interatomic interactions that they cannot be treated as individual entities (see Sec. 13.1C). Their conduction electrons are not bound to individual atoms; rather, they belong to the collection of atoms as a whole. The solution of the Schrödinger equation for the electron energy, in the periodic potential created by the collection of atoms in the crystal lattice, results in a splitting of the atomic energy levels and the formation of energy bands. Each band contains a large number of densely packed discrete energy levels that is well approximated as a continuum. As illustrated in Fig. 16.1-1, the valence and conduction bands are separated by the **bandgap energy** E_g, which plays an important role in determining the electrical and optical properties of the material.

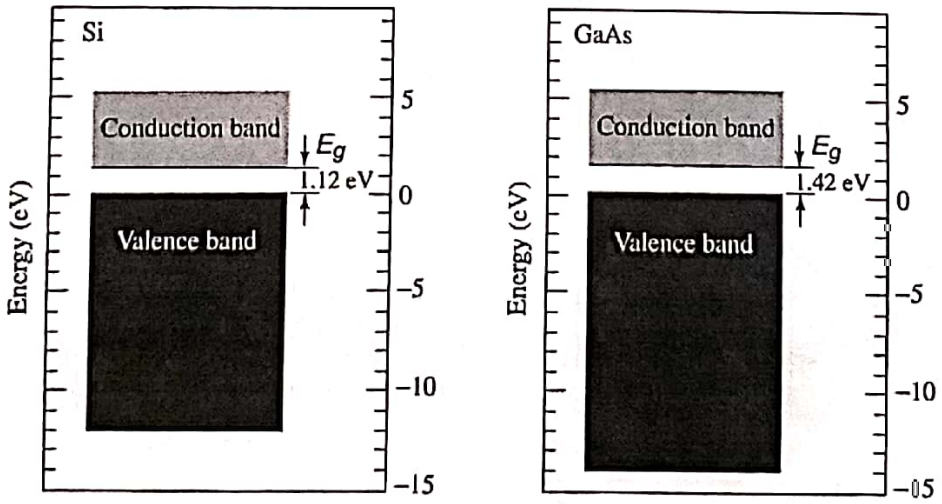


Figure 16.1-1 Energy bands in Si and GaAs. The bandgap energy E_g, which separates the valence and conduction bands, is 1.12 eV for Si and 1.42 eV for GaAs at room temperature.

The origin of the bandgap may be illustrated by means of the **Kronig–Penney model**. In this simple theory the crystal-lattice potential, a one-dimensional version of which is depicted in Fig. 16.1-2(a), is approximated by a 1D periodic rectangular-barrier potential, as shown in Fig. 16.1-2(b). The solution of the associated Schrödinger

equation (13.1-3) for this potential yields allowed energy bands with traveling-wave solutions, separated by forbidden bands with exponentially decaying solutions. It can be shown that the results are general and apply to three dimensions. This approach is similar to that used for analyzing the optics of one-dimensional periodic media, as set forth in Sec. 7.2. The traveling-wave eigenfunctions are **Bloch modes** with the periodicity of the crystal lattice [see (7.2-4)].

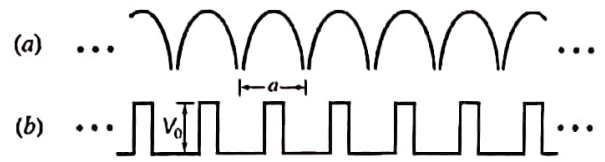


Figure 16.1-2 (*a*) Crystal-lattice potential associated with an infinite one-dimensional collection of atoms with lattice constant *a*. (*b*) Idealized rectangular-barrier potential (height V_0) used in the Kronig–Penney model.

Electrons and Holes

As discussed in Sec. 13.1C, the wavefunctions of the electrons in a semiconductor overlap so that the **Pauli exclusion principle** applies. This principle dictates that no two electrons may occupy the same quantum state and that the lowest available energy levels fill first. Elemental semiconductors, such as Si and Ge, have four valence electrons per atom that form covalent bonds. At $T = 0°$ K, the number of quantum states that can be accommodated in the valence band is such that it is completely filled while the conduction band is completely empty. The material cannot conduct electricity under these conditions.

As the temperature increases, however, some electrons can be thermally excited from the valence band into the empty conduction band, where unoccupied states are abundant (see Fig. 16.1-3). These electrons can then act as mobile carriers, drifting through the crystal lattice under the effect of an applied electric field, and thereby contributing to the electric current. Moreover, an electron departing from the valence band leaves behind an unoccupied quantum state, which in turn allows the remaining electrons in the valence band to exchange places with each other under the influence of an external field. The collection of electrons remaining in the valence band thus undergoes motion. This can equivalently be regarded as motion, in the opposite direction, of the hole left behind by the departed electron. The hole therefore behaves as if it has a positive charge $+e$.

The net result is that each electron excitation creates a free electron in the conduction band and a free hole in the valence band. The two charge carriers are free to drift under the effect of the applied electric field and thereby to generate an electric current. The material behaves as a *semi*conductor whose conductivity increases sharply with increasing temperature, as more and more mobile carriers are thermally generated.

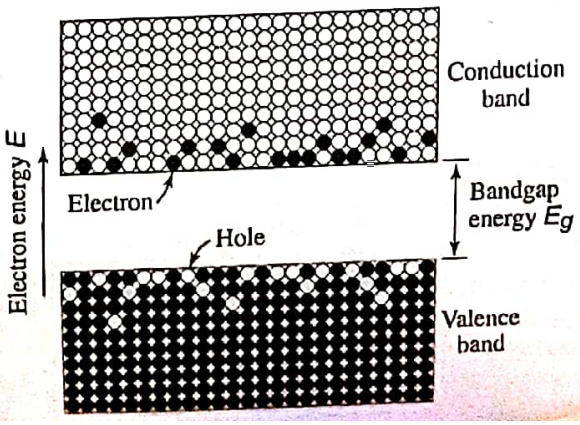


Figure 16.1-3 Electrons in the conduction band and holes in the valence band at $T >$ 0° K.

Energy–Momentum Relations

In accordance with wave mechanics, the energy E and momentum $\mathbf{p}$ of an electron in a region of constant potential, such as free space, are related by $E = p^2/2m_0 = \hbar^2 k^2/2m_0$, where p is the magnitude of the momentum, k is the magnitude of the wavevector $\mathbf{k} = \mathbf{p}/\hbar$, and m_0 is the electron mass (9.1×10^{-31} kg). The E–k relation for a free electron is thus a simple parabola.

EXERCISE 16.1-1

Energy–Momentum Relation for a Free Electron.

(a) Consider a one-dimensional version of the time-independent Schrödinger equation set forth in (13.1-3) for a free electron ($V = 0$) of mass m_0. Use a trial solution of the form $\psi(x) \propto \exp(-jkx)$ to show that the energy–momentum relation assumes the quadratic form

$$E = \frac{\hbar^2 k^2}{2m_0}, \tag{16.1-1}$$

so that the energy is not quantized in this ideal case.

(b) The free photon, in contrast, has a linear energy–momentum relation, as provided in (12.1-10):

$$E = pc = c\hbar k, \tag{16.1-2}$$

where c is the speed of light in the medium. What is the origin and significance of this distinction?

The motion of an electron in a semiconductor material is similarly governed by the Schrödinger equation, but with a potential generated by the charges in the periodic crystal lattice of the material. As discussed earlier, this construct results in allowed energy bands separated by forbidden bands, as predicted by the Kronig–Penney model. The ensuing E–k relations for electrons and holes, in the conduction and valence bands respectively, are illustrated in Fig. 16.1-4 for Si and GaAs. The energy E is a periodic function of the components (k_1, k_2, k_3) of the wavevector $\mathbf{k}$, with periodicities $(\pi/a_1, \pi/a_2, \pi/a_3)$, where a_1, a_2, a_3 are the crystal lattice constants. Figure 16.1-4 displays cross sections of this relation along two particular directions of the wavevector $\mathbf{k}$. The range of k values in the interval $[-\pi/a, \pi/a]$ defines the first **Brillouin zone**. The energy of an electron in the conduction band thus depends not only on the magnitude of its momentum, but also on the direction in which it is traveling in the crystal. The semiconductor E–k diagram bears some resemblance to the photonic-crystal ω–K diagram (see Fig. 7.3-5).

Effective Mass

It can be seen from Fig. 16.1-4 that near the bottom of the conduction band, the E–k relation may be approximated by the parabola

$$E = E_c + \frac{\hbar^2 k^2}{2m_c}, \tag{16.1-3}$$

where E_c is the energy at the bottom of the conduction band and k is measured from the wavevector where the minimum occurs. This relation tells us that a conduction-band electron behaves in a manner similar to that of a free electron, but with a mass m_c, known as the electron (conduction-band) **effective mass**, that differs from the free-electron mass m_0. The influence of the ions of the lattice on the motion of a conduction-band electron is thus contained in the effective mass m_c. This behavior is highlighted in Fig. 16.1-5.

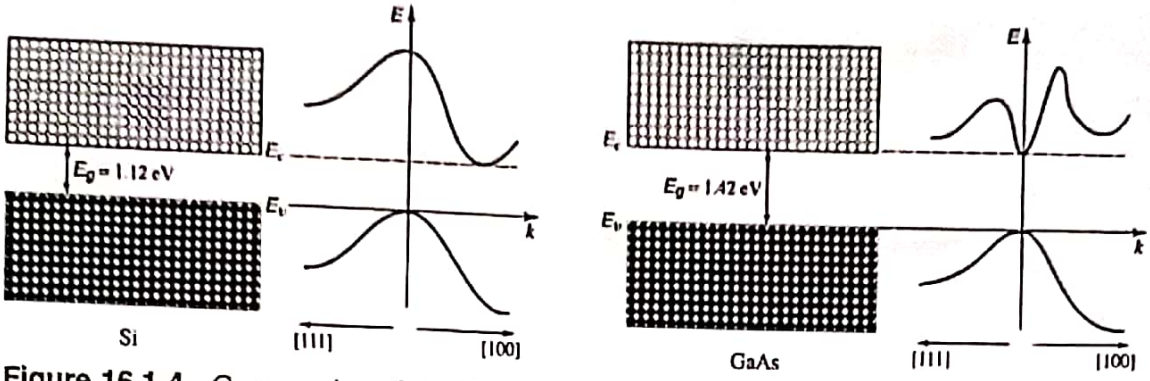


Figure 16.1-4 Cross section of the E–k function for Si and GaAs along two crystal directions: [111] toward the left and [100] toward the right.

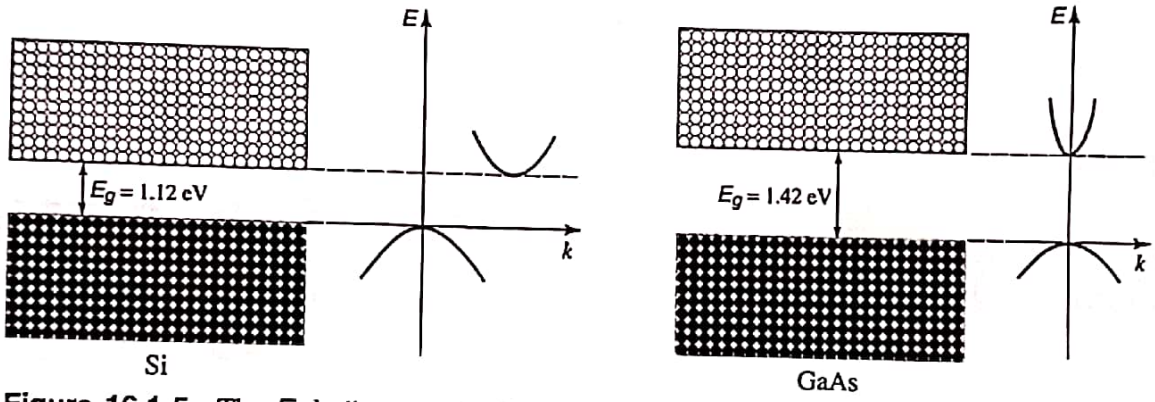


Figure 16.1-5 The E–k diagrams for Si and GaAs are well approximated by parabolas at the bottom of the conduction band and at the top of the valence band.

Similarly, near the top of the valence band, we have

$$E = E_v - \frac{\hbar^2 k^2}{2m_v}, \tag{16.1-4}$$

where $E_v = E_c - E_g$ is the energy at the top of the valence band and m_v is the hole (valence-band) effective mass, as portrayed in Fig. 16.1-5. The influence of the lattice ions on the motion of a valence-band hole is captured by the effective mass m_v. The effective mass depends on the crystal structure of the material and the direction of travel with respect to the lattice since the interatomic spacing varies with crystallographic direction. It also depends on the particular band under consideration. Indeed, several parabolas of different curvature often coexist near the top of the valence band; these correspond to so-called heavy holes, light holes, and holes associated with the split-off band. Typical ratios of the averaged effective masses to the mass of the free electron m_0 are provided in Table 16.1-1 for Si, GaAs, and GaN.

Table 16.1-1 Typical values of electron and hole effective masses in selected semiconductor materials.

	m_c/m_0	m_v/m_0
Si	0.98	0.49
GaAs	0.07	0.50
GaN	0.20	0.80

Direct- and Indirect-Bandgap Semiconductors

Semiconductors for which the conduction-band minimum energy and the valence-band maximum energy correspond to the same value of the wavenumber k (same momentum) are called **direct-bandgap** materials. Semiconductors for which this is not the case are known as **indirect-bandgap** materials. As is evident in Fig. 16.1-5, GaAs is a direct-bandgap semiconductor whereas Si is an indirect-bandgap semiconductor. The distinction is important because a transition between the bottom of the conduction band and the top of the valence band in an indirect-bandgap semiconductor must accommodate a substantial change in the momentum of the electron. It will be shown subsequently that direct-bandgap semiconductors such as GaAs are efficient photon emitters, whereas indirect-bandgap semiconductors such as Si cannot serve as efficient light emitters under ordinary circumstances.

B. Semiconductor Materials

Figure 16.1-6 reproduces the section of the periodic table that comprises most of the elements important in semiconductor electronics and photonics. Both elemental and compound semiconductors play crucial roles in these technologies.

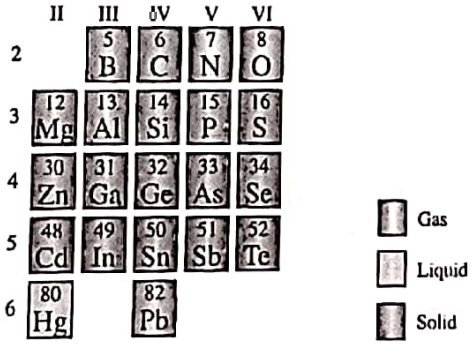


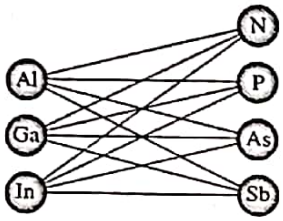
Figure 16.1-6 Section of the periodic table relating to semiconductors. Elements indicated in blue, yellow, and silver take the form of gases, liquids, and solids, respectively, at room temperature. The full periodic table is displayed in Fig. 13.1-3.

We proceed to discuss elemental, binary, ternary, and quaternary semiconductors in turn, and then consider doped semiconductors.

Elemental Semiconductors

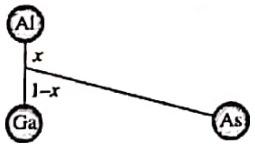
Silicon (**Si**) and germanium (**Ge**) are important elemental semiconductors in column IV of the periodic table. Virtually all commercial electronic integrated circuits and devices are fabricated using Si. Both Si and Ge also find widespread use in photonics, principally as photodetectors. These materials have traditionally not been used for the fabrication of light emitters because of their indirect bandgaps. However, some forms of Si are viable as light emitters and silicon photonics has come to the fore. The basic properties of Si and Ge are provided in Table 16.1-2.

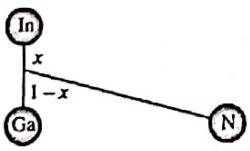
Binary III–V Semiconductors



Compounds formed by combining an element in column III, such as aluminum (**Al**), gallium (**Ga**), or indium (**In**), with an element in column V, such as nitrogen (**N**), phosphorus (**P**), arsenic (**As**), or antimony (**Sb**), are important semiconductors in photonics. These 12 III–V compounds are listed in Table 16.1-2, along with their crystal structure (zincblende or wurtzite), bandgap type (direct or indirect), bandgap energy E_g, and bandgap wavelength $\lambda_g = hc_o/E_g$ (the free-space wavelength of a photon of energy E_g). The bandgap energies and lattice constants of these compounds are also displayed in Fig. 16.1-7. Photon sources (light-emitting diodes and lasers) and detectors can be readily fabricated from many of these binary compounds. The first of the binary semiconductors to find use in photonics was gallium arsenide (**GaAs**), which is also sometimes used as an alternative to Si for fast electronic devices and circuits. Gallium nitride (**GaN**) plays a central role in photonics by virtue of its near-ultraviolet bandgap wavelength; it is also important in electronics because of its ability to withstand high temperatures. AlN, which is an insulator, has the highest bandgap of all III–V compounds and emits photons at the shortest wavelength, in the mid-ultraviolet region.

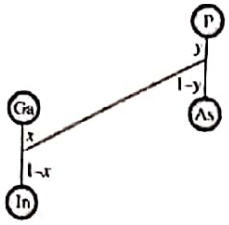
Ternary III–V Semiconductors





Compounds formed from two elements of column III with one element from column V (or one from column III with two from column V) are important ternary semiconductors. $(\mathbf{Al}_x\mathbf{Ga}_{1-x})\mathbf{As}$, for example, is a compound with properties that interpolate between those of AlAs and GaAs, depending on the compositional mixing ratio x (the fraction of Ga atoms in GaAs that are replaced by Al atoms). The bandgap energy E_g for this material varies between 1.42 eV for GaAs and 2.16 eV for AlAs, as x varies between 0 and 1 along the line connecting GaAs and AlAs in Fig. 16.1-7(a). Because this line is essentially vertical, $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ is lattice matched to GaAs; a layer of arbitrary composition of this material can therefore be grown on a layer of different composition without straining the lattice. Other useful III–V ternary compounds, such as $\mathbf{Ga}(\mathbf{As}_{1-x}\mathbf{P}_x)$, are also represented in the bandgap-energy versus lattice-constant diagram displayed in Fig. 16.1-7(a). $(\mathbf{In}_x\mathbf{Ga}_{1-x})\mathbf{As}$ is widely used for photon sources and detectors in the near-infrared region of the spectrum. Similarly, $(\mathbf{Al}_x\mathbf{Ga}_{1-x})\mathbf{N}$ and $(\mathbf{In}_x\mathbf{Ga}_{1-x})\mathbf{N}$ are important ternary semiconductors for photonic devices that operate in the ultraviolet, violet, blue, and green regions of the spectrum, as can be deduced from Fig. 16.1-7(b). In the domain of electronics, $(\mathrm{In}_x\mathrm{Ga}_{1-x})\mathrm{As}$/InP heterojunction bipolar transistors can be switched at speeds approaching 1 THz; indeed, various III–V compounds can be used to fabricate ultrafast transistors that emit light.

Quaternary III–V Semiconductors

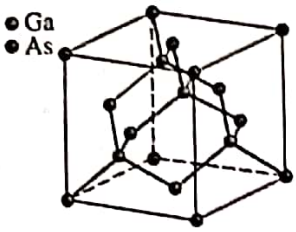


These compounds are formed by mixing two elements from column III with two elements from column V (or three from column III with one from column V). Quaternary semiconductors offer more flexibility for fabricating materials with desired properties than do ternary semiconductors by virtue of an additional degree of freedom. An example is provided by $\mathbf{In_{1-x}Ga_xAs_{1-y}P_y}$, whose bandgap energy varies between 0.36 eV (InAs) and 2.26 eV (GaP) as the compositional mixing ratios x and y vary between 0 and 1. The lattice constant usually varies linearly with the mixing ratio (Vegard's law). The stippled area in Fig. 16.1-7(a) indicates the range of bandgap energies and lattice constants spanned by this compound. For mixing ratios x and y that satisfy $y = 2.16(1-x)$, $In_{1-x}Ga_xAs_{1-y}P_y$ can be lattice matched to InP, which can therefore serve as a convenient template (substrate). This quaternary compound is used for fabricating light-emitting diodes, laser diodes, and photodetectors, particularly in the vicinity of the 1550-nm optical fiber communications wavelength (see Chapters 17, 18, and 24). Another example is provided by $\mathbf{(Al_xIn_yGa_{1-x-y})P}$, for which GaAs serves as a template; this compound offers high-brightness emission in the red, orange, and yellow spectral regions [see shaded region in Fig. 16.1-7(a)]. Yet another important quaternary material is the III–nitride compound $\mathbf{(Al_xIn_yGa_{1-x-y})N}$, which serves the green, blue, violet, and ultraviolet spectral regions in the same way [see Fig. 16.1-7(b)]. Convenient templates for the III–nitrides are sapphire and SiC.

Column IV elements can also be alloyed to form compound semiconductors. The binary alloy silicon carbide (SiC), also known as carborundum, has an indirect bandgap and is useful for fabricating ultraviolet photodetectors and as a template for III–nitride compounds. Silicon germanium ($Si_{1-x}Ge_x$) enjoys a variety of applications in electronics and photonics, including use as an infrared photodetector material. Ternary and quaternary column-IV semiconductor compounds include $Si_{1-x-y}Ge_xC_y$ and $Si_{1-x-y-z}Ge_xC_ySn_z$, respectively.

Binary II–VI materials, i.e., compounds formed from elements in column II (e.g., Zn, Cd, Hg) and column VI (e.g., S, Se, Te) of the periodic table are also useful semiconductors. This family includes ZnS, ZnSe, ZnTe, CdS, CdSe, CdTe, HgS, HgSe, and HgTe, as shown in Fig. 16.1-8. All of these materials have a zincblende structure and all are direct-bandgap semiconductors; the exceptions are HgSe and HgTe, which are semimetals with small negative bandgaps. A particular merit of ZnSe, is that it can be deposited on a GaAs substrate with a relatively low defect density since the lattice constants of the two materials are similar. Moreover, HgTe and CdTe are nearly lattice matched, so the ternary semiconductor $Hg_xCd_{1-x}Te$ can be grown without strain on a CdTe substrate. This material system is widely used for fabricating photon detectors, as are other II–VI compounds (see Chapter 18). Unlike the III–V alloys, the II–VI compounds are widely found in nature, but photon sources fabricated from these materials currently suffer from limited lifetimes. Nevertheless, binary II–VI semiconductor materials are readily fashioned into quantum dots with tunable photoluminescence emission wavelength (see, for example, Fig. 13.1-12). Ternary IV–VI semiconductor compounds, such as $Pb_xSn_{1-x}Te$ and $Pb_xSn_{1-x}Se$, have also been used as infrared photodetectors and laser diodes. However, these alloys have slower response times because of their large dielectric constants. They also have high thermal coefficients of expansion, so cycling between room and cryogenic temperatures can be problematic.

Zincblende and Diamond



Wurtzite

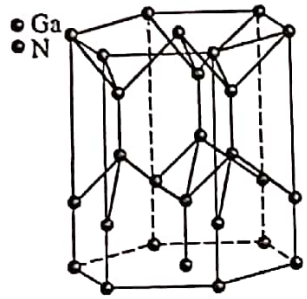


Table 16.1-2 Selected elemental and III–V binary semiconductors along with their crystal structures, bandgap types, bandgap energies, and bandgap wavelengths.

Material	Crystal Structure[a] (D/Z/W)	Bandgap Type[b] (I/D)	Bandgap Energy[c] E_g (eV)	Bandgap Wavelength[d] λ_g (μm)
Si	D	I	1.12	1.11
Ge	D	I	0.66	1.88
AlN	W	D	6.20	0.200
AlP	Z	I	2.45	0.506
AlAs	Z	I	2.16	0.574
AlSb	Z	I	1.58	0.785
GaN	W	D	3.39	0.366
GaP	Z	I	2.26	0.549
GaAs	Z	D	1.42	0.873
GaSb	Z	D	0.73	1.70
InN	W	D	0.65	1.91
InP	Z	D	1.35	0.919
InAs	Z	D	0.36	3.44
InSb	Z	D	0.17	7.29

[a]The crystal structure listed indicates the most commonly used form of the material: D = Diamond, Z = Zincblende, W = Wurtzite, as displayed at left. The zincblende structure comprises two interpenetrating face-centered-cubic lattices, one for each element, displaced from each other by $\frac{1}{4}$ of the body diagonal. The diamond lattice is the same as zincblende except that all atoms are identical. The Brillouin zone for these structures is illustrated in Fig. 7.3-4. The wurtzite structure consists of two hexagonal close-packed lattices, one for each element, displaced from each other along the three-fold c axis by $\frac{3}{8}$ of its length. All atoms are tetrahedrally bonded with their neighbors.

[b]I = Indirect bandgap; D = Direct bandgap.

[c]Data are provided at $T = 300°$ K.

[d]The bandgap wavelength λ_g is related to the bandgap energy E_g by $\lambda_g = hc_0/E_g$; when the bandgap energy is expressed in eV and the bandgap wavelength is expressed in μm, this relation becomes $\lambda_g \approx 1.24/E_g$.

Doped Semiconductors

The electrical and optical properties of semiconductors can be modified substantially by the controlled introduction into the material of small amounts of specially chosen impurities called **dopants**. The introduction of these impurities can alter the concentration of mobile charge carriers by many orders of magnitude. Dopants with excess valence electrons, called **donors**, replacing a small proportion of the normal atoms in the crystal lattice, create a predominance of mobile electrons. The material is then said to be an ***n*-type** semiconductor. Thus, atoms from column V (e.g., P or As) replacing column-IV atoms in an elemental semiconductor (e.g., Si or Ge), or atoms from column VI (e.g., Se or Te) replacing column-V atoms in a III–V binary semiconductor (e.g., As or Sb), produce an n-type material. Similarly, a ***p*-type** semiconductor is made by using dopants with a deficiency of valence electrons, called **acceptors**. The result is then a predominance of mobile holes. Column IV atoms in an elemental semiconductor replaced with column-III atoms (e.g., B or In), or column-III atoms in a III–V binary semiconductor replaced with column-II atoms (e.g., Zn or Cd), yield p-type material. Column-IV atoms act as donors for column III and as acceptors for column V, and therefore can be used to produce an excess of both electrons and holes in III–V materials. Of course, the charge neutrality of the material is not altered by the introduction of dopants.

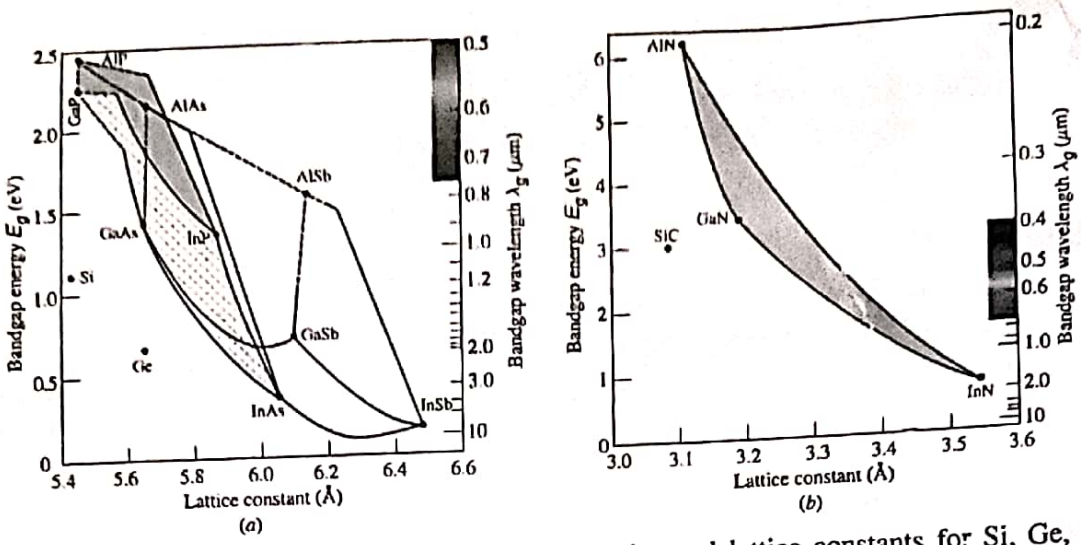


Figure 16.1-7 Bandgap energies, bandgap wavelengths, and lattice constants for Si, Ge, SiC, and 12 III–V binary compounds. Solid and dashed curves represent direct-bandgap and indirect-bandgap compositions, respectively. A material may have a direct bandgap for one mixing ratio and an indirect bandgap for a different mixing ratio. Ternary materials are represented along the line that joins two binary compounds. A quaternary compound is represented by the area formed by its binary components. (*a*) $In_{1-x}Ga_xAs_{1-y}P_y$ is represented by the stippled area with vertices at InP, InAs, GaAs, and GaP, while $(Al_xGa_{1-x})_yIn_{1-y}P$ is represented by the shaded area with vertices at AlP, InP, and GaP. Both are important quaternary compounds, the former in the near infrared and the latter in the visible. $Al_xGa_{1-x}As$ is represented by points along the line connecting GaAs and AlAs. As x varies from 0 to 1, the point moves along the line from GaAs and AlAs. Since this line is nearly vertical, $Al_xGa_{1-x}As$ is lattice matched to GaAs. (*b*) Although the III–nitride compound $In_xGa_{1-x}N$ can, in principle, be compositionally tuned to accommodate the entire visible spectrum, this material becomes increasingly difficult to grow as the composition of In becomes appreciable. $In_xGa_{1-x}N$ is principally used in the green, blue, and violet spectral regions, while $Al_xGa_{1-x}N$ and $Al_xIn_yGa_{1-x-y}N$ serve the ultraviolet region. All compositions of these III–Nitride compounds are direct-bandgap semiconductors.

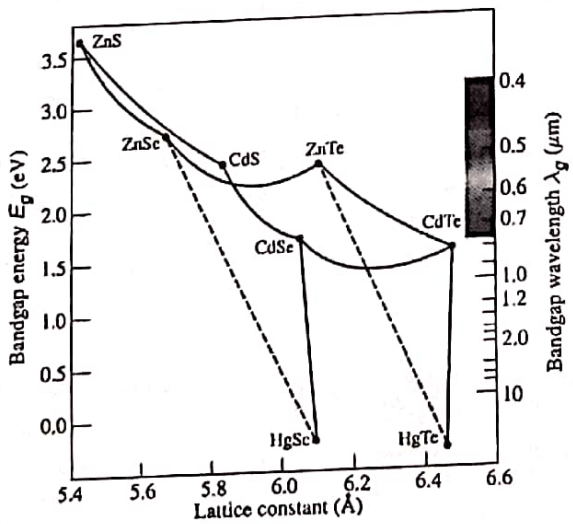


Figure 16.1-8 Bandgap energies, bandgap wavelengths, and lattice constants for various II–VI semiconductors (HgSe and HgTe are semimetals with small negative bandgaps). HgTe and CdTe are nearly lattice matched, as evidenced by the vertical line connecting them, so that the ternary semiconductor $Hg_xCd_{1-x}Te$ can be grown without strain on a CdTe template. It is an important mid-infrared photodetector material.

Undoped semiconductors (i.e., semiconductors devoid of intentional doping) are referred to as **intrinsic** materials, whereas doped semiconductors are called **extrinsic** materials. The concentrations of mobile electrons and holes are equal in an intrinsic semiconductor, $n = p = n_i$, where the intrinsic concentration n_i grows with increasing temperature at an exponential rate. On the other hand, the concentration of mobile electrons in an n-type semiconductor (**majority carriers**) is far greater than the concentration of holes (**minority carriers**), i.e., $n \gg p$. The opposite is true in

a p-type semiconductor, where holes are the majority carriers, and $p \gg n$. A doped semiconductor at room temperature typically has a majority-carrier concentration that is approximately equal to the doping concentration.

Single-ion implantation techniques can be used to fabricate semiconductor materials in which the number of dopant atoms, and their positions, are precisely controlled. The resulting materials exhibit properties that are more deterministic than those with random numbers of dopant atoms, which is useful in certain applications.

EXAMPLE 16.1-1. *Donor-Electron Ionization Energy.* Consider a germanium crystal of dielectric constant $\epsilon/\epsilon_o = 16$ (see Table 16.2-1), doped with arsenic donor atoms. The electron effective mass $m_c = 0.2\,m_0$, where m_0 is the free electron mass. The donor electron moves in the field of the singly charged arsenic ion (As^+), and has energy levels similar to those of an electron in the hydrogen atom. Choosing $n = 1$ and $Z = 1$ in (13.1-4), and replacing ϵ_o by ϵ, and M_r by m_c, to accommodate the polarization density and crystal lattice of the semiconductor material, respectively, the energy of the donor electron is given by

$$E_D = -\left(\frac{1}{4\pi\epsilon}\right)^2 \frac{m_c e^4}{2\hbar^2}. \tag{16.1-5}$$

Since the energy of the electron in the ground state of hydrogen is -13.6 eV (indicating that it is 13.6 eV below ionization), the energy of the arsenic donor electron is $E_D = -(m_c/m_0)(\epsilon_o/\epsilon)^2 \times 13.6\text{ eV} \approx -0.01$ eV. The donor electron thus resides in the forbidden band, at a level ≈ 0.01 eV below the conduction band. However, since the thermal energy $kT \approx 0.026$ eV at $T = 300°$ K, essentially all of the donors are ionized at room temperature and the donor electrons are elevated to the conduction band. The material thus has a conduction-band donor concentration that matches the impurity concentration.

Organic Semiconductors

Organic semiconductors are increasingly employed in a wide variety of fields. This includes photonics, where they are used to fabricate photovoltaic devices, light-emitting diodes, and displays. Although they offer neither the speed nor small size of conventional semiconductor structures, they can be inexpensively fabricated in the form of thin sheets, making low-cost, mechanically flexible optoelectronic components a reality. These materials come in a virtually unlimited array of variations that can be engineered to suit specific requirements and some can be printed on a suitable substrate using inkjet technology.

Organic semiconductors come in two principal varieties, as illustrated schematically in Fig. 16.1-9:

1. Small organic molecules such as pentacene, which consists of five linearly joined benzene rings [Fig. 16.1-9(a)].
2. Conjugated polymer chains such as polyacetylene, comprising hundreds or thousands of carbon atoms [Fig. 16.1-9(b)].

A hallmark of these amorphous materials, termed conjugation, is their alternating single and double carbon–carbon bonds. Although the double-bond electrons shown in Figs. 16.1-9(a) and (b) are portrayed as belonging to particular atoms, these electrons are actually delocalized and shared among multiple atoms, or along a segment of polymer comprising roughly 10 repeat units. The molecule, or polymer segment, behaves as a single system in which the allowed electron states form bands.

In its undoped state, the valence band of a conjugated polymer chain is typically full, and its conduction band empty, so that it behaves as an insulator. However, as

Each pair of quantum numbers (q_1, q_2) is associated with an energy subband that has a density of states $\varrho(k) = 1/\pi$ per unit length of the wire and therefore $1/\pi d_1 d_2$ per unit volume. The corresponding quantum-wire density of states (per unit volume), as a function of energy, is

$$\varrho_c(E) = \begin{cases} \dfrac{(1/d_1 d_2)(\sqrt{m_c}/\sqrt{2}\pi\hbar)}{\sqrt{E - E_c - E_{q1} - E_{q2}}}, & E > E_c + E_{q1} + E_{q2} \\ 0, & \text{otherwise,} \end{cases}$$

$$q_1, q_2 = 1, 2, 3, \ldots. \quad (16.1\text{-}40)$$

These are decreasing functions of energy, as illustrated in Fig. 16.1-29.

Quantum Dots

In a **quantum-dot** structure, the electrons are narrowly confined in all three directions within a region that we take to be a box of volume $d_1 d_2 d_3$. The energy is therefore quantized to

$$E = E_c + E_{q1} + E_{q2} + E_{q3}, \quad (16.1\text{-}41)$$

where

$$E_{q1} = \frac{\hbar^2(q_1\pi/d_1)^2}{2m_c}, \qquad E_{q2} = \frac{\hbar^2(q_2\pi/d_2)^2}{2m_c}, \qquad E_{q3} = \frac{\hbar^2(q_3\pi/d_3)^2}{2m_c},$$

$$q_1, q_2, q_3 = 1, 2, 3, \ldots. \quad (16.1\text{-}42)$$

The allowed energy levels are discrete and well separated so that the density of states is represented by a sequence of delta functions at the allowed energies, as illustrated in Fig. 16.1-29. Quantum dots are often called artificial atoms (see Sec. 13.1C). Even though they contain enormous numbers of strongly interacting natural atoms, the discrete energy levels of the quantum dot can, in principle, be chosen at will by proper design.

16.2 INTERACTIONS OF PHOTONS WITH CHARGE CARRIERS

We proceed to consider some of the basic optical properties of semiconductors, with an emphasis on the processes of absorption and emission important in the operation of photonic devices. This domain of study is known as **semiconductor optics.**

A. Photon Interactions in Bulk Semiconductors

A number of mechanisms can lead to the absorption and emission of photons in bulk semiconductors. The most important of these are:

- *Band-to-Band (Interband) Transitions.* An absorbed photon can result in an electron in the valence band making an upward transition to the conduction band, thereby creating an electron–hole pair [Fig. 16.2-1(*a*)]. Electron–hole recombination can result in the emission of a photon. Band-to-band transitions may be assisted by one or more phonons. A phonon is a quantum of the lattice vibrations associated with molecular or acoustic vibrations of the atoms in a material.

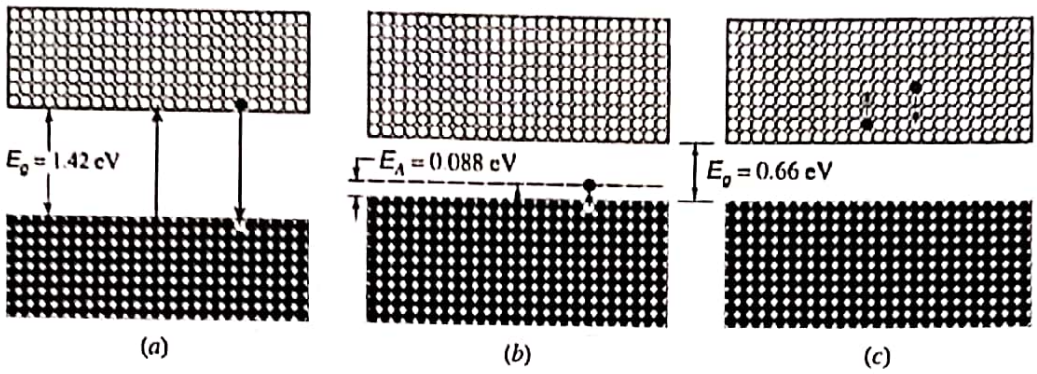


Figure 16.2-1 Examples of absorption and emission of photons in bulk semiconductors. (*a*) Band-to-band transitions in GaAs can result in the absorption or emission of photons of wavelength $\lambda_o < \lambda_g = hc_o/E_g = 0.87$ μm. (*b*) The absorption of a photon of wavelength $\lambda_A = hc_o/E_A = 14$ μm results in a valence-band to acceptor-level transition in Hg-doped Ge (Ge:Hg). (*c*) Free-carrier transitions within the conduction band of Ge.

- *Impurity-to-Band Transitions.* An absorbed photon can result in a transition between a donor (or acceptor) level and a band in a doped semiconductor. In a *p*-type material, for example, a low-energy photon can lift an electron from the valence band to the acceptor level, where it becomes trapped by an acceptor atom [Fig. 16.2-1(*b*)]. A hole is created in the valence band and the acceptor atom is ionized. Or a hole may be trapped by an ionized acceptor atom; the result is that the electron decays from its acceptor level to recombine with the hole. The energy may be released radiatively (in the form of an emitted photon) or nonradiatively (in the form of phonons). The transition may also be assisted by traps in defect states, as illustrated in Fig. 16.1-17.
- *Free-Carrier (Intraband) Transitions.* An absorbed photon can impart its energy to an electron in a given band, causing it to move higher within that band. An electron in the conduction band, for example, can absorb a photon and move to a higher energy level within the conduction band [Fig. 16.2-1(*c*)]. This is followed by thermalization, a process whereby the electron relaxes down to the bottom of the conduction band while releasing its energy in the form of phonons. The strength of free-carrier absorption is proportional to the carrier density; it decreases with photon energy as a power-law function.
- *Phonon Transitions.* Long-wavelength photons can release their energy by directly exciting lattice vibrations, i.e., by creating phonons.
- *Excitonic Transitions.* The absorption of a photon can result in the formation of an **exciton**. This entity is much like a hydrogen atom in which a hole plays the role of the proton. The hole and electron are bound together by their mutual Coulomb interaction. A photon may be emitted as a result of the electron and hole recombining, thereby annihilating the exciton.

These transitions all contribute to the overall absorption coefficient, which is displayed in Fig. 16.2-2 for Si and GaAs, and at greater magnification in Fig. 16.2-3 for a number of semiconductor materials. For photon energies greater than the bandgap energy E_g, the absorption is dominated by band-to-band transitions that form the basis of many photonic devices. The spectral region where the material changes from being relatively transparent ($h\nu < E_g$) to strongly absorbing ($h\nu > E_g$) is known as the **absorption edge**. Direct-bandgap semiconductors have a more abrupt absorption edge than indirect-bandgap materials, as is apparent from Figs. 16.2-2 and 16.2-3.

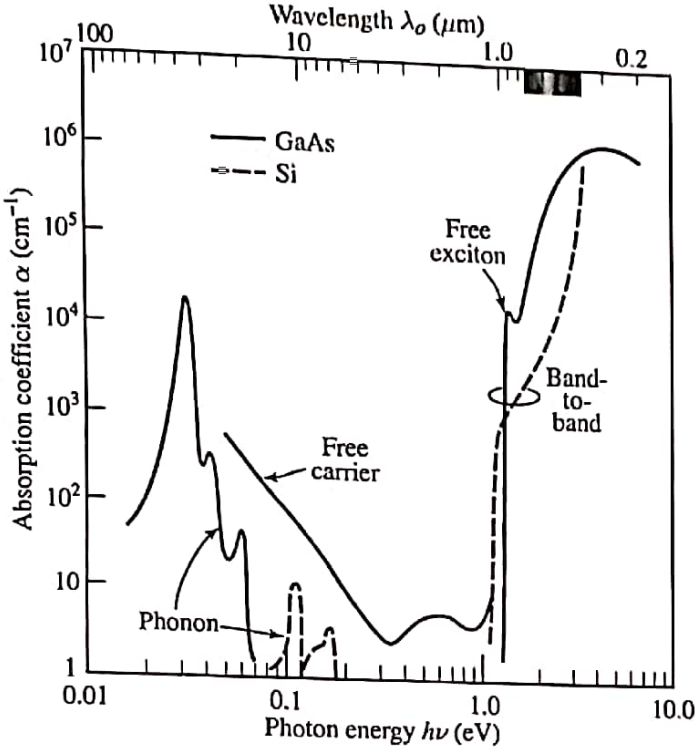


Figure 16.2-2 Observed optical absorption coefficient α versus photon energy and wavelength for Si and GaAs in thermal equilibrium at $T = 300°$ K. The bandgap energy E_g is 1.12 eV for Si and 1.42 eV for GaAs. Silicon is relatively transparent in the band $\lambda_o \approx 1.1$ to 12 μm, whereas intrinsic GaAs is relatively transparent in the band $\lambda_o \approx 0.87$ to 12 μm (see Fig. 5.5-1).

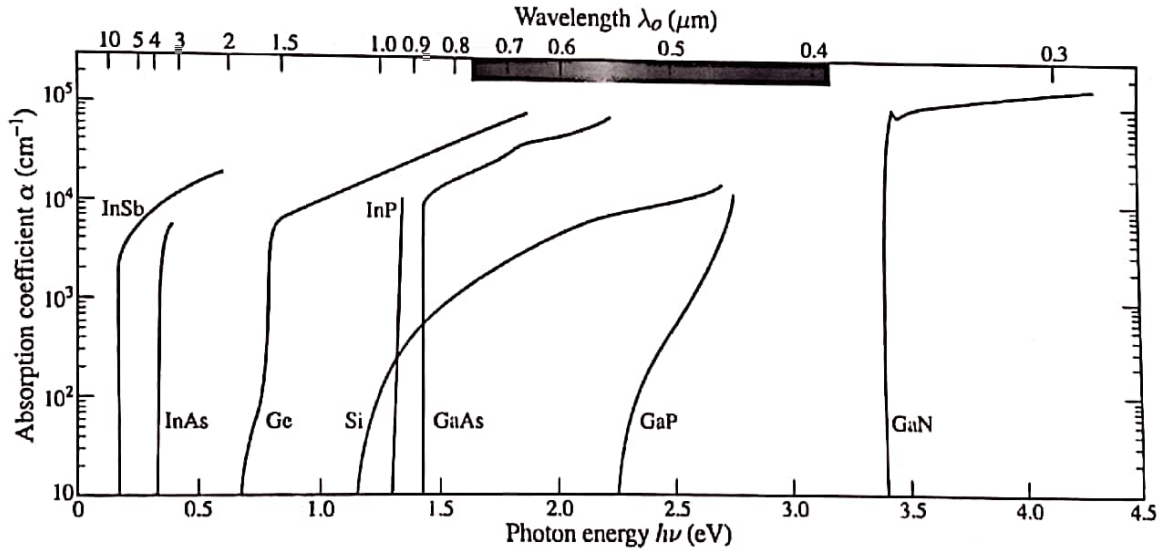


Figure 16.2-3 Absorption coefficient versus photon energy and wavelength for Ge, Si, GaAs, GaN and selected other III–V binary semiconductors at $T = 300°$ K, on an expanded scale.

B. Band-to-Band Transitions in Bulk Semiconductors

We proceed to develop a simple theory of direct band-to-band photon absorption and emission in bulk semiconductors, ignoring the other types of transitions.

Bandgap Wavelength

Direct band-to-band absorption and emission can take place only at frequencies for which the photon energy $h\nu > E_g$. The minimum frequency ν necessary for this to occur is $\nu_g = E_g/h$, so that the corresponding maximum wavelength is $\lambda_g = c_o/\nu_g = hc_o/E_g$. If the bandgap energy is given in eV (rather than in J), the bandgap wavelength

$\lambda_g = hc_o/eE_g$ in μm turns out to be

$$\lambda_g \approx \frac{1.24}{E_g}. \qquad (16.2\text{-}1)$$

Bandgap Wavelength
λ_g (μm) and E_g (eV)

The quantity λ_g is known as the **bandgap wavelength** (or **cutoff wavelength**).

The bandgap wavelength λ_g, and its associated bandgap energy E_g, are provided in Table 16.1-2, and in Figs. 16.1-7 and 16.1-8, for a number of semiconductor materials of importance in photonics. III–V ternary and quaternary semiconductors of different compositions span a substantial range of bandgap wavelengths, from the mid-infrared to the mid-ultraviolet, as is evident in Fig. 16.2-4.

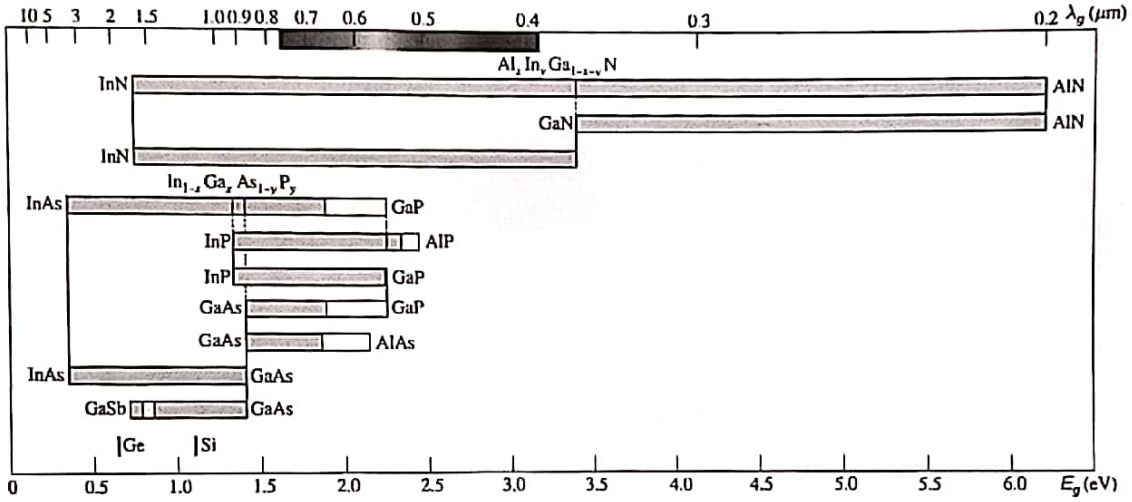


Figure 16.2-4 Bandgap wavelength λ_g, and corresponding bandgap energy E_g, for selected elemental and III–V binary, ternary, and quaternary semiconductor materials. Successive rows, starting at the top, represent AlInGaN, AlGaN, InGaN, InGaAsP, AlInGaP, InGaP, GaAsP, AlGaAs, InGaAs, and GaAsSb. The shaded regions indicate compositions for which the materials are direct-bandgap semiconductors.

Conditions for Absorption and Emission

Electron excitation from the valence to the conduction band may be induced by the absorption of a photon of appropriate energy ($h\nu > E_g$ or $\lambda < \lambda_g$). An electron–hole pair is generated [Fig. 16.2-5(*a*)]. This adds to the concentration of mobile charge carriers and increases the conductivity of the material. The material behaves as a photoconductor with a conductivity proportional to the photon flux. This effect is used to detect light, as discussed in Chapter 18.

Electron deexcitation from the conduction to the valence band (electron–hole recombination) may result in the spontaneous emission of a photon of energy $h\nu > E_g$ [Fig. 16.2-5(*b*)], or in the stimulated emission of a photon [Fig. 16.2-5(*c*)], provided that a photon of energy $h\nu > E_g$ is initially present (see Sec. 13.3). Spontaneous emission is the underlying phenomenon on which the light-emitting diode is based, as will be seen in Sec. 17.1. Stimulated emission is responsible for the operation of semiconductor optical amplifiers and laser diodes, as will be seen in Secs. 17.2, 17.3, and 17.4.

The conditions under which absorption and emission take place are summarized as follows:

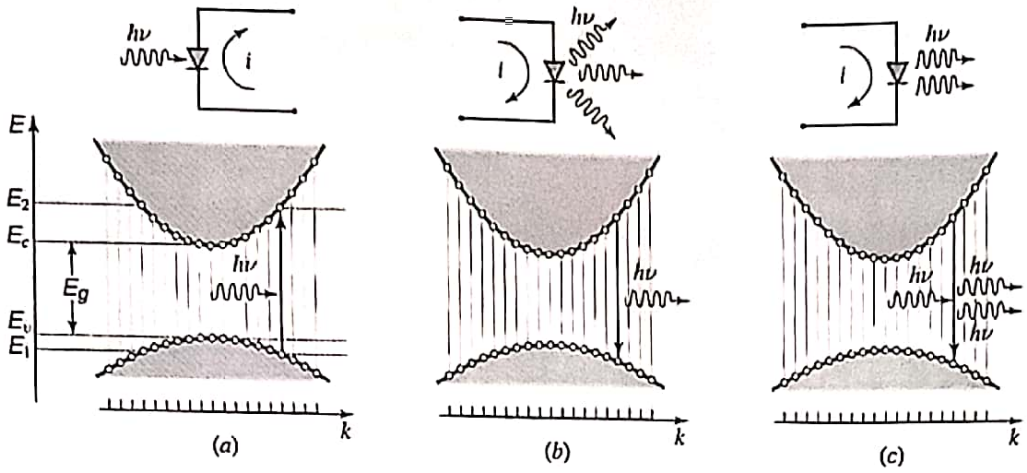


Figure 16.2-5 (*a*) The absorption of a photon results in the generation of an electron–hole pair. This process is used in the photodetection of light. (*b*) The recombination of an electron–hole pair results in the spontaneous emission of a photon. Light-emitting diodes (LEDs) operate on this basis. (*c*) Electron–hole recombination can be induced by a photon. The result is the stimulated emission of an identical photon. This is the underlying process responsible for the operation of semiconductor laser diodes.

- *Conservation of Energy.* The absorption or emission of a photon of energy $h\nu$ requires that the energies of the two states involved in the interaction (say E_1 and E_2 in the valence band and conduction band, respectively, as depicted in Fig. 16.2-5) be separated by $h\nu$. Thus, for photon emission to occur by electron–hole recombination, for example, an electron occupying an energy level E_2 must interact with a hole occupying an energy level E_1, such that energy is conserved, i.e.,

$$E_2 - E_1 = h\nu. \tag{16.2-2}$$

- *Conservation of Momentum.* Momentum must also be conserved in the process of photon emission/absorption, so that $p_2 - p_1 = h\nu/c = h/\lambda$, or $k_2 - k_1 = 2\pi/\lambda$. The photon-momentum magnitude h/λ is, however, very small in comparison with the range of momentum values that electrons and holes can assume. The semiconductor E–k diagram extends to values of k of the order $2\pi/a$, where the lattice constant a is much smaller than the wavelength λ, so that $2\pi/\lambda \ll 2\pi/a$. The momenta of the electron and the hole participating in the interaction must therefore be approximately equal. This condition, $k_2 \approx k_1$, is called the **k-selection rule**. Transitions that obey this rule are represented in the E–k diagram (Fig. 16.2-5) by vertical lines, indicating that the change in k is negligible on the scale of the diagram.
- *Energies and Momenta of the Electron and Hole with Which a Photon Interacts.* As is apparent from Fig. 16.2-5, conservation of energy and momentum require that a photon of frequency ν interact with electrons and holes of specific energies and momenta determined by the semiconductor E–k relation. Using (16.1-3) and (16.1-4) to approximate this relation for a direct-bandgap semiconductor by two parabolas, and writing $E_c - E_v = E_g$, (16.2-2) may be written in the form

$$E_2 - E_1 = \frac{\hbar^2 k^2}{2m_v} + E_g + \frac{\hbar^2 k^2}{2m_c} = h\nu, \tag{16.2-3}$$

from which

$$k^2 = \frac{2m_r}{\hbar^2}(h\nu - E_g), \tag{16.2-4}$$

where

$$\frac{1}{m_r} = \frac{1}{m_v} + \frac{1}{m_c}. \tag{16.2-5}$$

Substituting (16.2-4) into (16.1-3) provides that the energy levels E_1 and E_2 with which the photon interacts are

$$E_2 = E_c + \frac{m_r}{m_c}(h\nu - E_g) \tag{16.2-6}$$

$$E_1 = E_v - \frac{m_r}{m_v}(h\nu - E_g) = E_2 - h\nu. \tag{16.2-7}$$

In the special case when $m_c = m_v$, we obtain $E_2 = E_c + \frac{1}{2}(h\nu - E_g)$, as required by symmetry.

- *Optical Joint Density of States.* We now determine the density of states $\varrho(\nu)$ with which a photon of energy $h\nu$ interacts under conditions of energy and momentum conservation in a direct-bandgap semiconductor. This quantity incorporates the density of states in both the conduction and valence bands and is called the **optical joint density of states.** The one-to-one correspondence between E_2 and ν embodied in (16.2-6) permits us to readily relate $\varrho(\nu)$ to the density of states $\varrho_c(E_2)$ in the conduction band by use of the incremental relation $\varrho_c(E_2)\,dE_2 = \varrho(\nu)\,d\nu$, from which $\varrho(\nu) = (dE_2/d\nu)\varrho_c(E_2)$, so that

$$\varrho(\nu) = \frac{hm_r}{m_c}\varrho_c(E_2). \tag{16.2-8}$$

Using (16.1-7) and (16.2-6), we finally obtain the number of states per unit volume per unit frequency:

$$\varrho(\nu) = \frac{(2m_r)^{3/2}}{\pi\hbar^2}\sqrt{h\nu - E_g}, \qquad h\nu \geq E_g, \tag{16.2-9}$$

(16.2-9) Optical Joint Density of States

which is illustrated in Fig. 16.2-6. The one-to-one correspondence between E_1 and ν in (16.2-7), together with $\varrho_v(E_1)$ from (16.1-8), results in an expression for $\varrho(\nu)$ identical to (16.2-9).

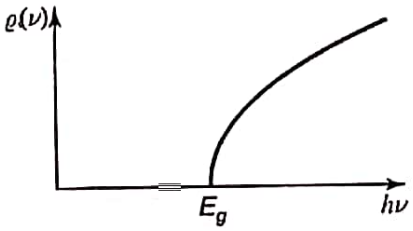


Figure 16.2-6 The density of states with which a photon of energy $h\nu$ interacts increases with $h\nu - E_g$ in accordance with a square-root law.

- *Photon Emission Is Unlikely in an Indirect-Bandgap Semiconductor.* Radiative electron–hole recombination is unlikely in an indirect-bandgap semiconductor. This is because transitions from near the bottom of the conduction band to near the top of the valence band (where electrons and holes, respectively, are most likely

to reside) requires an exchange of momentum that cannot be accommodated by the emitted photon. Momentum may be conserved, however, by the participation of phonons in the interaction. Phonons can carry relatively large momenta but typically have small energies (≈ 0.01–0.1 eV; see Fig. 16.2-2), so their transitions appear horizontal on the E–k diagram (see Fig. 16.2-7). The net result is that momentum is conserved, but the k-selection rule is violated. Because phonon-assisted emission involves the participation of three bodies (electron, photon, and phonon), the probability of its occurrence is quite low. Thus, Si, which is an indirect-bandgap semiconductor, has a substantially lower radiative recombination coefficient than does GaAs, which is a direct-bandgap semiconductor (see Table 16.1-4). Silicon is therefore not an efficient light emitter, whereas GaAs is.

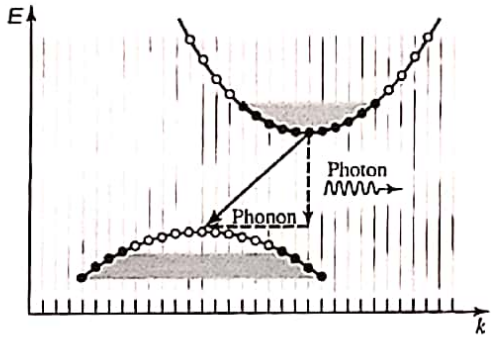


Figure 16.2-7 Photon emission in an indirect-bandgap semiconductor. The recombination of an electron near the bottom of the conduction band with a hole near the top of the valence band requires the exchange of energy *and* momentum. The energy may be carried off by a photon, but one or more phonons are also required to conserve momentum. This type of multiparticle interaction is therefore unlikely.

- *Photon Absorption Is Not Unlikely in an Indirect-Bandgap Semiconductor.* Although photon absorption also requires energy and momentum conservation in an indirect-bandgap semiconductor, this is readily achieved by means of a two-step process (Fig. 16.2-8). The electron is first excited to a high energy level within the conduction band by a k-conserving vertical transition. It then quickly relaxes to the bottom of the conduction band by a process called thermalization in which its momentum is transferred to phonons. The generated hole behaves similarly. Since the process occurs sequentially, it does not require the simultaneous presence of three bodies and is thus not unlikely. Silicon is therefore an efficient photon detector, as is GaAs.

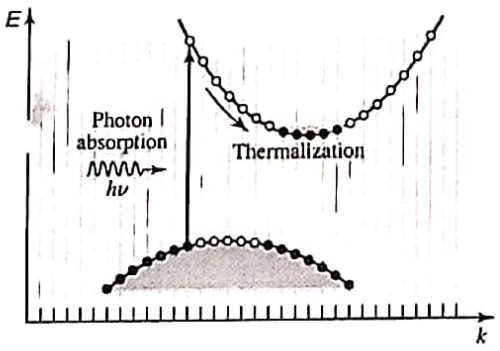


Figure 16.2-8 Photon absorption in an indirect-bandgap semiconductor via a vertical (k-conserving) transition. The photon generates an excited electron in the conduction band, leaving behind a hole in the valence band. The electron and hole then undergo fast transitions — to the lowest and highest possible levels in the conduction and valence bands, respectively, releasing their energy in the form of phonons. Since the process is sequential it is not unlikely.

C. Absorption, Emission, and Gain in Bulk Semiconductors

We now proceed to determine the probability densities of a photon of energy $h\nu$ being emitted or absorbed by a bulk semiconductor material in a direct band-to-band

3.1 ELECTRON-HOLE PAIR FORMATION AND RECOMBINATION

The operation of almost all optoelectronic devices is based on the creation or annihilation of electron-hole pairs. Pair formation essentially involves raising an electron in energy from the valence band to the conduction band, thereby leaving a hole behind in the valence band. In principle, any energetic particle incident on a semiconductor, which can impart an energy at least equal to the bandgap energy to a valence band electron, will create pairs. With respect to the bonding in the lattice, this process is equivalent to breaking a covalent bond. The simplest way to create electron-hole pairs is to irradiate the semiconductor. Photons with sufficient energy are absorbed, and these impart their energy to the valence band electrons and raise them to the conduction band. This process is, therefore, also called *absorption*. The reverse process, that of electron and hole recombination, is associated with the pair giving up its excess energy. Recombination may be *radiative* or *nonradiative*. In a nonradiative transition, the excess energy due to recombination is usually imparted to phonons and dissipated as heat. In a radiative transition, the excess energy is dissipated as photons, usually having energy equal to the bandgap (i.e., $\hbar\omega = \mathcal{E}_g$). This is the *luminescent* process, which is classified according to the method by which the electron-hole pairs are created. *Photoluminescence* involves the radiative recombination of electron-hole pairs created by injection of photons. *Cathodoluminescence* is the process of radiative recombination of electron-hole pairs created by electron bombardment. *Electroluminescence* is the process of radiative recombination following injection with a p-n junction or similar device.

In a semiconductor in equilibrium (i.e., without any incident photons or injection of electrons), the carrier densities can be calculated from an equilibrium Fermi level by using Fermi-Dirac or Boltzmann statistics outlined in Sec. 2.5.3. When excess carriers are created by one of the techniques described above, nonequilibrium conditions are generated and the concept of a Fermi level is no longer valid. One can, however, define nonequilibrium distribution functions for electrons and holes as

$$f_n(\mathcal{E}) = \frac{1}{1 + \exp\left(\frac{\mathcal{E} - \mathcal{E}_{fn}}{k_B T}\right)} \tag{3.1}$$

$$1 - f_p(\mathcal{E}) = \frac{1}{1 + \exp\left(\frac{\mathcal{E} - \mathcal{E}_{fp}}{k_B T}\right)} \tag{3.2}$$

These distribution functions define $\mathcal{E}_{fn}$ and $\mathcal{E}_{fp}$, the *quasi-Fermi levels* for electrons and holes, respectively. In some texts they are referred to as IMREFs (Fermi spelled backward). When the excitation source creating excess carriers is removed, $\mathcal{E}_{fn} = \mathcal{E}_{fp} = \mathcal{E}_F$. The difference $(\mathcal{E}_{fn} - \mathcal{E}_{fp})$ is a measure of the deviation from equilibrium. As with equilibrium statistics, we obtain for the nondegenerate case

$$f_n(\mathcal{E}) \cong \exp\left(\frac{\mathcal{E}_{fn} - \mathcal{E}}{k_B T}\right) \tag{3.3}$$

$$f_p(\mathcal{E}) \cong \exp\left(\frac{\mathcal{E} - \mathcal{E}_{fp}}{k_B T}\right) \tag{3.4}$$

and the nonequilibrium carrier concentrations are given by

$$n = N_C \exp\left(\frac{\mathcal{E}_{fn} - \mathcal{E}_C}{k_B T}\right) \tag{3.5}$$

$$p = N_V \exp\left(\frac{\mathcal{E}_V - \mathcal{E}_{fp}}{k_B T}\right) \tag{3.6}$$

The concept of quasi-Fermi levels is extremely useful, since it provides a means to take into account changes of carrier concentration as a function of position in a semiconductor. As we shall see in Chapter 4, in a p-n junction under forward bias a large density of excess carriers exist in the depletion region and close to it on either side. The concentration of these carriers can be determined from the appropriate quasi-Fermi levels. A junction laser is operated under such forward bias injection conditions to create population inversion. To consider a simple example, assume that an n-type semiconductor with an equilibrium electron density n_o ($\equiv N_D$, the donor density) is uniformly irradiated with intrinsic photoexcitation (above-bandgap light) so as to produce Δn electron-hole pairs with a generation rate G. The nonequilibrium electron and hole concentrations are given by

$$n = \Delta n + n_o \tag{3.7}$$

$$p = \Delta n + n_i^2/n_o \tag{3.8}$$

Using Eqs. 3.3–3.8, Fig. 3.1 illustrates the change in the energy position of the quasi-Fermi levels in GaAs as the generation rate changes.

The excess carriers created in a semiconductor must eventually recombine. In fact, under steady-state conditions the recombination rate must be equal to the generation rate

$$G = R \tag{3.9}$$

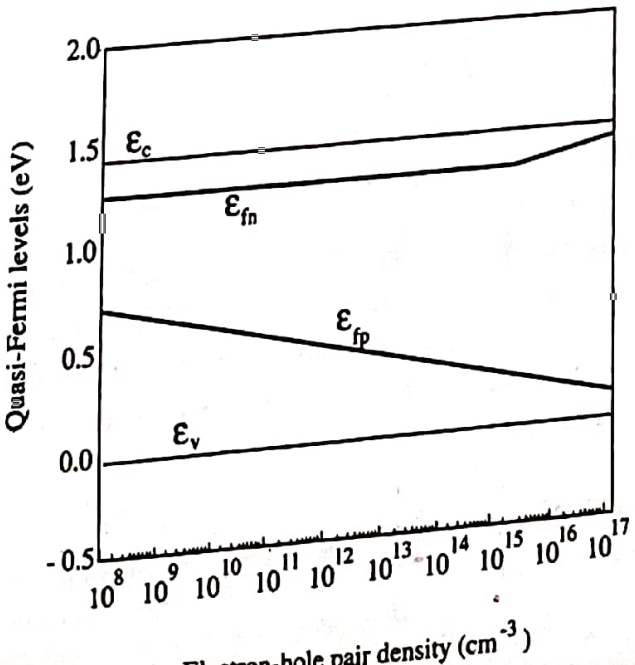
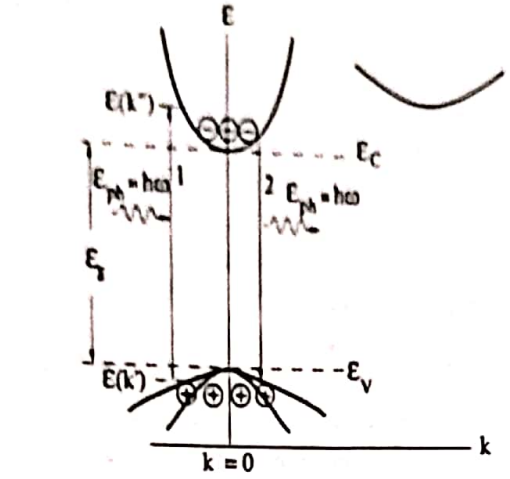
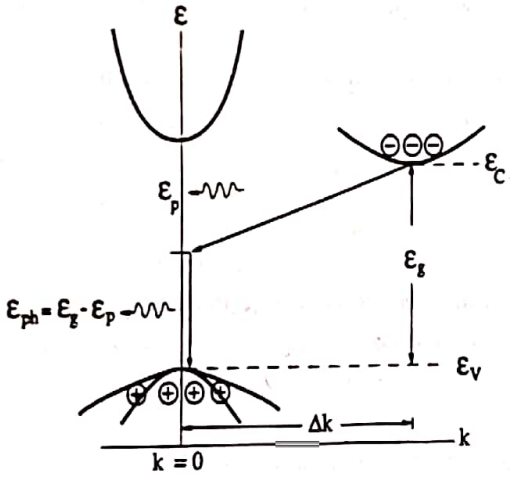


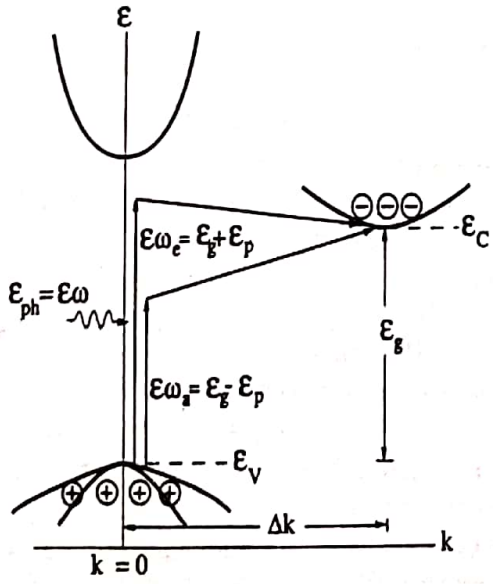
Figure 3.1 Energy position of the electron and hole quasi-Fermi levels as a function of pair density in GaAs at room temperature. It is assumed that the sample is n-type with $N_D = 10^{15}\ cm^{-3}$ (from M. Shur, *Physics of Semiconductor Devices*, © 1990. Reprinted by permission of Prentice Hall, Englewood Cliffs, New Jersey).



(a)



(b)



(c)

Figure 3.2 Illustration of band-to-band of absorption and recombination processes in (a) direct bandgap semiconductor and (b) and (c) indirect bandgap semiconductor.

Generation and recombination processes involve transition of carriers across the energy bandgap and are therefore different for direct and indirect bandgap semiconductors, as illustrated in Fig. 3.2. In a direct bandgap semiconductor, as shown in Fig. 3.2(a), the valence band maximum and the conduction band minimum occur at the zone center ($\mathbf{k} = 0$) and an upward or downward transition of electrons does not require a change in momentum or the involvement of a phonon. Therefore, in direct bandgap semiconductors such as GaAs, an electron raised to the conduction band, say, by photon absorption, will dwell there for a very short time and recombine again with a valence band hole to emit light of energy equal to the bandgap. Thus, the probability of *radiative recombination* is very high in direct bandgap semiconductors. The processes are quite different in an indirect bandgap semiconductor. Considering the band diagrams shown in Fig. 3.2(b) and (c), since the conduction band minima are not at $\mathbf{k} = 0$, upward or downward transition of carriers requires a change in momentum, or the involvement of a phonon. Thus, an electron dwelling in the conduction band minimum, at $\mathbf{k} \neq 0$, cannot recombine with a hole at $\mathbf{k} = 0$ until a phonon with the right energy and momentum is available. Both phonon emission and absorption processes can assist the downward transition. In order for the right phonon collision to occur, the dwell time of the electron in the conduction band increases. Since no crystal is perfect, there are impurities and defects in the lattice that manifest themselves as traps and recombination centers. It is most likely that the electron and hole will recombine nonradiatively through such a defect center, and the excess energy is dissipated into the lattice as heat. The competing nonradiative processes reduce the probability of radiative recombination in indirect bandgap materials such as Si, Ge, or GaP. These semiconductors are therefore, in general, not suitable for the realization of light sources such as light-emitting diodes and lasers.

3.1.1 Radiative and Nonradiative Recombination

For continuous carrier generation by optical excitation or injection, a quasi equilibrium or steady state is produced. Electrons and holes are created and annihilated in pairs and, depending on the injection level, a steady-state excess density $\Delta n = \Delta p$ is established in the crystal. This equality is also necessary for the maintenance of overall charge neutrality. When the excitation source is removed, the density of excess carriers returns to the equilibrium values, n_o and p_o. The decay of excess carriers usually follows an exponential law with respect to time $\sim \exp(-t/\tau)$, where τ is defined as the lifetime of excess carriers. The lifetime is determined by a combination of intrinsic and extrinsic parameters, and the performance characteristics of most optoelectronic devices depend on it. In the discussion that follows, we will be concerned mainly with *bulk* recombination processes. It is important to remember that, depending on the semiconductor sample and its surface, there can be a very strong *surface* recombination component which depends on the density of surface states.

In general, the excess carriers decay by radiative and/or nonradiative recombination, in which the excess energy is dissipated by photons and phonons. The former is of importance for the operation of luminescent devices. Nonradiative recombination usually takes place via surface or bulk defects and traps (Fig. 3.3), as discussed in Chapter 2, and reduces the radiative efficiency of the material. Therefore the total lifetime τ can be expressed as

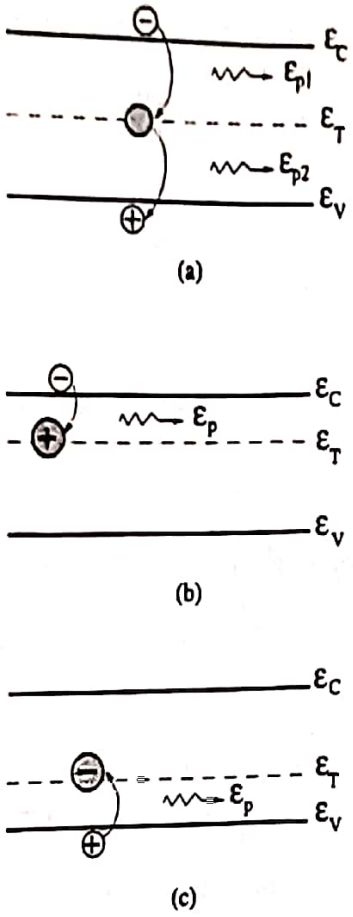


Figure 3.3 Nonradiative recombination at (a) recombination center, (b) electron trap, and (c) hole trap. The excess carrier energy in all cases is dissipated by single or multiple phonons.

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \tag{3.10}$$

where τ_r and τ_{nr} are the radiative and nonradiative lifetimes, respectively. Also, the total recombination rate R_{total} is given by

$$R_{total} = R_r + R_{nr} = R_{sp} \tag{3.11}$$

where R_r and R_{nr} are radiative and nonradiative recombination rates per unit volume, respectively, and R_{sp} is called the *spontaneous* recombination rate, to distinguish R_{total} from the *stimulated* recombination rate to be defined later in Chapter 6. The internal quantum efficiency or radiative recombination efficiency is defined as

$$\eta_r = \frac{R}{R_r + R_{nr}} \tag{3.12}$$

For an exponential decay process, $\tau_r = \Delta n/R_r$ and $\tau_{nr} = \Delta n/R_{nr}$ where Δn is the excess electron concentration. Therefore,

$$\eta_r = \frac{1}{1 + \tau_r/\tau_{nr}} \tag{3.13}$$

To achieve high internal quantum efficiency, the ratio τ_r/τ_{nr} should be as small as possible, or τ_{nr} should be as large as possible. The value of τ_{nr} is controlled by the properties of defects, which produce levels in the bandgap of a semiconductor. The excess

energy of carriers recombining at these levels is dissipated by phonons. Another nonradiative process is *Auger recombination*, to be discussed in Sec. 3.8. It also follows from Eq. 3.9 that under steady-state conditions $\Delta n = G\tau_r$.

3.1.2 Band-to-Band Recombination

The simplest carrier decay process is spontaneous band-to-band recombination, whose rate, without momentum conservation, is given by

$$R_{sp} = B_r np \tag{3.14}$$

where B_r is defined as the coefficient for band-to-band recombination in units of $cm^3.s^{-1}$. B_r is related to the transition probability P to be discussed in the next section. In terms of the equilibrium and excess carrier densities,

$$R_{sp} = B_r(n_o + \Delta n)(p_o + \Delta p) \tag{3.15}$$

where $\Delta n = \Delta p$. The spontaneous radiative recombination rate for excess carriers can be expressed as

$$R_{sp}^{ex} = \frac{\Delta n}{\tau_r} \tag{3.16}$$

and therefore

$$R_{sp} = R_{sp}^{o} + R_{sp}^{ex} \tag{3.17}$$

where

$$R_{sp}^{o} = B_r n_o p_o \tag{3.18}$$

is the spontaneous recombination rate in thermal equilibrium. From Eqs. 3.15 and 3.18,

$$R_{sp} = B_r[n_o p_o + \Delta n(n_o + p_o) + (\Delta n^2)] \tag{3.19}$$

$$R_{sp}^{ex} = B_r \Delta n[n_o + p_o + \Delta n] \tag{3.20}$$

and

$$\tau_r = \frac{1}{B_r(n_o + p_o + \Delta n)} \tag{3.21}$$

When $\Delta n \gg n_o, p_o$, which is relevant to laser operation,

$$\tau_r \cong \frac{1}{B_r \Delta n} \tag{3.22}$$

This is the *bimolecular* recombination regime, when the lifetime changes with Δn. At low injection levels, such that $\Delta n < n_o, p_o$

$$\tau_r \cong \frac{1}{B_r(n_o + p_o)} \tag{3.23}$$

which remains constant, being determined by the background carrier concentrations. For an intrinsic semiconductor under low-level injection, since $n_o = p_o = n_i$,

$$\tau_r = \frac{1}{2B_r n_i} \tag{3.24}$$

Eq. 3.22 is valid for $10^{17} < \Delta n \leq 10^{18}$ cm^{-3}. For higher values of Δn,

$$\tau_r \cong \tau_o \tag{3.25}$$

which is usually constant for any material. For example for GaAs, $\tau_o \cong 0.5$ ns.

The value of the recombination coefficient depends on the bandgap and whether the semiconductor has a direct or an indirect bandgap. Direct bandgap semiconductors usually have values of B_r ranging from 10^{-11} to 10^{-9} cm^3.s^{-1} and indirect bandgap semiconductors have values of B_r ranging from 10^{-15} to 10^{-13} cm^3.s^{-1}.

EXAMPLE 3.1

Objective. To calculate τ_r in GaAs having $n_o = 10^{14}$ cm^{-3} under high- and low-level injections for $B_r = 7 \times 10^{-10}$ cm^3/s.

At a high injection level of 10^{18} cm^{-3}, $\tau_r = (7 \times 10^{-10} \times 10^{18})^{-1} s = 1.43$ns.

At a low injection level of 10^{16} cm^{-3}, $\tau_r = 143$ ns for the same value of B_r. This value of τ_r is almost an order of magnitude larger than that measured in pure GaAs samples. Therefore, the value of B_r used here is only valid for the high-level injection case and should be larger for the case of low-level injection.

3.2 ABSORPTION IN SEMICONDUCTORS

3.2.1 Matrix Elements and Oscillator Strength for Band-to-Band Transitions

The operation of optical devices that we will describe and discuss in this text depends on the upward and downward transitions of carriers between energy bands. These transitions result in absorption or emission of light, which is electromagnetic energy. The measurement of absorption and emission spectra in semiconductors constitutes an important aspect of materials characterization. They not only provide information on the bandgap, but the measurements also provide information on direct and indirect transitions, the distribution of states, and the energy position of defect and impurity levels. The absorption spectrum spans a wide energy (or wavelength) range, extending from the near-bandgap energies to the low-energy transitions involving free carriers and lattice vibrations. In the context of this text the more important ones are the near-bandgap transitions.

The process of photon absorption results in the transition of an electron from a lower energy state to a higher energy state, the simplest form of which may be a direct transition from the valence to the conduction band. The different possible transitions are outlined in this chapter. In what follows, the process of band-to-band transitions in semiconductors, in which photons are absorbed, is analyzed.

The energy–momentum diagrams of a direct and an indirect semiconductor were shown in Figs. 3.2(a) and (b), respectively. Considering the case of an electron raised from the top of the valence band to the bottom of the conduction band due to absorption of a photon in a direct transition, there is no change in momentum. Strictly, there

is a small change in k due to the finite momentum of the photon, which is equal to h/λ. For most III–V semiconductors $\lambda \sim 1\ \mu$m, and the resultant momentum change is very small. An indirect transition due to the absorption of a photon is illustrated in Fig. 3.2(c). Since a large change in momentum is involved in this case, the transition can occur only by the emission or absorption of a phonon. The process can be described by the equation

$$\mathcal{E}_{ph} \pm \mathcal{E}_p = \mathcal{E}_g \tag{3.26}$$

and the change in momentum is given by

$$\Delta \mathbf{k} = \mathbf{k}_p \tag{3.27}$$

where $\mathbf{k}_p$ is the wavevector of the phonon and $\mathcal{E}_{ph}$ and $\mathcal{E}_p$ are the photon and phonon energies, respectively. Therefore, an optical or acoustic phonon with the right energy and momentum must be involved in an indirect transition.

EXAMPLE 3.2

Objective. To calculate the momentum change due to photon absorption in InP.

InP is a direct bandgap semiconductor with $\mathcal{E}_g = 1.35$ eV at room temperature. The corresponding wavelength is $\lambda_g = 0.92\ \mu$m. Therefore the momentum of the absorbed photon is

$$k_{ph} = \frac{2\pi}{\lambda_{ph}} \cong \frac{2\pi}{\lambda_g} = 6.83 \times 10^4\ \text{cm}^{-1}$$

which is negligible on the scale of momentum used in describing the bandstructure of a semiconductor (see Fig. 1.1).

The wavelength dependence of direct transitions, for the case of absorption, is illustrated in Fig. 3.2(a), where we consider a transition away from the zone center. The top of the valence band is taken as the zero of energy. The transition occurs in energy from $\mathcal{E}(\mathbf{k}')$ to $\mathcal{E}(\mathbf{k}'')$ where

$$\mathcal{E}(\mathbf{k}') = -\frac{\hbar^2 \mathbf{k}'^2}{2m_h^*} \tag{3.28}$$

and

$$\mathcal{E}(\mathbf{k}'') = \mathcal{E}_g + \frac{\hbar^2 \mathbf{k}''^2}{2m_e^*} \tag{3.29}$$

Here $\mathcal{E}_g$ is the direct bandgap at $\mathbf{k} = 0$. The energy of the absorbed photon is $\mathcal{E}_{ph} = \mathcal{E}(\mathbf{k}'') - \mathcal{E}(\mathbf{k}')$, which is the requirement for energy conservation. Remember that such a transition will take place only if the level at $\mathcal{E}(\mathbf{k}')$ is filled and that at $\mathcal{E}(\mathbf{k}'')$ is empty. Also, for momentum conservation $\mathbf{k}'$ must be nearly equal to $\mathbf{k}''$. This is called the *k-selection rule.*

An important, and often overlooked, point regarding photon absorption is that if a photon with energy larger than the bandgap energy is absorbed in a semiconductor, the electron and hole are generally not created with the same energy. Usually the electron, with its lower mass in compound semiconductors, is created with a larger energy than the hole. Example 3.3 below illustrates the point.

EXAMPLE 3.3

Objective. To calculate the energy of the electron and heavy hole produced by absorbing a 1.5 eV photon in InP.

From Eqs. 3.28 and 3.29 and knowing that the reduced effective mass associated with an optical transition is given by

$$\frac{1}{m_r^*} = \frac{1}{m_e^*} + \frac{1}{m_h^*}$$

we get

$$\mathcal{E}(\mathbf{k}'') = \mathcal{E}_g + \frac{m_r^*}{m_e^*}(\mathcal{E} - \mathcal{E}_g)$$

and

$$\mathcal{E}(\mathbf{k}') = -\frac{m_r^*}{m_{hh}^*}(\mathcal{E} - \mathcal{E}_g)$$

where $\mathcal{E}$ is the energy of the photon absorbed and it is assumed that $\mathbf{k}' = \mathbf{k}'' = \mathbf{k}$. For InP $\mathcal{E}_g$ = 1.35 eV, $m_e^* = 0.082m_0$, $m_{hh}^* = .085m_0$ and $m_r^* = 0.075m_o$. In a direct transition the reduced effective mass and the joint density of states, which we will study in Chapter 5, are the consequence of equal k-values of the electron and hole involved in the transition. We therefore get $\mathcal{E}(\mathbf{k}'') - \mathcal{E}_g = 0.137\ eV$ and $\mathcal{E}(\mathbf{k}') = -0.013\ eV$.

The matrix element and probability of an optical transition from x$\mathcal{E}(\mathbf{k}')$ to $\mathcal{E}(\mathbf{k}'')$ can be calculated by considering first-order time-dependent perturbation theory. The time-independent form of the Schrödinger equation is

$$\mathbf{H}_0\Psi = \mathcal{E}(\mathbf{k}')\Psi \tag{3.30}$$

where $\mathbf{H}_0$ is the Hamiltonian of the unperturbed system. In the case of a perturbation $\mathbf{H}_1$, which in our context is light or electromagnetic radiation, causing a carrier transition from a state at $\mathcal{E}(\mathbf{k}')$to a state at $\mathcal{E}(\mathbf{k}'')$, the time-dependent Schrödinger equation can be expressed as

$$(\mathbf{H}_0 + \mathbf{H}_1)\Psi = j\hbar\frac{d\Psi}{dt} \tag{3.31}$$

with

$$\Psi = \sum_m A_m(t)\Psi_m e^{-j\mathcal{E}_m(\mathbf{k}')t/\hbar} \tag{3.32}$$

It may be noted that $|A_m(t)|^2$ is the transition probability. The calculation of the matrix element of an optical transition has been described in detail in a few texts and is not repeated here. The matrix element for direct transitions, where the condition

$$\mathbf{k}'' - \mathbf{k}' = \mathbf{k} \cong 0 \tag{3.33}$$

is satisfied, is given by†

†R. H. Bube, *Electronic Properties of Crystalline Solids*, Academic Press, New York, 1974.

$$H_{i'i} = \int \Psi_{i'}^{*} H_1 \Psi_i \, d\mathbf{r}$$

$$= \frac{jq\hbar A}{2m_0} \int u_C^{*}(\mathbf{r}, \mathbf{k}'')[\mathbf{a}_0 \cdot \nabla u_V(\mathbf{r}, \mathbf{k}') + j(\mathbf{a}_0 \cdot \mathbf{k}')u_V(\mathbf{r}, \mathbf{k}')]d\mathbf{r} \qquad (3.34)$$

where u_V and u_C are the Bloch functions corresponding to the valence and conduction bands, respectively, A is the magnetic vector potential of the electromagnetic wave and $\mathbf{a}_0$ is a polarization unit vector. The first term represents the matrix element for *allowed* direct transitions and is usually much larger than the second term, which represents *forbidden* transitions. If $\mathbf{k}' = \mathbf{k}''$, the matrix element of the forbidden transition is zero. However, because of the small change in momentum due to the small but finite momentum of the photon, the matrix element of the forbidden transition has a finite value. It is shown in Appendix 4 that the transition probability per unit volume per unit time for an allowed direct transition is given by

$$P(\hbar\omega) = \frac{q^2|\mathbf{A}|^2(2m_r^{*})^{3/2}p_{CV}^2}{4\pi m_0^2\hbar^4}(\hbar\omega - \mathcal{E}_g)^{1/2} \qquad (3.35)$$

where p_{CV} is the matrix element of the momentum operator (or the momentum matrix element) and m_r^{*} is the reduced mass given by

$$m_r^{*} = \frac{m_e^{*}m_h^{*}}{m_e^{*} + m_h^{*}} \qquad (3.36)$$

Equation 3.35 contains the *joint density of states* to be discussed in Chapter 5. When the k-selection rule is obeyed, $|p_{CV}|^2 = 0$ unless $\mathbf{k}' = \mathbf{k}''$. Equation 3.35 is an important relationship and shows that the transition probability of a direct allowed transition varies as $(\hbar\omega - \mathcal{E}_g)^{1/2}$. The transition probability includes the summation over all filled valence band states and empty conduction band states and over all $\mathbf{k}'$ and $\mathbf{k}''$ values that satisfy energy and momentum conservation. In the form expressed in Eq. 3.35, it is assumed that the semiconductor is at 0°K when the valence band is completely filled and the conduction band is empty.

Since the absorption of a photon of energy $\hbar\omega$ is involved in a direct transition, it is important to calculate the absorption coefficient α. Assume that a monochromatic photon flux $\mathfrak{I}_i$, given by

$$\mathfrak{I}_i = \frac{|\mathbf{S}|}{\hbar\omega} \quad (\text{photons/cm}^2\text{.s}) \qquad (3.37)$$

is incident on the crystal. Here |S| is the radiation energy crossing unit area in unit time, or the *Poynting vector*. The transmitted intensity $\mathfrak{I}_d$ is then

$$\mathfrak{I}_d = \frac{|\mathbf{S}|}{\hbar\omega} - P(\hbar\omega)d \qquad (3.38)$$

where d is the thickness of the sample. The second term represents the number of photons absorbed per unit time per unit area, normal to the incident light in a thickness d. Equation 3.38 can be written as

$$\Im_d = \frac{|\mathbf{S}|}{\hbar\omega} e^{-\alpha d}$$

$$\cong \frac{|\mathbf{S}|}{\hbar\omega}(1 - \alpha d) \tag{3.39}$$

for small αd. Thus,

$$\alpha(\hbar\omega) = \frac{P\hbar\omega}{|\mathbf{S}|} \tag{3.40}$$

The average value of the Poynting vector over a period of the electromagnetic wave can be expressed as

$$|\mathbf{S}| = \frac{1}{2} n_r \epsilon_0 c \omega^2 |\mathbf{A}^2| \tag{3.41}$$

where n_r is the refractive index of the crystal. Substitution of Eqs. 3.35 and 3.41 into Eq. 3.40 leads to

$$\alpha(\hbar\omega) = C_1 n_r^{-1} \left(\frac{2m_r^*}{m_0}\right)^{3/2} \frac{f_{CV}}{\hbar\omega} (\hbar\omega - \mathcal{E}_g)^{1/2} \tag{3.42}$$

where

$$C_1 = \frac{q^2 m_0^{1/2}}{4\pi\hbar^2\epsilon_0 c} \tag{3.43}$$

and

$$f_{CV} = \frac{2p_{CV}^2}{m_0} \tag{3.44}$$

Expressing $\hbar\omega$ and $\mathcal{E}_g$ in eV,

$$\alpha(\hbar\omega) = 2.64 \times 10^5 n_r^{-1} \left(\frac{2m_r^*}{m_0}\right)^{3/2} \frac{f_{CV}}{\hbar\omega} (\hbar\omega - \mathcal{E}_g)^{1/2} \, (cm^{-1}) \tag{3.45}$$

f_{CV} is a measure of the *oscillator strength* for the transition. It has a value approximately equal to 20 eV in most semiconductors. Therefore, for GaAs (f_{CV} = 23 eV), we get from Eq. 3.45,

$$\alpha(\hbar\omega) = 5.6 \times 10^4 \frac{(\hbar\omega - \mathcal{E}_g)^{1/2}}{\hbar\omega} \, (cm^{-1}) \tag{3.46}$$

The value of α expressed in Eqs. 3.42 or 3.45 corresponds to a fixed photon energy $\hbar\omega$ when the semiconductor is at 0°K and the values of the Fermi-Dirac distribution functions in the conduction and valence bands are zero and unity, respectively. To express the temperature dependence of α, one must include the Fermi functions in the summation over all energies used to calculate the probability function $P(\hbar\omega)$. As a result, the

absorption coefficient expressed by Eq. 3.42 or 3.45 must include a factor $[f_p(\mathcal{E}(\mathbf{k}')) - f_n(\mathcal{E}(\mathbf{k}''))]$, where f_n and f_p are the Fermi functions in the respective bands. It is also assumed that the semiconductor is very pure. Impurity atoms will induce scattering, which will relax the momentum conservation requirements. In addition, impurity levels will give rise to bandtail states that will result in a finite value of α for photon energies $\hbar\omega < \mathcal{E}_g$. The absorption coefficient α expressed in Eq. 3.46 corresponds to a measured value for photon energies $\hbar\omega > \mathcal{E}_g$ under the ideal conditions described above.

Similarly, starting with the matrix element for the direct forbidden transition in Eq. 3.34, it can be shown that the transition probability is given by

$$P(\hbar\omega) = \frac{q^2|\mathbf{A}|^2}{12\pi m_0^2\hbar^4}(2m_r^*)^{5/2} f'_{CV}(\hbar\omega - \mathcal{E}_g)^{3/2} \tag{3.47}$$

It is important to note that the probability is proportional to $(\hbar\omega - \mathcal{E}_g)^{3/2}$. The absorption coefficient is given by

$$\alpha(\hbar\omega) = C_2 n_r^{-1}\left(\frac{2m_r^*}{m_0}\right)^{5/2}\frac{f'_{CV}}{\hbar\omega}(\hbar\omega - \mathcal{E}_g)^{3/2} \tag{3.48}$$

where

$$C_2 = \frac{q^2 m_0^{1/2}}{6\pi\hbar^2\epsilon_0 c} \tag{3.49}$$

f'_{CV} is the oscillator strength for the forbidden transition, and its value is much less than unity. Again, if with $\hbar\omega$ and $\mathcal{E}_g$ are expressed in eV,

$$\alpha(\hbar\omega) = 1.76 \times 10^5 \frac{n_r^{-1}}{\hbar\omega}\left(\frac{2m_r^*}{m_0}\right)^{5/2} f'_{CV}(\hbar\omega - \mathcal{E}_g)^{3/2} \ (cm^{-1}) \tag{3.50}$$

From experimental results it is evident that the direct optical transition corresponding to the absorption of a photon with $\hbar\omega > \mathcal{E}_g$ is dominantly observed in most direct-bandgap semiconductors.

EXAMPLE 3.4

Objective. To calculate α for the allowed transitions in GaAs at a photon energy $\hbar\omega$ = 1.52 eV.

Assuming $\mathcal{E}_g$ = 1.5 eV at 0°K, from Eq. 3.46 $\alpha = 5.2 \times 10^3$ cm^{-1}.

The following general comments may be made regarding the relations given above. For very small values of $(\hbar\omega - \mathcal{E}_g)$ and in a pure semiconductor, *excitons* are generally formed due to the Coulomb interaction between electrons and holes. The expressions for the matrix element, transition probability, and absorption coefficient given above are only for band-to-band transitions and do not consider exciton-related processes. Second, the equations are valid insofar as the parabolic band approximation is true, and are therefore generally not valid for higher-lying regions of the direct band. For large values of photon energy, contributions from the satellite valleys can become

important. In most III–V semiconductors, the valence band is degenerate at the zone center. Strictly, the effect of such degenerate bands will contribute to the absorption. For a heavily doped semiconductor, the absorption edge moves to higher energies, as illustrated in Fig. 3.4. In addition, bandtail states are formed. It is important to note that the matrix element for transitions between bandtail states is different from that involving free-electron and hole states. Consequently, the k-selection rule does not apply.

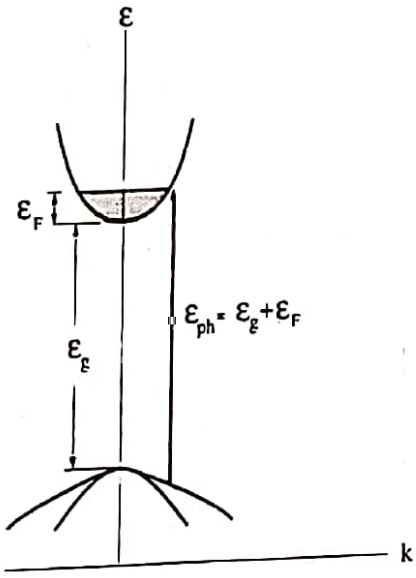


Figure 3.4 Simplified illustration of absorption of photons of energy larger than the bandgap in a degenerately doped n-type semiconductor.

With variation of temperature, the variation of the absorption coefficient follows the variation of the bandgap with temperature given by Eq. 2.15. Finally, it should be remembered that the process of radiation is complementary to absorption and is governed by similar equations for transition probability and oscillator strength.

3.2.2 Indirect Intrinsic Transitions

The momentum or wavevector change required in an indirect transition may be provided by single or multiple phonons, although the probability of the latter to occur is very small. As seen in Chapter 2, there are optical and acoustic phonons. Each of these has transverse and longitudinal modes of vibrations, with characteristic energy and momentum. The indirect transition process is illustrated in Fig. 3.2(c). Conservation of momentum requires

$$\mathbf{k}'' \pm \mathbf{k}_p = \mathbf{k}' + \mathbf{k}_{ph} \tag{3.51}$$

where $\mathbf{k}''$ and $\mathbf{k}'$ are the electron wavevectors for the final and initial states, $\mathbf{k}_p$ is the wavevector of the phonon, and $\mathbf{k}_{ph}$ is the wavevector of the absorbed photon. Since the latter is small, the conservation of momentum for an indirect transition can be expressed as

$$\mathbf{k}'' - \mathbf{k}' = \mp\mathbf{k}_p \tag{3.52}$$

Similarly, the conservation of energy for the two cases of phonon emission and absorption can be expressed as

$$\hbar\omega_e = \mathcal{E}_C - \mathcal{E}_V + \mathcal{E}_p \tag{3.53}$$

$$\hbar\omega_a = \mathcal{E}_C - \mathcal{E}_V - \mathcal{E}_p \tag{3.54}$$

where the left-hand side represents the energy of the photon absorbed. Note that in the first case of phonon emission, the energy of the absorbed photon could be equal to the direct gap at or very near k = 0. From this energy state the electron finally reaches the indirect valley by phonon scattering. The intermediate energy state of the electron is termed a *virtual* state, in which the carrier resides until a phonon of the right energy and momentum is available for the scattering process. Indirect transition probabilities involving virtual states can be calculated using a second-order time-dependent perturbation theory. However, there is a process to slightly counterbalance the low transition probability, which is often overlooked. From Eqs. 3.53 and 3.54 it is evident that the initial and final states of the electron in the valence and conduction bands, respectively, can have an energy range given by $\hbar(\omega_{ph} \pm \omega_p)$ where ω_p and ω_{ph} correspond to the angular frequencies of the phonon and photon, respectively. The total probability is obtained by a summation over these energy states, as long as each particular transition conserves energy between initial and final states.

For a transition with phonon absorption,

$$\alpha_a(\hbar\omega) \propto \frac{(\hbar\omega - \mathcal{E}_g + \mathcal{E}_p)^2}{e^{\mathcal{E}_p/k_BT} - 1} \tag{3.55}$$

for a photon energy $\hbar\omega > (\mathcal{E}_g - \mathcal{E}_p)$. Similarly, for a transition with phonon emission the absorption coefficient is given by

$$\alpha_e(\hbar\omega) \propto \frac{(\hbar\omega - \mathcal{E}_g - \mathcal{E}_p)^2}{1 - e^{-\mathcal{E}_p/k_BT}} \tag{3.56}$$

for $\hbar\omega > (\mathcal{E}_g + \mathcal{E}_p)$. Since for $\hbar\omega > (\mathcal{E}_g + \mathcal{E}_p)$ both phonon emission and absorption are possible, under these conditions

$$\alpha(\hbar\omega) \cong \alpha_a(\hbar\omega) + \alpha_e(\hbar\omega) \tag{3.57}$$

The temperature dependence of the absorption coefficient is illustrated in Fig. 3.5. At very low temperatures, the density of phonons available for absorption becomes small and therefore α_a is small. With increase of temperature, α_a increases. The shift of the curves to lower energies with increase of temperature reflects the temperature dependence of $\mathcal{E}_g$. In fact, the plots of $\sqrt{\alpha_e}$ and $\sqrt{\alpha_a}$ extrapolate to the energy axis at $(\mathcal{E}_g +$

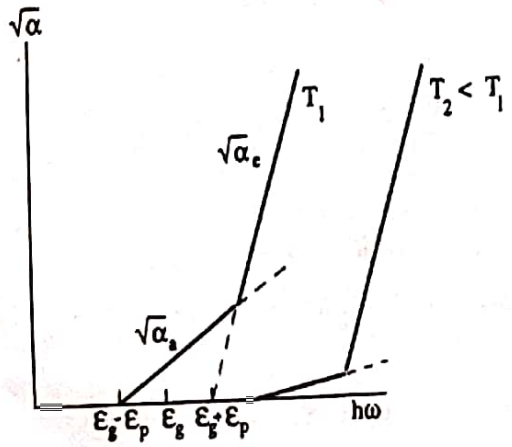


Figure 3.5 Energy-dependent absorption coefficient due to phonon emission and absorption as a function of temperature.

$\mathcal{E}_p$) and ($\mathcal{E}_s - \mathcal{E}_p$), respectively. This is a convenient technique to experimentally determine the bandgap.

EXAMPLE 3.5

Objective. To estimate the momentum change involved in the absorption of a photon in silicon.

Silicon is an indirect bandgap semiconductor. From Fig. 1.1 it is evident that the valence band maximum is at k = 0 and the conduction band minimum is at $k = (0.85\frac{2\pi}{a})$. Actually this band edge is one of the six equivalent (k, 0, 0) minima in the three-dimensional Brillouin zone. The change in momentum required is

$$\hbar\Delta k = \frac{1.05 \times 10^{-34}(J.s) \times 0.85 \times 2\pi}{5.43(\mathring{A})}$$

$$= 1.03 \times 10^{-22} kg.cm.s^{-1}$$

$$= \hbar k_p$$

3.2.3 Exciton Absorption

In very pure semiconductors, where the screening effect of free carriers is almost absent, electrons and holes produced by the absorption of a photon of near-bandgap energy pair to form an *exciton*. This is the free exciton. The binding energy of the exciton, $\mathcal{E}_{ex}$, is calculated by drawing analogy with the Bohr atom for an impurity center, and is quantized. It is therefore expressed as

$$\mathcal{E}_{ex}^{l} = \frac{-m_r^* q^4}{2(4\pi\epsilon_r\epsilon_o\hbar)^2} \cdot \frac{1}{l^2}, \quad l = 1, 2, 3, \cdots$$

$$= \frac{-13.6}{l^2}\frac{m_r^*}{m_0}\left(\frac{1}{\epsilon_r}\right)^2 (eV). \tag{3.58}$$

Here m_r^* is the reduced effective mass of the exciton given by Eq. 3.36 and l is an integer. Such excitons are also known as *effective mass* or *Wannier excitons*.

The optical excitation and formation of excitons usually manifest themselves as a series of sharp resonances (peaks) at the low energy side of the band edge in the absorption spectra of direct bandgap semiconductors. The total energy of the exciton is given by

$$\mathcal{E}_{ex} = \frac{\hbar^2 k_{ex}^2}{2(m_e^* + m_h^*)} - \mathcal{E}_{ex}^{l} \tag{3.59}$$

where the first term on the right is the kinetic energy of the exciton. The kinetic energy contributes to a slight broadening of the exciton levels. For a direct transition conservation of momentum requires that $k_{ex} \cong 0$. This is because the electron and hole must move in the same direction. Usually a sharp line transition is observed for direct excitonic transitions, which broadens with increase of temperature. Data on the excitonic absorption in pure GaAs are shown in Fig. 3.6.

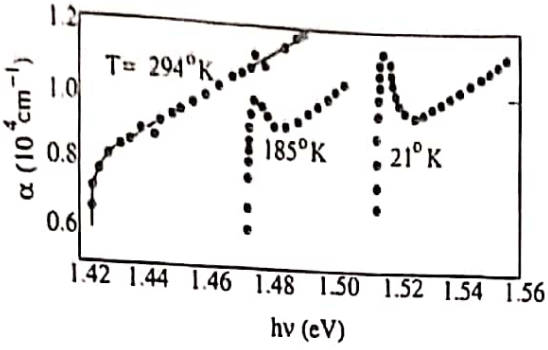


Figure 3.6 Absorption coefficient measured near the band edge of GaAs at $T = 294$, 185, and 21°K. The two absorption peaks at $h\nu$ slightly below the respective band gap at 185° K and 21° K are due to bound excitons (from M. D. Sturge, *Physical Review*, 127, 768, 1962).

In indirect bandgap semiconductors excitons may also be formed with the absorption or emission of a phonon. In this case the center of gravity of the exciton may have a finite momentum $\hbar k_{ex}$, conserved by an interacting phonon. Again, transverse and longitudinal acoustic and optical phonons may participate. An increase in absorption coefficient is obtained near the band edge due to exciton absorption, given by

$$\hbar\omega = \mathcal{E}_g \mp \mathcal{E}_p - \mathcal{E}_{ex} \tag{3.60}$$

where the two signs of the second term on the right hand side correspond to the cases of phonon absorption or emission. Exciton-related transitions are seen in the absorption spectra of an indirect bandgap semiconductor as a large number of steps near the absorption edge. Note that steps are observed instead of peaks, as in direct bandgap semiconductors, because the interacting phonons allow the carrier transition between states with equal $d\mathcal{E}/dk$ in the valence and conduction bands at energies greater than in direct band gap semiconductors where usually excitons are formed at the zone center ($d\mathcal{E}/dk = 0$).

Excitons are formed in very pure semiconductors at low temperatures. In fact, excitons were not observed in semiconductors until epitaxial techniques enabled the growth of very pure crystals. In such crystals, the very few unintentional impurities that are present—donors and acceptors—are neutral. If an electric field is applied, it can ionize these impurities, and the additional charge modifies the bandedge potential. This is seen in the experimental absorption spectra as a change in the slope of the absorption edge. In addition, the ionized carriers screen the Coulomb interaction between the electrons and holes, thereby inhibiting or preventing the formation of excitons. This is observed in the experimental absorption spectra as a disappearance of the excitonic resonances, peaks or steps, as the case may be.

3.2.4 Donor-Acceptor and Impurity-Band Absorption

Intentionally or unintentionally, both donors and acceptors are simultaneously present in a semiconductor, and any semiconductor is usually always compensated to some degree. Depending on the temperature and the state of occupancy of the impurity levels, it is possible to raise an electron from the acceptor to the donor level by absorbing a photon. This process is shown in Fig. 3.7. The energy of the photon absorbed is given by

$$\hbar\omega = \mathcal{E}_g - \mathcal{E}_D - \mathcal{E}_A + \frac{q^2}{\epsilon_0 \epsilon_r r} \tag{3.61}$$

where the last term on the right-hand side accounts for the Coulomb interaction between the donor and acceptor atoms in substitutional sites, which results in a lower-

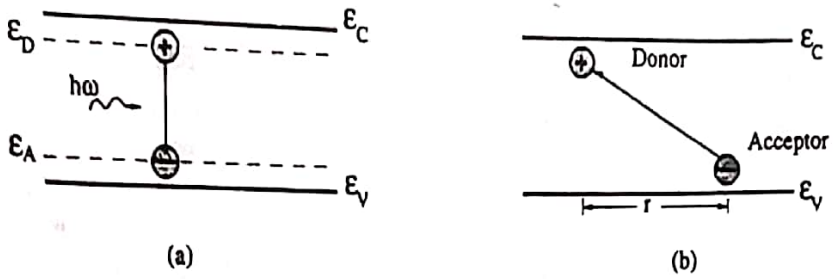


Figure 3.7 Illustration of photon absorption due to a donor-acceptor transition. The separation between the impurity centers, r, is shown in (b).

ing of their binding energies. This can be understood as follows. Assume that at very low temperatures the donor and acceptor atoms are neutral. If they are brought closer together, the additional orbiting electron of the donor becomes "shared" by the acceptor, as in a covalent bond, and both become more ionized, resulting in a lowering of their binding energy. Also, it is important to remember that since the donor and acceptor atoms are located at discrete substitutional sites in the lattice, r varies in finite increments, being the smallest for nearest neighbors. Therefore, for the ground state of the impurities, the energies $\mathcal{E}_D$ and $\mathcal{E}_A$ correspond to the most distant pairs and $\hbar\omega \cong \mathcal{E}_g - \mathcal{E}_D - \mathcal{E}_A$. For fully ionized impurities, such as for nearest neighbors, the excited states may lie within the respective band and it is possible that $(q^2/\epsilon_0 \epsilon_r r) > \mathcal{E}_D + \mathcal{E}_A$. At low temperatures the absorption resonances modify the bandedge absorption, with the lowest energy transitions for the most distant pairs and higher-energy transitions for nearer pairs. However, because the resonances occur so close to the absorption edge, they are not always very clearly defined. The pair transitions are more clearly identified in emission experiments.

High-energy (near-bandgap) transitions can occur between ionized impurity levels and the opposite bandedge, as illustrated in Fig. 3.8. The photon energy absorbed is $\hbar\omega \cong \mathcal{E}_g - \mathcal{E}_b$, where $\mathcal{E}_b$ is the binding energy of the donor or acceptor level. It should be noted that the impurity levels need to be ionized. Since the transition occurs between a discrete impurity level and a band of energies, the transitions are observed as shoulders on the low-energy side of the absorption edge. In the emission spectra, these transitions are observed as peaks. As in a band-to-band transition, phonons need to be involved in impurity-band transitions in an indirect bandgap semiconductor for momentum conservation.

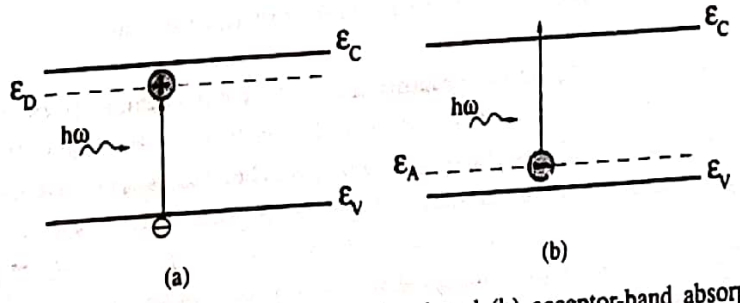


Figure 3.8 Illustration of (a) donor-band and (b) acceptor-band absorption transitions.

The absorption spectrum is largely altered if the doping level is increased and gradually taken to the point of degeneracy. For example, in a degenerately doped n-type semiconductor, the Fermi level $\mathcal{E}_{fn}$ is above the conduction bandedge. If the semiconductor is direct bandgap, as shown in Fig. 3.4, then, for the conservation of momentum, the transition resulting from the absorption of a photon will involve states in the conduction band that are at or higher than $\mathcal{E}_g + \mathcal{E}_{fn}$. This shift of the absorption to higher energies due to doping-induced band-filling is called the *Burstein-Moss shift*. An indirect bandgap semiconductor will be similarly affected, except that phonons need not be involved in the transition. Momentum is conserved by impurity scattering.

Degeneracy in semiconductors not only pushes the Fermi level into the band, but also results in a shrinkage of the bandgap. This effect is more commonly known as *bandtailing*, which results in an exponentially increasing absorption edge with photon energy as shown in Fig. 3.9 for GaAs.

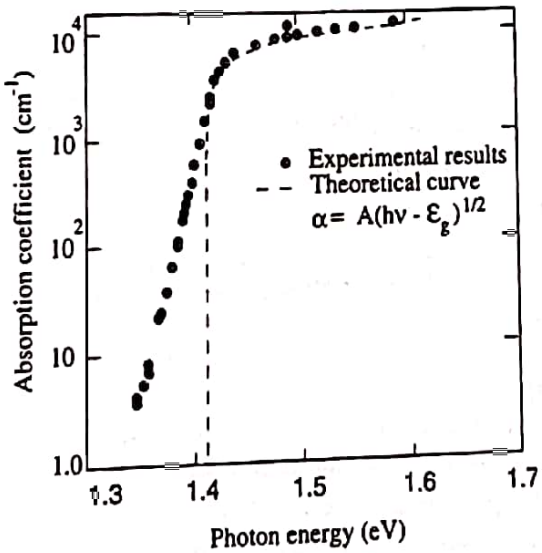


Figure 3.9 Absorption edge of GaAs at room temperature (from T. S. Moss, *Journal of Applied Physics*, 32, 2136, 1961).

3.2.5 Low-Energy (Long-Wavelength) Absorption

Several types of transitions involving shallow impurity levels, bandedges, split bands, and free carriers give rise to resonances at very small energies in the absorption spectra. These are observed as steps or peaks in the long-wavelength region of the absorption spectra. The different processes are briefly described below.

3.2.5.1 Impurity-Band Transitions. We have seen impurity-band transitions that have energies close to the bandgap. These higher-energy impurity-band transitions usually require that the impurity levels are ionized (or empty). At low temperatures, when these shallow impurity levels are usually filled with their respective carriers, these carriers can be excited to the respective bandedge by a photon (Fig. 3.10). For this absorption process the energy of the photon must be at least equal to the ionization energy of the impurity. This energy usually corresponds to the far infrared region of

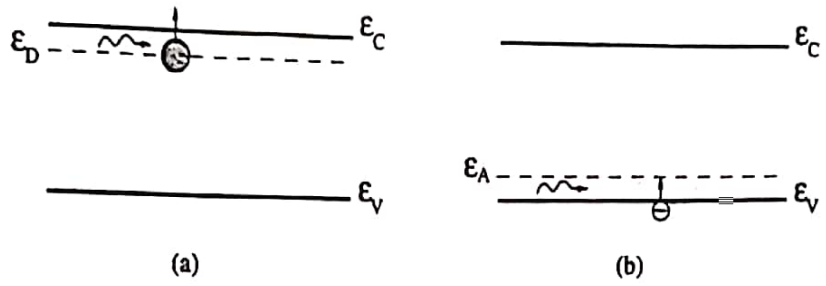


Figure 3.10 Low-energy (a) donor-band and (b) acceptor-band absorption transition.

the optical spectrum. Peaks corresponding to such impurity-band transitions have been observed in many semiconductors.

3.2.5.2 Intraband Transitions At the zone center the valence band structure of most semiconductors consists of the light-hole (LH), the heavy-hole (HH) bands, and the split-off (SO) band. The three subbands are separated by spin-orbit interaction. In a p-type semiconductor the valence band is filled with holes and the occupancy of the different bands depend on the degree of doping and the position of the Fermi level. Absorption of photons with the right energy can result in transitions from LH to HH, SO to HH, and SO to LH bands, depending on the doping and temperature of the sample. These transitions have been observed experimentally. They are normally not observed in n-type semiconductors.

3.2.5.3 Free-Carrier Absorption As the name suggests, this mechanism involves the absorption of a photon by the interaction of a free carrier within a band, which is consequently raised to a higher energy. The transition of the carrier to a higher energy within the same valley must conserve momentum. This momentum change is provided by optical or acoustic phonons or by impurity scattering. Free-carrier absorption usually manifests in the long-wavelength region of the spectrum as a monotonic increase in absorption with a wavelength dependence of the form λ^p, where p ranges from 1.5 to 3.5. The value of p depends on the nature of the momentum-conserving scattering (i.e., the involvement of acoustic phonons, optical phonons, or ionized impurities). The absorption coefficient due to free-carrier absorption can be expressed as

$$\alpha = \frac{Nq^2\lambda^2}{4\pi^2 m^* n_r c^3 \epsilon_0} \left\langle \frac{1}{\tau} \right\rangle \tag{3.62}$$

where N is the free-carrier concentration, n_r is the refractive index of the semiconductor, and $\left\langle \frac{1}{\tau} \right\rangle$ is the average value of the inverse of the relaxation time of the scattering process.

As a concluding note to this section, the absorption coefficients for different elemental and III–V compound semiconductors at room temperature are shown in Fig. 3.11.

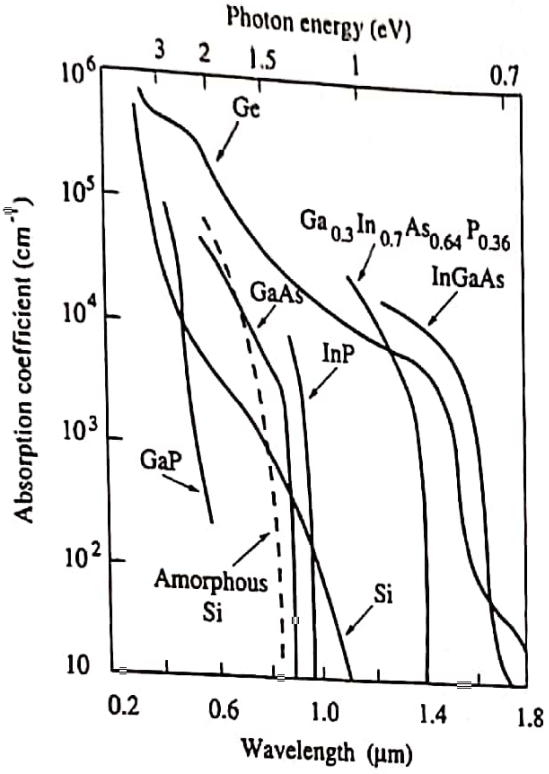


Figure 3.11 Near-bandgap absorption spectra of different semiconductors (from M. Shur, *Physics of Semiconductor Devices*, ©1990. Reprinted by permission of Prentice Hall, Englewood Cliffs, New Jersey).

3.3 EFFECT OF ELECTRIC FIELD ON ABSORPTION: FRANZ-KELDYSH AND STARK EFFECTS

The change in absorption in a semiconductor in the presence of a strong electric field is the *Franz-Keldysh effect,* which results in the absorption of photons with energies less than the bandgap of the semiconductor. The energy bands of a semiconductor in the presence of an electric field E and with an incident photon of energy $\hbar\omega < \mathcal{E}_g$ are shown in Figs. 3.12(a) and (b). It is important to note that at the classical turning points marked A and B, the electron wavefunctions change from oscillatory to decaying behavior. Thus, the electron in the energy gap is described by an exponentially decaying function $u_k e^{jkx}$, where $\mathbf{k}$ is imaginary. With increase of electric field, the distance AB decreases and the overlap of the wavefunctions within the gap increases. In the absence of a photon, the valence electron has to tunnel through a triangular barrier of height $\mathcal{E}_g$ and thickness d, given by $d = \mathcal{E}_g/q\,E$. With the assistance of an absorbed photon of energy $\hbar\omega < \mathcal{E}_g$, it is evident that the tunneling barrier thickness is reduced to $d' = (\mathcal{E}_g - \hbar\omega)/qE$, the overlap of the wavefunctions increases further, and the valence electron can easily tunnel to the conduction band. The net result is that a photon with $\hbar\omega < \mathcal{E}_g$ is absorbed. One has to keep in mind, of course, the conservation of momentum in these transitions and in this case the transverse component of momentum is conserved. The Franz-Keldysh effect is therefore, in essence, photon assisted tunneling.

It can be shown that the electric-field dependent absorption coefficient is given by†

†S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, Prentice Hall, Englewood Cliffs, NJ, 1989.

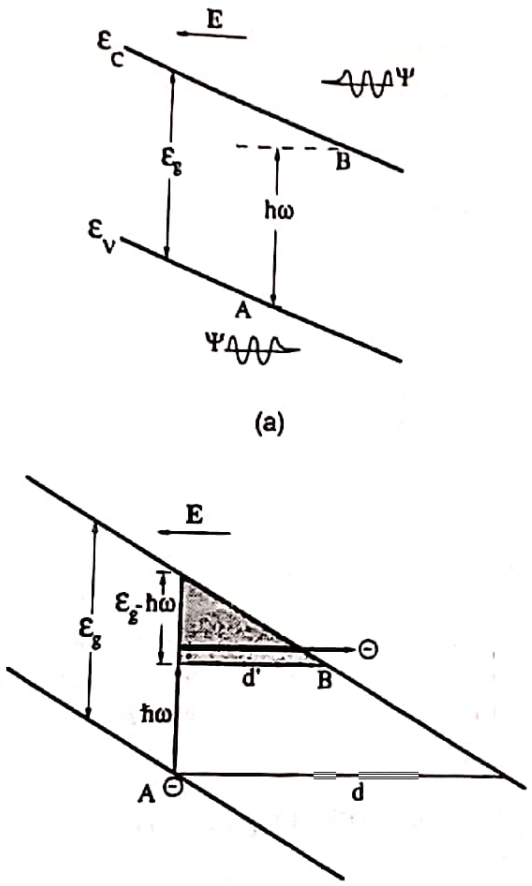


Figure 3.12 (a) Bending of bands due to an applied electric field and (b) absorption of photon with $\hbar\omega < \mathcal{E}_g$ due to carrier tunneling (Franz-Keldysh effect).

$$\alpha = K(E')^{1/2}(8\beta)^{-1} \exp\left(-\frac{4}{3}\beta^{3/2}\right) \tag{3.63}$$

Here $E' = \left(\frac{q^2E^2\hbar^2}{2m_r^*}\right)^{1/3}$, $\beta = \frac{\mathcal{E}_g - \hbar\omega}{E'}$, and K is a material-dependent parameter that has a value of 5×10^4 cm^{-1} (eV)$^{-1/2}$ in GaAs. Although not derived here, the various terms in Eq. 3.63 can be examined qualitatively. The exponential term is the transmission coefficient (or tunneling probability) of an electron through a triangular barrier of height $(\mathcal{E}_g - \hbar\omega)$ and can be obtained from the well-known Wentzel-Kramers-Brillouin (WKB) approximation. The other factors are related to the upward transition of an electron due to photon absorption. Substituting appropriate values for the different parameters, it is seen that in GaAs $\alpha = 4$ cm^{-1} at a photon energy of $\mathcal{E}_g - 20$ meV with electric field $E \sim 10^4$ V/cm. This value of absorption coefficient is much smaller than the values of α at the band edge at zero field. Therefore, the Franz-Keldysh effect will be small unless $E \geq 10^5$ V/cm.

The *Stark effect* refers to the change in atomic energy upon the application of an electric field. The electric field affects the higher-order, or outer, orbits of the precessing electrons so that the center of gravity of the elliptical orbit and the focus are displaced with respect to each other and linearly aligned in the direction of the electric field. As a result, there is a splitting of the energy of the outer $2s$ or $2p$ states, and the energy shift is simply given by $\Delta\mathcal{E} = qdE$, where d is the eccentricity of the orbit. This

is the *linear Stark effect*. The effect of the electric field on ground state orbits also leads to an energy shift of the state, and this is the *quadratic or second-order Stark effect*.

3.4 ABSORPTION IN QUANTUM WELLS AND THE QUANTUM-CONFINED STARK EFFECT

In a bulk semiconductor the exciton binding energy is given by Eq. 3.58. For example in GaAs, upon substitution of the effective mass and dielectric constant, a ground-state binding energy of 4.4 meV is obtained. This is comparable to the thermal broadening of – 4 meV produced by optical phonon scattering and inhomogeneous broadening, to be discussed in Chapter 6. In other words, the exciton dissociates in a very short time (a few hundred femtoseconds) and is hardly detected in the absorption spectra at room temperature, except in very pure samples.

The situation is drastically altered in a quantum well. In a single-quantum well (SQW) or multiple-quantum well (MQW) with thick barriers (≥ 100 Å), electrons and holes are confined in the region defined by the well width, and the overlap of their wavefunctions is increased. This results in an increase in the oscillator strength of the interband transitions between the discrete electron and hole energy bound states, which are produced by the size quantization. Consequently, strong resonances corresponding to the heavy-hole and light-hole transitions are seen near the bandedge of the well material even at room temperature. Shown in Fig. 3.13 are the measured and calculated absorption spectra of an $In_{0.53}Ga_{0.47}As$ (100 Å)/$In_{0.52}Al_{0.48}As$ (100 Å) MQW lattice matched to InP, in which the resonances for $l = 2$ and 3 are also clearly seen. It may be noted that distinct resonances are seen for heavy-hole and light-hole transitions. This is because a splitting between the HH and LH bands occurs at the zone center due to the difference in the energy eigenvalues resulting from different hole masses.

In a two-dimensional quantum well the exciton is compressed like a pancake. However, since typical well dimensions are ~ 100 Å and the exciton diameter is ≥ 300 Å, there is some penetration of the exciton wavefunction into the barrier material. This is depicted in Fig. 3.14. In the limit, for small well widths, the situation becomes similar to a three-dimensional solid. It has been shown that for a purely two-dimensional exciton, its binding energy is four times the bulk value, given by Eq. 3.58. However, because of the extension of its wavefunction into the barrier, the binding energy in practical SQW and MQW structures ranges between $2\mathcal{E}_{ex}$ and $3\mathcal{E}_{ex}$. Since this binding energy is much larger than the thermal broadening, the exciton resonances are clearly seen in the absorption spectra even at room temperature. The coefficient for intersubband absorption in a quantum well can be expressed in cgs units as[†]

$$\alpha_{2D}(\hbar\omega) = \frac{4\pi^2 q^2 \hbar}{n_r c m_0^2 \hbar\omega} p_{CV}^2 a_p \frac{N_{2D}(\hbar\omega - \mathcal{E}_g)}{L_z}$$

$$= 1.77 \times 10^{-28} \frac{a_p}{n_r} \left(\frac{f_{CV}}{\hbar\omega}\right) \frac{N_{2D}}{L_z} \qquad (3.64)$$

[†]J. Singh, *Physics of Semiconductors and Their Heterostructures*, McGraw-Hill, New York, 1993.

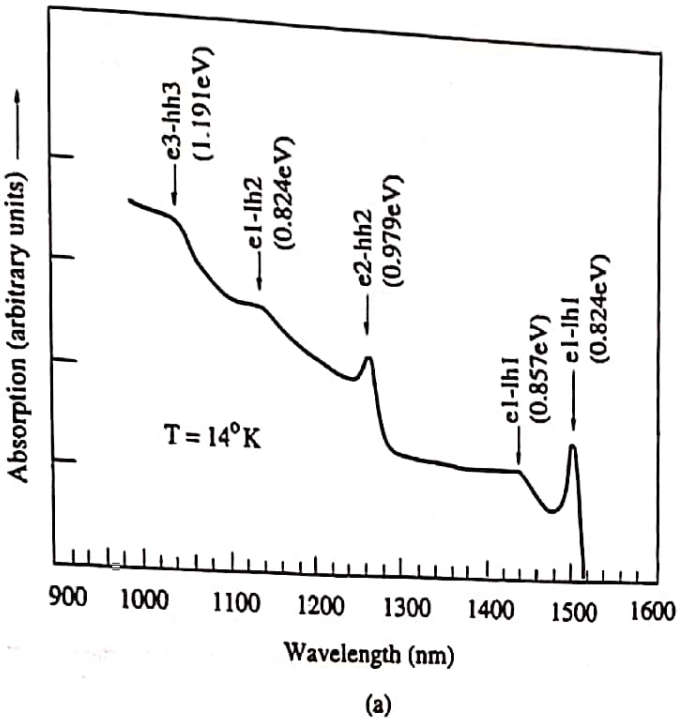


Figure 3.13 (a) Absorption spectrum of 40-period lattice-matched InGaAs/InAlAs MQW (measured by the author and co-workers) at 14°K; and (b) the calculated transitions based on a finite square well model at room temperature (from S. Gupta et al., *Journal of Applied Physics*, **69**, 3219, 1991).

where L_z is the width of the well, $N_{2D} = m_r^*/\pi\hbar^2$ is the 2-D density of states, a_p is a factor due to polarization dependence of the matrix elements (in GaAs $a_p = 1/2$). This equation is valid for a single pair of conduction and valence subbands. Note also that the value of α is constant and will remain so till the next subband energies are reached. This is evident in Fig. 3.13. It may seem from Eq. 3.64 that the absorption coefficient can be increased by decreasing L_z. However, as L_z is decreased, the electron and hole wavefunctions spread outside the well and the overlap decreases. Therefore, for every material system there is an optimum well size. For example, for GaAs this $L_z \sim 50$Å and for $In_{0.53}Ga_{0.47}As$ $L_z \sim 80$Å.

The ground-state wavefunctions of the electron and hole subband with no applied transverse field (in the direction perpendicular to the layers) are shown in Fig. 3.15(a). With the application of an electric field several things happen. The electron and hole

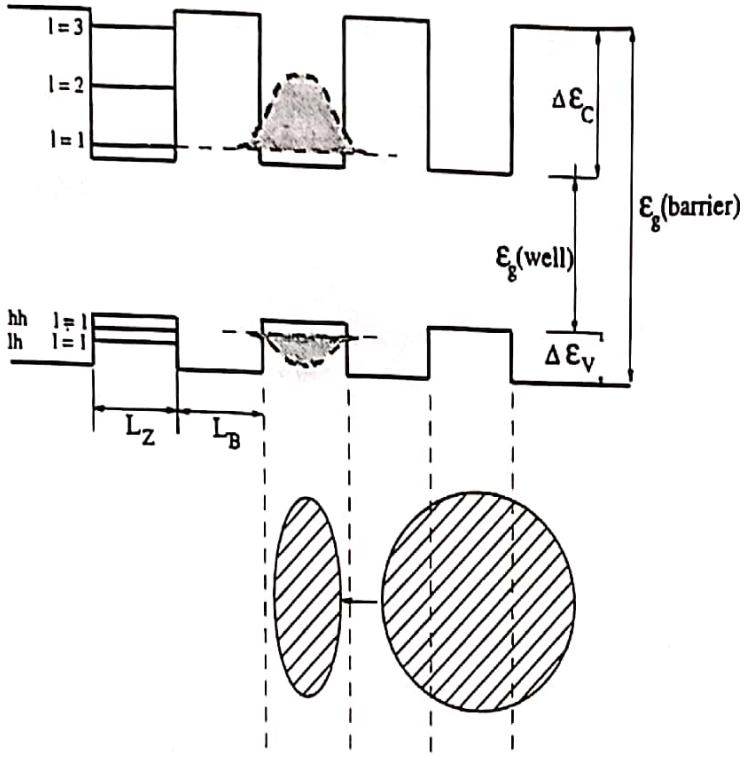
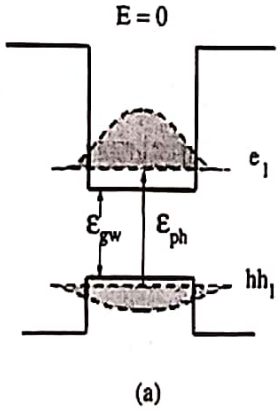


Figure 3.14 A typical multiquantum well and the compression of the bulk exciton in the well region.



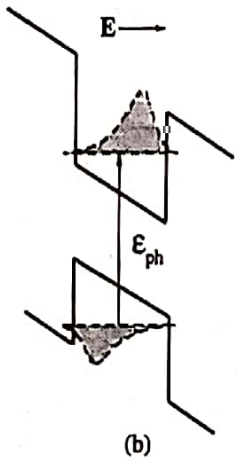


Figure 3.15 Absorption in a quantum well in the (a) absence and (b) presence of a transverse electric field.

wavefunctions are separated and pushed toward the opposite sides of the well, as shown in Fig. 3.15(b). The reduced overlap results in a corresponding reduction in absorption and in luminescence. The probability of carriers tunneling out of the wells

also increases, resulting in a decrease in carrier lifetimes and a broadening of the absorption spectra. The transition energy is given by

$$\mathcal{E}_{ph} = \mathcal{E}_e + \mathcal{E}_h + \mathcal{E}_{gw} - \mathcal{E}_{ex} \tag{3.65}$$

where $\mathcal{E}_e$ and $\mathcal{E}_h$ are the electron and hole subband energies. With the application of moderate electric fields ($10^4 - 10^5$ V/cm), there is little change in $\mathcal{E}_{ex}$ and a very small change in $\mathcal{E}_{gw}$ due to the Stark effect in the well material. However, due to the modification of the envelope functions, there is a reduction in $\mathcal{E}_e$ and $\mathcal{E}_h$, the subband energies. This results in a shift of the absorption spectrum to lower energies, including the heavy- and light-hole resonances. The energy shift of the resonances has a quadratic dependence on the applied transverse electric field, as will be described in Chapter 11. The shift is much larger than the Stark shift in bulk materials and is ~ 20 meV for $E = 10^5$ V/cm in a 100 Å $GaAs/Al_{0.3}Ga_{0.7}As$ quantum well. Experimental data are shown in Fig. 3.16 and the phenomenon is known as the *quantum confined Stark effect* (QCSE). As we shall see in Chapter 11, the effect can be used for the design and realization of very efficient light modulators.

3.5 THE KRAMERS-KRÖNIG RELATIONS

The complex refractive index of a semiconductor is given by

$$\mathbf{n}_c = \mathbf{n}_r + j\mathbf{k}_a \tag{3.66}$$

where $\mathbf{n}_r$ and $\mathbf{k}_a$ are the real and imaginary parts of $\mathbf{n}_c$. It may be noted that $\mathbf{k}_a$ is an attenuation or damping factor and is a measure of the loss in power of a wave propagating through the semiconductor. Therefore, $\mathbf{k}_a$ is directly proportional to the absorption coefficient α of the material. If a plane wave described by

$$\mathbf{E} = \mathbf{E}_0 \exp\left[j\omega\left(\frac{x}{\vartheta} - t\right)\right]$$

$$= \mathbf{E}_0 \exp\left[j\omega\left(\frac{n_r x}{c} - t\right)\right] \exp\left(-\frac{\omega \mathbf{k}_a x}{c}\right) \tag{3.67}$$

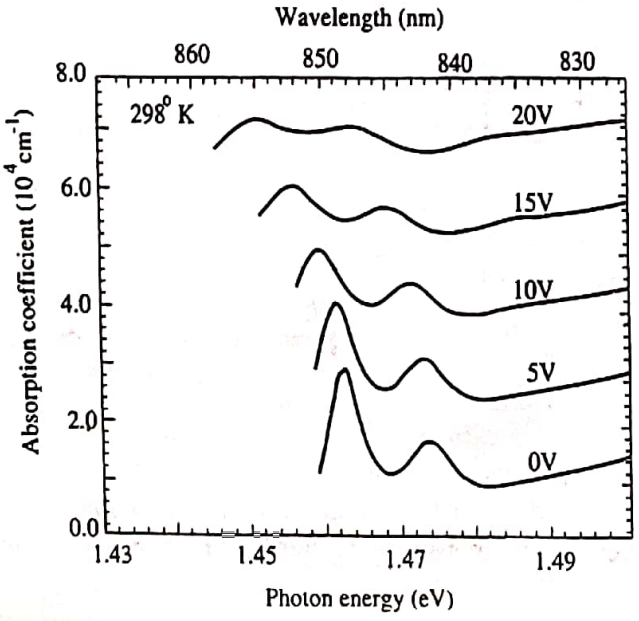


Figure 3.16 Room temperature optical absorption coefficients measured by the author and co-workers in a 100-Å $GaAs/Al_{0.3}Ga_{0.7}As$ MQW structure at various bias levels (from J. Singh et al., *Journal of Lightwave Technology*, **6**, 818, ©1988 IEEE).

is propagating in the x-direction in a semiconductor with a velocity $\vartheta = \frac{c}{n_r}$, then it can be shown (Problem 3.7) that the absorption coefficient is given by

$$\alpha = \frac{2\omega \mathbf{k}_a}{c} = \frac{4\pi\nu \mathbf{k}_a}{c} \tag{3.68}$$

In a material whose conductivity $\sigma \to 0$, the refractive index is related to the dielectric constant by the relation

$$n_r \cong \sqrt{\epsilon_r} \tag{3.69}$$

where ϵ_r is the static dielectric constant ϵ_s and

$$k_a \cong 0 \tag{3.70}$$

It is also useful to know that the refractive index is inversely related to the bandgap of a semiconductor.

The complex dielectric constant of a material is given by

$$\epsilon_r(\omega) = \epsilon_r'(\omega) + j\epsilon_r''(\omega) \tag{3.71}$$

In time-invariant form, the electric field $\mathbf{E}$ and the electric flux density $\mathbf{D}$ are related by

$$\begin{aligned} \mathbf{D} &= \epsilon_0 (1 + \chi^e)\, \mathbf{E} \\ &= \epsilon_r \epsilon_o \mathbf{E} \end{aligned} \tag{3.72}$$

where χ^e is the electric *susceptibility*. The temporal response of $\mathbf{D}$ to a change or switching of $\mathbf{E}$ must include the change of polarization with time, and can be expressed by a causality relation of the type

$$\mathbf{D}(t) = \epsilon_o \epsilon_\infty' \delta(t)\mathbf{E}(t) + \int_{-\infty}^{t} \epsilon_o f(t - t')\mathbf{E}(t')dt' \tag{3.73}$$

in which the integral represents the response of the system at time t to the applied field $\mathbf{E}$ at a previous time t'. Note that $\epsilon(t)$ has a δ-function singularity at $t = 0$. ϵ_∞' is the high-frequency dielectric constant, and $\epsilon_\infty'' = 0$. Now the following Fourier transforms can be written for $\mathbf{D}$ and $\mathbf{E}$:

$$\mathbf{D}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{D}(\omega) e^{-j\omega t} d\omega \tag{3.74}$$

and

$$\mathbf{E}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{E}(\omega) e^{-j\omega t} d\omega \tag{3.75}$$

Substituting these in Eq. 3.73 one gets

$$\int_{-\infty}^{\infty} [\mathbf{D}(\omega) - \epsilon_o(\epsilon'_\infty + f(\omega))\mathbf{E}(\omega)] e^{-j\omega t} d\omega = 0 \tag{3.76}$$

with

$$f(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) e^{j\omega t} dt \tag{3.77}$$

Equation 3.76 must be valid for all values of t. Therefore, the relation

$$\mathbf{D}(\omega) = \epsilon_o(\epsilon'_\infty + f(\omega))\,\mathbf{E}(\omega) \tag{3.78}$$

is valid between the Fourier components, so that

$$\epsilon_r(\omega) = \epsilon'_\infty + \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{j\omega t}dt$$

$$= \epsilon'_\infty + f(\omega) \tag{3.79}$$

From Eq. 3.79, by application of the Cauchy theorem to the function $[\epsilon_r(\omega) - \epsilon'_\infty]/(\omega' - \omega)$, the following relations can be derived (Appendix 5):

$$\epsilon'_r(\omega) = \epsilon'_\infty + \frac{2}{\pi}P\int_0^\infty \frac{\omega'\epsilon''_r(\omega)d\omega'}{\omega'^2 - \omega^2} \tag{3.80}$$

$$\epsilon''_r(\omega) = -\frac{2\omega}{\pi}P\int_0^\infty \frac{[\epsilon'_r(\omega') - \epsilon'_\infty]d\omega'}{\omega'^2 - \omega^2} \tag{3.81}$$

where P is the principal value of the Cauchy integrals. These integrals are known as the Kramers-Krönig relations. A more relevant form is the relation between refractive index and absorption coefficient. By analogy with Eq. 3.80,

$$n_r(\mathcal{E}) - 1 = \frac{2}{\pi}P\int_0^\infty \frac{\mathcal{E}'k_a(\mathcal{E}')}{\mathcal{E}'^2 - \mathcal{E}^2}d\mathcal{E}' \tag{3.82}$$

and, by virtue of the relation in Eq. 3.68,

$$n_r(\mathcal{E}) - 1 = \frac{ch}{2\pi^2}P\int_0^\infty \frac{\alpha(\mathcal{E}')}{\mathcal{E}'^2 - \mathcal{E}^2}d\mathcal{E}' \tag{3.83}$$

which enables the determination of the refractive index from the absorption spectrum. The dielectric constant and refractive index of some important binary III–V compounds are given in Table 3.1.

TABLE 3.1 DIELECTRIC CONSTANT AND REFRACTIVE INDEX IN SOME BINARY MI-V COMPOUNDS.

Material	Static Dielectric Constant (ϵ_s)	High-Frequency Dielectric Constant (ϵ_∞)	Refractive Index (n_r) at Bandgap Energy
AlAs	10.06	8.5	3.17
GaP	11.11	9.11	3.45
GaAs	13.18	10.89	3.66
InP	12.56	9.61	3.45
InAs	15.15	12.3	3.52

3.6 RADIATION IN SEMICONDUCTORS

3.6.1 Relation Between Absorption and Emission Spectra

In the last few sections of this chapter we have studied the various absorption processes. In these a photon is absorbed in the semiconductor, as a result of which an electron

is usually raised from a lower-energy filled state to a higher-energy empty state, and in most cases the energy difference between the two states is equal to the energy of the absorbed photon. If the higher-energy level to which the electron is raised is not the equilibrium state, then it will make a downward transition to the lower-energy empty state and emit electromagnetic radiation in the process. The energy of the radiation is very close to the energy difference between the higher and lower energy states. These are radiative transitions.

In principle, the reverse of all the absorption processes we have considered can occur to produce radiation. However, there is an important difference between absorption and emission spectra. While the absorption process can couple a broad energy range of filled and empty states (with momentum conservation) to produce a broad absorption spectrum, the emission process usually couples a narrow band of nonequilibrium filled states with a narrow band of empty states, to give a narrow emission spectrum. Therefore, a shoulder in the absorption spectrum can very well correspond to a narrow emission peak. Also, it is essential for the semiconductor to have a nonequilibrium population in the higher-energy states to produce a spontaneous emission spectrum. Depending on how the nonequilibrium state is produced, we defined different types of luminescence in Sec. 3.1.

The absorption and spontaneous emission spectra are related by the principle of detailed balance as calculated by van Roosbroeck and Shockley. At thermodynamic equilibrium, the rate of spontaneous photon emission $R_{sp}(\nu)$ at frequency ν in an interval $d\nu$ is given by

$$R_{sp}(\nu)d\nu = P_{abs}(\nu)\varphi(\nu)d\nu \tag{3.84}$$

where $P_{abs}(\nu)$ is the probability of absorbing a photon of energy $h\nu$ per unit time, and $\varphi(\nu)d\nu$ is the radiation density of frequency ν in an interval $d\nu$. This is obtained from Planck's radiation law (Appendix 6) as:

$$\varphi(\nu)d\nu = \frac{8\pi\nu^3 n_r^3}{c^3}\frac{1}{\exp\left(\frac{h\nu}{k_BT}\right)-1}d\nu \tag{3.85}$$

The absorption probability $P_{abs}(\nu)$ can be calculated in the following way. If the absorption coefficient of the photon is $\alpha(\nu)$ and it travels with a velocity $\vartheta \equiv c/n_r$ in the material with refractive index n_r, then the mean lifetime of the photon is given by $\tau(\nu) = 1/\alpha(\nu)\,\vartheta$ and the absorption probability is given by

$$P_{abs}(\nu) \equiv \frac{1}{\tau(\nu)} \equiv \alpha(\nu)\vartheta = \alpha(\nu)\frac{c}{n_r} \tag{3.86}$$

Substituting Eqs. 3.85 and 3.86 into Eq. 3.84, we get

$$R_{sp}(\nu)d\nu = \frac{\alpha(\nu)8\pi\nu^3 n_r^2}{c^2[\exp(h\nu/k_BT)-1]}d\nu \tag{3.87}$$

which expresses the desired relation between absorption and emission spectra. Substitution of Eq. 3.68 leads to

$$R_{sp}(\nu)d\nu = \frac{32\pi^2 k_a(\nu)n_r^2\nu^4}{c^2[\exp(h\nu/k_BT)-1]}d\nu \tag{3.88}$$

The total emission rate per unit volume is obtained by integrating Eq. 3.87 over all frequencies, or energies, as

$$R_{sp} = \frac{8\pi n_r^2 (k_B T)^4}{c^2 h^4} \int_0^\infty \frac{\alpha(\nu) u^3}{e^u - 1} du \tag{3.89}$$

where $u = h\nu/k_B T$. Although derived for thermodynamic equilibrium, Eqs. 3.87 and 3.89 express the fundamental relation between absorption and emission spectra for any means of excitation. This formulation is valid for any transition between a higher-energy and a lower-energy state. The relation between emission and absorption spectra is schematically illustrated in Fig. 3.17.

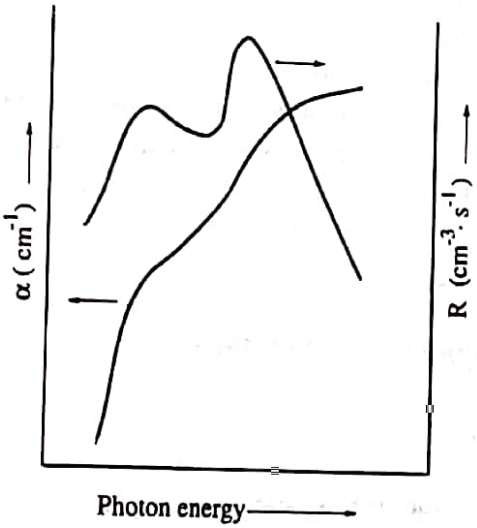


Figure 3.17 Schematic illustration of relation between emission and absorption spectra.

3.6.2 Stokes Shift in Optical Transitions

The Stokes shift is a difference in transition energy of the emission and absorption spectra resulting from defects in the material and, in general, partial nonradiative decay. The process can be understood with respect to the configuration coordinate diagram shown in Fig. 3.18. The lower and upper curves represent the energies of the

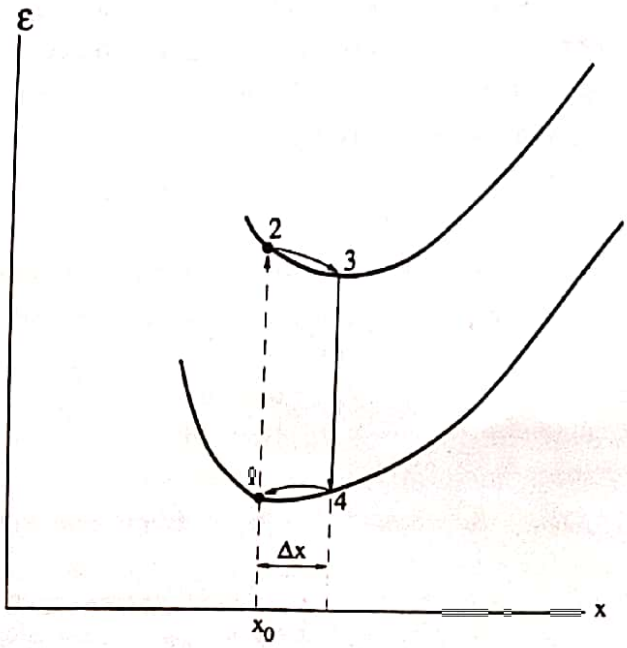


Figure 3.18 Configuration coordinate diagram illustrating the Stokes shift in a semiconductor.

lower and upper states of an optical transition as a function of distance that could be the ground and excited states of an impurity atom, a host lattice atom, or a deep-level trap. The point labeled 1 in the lower curve is the minimum energy, or equilibrium, position in the ground state. Due to photon absorption of energy $\hbar\omega_1$, an electron may be raised to the point 2 in the upper state ($\Delta k = 0$), which is not the minimum energy configuration. The displacement Δx of the excited state may be caused by the defect or impurity potential. Therefore, the system relaxes to the state 3, which is at the lowest energy. This phenomenon is usually termed *lattice relaxation* and a phonon is involved. After living a mean lifetime the excited carrier returns to the ground state by radiative recombination at the point 4, which is again not the minimum energy configuration. The energy of the emitted photon is $\hbar\omega_2$. Therefore, the system relaxes again with phonon participation to the point 1. Both optical and/or acoustic phonons may be involved but usually optical phonons produce the largest change in energy per unit displacement. The energy difference of the photon absorbed and emitted, $(\hbar\omega_1 - \hbar\omega_2)$ is called the Stokes shift or Franck-Condon shift. This degradation of optical energy arises directly from imperfections in the material or interfaces, such as in a heterostructure or quantum well.

3.6.3 Near-Bandgap Radiative Transitions

3.6.3.1 Exciton Recombination. We have seen in Sec. 3.2.3 that electrons and holes produced by the absorption of a photon of near-bandgap energy can pair to form an exciton. Recombination of the electron-hole pair results in a narrow and sharp peak in the emission spectra. The energy of the emitted photon is

$$\hbar\omega = \mathcal{E}_g - \mathcal{E}_{ex} \tag{3.90}$$

where $\mathcal{E}_{ex}$ is quantized. In other words, in very pure crystals emission lines corresponding to the ground state and higher-order states may be seen. The process is shown in Fig. 3.19(a). In indirect bandgap semiconductors, a phonon needs to be involved, in the transition for momentum conservation, as shown in Fig. 3.19(b).

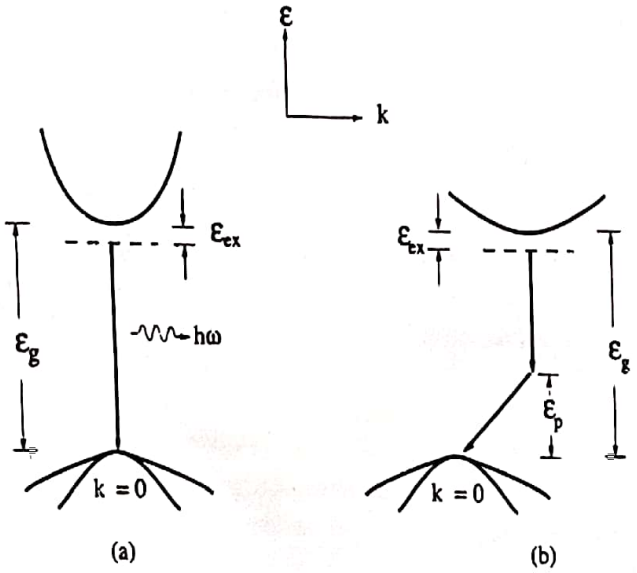


Figure 3.19 Exciton recombination in (a) direct bandgap and (b) indirect bandgap semiconductor.

Therefore, the probability of exciton recombination transitions is very low in indirect bandgap materials.

In semiconductors with impurities—donors and/or acceptors—present, the free exciton couples with the impurity atoms to produce *bound excitons*. Bound excitons produce sharp peaks at photon energies lower than that of the free exciton. Also, the linewidth of the bound exciton resonances are much smaller than that of the free excitons—almost by a factor of 10. In most semiconductors, free and bound exciton resonances are seen simultaneously in the emission spectra. Figure 3.20 shows the low-temperature photoluminence spectrum of high-purity GaAs grown by VPE. In addition to the free-exciton peak, a host of peaks attributed to excitons bound to impurities and electrically active defects are seen in the energy range 1.4–1.51 eV.

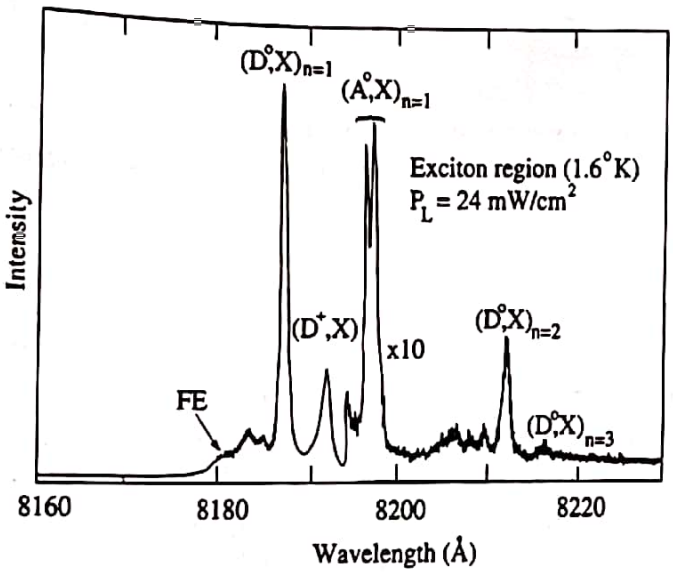


Figure 3.20 Exciton-related recombinations seen in the photoluminescence spectrum of a high-purity GaAs sample grown by the hydride VPE process. FE denotes free exciton and $D^{\circ} - X$ and $A^{\circ} - X$ are neutral donor-bound and acceptor-bound excitonic transitions, respectively. $D^{+} - X$ denotes an ionized donor-bound excitonic transition (from B. J. Skromme et al., *Journal of Electronic Materials*, **12** (2), 433, © 1983 IEEE).

3.6.3.2 Band-to-Band Recombination. If the temperature of the sample is high enough so that $k_B T > \mathcal{E}_{ex}$, or if there are sufficient number of free carriers in the semiconductor producing local fields to dissociate the exciton, then most photogenerated carriers exist as separate electrons and holes in the bands. Most of these free carriers live a mean lifetime and then recombine radiatively. In direct gap semiconductors the process is complementary to the absorption process and electrons recombine with holes with momentum conservation. The energy position of the emission peak depends on the temperature and intensity of excitation. At low temperature and low excitation intensity the recombination is characterized by a single peak with the peak energy or the low energy cut-off at $\hbar\omega = \mathcal{E}_g$. As the temperature or excitation energy is increased, electrons and holes are filled at higher energies in the respective bands and these recombine to produce photons of higher energy. In the emission spectrum this is seen as a temperature or intensity-dependent tail on the high-energy side. Similarly, as the doping of the sample is increased, the whole curve may move to lower energies due to bandtailing effects. Some experimental curves obtained for GaAs are shown in Fig. 3.21. One may wonder at this point about radiative direct transitions involving the light-hole band that is at a higher (hole) energy for $\mathbf{k} \neq 0$. Since the hole mass is considerably lower in this band, the density of states and transition probability is much lower, and therefore transitions involving light holes are normally not seen in emission spectra of bulk semiconductors.

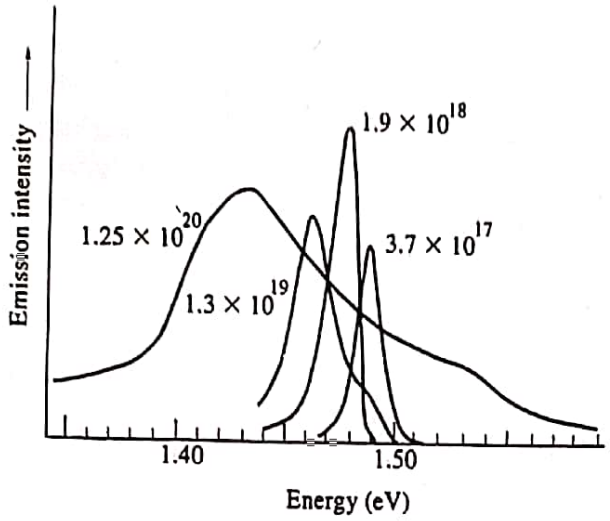


Figure 3.21 Cathodoluminescence spectra of Zn-doped GaAs at 4.2°K (from J. I. Pankove, Proceedings of the International Conference on the Physics of Semiconductors, Kyoto, 1966, *Journal of the Physical Society of Japan*, **21**, Supplement, 1966).

In indirect bandgap semiconductors radiative transitions from the conduction band to the valence band take place with the help of phonons to conserve momentum. Usually the process of phonon emission is most likely and the probability of phonon emission remains high even at very low temperatures.

3.6.3.3 Donor-Acceptor and Impurity-Band Transitions. Intentional and unintentional donor and acceptor levels in semiconductors give rise to radiative transitions. In this section we will restrict the discussion to *shallow*, hydrogenic donors and acceptors, for which the energy of the emitted photons is close to the fundamental bandgap. It may be remembered that in GaAs typical donor energies are between 4 and 8 meV and typical acceptor energies are between 25 and 40 meV. As in the case of absorption, the energy of the emitted photon is given by Eq. 3.61. In general, the donor-acceptor (D-A) transition gives rise to a broad peak in the emission spectrum.

The other important near-bandgap transitions are band-to-impurity transitions, which are complementary to the absorption process. In semiconductors where the donor and acceptor binding energies are nearly equal (due to equality of effective masses), it is not easy to distinguish between the two types of impurity-band transitions (donor-band and acceptor-band). In this case the conductivity type of the material has to be known and the temperature of the sample is an important factor. For indirect impurity-related transitions a phonon emission process is involved, and in this case the emitted photon energy is given by $\hbar\omega = \mathcal{E}_g - \mathcal{E}_i - \mathcal{E}_p$ where $\mathcal{E}_i = \mathcal{E}_D$ or $\mathcal{E}_A$.

It is important to know the difference in transition probability between the D-A and impurity-band transitions, provided there are carriers available for both to occur. These transition probabilities can be determined from quantum mechanical calculations, as outlined for the case of photon absorption in Sec. 3.2. The important parameter is the carrier lifetime for the relevant process, which is crucial for the operation of injection lasers. From these calculations one gets the impurity-band carrier lifetimes of the order of several nanoseconds, while that for the band-to-band transitions varies in the range of several hundred picoseconds to one nanosecond. What it amounts to roughly is that if there are electrons in the conduction band and donor level, and there are holes in the valence band and acceptor level, the probability of

the band-to-band transition is approximately four times that of the impurity-band transitions.

Donor-acceptor or impurity-band transitions can be selectively observed in the luminescence spectra by causing selective occupation of the bandedge or impurity levels. This can be done by altering the temperature of the sample or by changing the excitation intensity. Consider a GaAs sample that has both donor ($\mathcal{E}_D \simeq 5$ meV) and acceptor ($\mathcal{E}_A = 30$ meV) impurities. At very low temperatures (~ 4°K) both donor and acceptor levels are occupied with electrons and holes, respectively, and the prominent peak seen in the luminescence spectrum is due to a D-A transition. If the temperature is raised slightly, to 20°K, some of the donor atoms are ionized and the electrons from these levels are raised to the conduction band. The acceptor level will still remain filled with holes. A shoulder develops to the high-energy side of the D-A peak in the luminescence spectrum, which corresponds to band-to-acceptor te (B-A) transitions. As the temperature is raised, the B-A transition becomes more prominent and the D-A transition is quenched. In fact, the energy separation between the two peaks corresponds to the donor ionization energy. The D-A and B-A peaks are schematically shown in Fig. 3.22. Selective occupation of donor levels and the conduction bandedge can also be achieved at a fixed (low) temperature by varying the excitation intensity. At low excitation (photon density) levels the D-A transition is the prominent one. As the excitation intensity is increased, the conduction band is filled with electrons and the B-A transition becomes more prominent.

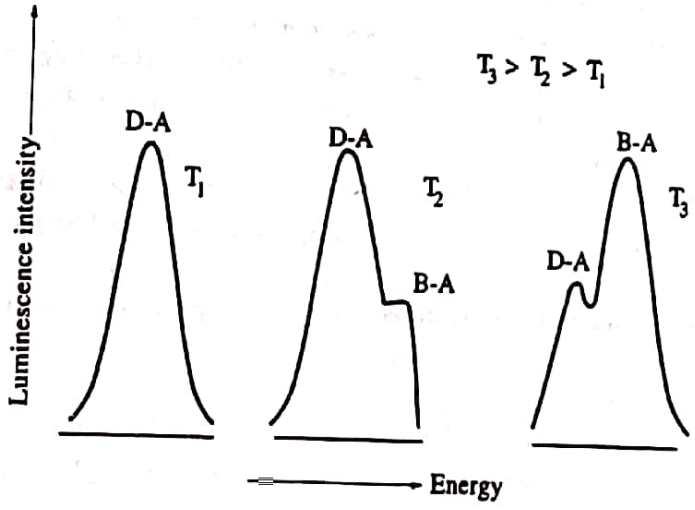


Figure 3.22 Schematic illustration of the evolution of band-acceptor and donor-acceptor transitions in the PL spectra of a semiconductor at varying temperatures.

As a concluding note, it should be mentioned that impurity-band transitions, involving the impurity level *closest* to the bandedge, can give rise to low-energy transitions. The transition energy corresponds to the impurity binding energy. Peaks believed to be due these transitions have been observed in the far-infrared region of the emission spectra of semiconductors. However, considering the transition probabilities it is debatable whether such transitions are radiative or nonradiative. In other words, instead of emitting a photon, single or multiple phonons may be emitted. In the latter case the excess energy is dissipated as heat in the lattice.

3.7 DEEP-LEVEL TRANSITIONS

Deep levels in the forbidden energy gap of semiconductors essentially act as carrier recombination or trapping centers and adversely affect device performance. Native