1

ELECTRONIC MATERIALS AND CHARGE TRANSPORT

This chapter presents an overview of the quantum mechanical nature of electrons in a *solid*. Following this discussion, through adoption of a relatively simple language, we will attempt to tackle a number of very complicated phenomena (such as *charge transport*). One of the remarkable achievements of the pioneers of semiconductor electronics has been their success in proposing an intuitive set of simplifying assumptions for reducing the mathematically complex language of quantum mechanics to the closed-form descriptions of the *effective* charged particles. While oftentimes we find it hard to justify these assumptions, they have proven their capabilities for large devices and under low electric fields.

1.1 WAVE/PARTICLE ELECTRONS IN SOLIDS

Early in the twentieth century, a number of important experiments revealed that electrons are not just simply particles in the *Newtonian* sense. Electrons in some of the experiments demonstrated a wavelike nature. These experiments were very similar to the *double-slit* optical experiments. It was demonstrated that electrons coming from the two slits, if not *observed*, produce interference patterns expected from wave propagation only. However, if we try to *observe* the electrons, by means such as the study

Field Effect Transistors, A Comprehensive Overview: From Basic Concepts to Novel Technologies, First Edition. Pouya Valizadeh.

of absorption and emission of light, the *uncertainty*[1] created by these interactions reduces electrons to almost *Newtonian* particles. The same argument can be applied to the movement of electrons in a perfectly crystalline lattice (also known as *monocrystalline* lattice). As a result of these experiments, *wave/particle* duality is attributed to electrons. In this duality, the *de Broglie wavelength*[2] is the *wavelength* assigned to a particle of momentum p (given by $\lambda = h/p$).

Electronic materials, which are suitable for fabrication of high-performance semiconductor devices, come in *monocrystalline solid* forms. The presence of both *long-* and *short-range* orders in the structure of these crystalline *solids* extends many important properties to *charge transport* through these media. However, the assumption of periodicity in crystals is always relative. In real crystals a number of intentional and unintentional mechanisms (such as introduction of impurities, crystal defects, and thermal vibrations) result in *scattering* of electrons after a typical travel distance on the order of 100 Å. Such *scattering* processes can be seen as mechanisms of *observation* of electrons. Of course, the distance paced between successive *scattering* events is dependent on the presence and dominance of different *scattering* mechanisms. The most unavoidable crystal imperfection is rooted in *lattice vibrations*, which are even present in a defectless crystal devoid of impurities at temperatures above 0 K. As a result of these *scattering* events, instead of seeing a three-dimensional large crystal, electrons are only exposed to a small volume on the order of $10^{-18}\,\text{cm}^3$ containing only about 3000 atoms, before they are *dephased* by *lattice vibrations*.

In a flawless *monocrystalline* lattice, electrons are predicted by quantum mechanics to travel as *propagating waves*. This results in either a *persistent current* or in *oscillations*.[3] Electron interactions with the vibrations of the lattice (expressed in terms of *observation* by quantum particles known as *phonons*) and the resulting generation of *Joule heat* prevent electrons from acting in such a fashion. The emission of *phonons* with a broad range of energies and *wave vectors* results in *dephasing* of electrons and loss of *coherence*.

Although the *electron–phonon* interactions do not result in total elimination of quantic nature of electrons, they considerably weaken these properties. This reduced degree of quantum nature is often expressed in terms of *perturbation theory*, through employing the so-called *Fermi golden* rule.[4] Throughout this chapter, with the help of analytical models developed on the basis of this theory, familiar characteristics of semiconductors (e.g., their sometimes *Ohmic* behavior[5]) are evaluated. Rivaling the results of this approach, the only way that quantum mechanics can produce an

[1] *Heisenberg*'s *uncertainty principles* (which are named after *Werner Heisenberg*) state that the *uncertainties* of two conjugate variables such as momentum and position (i.e., Δp and Δx, respectively) cannot be both reduced to 0, since $\Delta p \Delta x \geq h/2\pi$. This is also true between uncertainty of energy and time. Where h, named after *Max Planck*, is the *Planck*'s constant, which is equal to 6.63×10^{-34} J s.

[2] Named after *Louis de Broglie*.

[3] These phenomena will be further elaborated in Section 1.7.

[4] Named after *Enrico Fermi*.

[5] Named after *Georg Ohm*.

Ohmic characteristic is when the complicated *many-body Schrödinger equations*[6] for electrons and *phonons* are solved.

Since electrons in *solids* are subject to the laws of quantum mechanics, any discussion of *electron transport* through *solids* requires at least a rudimentary description of the quantum mechanical wave nature of electrons. Although the present text does not seek to present this picture through explicitly invoking the laws of quantum mechanics, a number of important outcomes of such studies are reviewed in this section.

1.1.1 Quantum Description of Electrons

In order to formulate the *wave function* of electrons (which is often represented in form of function $\psi\left(\vec{r}\right)$ in space) the *time-independent* 3-D *Schrödinger equation* should be solved,

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + U\left(\vec{r}\right)\right]\psi\left(\vec{r}\right) = E\psi\left(\vec{r}\right). \tag{1.1}$$

This equation can be presented equivalently through invoking the concept of *Hamiltonian*[7] (i.e., *H*),

$$H\psi\left(\vec{r}\right) = E\psi\left(\vec{r}\right). \tag{1.2}$$

In (1.1) and (1.2), $\hbar$ is the *modified Planck's constant*,[8] m is the electron mass,[9] E is the energy, and vector $\vec{r}$ represents the spatial coordinates. In these partial differential equations, the potential function $U\left(\vec{r}\right)$ is assumed to be time independent.

For a constant potential (i.e., the case of an electron in *free space*), the solution is rendered in the form of plane waves,

$$\psi\left(\vec{r}\right) = \frac{1}{\sqrt{\Omega}}\exp\left(j\vec{k}\cdot\vec{r}\right) \tag{1.3}$$

where Ω represents the *normalization volume*, which is defined using the square of the amplitude of the *wave function* as the *probability density function*. Evidently, $\vec{k}$ in (1.3) is the *wave vector*. In quantum mechanics, the eigenvalue of momentum is given by $\hbar\vec{k}$.

Oftentimes, semiconductor devices such as *field-effect transistors* (*FETs*) are realized as a 2-D plane through which *charge transport* takes place. Assuming a semiconductor slab in the x–y plane with an infinitesimal thickness W in the z-direction,

[6] Named after *Erwin Schrödinger*.

[7] Named after *William Hamilton*.

[8] $\hbar = h/2\pi$, where $h = 6.62\times10^{-34}$ J s.

[9] In here, intentionally the notions of effective mass (m^*, which is to be defined in Section 1.1.2) and the rest mass of electron (i.e., $m_0 = 9.1\times10^{-31}$ kg) are not used. m will be later replaced by m^*.

quantum mechanics provides a picture for electrons confined in the z-direction while free to move in the x–y plane. For these quasi 2-D electrons, separation of variables of the 3-D *time-independent Schrödinger equation* results in

$$\psi(\vec{r}) = \phi(z)\cdot\varphi(x,y) = \phi(z)\cdot\frac{1}{\sqrt{A}}\exp\left[j\left(k_x x + k_y y\right)\right] = \phi(z)\cdot\frac{1}{\sqrt{A}}\exp\left(j\,\vec{k_{\|}}\cdot\vec{\rho}\right) \tag{1.4}$$

in which A is the *normalization area* and $\vec{\rho}$ is a vector in the x–y plane. In the case of a confining potential, which imposes an infinite barrier against the movement of electrons normal to x–y plane, quantization of energy will yield

$$\phi(z) = \sqrt{\frac{2}{W}}\sin(k_n z) = \sqrt{\frac{2}{W}}\sin\left(\frac{n\pi z}{W}\right), \quad n = 1,2\ldots \tag{1.5}$$

This is the case of the so-called infinite *potential well*.

If the confinement is further extended to the y-direction and the confining potential is turned into a simple square well of infinitely high barriers, one will have

$$\psi(\vec{r}) = \phi(y,z)\cdot\varphi(x) = \phi(y,z)\cdot\frac{1}{\sqrt{L}}\exp[j(kx)] \tag{1.6}$$

where

$$\phi(y,z) = \frac{2}{W}\sin\left(\frac{m\pi y}{W}\right)\sin\left(\frac{n\pi z}{W}\right), \quad m,n = 1,2\ldots \tag{1.7}$$

in which L is the *normalization length*. This is the case of the so-called quantum *wire*.

Figure 1.1 provides schematics for a *quantum well* and a *quantum wire*.

What is common between the cases of electron in *free space*, confinement in a *quantum well*, and confinement in a *quantum wire* is that according to (1.3), (1.4), and (1.6) for all three cases, electrons assume a *propagating-wave* nature. What this means is that electrons feel almost free to move, either in 3-D, 2-D, or 1-D space.

As mentioned earlier, confinement of electrons in one and two dimensions results in energy quantization. This is expressed through integer values of n and m.

Assuming a *potential well* in which an electron is only allowed to move between the barriers and not parallel to them, energy quantization is expressed by

$$E_{\mathrm{n}} = \frac{\hbar^2\pi^2}{2m^* W^2}n^2, \quad n = 1,2,\ldots \tag{1.8}$$

where, as shown in Figure 1.1, W is the distance between the two barriers.

In this equation, each of the values of energy expressed by E_{n} represents an allowed *energy level* (or overlooking the *spin* of an electron: *energy state*). Electrons are prohibited to assume all other values of energy. In the case of a finite *potential well*,

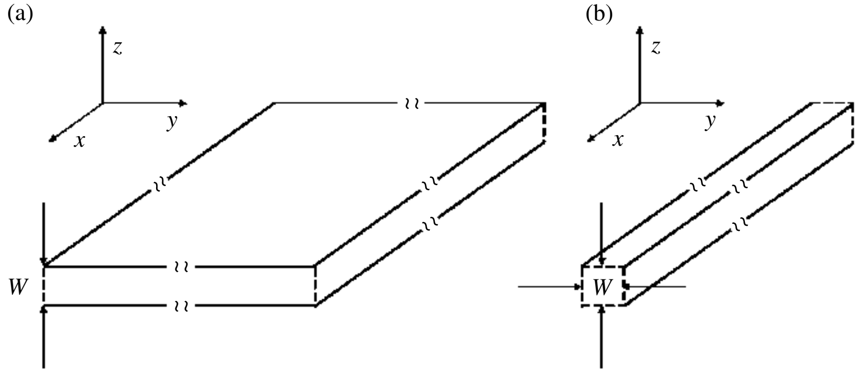


FIGURE 1.1 (a) A 2-D confining quantum well. (b) A 1-D confining quantum wire.

quantization is present to a lesser degree, and a recursive equation should be solved to calculate the allowed energy values.

Extending the degrees of freedom of electrons in the *potential well* results in overlapping of the *wave functions* of the neighboring electrons residing on each of these *energy states*. As a result, due to the restrictions of *Pauli's exclusion principle*,[10] each quantized *energy state* will split into a number of very closely packed *energy states*. As will be seen later on in this section, these split *energy levels*, depending on the degree of confinement, can grow into *bands* or the so-called subbands of energy.

In the form of confinement in a 2-D plane, the *quantum-well* situation especially resembles that of the thin silicon body of a *silicon-on-insulator*[11] *MOSFET*. This device will be seen in Chapter 3. In this context, each of the *energy levels* expressed in (1.8) stands for a *subband* with many allowed *momentum states* (also known as *k-states*) in the x–y silicon plane. As shown by (1.8), increasing the confinement (i.e., reducing W) increases the separation of the *subbands*. As will be observed later on in this chapter, this is important to the reduction of the chance of electron *scattering* from one *subband* to the next. In this system, through assigning an *effective* value of mass to electrons (i.e., m^*), the total energy written as the sum of kinetic energy and confining energy is given by

$$E\left(\vec{k}\right)=E_n+\frac{\hbar^2k_\parallel^2}{2m^*},n=1,2,3,\ldots \tag{1.9}$$

where $k_\parallel^2=k_x^2+k_y^2$. The relationship between the energy and momentum (which is represented by $\hbar k$) is known as the dispersion relationship (also known as E–k).

As observed through our brief encounter with the *Schrödinger equation*, knowledge of the *Hamiltonian* and the potential function is pivotal to understanding the

[10] Named after *Wolfgang Pauli*.

[11] Also known as *SOI*.

behavior of electrons in different *solid* structures. The following laws of mechanics govern the time evolution of position (i.e., x) and momentum (i.e., p) in terms of H:

$$\frac{\partial p_i}{\partial t} = -\frac{\partial H(p_i, x_i)}{\partial x_i} \tag{1.10}$$

$$\frac{\partial x_i}{\partial t} = \frac{\partial H(p_i, x_i)}{\partial p_i}. \tag{1.11}$$

The vector of average velocity for electrons is also identified as

$$\vec{v} = \frac{1}{\hbar} \nabla_k E\left(\vec{k}\right). \tag{1.12}$$

1.1.2 Band Diagram and Effective-Mass Formalism

The most basic outcome of such a quantum mechanical entity (i.e., electrons in a *solid*) appears in the form of *energy bands* and *forbidden gaps*. The evolution of the *band diagram* composed of these *energy bands* and *forbidden gaps* (and its more detailed version known as the *E–k* diagram) is deeply rooted in the interaction between the wave-natured electrons and the periodic potential function of the crystalline *solid*. Within each of the allowed *bands*, the variation of energy with momentum is calculated and represented in the form of energy versus momentum (i.e., *E–k*) diagrams.[12]

A major component of the simplified theory representing electrons in *solids* is the *band theory*, which is the outcome of *effective-mass theory*. *Effective-mass theory* is based on the resemblance of the form of the *electron wave dispersion diagram* at the energy range of interest[13] to that of electrons in *free space*. For an electron in *free space*, since the *effective mass* (i.e., m^*) is clearly a constant (i.e., the rest mass of the electron: m_0), Equation (1.9) reduces to a parabola.

Although in a real semiconductor *charge carriers* are not truly traveling as *propagating waves*, under many important circumstances, they almost behave that way. These almost free *charge carriers*, however, feel a different mass than that of the free electron. Based on the almost parabolic form of the *dispersion diagram*, *effective-mass theory* uses the curvature of this diagram to assign a value of mass to *charge carriers*,

$$m^* = \frac{\hbar^2}{\partial^2 E / \partial k^2}. \tag{1.13}$$

While this simple picture successfully reduces the *wave/particle* electrons to simple particles, it fails to hold when *charge carriers* receive substantial amounts of kinetic energy. Under those circumstances, the resemblance in form to the *dispersion diagram* of a free electron vanishes.

[12] Also referred to as *dispersion diagram*.
[13] This will be clarified shortly.

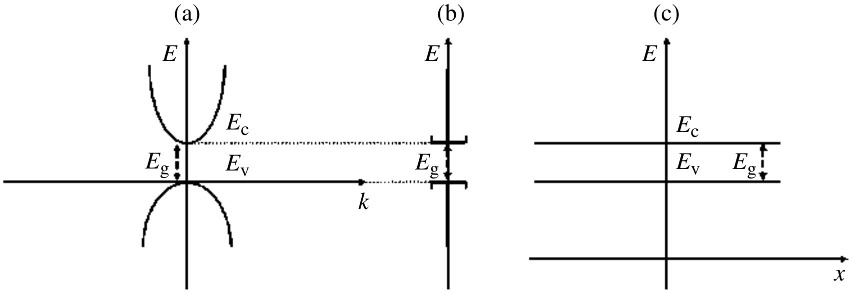


FIGURE 1.2 (a) A simplified typically observed *E–k* diagram among a number of semiconductors. As will be indicated shortly, in this diagram E_c and E_v mark the bottom of the conduction and the top of the valence band and E_g represents the size of the forbidden gap of energy. (b) A 1-D representation of the bandgap. (c) The 2-D band diagram.

In some situations the complete set of information offered in an *E–k* diagrams is needed; however, in most cases of interest to the present volume, we only adopt a selective portion of this information. As will be pointed out shortly, this selection of information in the *E–k* diagram comes from calculating the *E–k* curvature given specific values of momentum matched to the edges of two of the *bands* resulting from splitting of *energy levels*.

Localization of electron *states*, caused by unavoidable imperfections in real semiconductors, can be incorporated into this model through the corrective measure of *scattering* processes. Incorporation of such small perturbations to the perfect crystal is implemented through determining the *scattering* rates using *Fermi golden rule* (or *Born approximation*[14]).

Under these presented conditions, the *E–k* diagram is reduced to merely a one-dimensional energy diagram, which, at least for the sake of presentation, is often expanded into a second dimension (i.e., position). Figure 1.2 illustrates a typical *E–k* diagram and the resulting 1-D and 2-D *band* diagrams. As pointed out in these illustrations, *forbidden gaps* of energy are present among allowed energy *bands*. As shown, the momentum information is not presented in an energy-*band* diagram. Later on in this chapter, through invoking the concept of *momentum relaxation time constant*, we will discuss the significance of this lost bit of information.

The curvatures of the *E–k* diagram and therefore the values of the *effective mass* are different along different directions in a crystal. At this point in our discussion, it is important to point out that in calculating the *subband* energies in (1.8), *effective mass* in the direction of the confining potential should be employed.

1.1.3 Density of States Function

As indicated earlier, the *bands* of energy are composed of very narrowly packed sets of individual *energy states*. However, since at normal temperatures of operation of an

[14] Named after *Max Born*.

electronic device the separation between the individual *states* is much smaller than the average random thermal energy of an electron in a 3-D *solid* (i.e., $3/2kT$[15]), electrons tend to see the *band* as a continuum of *energy states*. Since the occupancy of each individual *energy state* is governed by *Pauli's exclusion principle* (i.e., maximum accommodation for only two electrons that must have opposite spins at each *energy level*), each *band* has a certain maximum capacity for electrons, which is identified by a *density of states* (i.e., *DOS*) *function*. This function is a double-density function, and it essentially expresses the *density of states* per unit energy, per unit volume.

For three-dimensionally moving *charge carriers*, the *DOS* at the bottom of a *band* (i.e., identified by E_c) can be approximated by

$$D_{3D}(E) = \frac{(2m^*)^{\frac{3}{2}}}{2\pi^2\hbar^3}\sqrt{E-E_c}. \tag{1.14}$$

However, in the case of 2-D *charge-carrier* confinement, the two-dimensional *density of states* per unit energy and area is given by

$$D_{2D}(E) = \frac{m^*}{\pi\hbar^2}. \tag{1.15}$$

This is only with the assumption of the first *subband*. Finally, with the same assumption the *DOS* in the one-dimensional case is given by

$$D_{1D}(E) = \frac{\sqrt{2m^*}}{\pi\hbar}\frac{1}{\sqrt{E-E_c}}. \tag{1.16}$$

Figure 1.3 illustrates the dependence of the *DOS* function on energy for the three cases. Due to the direction dependence of *effective mass*, the value of the *effective mass* used in calculation of the *DOS* is very often different than that used in the *transport* problem.

1.1.4 Conduction and Valence Bands

Due to electrons' natural tendency to arrive at minimum enthalpy, electrons occupy *states* of lower energy before filling up *states* of higher energy. In the presence of thermal energy, however, this picture gets slightly distorted. This is caused by thermal excitation of electrons from lower *energy states* to higher *energy states*. In the case of a semiconductor, this renders lower *energy bands* to be full, renders higher *energy bands* to be empty, and also causes two of the adjacent *bands* to be only partially occupied at temperatures above 0 K. At 0 K, however, a semiconductor behaves like an insulator and is composed only of full and empty *bands*. Figure 1.4 illustrates this

[15] This is based on statistical mechanics. According to *equipartition* theorem in 1-D and 2-D cases, this energy is equal to $1/2kT$ and kT, respectively. k is the *Boltzmann* constant, which is equal to 1.38×10^{-23} J/K.

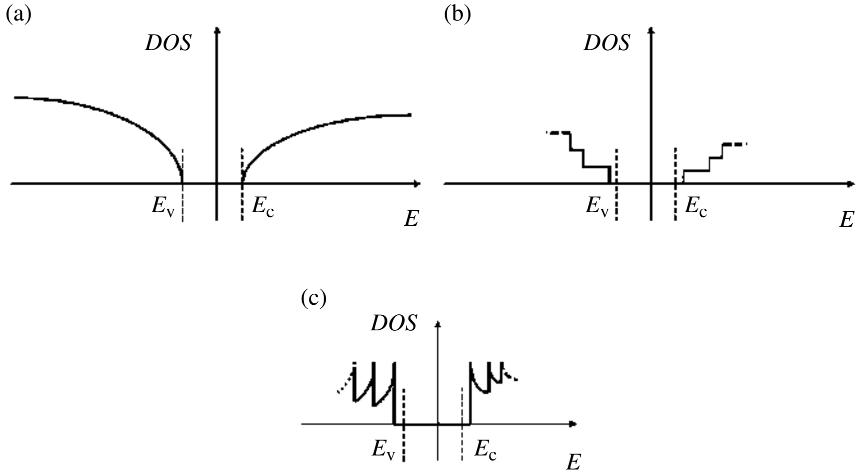


FIGURE 1.3 (a) Typical form of the density of states function in a 3-D semiconductor. (b) Typical form of the *DOS* function in a quantum well. (c) Typical form of the *DOS* function in a quantum wire.

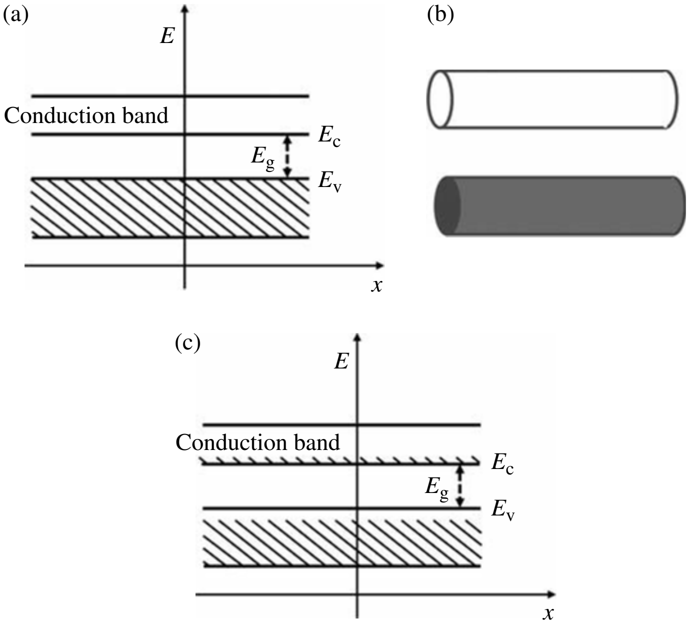


FIGURE 1.4 (a) Occupancy status of the conduction and valence band of a semiconductor at 0 K, where the full portion of a band is hash marked. (b) The water pipe analogous to the band occupancy presented in (a). In this analogy the full portion of the pipe is presented in gray, while the empty portion is white. (c) Occupancy status of the conduction and valence band of a semiconductor at a finite temperature, where the full portion of a band is hash marked. One can see the electron transfer from valence to the conduction band in analogy with water transfer from the lower to the upper pipe.

picture at 0 K and at a temperature T, in analogy with doubly sealed water pipes. According to this picture, since neither a full nor an empty *band* can contribute to net movement of *charge carriers* (i.e., in water pipe analogy: water), among all of the *bands* resulting from the aforementioned quantum mechanical description, only those *bands* that at normal temperature of operation of a device are partially full/partially empty are of prime value.[16]

A partially full *band* is created by means such as thermal excitation of electrons from the top of a lower *band* to the unoccupied *states* at the bottom of a higher *band*. In semiconductors we therefore encounter two partially full/partially empty *bands*. All other *bands* below these remain completely full, and those above remain completely empty. As a result, only two *bands* are worth mentioning: the *valence* and the *conduction band*. The *valence band* is the highest *band* that is full at 0 K, and the *conduction band* is the lowest *band* that is empty at 0 K. The energy difference between the bottom of the *conduction band* (i.e., E_c) and the top of the *valence band* (i.e., E_v) is the *bandgap* (i.e., E_g).

Since the size of the *forbidden gap* is much larger than the average amount of thermal energy acquired by an electron in a three-dimensional *solid*, one can expect such a *charge-carrier* transfer to be far more probable between the top of the *valence band* and the bottom of the *conduction band*. These values of energy are therefore much more important to study than the other *energy levels* in the rest of *conduction* and *valence bands*.

Although approximate equations such as (1.14) present the *DOS* of *conduction* and *valence band*[17] in the form of a square-root law, higher in the *conduction band* (and also lower into the *valence band*) the proportionality of *DOS* to $\sqrt{E-E_c}$ (and $\sqrt{E_v-E}$ in the *valence band*) vanishes. This is due to the more complicated *E–k* variation at higher values of energy. For these higher energies, numerical methods are needed to evaluate the *DOS*. As will be observed shortly, even for the case of lower energies, in the presence of high concentrations of impurities, there is a need to reevaluate this square-root law.

1.1.5 Band Diagram and Free Charge Carriers

The *band diagram* representation of Figure 1.5, through extending the *bandgap* information in space, substantially helps in implying the relatively free nature of electrons in the *conduction* and *valence band* (i.e., those that are not associated with the nuclei of the atoms in the crystal). This relative freedom is implied since the electrons are now free to be at different positions in the semiconductor without any need for changing their energy value. The energy axis of the *band* diagram imports only the energy difference between the lowest energy value of the *conduction band* (i.e., E_c) and the highest energy value of the *valence band* (i.e., E_v) of the *E–k* diagram. As a result, as mentioned earlier, momentum and curvature information of the *E–k* diagram are simply omitted (of course with the exception of the curvature used in calculation of the

[16] Just like the case of a *zero* net movement of liquid in a full, or an empty, doubly sealed water pipe.
[17] In this equation, in case of *valence band* $\sqrt{E-E_c}$ changes to $\sqrt{E_v-E}$.

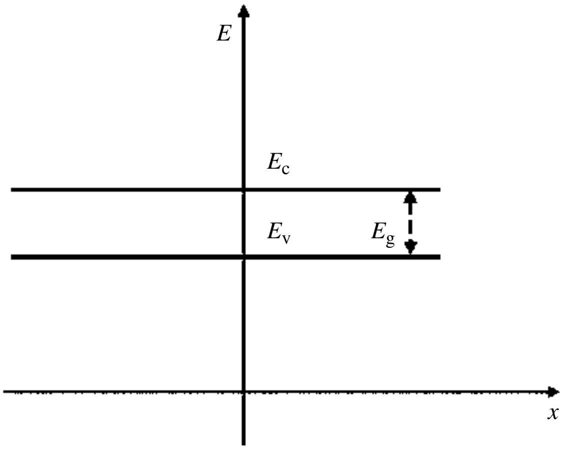


FIGURE 1.5 Band diagram in a semiconductor.

TABLE 1.1 Energy Range of Interest in the Band Structure of Semiconductor for Understanding the Behavior of Various Electronic Devices and under Different Conditions

Electronic Device	Energy Range of Interest within the Bandgap
Low electric-field regions of a transistor	Approximately 2*kT* from the band edges
High electric-field regions of a transistor	Approximately 0.5 eV
Power transistors and devices operating close to breakdown conditions or operating on those basis	The full size of the bandgap

k, is the Boltzmann constant; *T*, is the temperature in Kelvin.

effective electron mass). While in reality in *charge transport* through electronic devices larger than tens of *nanometers*, the presence of many *scattering* processes results in the loss of *coherence* in *transport* and the randomization of momentum, this simplifying omission of momentum information comes at a very affordable cost. As mentioned earlier, this cost is paid through adoption of the notion of *effective* electron mass.

Knowing that the *band diagram* provides only a selective set of information about the energy and momentum, which is sufficient only for specific applications, Table 1.1 provides a summary of links between the considered range of energy in the *band* structure and various electronic devices. This table can provide a sense of the limitations inherent to the *band diagram*.

1.1.6 Supplementary Notes on Band Diagram

A two-dimensional *band diagram*, such as the one shown in Figure 1.5, is often used while discussing the behavior of electrons in *bulk* semiconductors. However, the

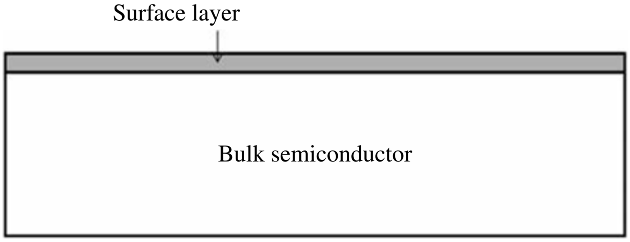


FIGURE 1.6 Surface layer is often treated as a separate layer envisioned at the surface, with a certain concentration of energy levels, known as surface states, within the bandgap of the semiconductor.

one-sided potential imposed on the electrons at the surfaces and interfaces disrupts the periodicity of potential applied on electrons and as a result the formation of the *bandgap*. Nevertheless, with the goal of simplifying the complicated reality of these surfaces and interfaces, oftentimes the aforementioned notion of a *band diagram* is readily extended to the surface. The presence of a surface is then factored into the *band diagram* through improvising an interfacial layer known as *surface layer*. Since most FETs operate through *charge transport* in the vicinity of these surfaces and interfaces, later on in the chapter, we will deal with these nonidealities at greater length. Figure 1.6 shows an example of incorporation of such a layer.

In the operation of semiconductor devices, the presence of *surface states* (which as implied in Fig. 1.6 are present in the *bandgap* of the *surface layer*) often results in hysteresis of the devices' characteristics. In order to avoid this, in high-speed semiconductor devices, bare surfaces should always be avoided. This can be done through implementing the so-called *self-aligned* technologies or even by passivating these surfaces through *deposition* of an insulator layer over them. In later chapters, these technologies will be covered in greater depth.

Additionally, while periodicity of crystal structure is key in developing a *band diagram*, in the description of the behavior of electronic devices, the notion of a *band diagram* is extended to *amorphous* materials (such as dielectrics used in *FET* technologies), which clearly do not possess a periodic structure. In strictest sense such an extension is not allowable. However, in these materials the presence of *short-range* order induces effects that are approximately explainable with the aid of a *band diagram*. One of the most important assets of *silicon* technology has been the possibility of achieving an almost perfect interface between the crystalline *silicon* and *amorphous* SiO_2. Thermal growth of SiO_2 on *silicon* results in a few *dangling bonds* with very little *bond-angle* distortion at the Si/SiO_2 interface. Even the few present *dangling bonds* can be saturated by the presence of *hydrogen* or *fluorine* throughout the *thermal oxidation* process. While SiO_2 is an *amorphous solid*, it exhibits an almost tetrahedral arrangement at short range.

As already pointed out, *bandgap* is an important property of semiconductors. This property is itself dependent on a number of factors including temperature. Table 1.2 summarizes the temperature dependence of the size of the *bandgap* of a number of important semiconductors.

TABLE 1.2 The Temperature Dependence of the Size of the *Bandgap* of a Number of Important Semiconductors

Semiconductor	Temperature Dependence of the Size of the Energy Bandgap
Si	$E_g(\text{eV}) = 1.17 - 4.73 \times 10^{-4} \cdot T^2/(T+636)$
Ge	$E_g(\text{eV}) = 0.742 - 4.8 \times 10^{-4} \cdot T^2/(T+235)$
GaAs	$E_g(\text{eV}) = 1.519 - 5.405 \times 10^{-4} \cdot T^2/(T+204)$ for $0 < T < 1000$
InAs	$E_g(\text{eV}) = 0.415 - 2.76 \times 10^{-4} \cdot T^2/(T+83)$ for $0 < T < 300$
InP	$E_g(\text{eV}) = 1.421 - 4.9 \times 10^{-4} \cdot T^2/(T+327)$ for $0 < T < 800$
InSb	$E_g(\text{eV}) = 0.24 - 6 \times 10^{-4} \cdot T^2/(T+500)$ for $0 < T < 300$
GaN	$E_g(\text{eV}) = 3.47 - 7.7 \times 10^{-4} \cdot T^2/(T+600)$
AlN	$E_g(\text{eV}) = 6.21 - 1.799 \times 10^{-4} \cdot T^2/(T+1462)$ for $0 < T < 300$

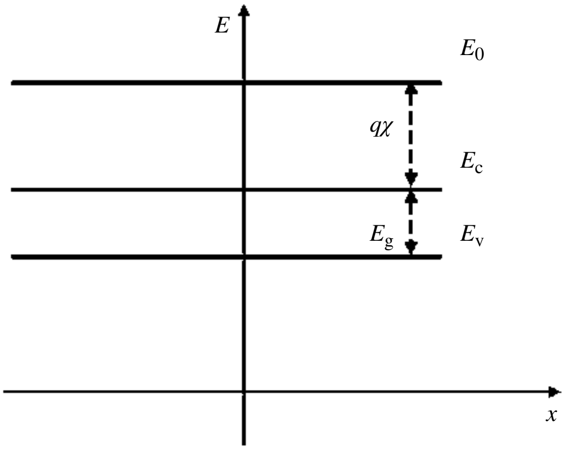


FIGURE 1.7 Identification of electron affinity (i.e., $q\chi$) of a given semiconductor on band diagram.

Another piece of information that is present in a *band diagram* is referred to as *electron affinity* (i.e., $q\chi$). This quantity represents the amount of energy required to set a *conduction-band* electron free from the *bulk* of the semiconductor. This amount of energy is denoted by the difference between the bottom of *conduction band* (i.e., E_c) and *vacuum energy level* (i.e., E_0). Figure 1.7 illustrates this in a *band diagram*. The definition of *electron affinity* is only loosely connected to measurable quantities such as *photothreshold*. This is due to the contribution of other factors including *image forces*[18] and the formation of *surface layer* to these measurable quantities. As a result, *electron affinity* remains as a quantity that can only be used in relative terms with a certain attributed approximation.

[18] These will be dealt with in Chapter 2.

1.1.7 Bond Model

A parallel description often used in explaining the electronic behavior of semiconductors is the *bond model*. The analog of migration of electrons across the *bandgap* in the *bond model* is the breaking away of electrons from covalent bonds. In other words, electrons in the *valence band* represent the electrons in covalent bonds with the atoms in the crystal, while those in the *conduction band* are the ones that have broken away from the covalent bonds and are almost free to roam within the confines of the crystal. While the *bond model* provides a simple picture for the electrons in the *conduction* and *valence bands*, it fails to satisfactorily describe the quantic nature of the *wave/particle* electrons and the role of the *density of states function* in each *band*.

1.2 ELECTRONS, HOLES, AND DOPING IN SEMICONDUCTORS

According to the *energy-band* and *bond* models, in an ideal semiconductor the number of electrons missing from the covalent *bonds* (i.e., missing from the *valence band*) and the number of free electrons (i.e., *conduction-band* electrons) are equal to one another. Such a semiconductor is referred to as *intrinsic* semiconductor. In the operation of electronic devices, however, we deal with semiconductors in which electron concentration in the *conduction band* (i.e., n_0 in cm^{-3})[19] and concentration of energy *states* devoid of electrons in the *valence band* (i.e., p_0 in cm^{-3}) are often not equal to one another. These semiconductors are referred to as *extrinsic* semiconductors.

The process of forming of an *extrinsic* semiconductor is referred to as *doping*. *Doping* occurs when a limited concentration of impurities is introduced into the semiconductor, and impurities are driven to replace some of the compositional atoms of the crystal. There are a few different *doping* processes. In these processes, impurities (also referred to as *dopants*) are either *in situ* incorporated into the crystal structure, as the crystal is being grown, or added in after crystallization. In the latter case, dopants are either *thermally diffused* into the crystal at the gas/solid interface (i.e., due to their higher concentration at the gaseous side) or *implanted* into the crystal in the form of highly energized ions. Each of these techniques has its own advantages in terms of cost/accuracy balance in the creation of the *dopant* profiles. The creation of a certain *dopant* profile (i.e., with precise control on the concentration, depth, and lateral extent of the distribution of impurities) and maintenance of their form throughout the operation of the device (i.e., when the device is exposed to high electric fields and temperatures) are keys to the success of a *doping* process.

1.2.1 Electrons and Holes

In both *doped* and *undoped* semiconductors, empty *energy states* of the *valence band* comprise only a small fraction of the *states* available in that *band*. Likewise, only a

[19] Whereas electron density in metals is in the order of 10^{23} cm^{-3}, in semiconductors this value is at least *two* orders of magnitude smaller.

small fraction of the *states* of the *conduction band* are filled by electrons. Based on the knowledge of *Pauli's exclusion principle* and the limited capacity of each of these *bands* for electrons, *charge transfer* can be envisioned either through movement of electrons or equally well through the transfer of empty *states* (of course in the opposite direction). As previously done in Figure 1.4, if we imagine a doubly sealed partially full water pipe, we can analogously say that the net movement of liquid caused by tilting the pipe is also representable through the movement of bubble in the opposite direction. In dealing with *charge transfer* in the *conduction band*, electrons are fewer in number than the empty *states*. Between the two equivalent presentations of *charge transfer*, it will be much simpler to focus on electrons. The opposite of that happens in the *valence band*. In the *valence band*, instead of electrons, charge transfer is studied through the movement of their complementary profile (i.e., profile of empty *states*). Of course, it will be easier to treat this complementary profile if we use a more addressable name for the empty *states*. Traditionally we call these empty *states holes*.[20]

Increasing the temperature of an *undoped* semiconductor causes the generation of electrons and *holes* in pairs (known as *electron–hole pairs* (EHP)). Caused by a number of events including collisions, these *charge carriers* can also go through the inverse of this *generation* process (which is known as *recombination*). During the *recombination* process, electrons fall back into the *valence band* and effectively annihilate an equal number of *holes*. These processes can either take the form of direct *band-to-band* transitions or be *assisted* transitions involving *impurity states* or *excitonic states*.[21] Through the process of *recombination*, the energy difference (i.e., almost equal to E_g) is either emitted in form of *photons* or is nonradiatively passed onto other particles such as *phonons*.

Holes have an equal amount of charge but are opposite in polarity to electrons. *Hole effective mass* is calculated according to the quantum mechanical information expressed in the *E–k* diagram of the *valence band*. *E–k diagrams* or *band diagrams* are developed for a negatively charged particle (i.e., electron). The opposite charge polarity of a *hole* requires us to look at these diagrams upside down while studying *holes*. Whereas the basis of calculation of the *effective mass* is already presented (1.13), it is important to point out that often the *E–k* diagrams are not as simple as illustrated in Figure 1.2. Among the intricacies present in these diagrams are the overlapping profiles of branches known as *degeneracy*. In terms of the presence of *degeneracies*, the structure of the *E–k* diagram of *conduction band*, however, is much simpler than that of the *valence band*.

Among the elements of the *periodic table* of which semiconductors are made of, valence electrons either occupy s-type or p-type orbitals. Even in the crystalline semiconductor made out of these elements, *charge carriers* in the *conduction* and *valence*

[20] These are imaginary charged particles in the *valence band*.

[21] An *exciton* is a *hydrogen* atom-like entity, in which the role of the nucleus is replaced by a *hole*. In the case of an *exciton* in the form of a bond between an electron and a very heavy *hole*, the *exciton* acts not unlike an *impurity level*. However, the picture is more complicated if the *hole* mass is comparable to that of the electron.

band retain much of these s- and p-type characters. This is especially remarkable, when we remind ourselves of the fact that these *charge carriers* are now free *Bloch electrons*[22] and not bound by the nucleus of an atom. This s- and p-type nature plays a very important role in *charge transport* in semiconductors. This nature also extensively simplifies the quantum mechanical description of semiconductors, as now the shape of s- and p-type orbitals can be used as eigenvectors in the matrix representation. This simplification is used in the implementation of *tight-binding* technique.

According to the *tight-binding* technique, the top of the *valence band* is explained in terms of a threefold *degeneracy*, corresponding to that of *p*-orbitals (i.e., p_x, p_y, p_z). With incorporation of *spin degeneracy*, this threefold *degeneracy* grows to a sixfold. The threefold *degeneracy*, or sixfold with reference to *spin degeneracy*, is observed in the form of a twofold *degeneracy* of two bands (one *heavy hole* and one *light hole*) of equal energy at the top of the *valence band* and another *band* with slightly lower electron energy referred to as *split-off band*. The names *heavy* and *light hole* are given in terms of *effective mass* of electrons and *holes* at the top of the *valence band*. In terms of the definition of the *effective mass*, the wider *band* (i.e., of smaller curvature) poses a larger *effective mass* on the *holes*, hence the name *heavy-hole band*. The effects of these *bands* can be incorporated only in the case of incorporation of *relativistic* effects of the problem, which are referred to as *spin–orbit coupling*.

Figure 1.8 provides a schematic representation of the nature of the *valence* and *conduction band* of most semiconductors. On the basis of the typical concave and convex *E–k* diagrams presented in Figure 1.8 (i.e., for *conduction* and *valence bands*, respectively) and considering the opposite polarity of the charge of a *hole*, it can be observed that both electrons and *holes* (i.e., *charge carriers* of the *conduction* and *valence bands*, respectively) possess positive *effective mass*. In spite of commonalities between semiconductors, the curvature at the top of the *E–k* diagram of a *band* is not always negative. In the case of *tellurium*, the *E–k* diagram is of the form shown in Figure 1.9, which prohibits the definition of negative mass for electrons, and as result prohibits the definition of *imaginary* positively charged particles (i.e., *holes*) for missing electrons.

Semiconductors are divided into two groups according to their value of momentum at the *conduction-band* minima: *direct bandgap* and *indirect bandgap*. The *E–k* diagrams for these two groups are illustrated in Figure 1.8. While in the *E–k* diagram of *direct-bandgap* semiconductors the top of the *valence band* and the bottom of *conduction band* coincide at a momentum of 0,[23] *indirect* semiconductors only have the top of their *valence band* defined there. As a result of the difference between the momentum at the top of the *valence band* and the bottom of *conduction band* of *indirect* semiconductors, these semiconductors are not efficient producers of *photons*. This is because another momentum changing collision will be required

[22] Electrons expressed in terms of *propagating waves* belonging to quantized *energy states*. The quantized *energy states* are referred to as *Bloch states*. Named after *Felix Bloch*.

[23] We will identify later in the chapter that this coincidence happens at the Γ-*point* in the so-called *Brillouin zone*.

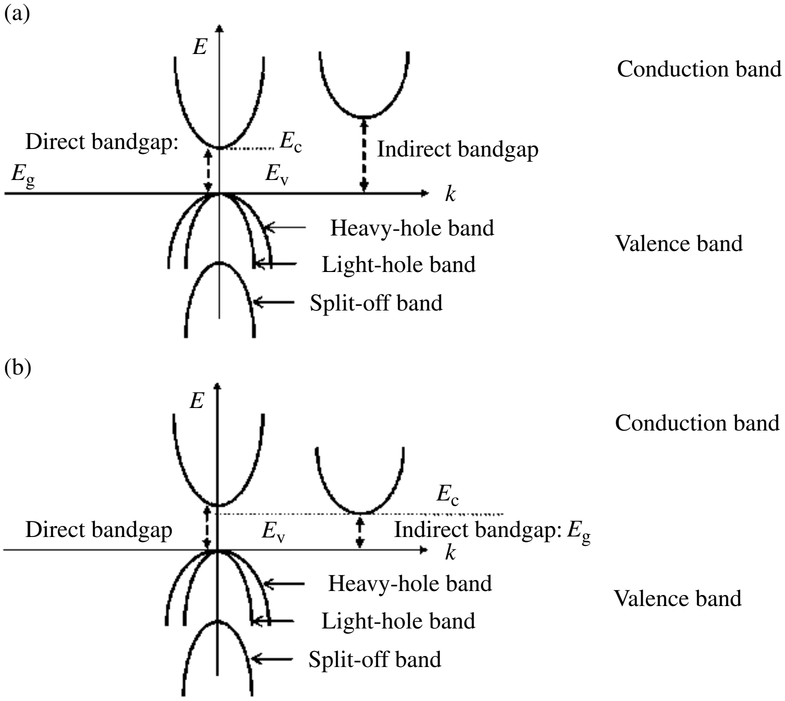


FIGURE 1.8 (a) *E–k* diagram of a direct-bandgap semiconductor; in this case the direct bandgap is smaller than the indirect bandgaps. (b) *E–k* diagram of an indirect-bandgap semiconductor; in this case the direct bandgap is larger than one or more of the indirect bandgaps. Top of the valence band is defined at $k = 0$.

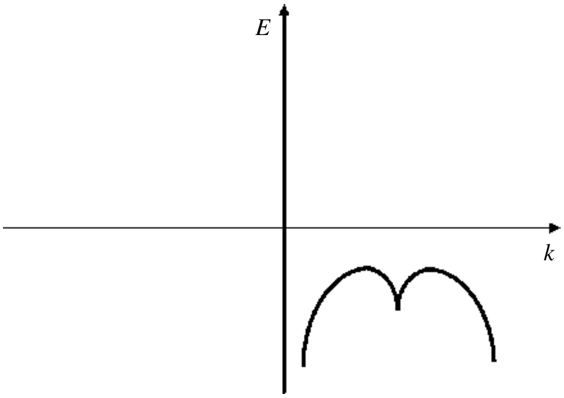


FIGURE 1.9 Top of the valence band in some materials such as tellurium present a positive effective mass for electrons hence at this position prohibits the definition of the positive charge carrier (i.e., hole).

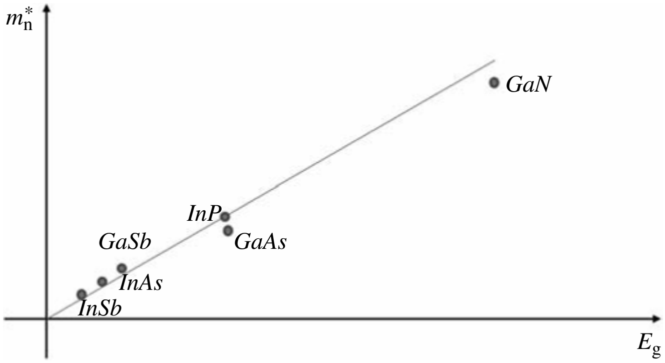


FIGURE 1.10 The prevalent trend between the effective electron mass and the size of the bandgap demonstrated among a few direct-bandgap *III–V* semiconductors. This approximate diagram is presented in linear scales.

besides the *electron–hole* interaction involved in the *recombination process*. In this case, the transitions involve a *photon* (which is responsible for energy change) and a *phonon* (which is responsible for momentum change). This division of responsibilities is caused by the very small momentum of *photons* and very small energies of *phonons*.

Silicon has an *indirect bandgap*,[24] with a sixfold *degenerate conduction-band edge*. *Germanium* is another well-known *indirect* semiconductor.[25] In both of these cases, a strong anisotropy of electron *wave function* persists near the *band edge*, which is caused by the mixing of p-type and s-type orbitals. The *conduction-band* minima of *direct-bandgap* semiconductors, being made of s-type orbitals, have spherically symmetric central cells. However, in the *states* further into the *conduction band*, this spherical symmetry, as the result of increasing contribution of p-type orbitals, fades. This is an important issue to be considered under high kinetic energy conditions.

In contrast to the *conduction-band* edge, the *valence-band* edges of most semiconductors are quite similar. In the case of the *valence band*, the central part of the electron *wave functions* is primarily p-type.[26]

As shown in Figure 1.10, there exists a declining trend between the electron *effective mass* and the size of the *bandgap* of *direct-bandgap* semiconductors.

1.2.2 Doping

Now that we know about *holes*, we can say that in an *intrinsic* semiconductor electron concentration of the *conduction band* (i.e., n_0) is equal to *hole* concentration of the *valence band* (i.e., p_0). In an *extrinsic* semiconductor, this balance is either tilted in

[24] Which, as will be seen later in this chapter, has a *conduction-band* minimum near the *X point* of the *Brillouin zone.*

[25] Whose *band* edge is defined near *L point* of the *Brillouin zone.*

[26] This nature induces a strong *spin–orbit* interaction.

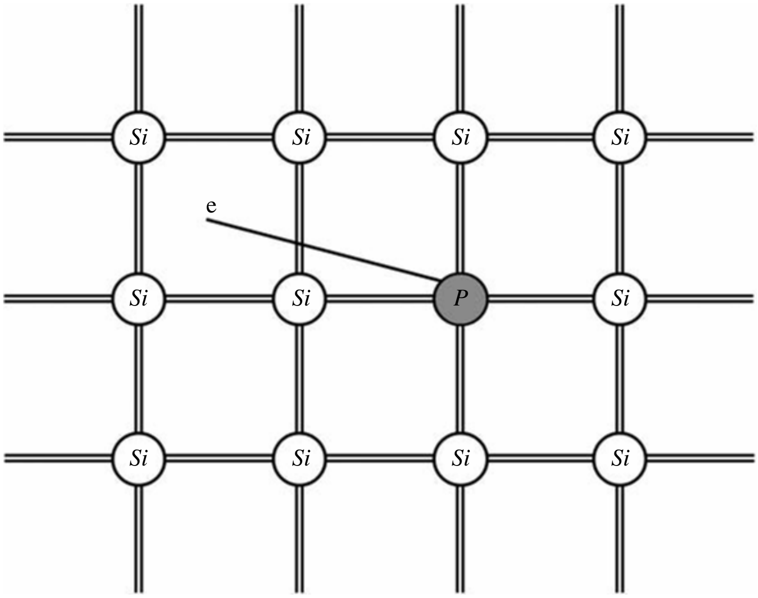


FIGURE 1.11 A 2-D schematic representation of *silicon* crystal upon substitutional doping with phosphorus. Covalent bonds are represented by the double lines, while the circles are representative of the silicon and phosphorus atoms.

favor of electrons (i.e., $n_0 > p_0$ in an n-type semiconductor) or holes (i.e., $p_0 > n_0$ in a p-type semiconductor).

Tilting the intrinsic balance of electron and *hole* concentration upon *doping* can be explained in terms of either the *energy-band* or *bond* models. In this review, we will address the process of *doping* in a specific group *IV* semiconductor (of the *periodic table*): *silicon*.

In an n-type semiconductor: $n_0 > p_0$. This means that not all of the electron population of the *conduction band* has originated from the *valence band*. In this case, the source of the *excess* population of electrons is the loosely bound electrons to the nucleus of *dopants* added to the semiconductor crystal. In *silicon* technology, the *dopants* used for this purpose are group *V* atoms of *phosphorus* and *arsenic*. These group *V* atoms possess one electron in excess of the four that they should share with the nearest neighboring *silicon* atoms while substituting one. This is schematically shown in Figure 1.11. The *ionization energy* of this *excess* electron is much smaller than the *ionization energy* of the *dopant* atom outside the confines of the crystal. This is caused by the overriding effect of the other atoms surrounding the *dopant* in the crystal, which renders an *ionization energy* quite close to that of the *hydrogen* atom. The *effective mass* of this electron and the value of the *permittivity* needed in the calculation of *ionization energy* are, however, modified to those values imposed by the *silicon* crystal. For offering an effective impurity, this *ionization energy* should be small compared to the average thermal energy of electrons at operating temperature

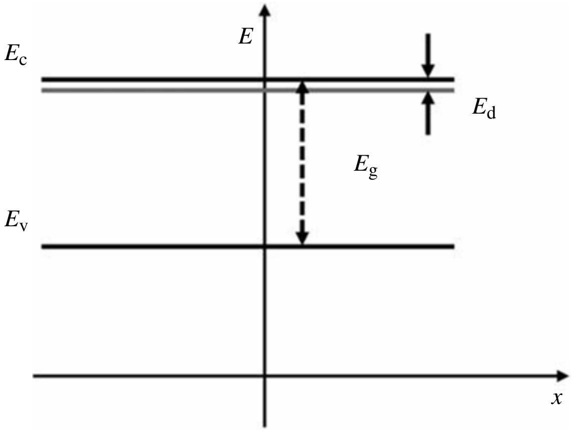


FIGURE 1.12 Band diagram of a semiconductor doped with a shallow donor.

of the device. Under such a condition, *dopants* can be effectively *activated* (i.e., contribute their *excess* electrons to the *conduction band*). This is one of the reasons *P* and *As* are chosen as suitable n-type *dopants* in *silicon* technology as their *ionization energies* are within the range of 20–40 meV.[27] These dopants are referred to as *shallow dopants*. This is in contrast to the *dopants* with higher *ionization energies*, which are known as *deep dopants*.

The other reason for the choice of *As* and *P* in *silicon* technology is the relatively high concentrations at which these atoms can be incorporated into the *silicon* structure under *thermodynamic equilibrium*. Under *thermodynamic equilibrium* there is an upper limit to the concentration of impurities that can be incorporated into a *solid* at a given temperature, known as *solid solubility*. Because the earliest technique of *doping* (i.e., *diffusion doping*) was a *thermal-equilibrium* process, this property played an important role in the choice of these *dopants*. However, in *ion implantation doping*, which is not performed under *thermodynamic equilibrium*, the *solid solubility* is not as important.

While in *silicon* crystal the concentration of *Si* atoms is on the order of 10^{22} cm^{-3} at room temperature, the maximum level of *doping* is only slightly above 10^{20} cm^{-3}. As a result, to a first approximation we can assume that semiconductor properties such as the size of the *bandgap* are not much altered by the *doping* process. In reality, the random nature of *substitutional doping* results in a disrupted periodicity of the crystal (and also a disruption in the formation of *forbidden* energy *gaps*). Due to this disruption, the *doping* process results in an introduction of *energy levels* within the *forbidden* energy gap. If an n-type *dopant* is shallow, the *energy levels* will form closer to the *conduction-band* edge (i.e., E_c). This is the *band diagram* explanation of the n-type *doping* process, which is illustrated in Figure 1.12. With an increasing *doping* level

[27] Since we deal small amounts of energy, in semiconductor electronics instead of using *Joule* as the unit of energy, we use *electron Volt* (i.e., eV). One electron Volt is almost equal to 1.6×10^{-19} J.

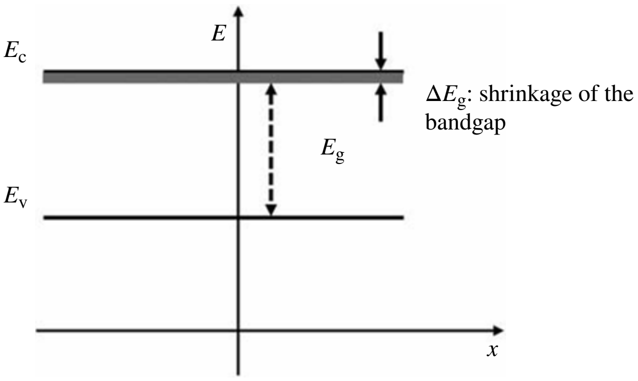


FIGURE 1.13 Schematic depiction of the role of heavy doping in shrinking the bandgap of a semiconductor. The wide gray band of energy stands for the impurity band resulting from the splitting of donor energy levels.

(i.e., number of incorporated *substitutional dopants* per unit volume), these *shallow energy levels* will split into *bands* of energy (often referred to as an *impurity band*). The splitting is caused by the overlapping *wave function* of the electrons residing in the impurity *energy levels*. As shown in Figure 1.13, doping at very high levels, through merging of these *bands* with the *conduction band* (or *valence band* in the case of p-type *doping*), results in *shrinkage of the bandgap*. If the *energy levels* introduced by *dopants* are separated from the *conduction band* by only tens of milli-electron volt, at room temperature electrons residing in those *energy states* (i.e., loosely bound electrons to the nuclei of the n-type impurity) can readily leave these *levels* and jump to the *conduction band* without adding a *hole* to the *valence band*. This is how the balance between the electron and *hole* concentration is broken in an *extrinsic* n-type semiconductor.

The more appropriate name for an n-type *dopant* is *donor* because it donates electron to the *conduction band*. In the *energy-band* model, the *energy levels* provided by these *donors* to the previously *forbidden bandgap* are also known as *donor levels*. Through the process of electron donation to the *conduction band*, the *donor* atom becomes a positively charged ion, and as a result, charge neutrality prevails in the semiconductor. Using the *energy-band* model, we can describe this behavior in terms of the following definition for the *donor level*:

An *energy level* is referred to as *donor*, if it were to be neutral when full and positively charged when empty.

For very highly *doped* semiconductors,[28] when the *impurity band* developed by the *donors* merges with the *conduction band*, the semiconductor adopts a *metallike* behavior and remains highly conductive even at very low temperatures. This formation is also known as an *impurity band tail*. As shown in Figure 1.14, the choice of the name comes from the presence of finite *DOS* at the *band edge* and the gradual, and not sharp, increase of the *DOS* function. This is unlike the predictions of (1.14). Later in

[28] For which a more quantitative definition is to follow in Section 1.3.5.

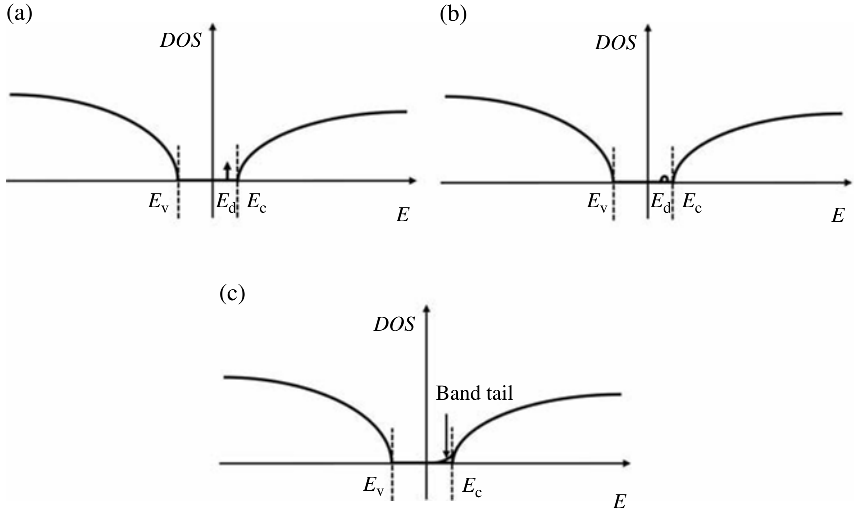


FIGURE 1.14 Schematic depiction of the evolution of the conduction band's *DOS* function upon increase of the donor concentration, from (a) to (c). Scales are linear.

this chapter, we will talk about formation of *band tails* also in the case of the so-called *alloyed* semiconductors.

It is worthwhile mentioning that *arsenic*, due to its larger atomic size than *phosphorus*, has a smaller *thermal diffusion constant* through the *silicon* crystal. Since many processing steps run at very high temperatures, achieving the small n-type *dopant* profiles that are required in modern devices is much easier with the use of *As*. For that reason, in many applications of *Si* technology, *As* has already replaced *P* as the proper *donor*.

So far, with exception of a few points, we have avoided discussing p-type *doping*. A slightly different version of events can explain the intricacies of p-type *doping*. The suitable p-type *dopant* in *Si* technology is the group *III* atom of *boron*. Substitution of *Si* with *B* (which has only three electrons in its valence shell) leaves one out of four broken covalent bonds, originally formed between nearest neighboring *Si* atoms, in need of an electron. The missing electron, however, can be acquired by breaking one electron away from an existing covalent bond in the vicinity of the impurity. In terms of *energy-band* model, this can be explained through the addition of an *energy level* closer to *valence band* in the *forbidden gap*, which is keen to accept electrons. Through receiving electrons from the *valence band*, a *hole* is left behind in that *band*. For this reason, this *energy level* is called an *acceptor level*, and the p-type *dopants* are referred to as *acceptors*. Quite similar to the definition of the *donor energy level*, an *acceptor energy level* is neutral when empty and negatively charged when full. *Boron*, like *arsenic* and *phosphorus*, provides *silicon* with a *shallow impurity level*.

Doping process can take place through substitution of structural atoms of the crystal with *dopants* (also referred to as *direct doping*) or indirectly through incorporation

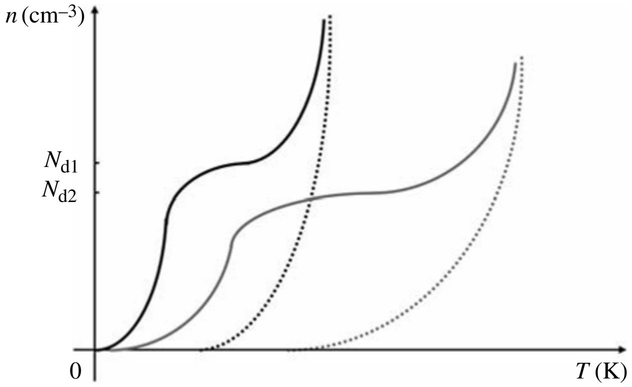


FIGURE 1.15 Variation of the electron concentration versus temperature for two n-doped semiconductors of different bandgaps. The wider bandgap semiconductor is assumed to have been doped more lightly. The levels of doping of the two semiconductors, in increasing order of the size of the bandgap, are indicated by N_{d1} and N_{d2}, respectively. The dashed lines represent the variation of electron concentration in each semiconductor when undoped (i.e., n_i). Diagrams in black represent the smaller bandgap semiconductor, while the gray diagrams represent the wider bandgap semiconductor. Within the temperature range for which the electron concentration reaches a plateau, semiconductor behaves extrinsically. Scales are linear.

of *dopants* into the *interstitial* sites. For *dopants* to contribute to electron and *hole* concentration, however, they need to take *substitutional* positions. As already indicated implicitly, *dopants* also need to be *activated* through *ionizations*. The movement of *dopants* in the crystal and their later *activation* is a process that requires exposure to high temperatures. The process of movement of *dopants* within a crystal can happen through thermal diffusion of *dopant* in the form of *interstitial diffusion*, *substitutional diffusion*, or a mixture of the two (*dopants* sometimes taking *substitutional* positions and sometimes *interstitial* throughout the diffusion process), which is referred to *interstitialcy*.

A *doped* semiconductor under a number of conditions will retain its *intrinsic* characteristics:

1. When the *lattice temperature*[29] is very low and *dopants* are under *carrier freeze-out* (i.e., *not ionized*).
2. When the concentration of *activated donors* and *acceptors* are equal to one another. This case is referred to as *compensative doping*.
3. When *lattice temperature* is high and the number of *thermally induced charge carriers* to *conduction* and *valence band*, in the form of *EHP*, well exceeds the concentration of electrons and *holes* contributed by *donors* and *acceptors*, respectively. As shown in Figure 1.15, the characteristic temperature of this behavior is determined by the size of the *bandgap* of the semiconductor.

[29] It will be identified in Section 1.4.4 why instead of simply speaking of temperature, we are talking about a temperature quantity called *lattice temperature*.

Realization of *extrinsic* semiconductors is a prime requirement in formation of electronic devices. In these devices, the interaction between regions of different *charge-carrier* concentrations provides the chance for enforcing selective control over the movement of *charge carriers*. Such junctions between regions of different *dopant* concentration can be realized through simply *doping* the semiconductor and/or *growing* semiconductors of different *bandgaps* and/or *electron affinities* on top of one another.

1.2.3 Calculation of Ionization Energies in Semiconductors

With regard to the process of *doping*, it is quite important to be able to have means to approximately calculate the energy of *donor* and *acceptor states*. For the loosely bound fifth electron of ordinary *donors* such as *P*, *As*, and *Sb* in *Ge* or *Si* technologies, the *ionization energy*, referenced to the *conduction-band* edge (i.e., E_c), calculated in terms of *hydrogenic* models provides us with

$$E_d = -13.6\left(\frac{m^*}{m_0}\right)\left(\frac{\epsilon_0}{\epsilon}\right)^2\frac{1}{n^2}\ \mathrm{eV} \tag{1.17}$$

where n is the principal *quantum number*, ϵ is the dielectric constant, and ϵ_0 is the permittivity of the vacuum.

In the above equation, the differences between the *effective mass* of an electron and the free electron mass, and also the *dielectric constant* of a semiconductor and that of the free space, cause the electron orbit for the impurity atom to be much larger than the value in a free atom. As a result, *ionization energy* is much smaller for an impurity in a semiconductor than that of the *hydrogen* atom.

A similar relationship can be developed for *acceptor states*. This approximate framework is, however, not as valid due to the more complicated nature of the *E–k* diagram of the *valence band*. Here the *effective mass* of the electron should be replaced with that of the *hole*. Because of the much higher *effective mass* of *holes* compared to electrons in most semiconductors (with the exception of *Ge*), realizing the full *activation* of *acceptors* requires higher temperatures than those required for *activating donors*. Between *Si* and *Ge*, the *ionization energies* in *Si* for both types of *dopants* are larger. This is due to the smaller dielectric constant and larger *effective mass* in *Si*.

Despite their usefulness, it should be pointed out that the aforementioned formalism of calculation of *donor* and *acceptor energy levels* is only sufficient for evaluation of *ionization energies* within an order of magnitude approximation. However, since more accurate estimations require more intensive quantum mechanical mathematics, this framework is often used for providing the first-order estimate. Although for the considered case of a *hydrogen* atom the potential and, as a result, *effective mass* are spherically symmetric, not all semiconductor *energy bands* have this degree of symmetry. With regard to this case, it is necessary to use an appropriate average of the *effective mass* in different directions. The other simplifying assumption used in the development of the *hydrogenic* model of (1.17) is that the position of the *conduction band* is taken to coincide with E_d when the *quantum number* tends to infinity. While not true in a strict sense, this has been deemed to provide a good estimate. The last limitation that this model suffers from is that it does not distinguish between the *ionization energies* of different *dopants*. Although the impact of the ion core on the *ionization energy* is reduced in a *dopant*-in-semiconductor scenario, this reduced impact still can be felt in terms of the difference in *ionization energy* of different *dopants*.

Table 1.3 provides a detailed list of ionization energies for impurities in *Si*, *Ge*, *GaAs*, *InP*, and *GaN*.

1.3 THERMAL-EQUILIBRIUM STATISTICS

Considering the large number of *energy states* in the *conduction* and *valence band*, studying the state of the electron population of these *bands* (i.e., electron concentration and *hole* concentration in the *conduction* and *valence band*, respectively) can be only addressed statistically. However, before presenting the suitable statistical framework of this problem, it should be mentioned that such a statistical framework loses its validity when statistical fluctuation of the number of *dopants* in an extremely small-size device grows close to the small number of *dopants* incorporated in its volume. As a result, the framework presented in this section faces difficulties in application to *nanoscale* devices.

Evidently, absence of time variation yields a more manageable mathematical framework. The condition set in the study of this time-invariant situation is referred to as *thermal-equilibrium* condition. Under *thermal equilibrium*, the semiconductor is not exposed to any external source of excitation, and at a given temperature (which we will call from now on *lattice temperature*), all thermal processes are counterbalanced. As a result, electron and *hole* concentration among other properties of the semiconductor will remain time independent.

1.3.1 *Fermi–Dirac* Statistics

In Appendix 1.A, with special attention to the restrictions of *Pauli's exclusion principle*, the mathematical derivation of the fundamental outcome of this *thermal-equilibrium* statistics, which is referred to as the *Fermi–Dirac* distribution function,[30] is provided. This temperature-dependent statistical model, irrelevant of the presence of an *energy state* at a given *energy value*, determines the chance of having an electron of that value of energy. As a result of the definition of *Fermi–Dirac* statistics, one can calculate the spatial concentration of electrons within an infinitesimally small range of energy by multiplying the *Fermi–Dirac* distribution function by the known *density of states function* of the *conduction band*.

The process of establishing a state of *thermal equilibrium* is not instantaneous. Instead, for a given crystal at a set temperature, *thermal equilibrium* evolves spontaneously over a certain period of time.

[30] Named after *Enrico Fermi* and *Paul Dirac*.

TABLE 1.3 List of Ionization Energies for a Number of Important Impurities in Si, Ge, GaAs, InP, and GaN

Semiconductor	Ionization Energies of a Number of Important Impurities
Si	Li: $E_c - 0.034$ eV (D)
	Sb: $E_c - 0.043$ eV (D)
	P: $E_c - 0.046$ eV (D)
	As: $E_c - 0.054$ eV (D)
	Be: $E_V + 0.42$ eV (A), $E_v + 0.17$ eV (A)
	Au: $E_c - 0.54$ eV (A), $E_v + 0.35$ eV (D), $E_v + 0.29$ eV (D)
	Cu: $E_v + 0.53$ eV (A), $E_v + 0.4$ eV (A), $E_v + 0.24$ eV (A)
	B: $E_v + 0.044$ eV (A)
	Al: $E_v + 0.069$ eV (A)
	Ga: $E_v + 0.073$ eV (A)
Ge	Li: $E_c - 0.0095$ eV (D)
	Sb: $E_c - 0.0096$ eV (D)
	P: $E_c - 0.012$ eV (D)
	As: $E_c - 0.013$ eV (D)
	Pt: $E_c - 0.23$ eV (A), $E_v + 0.04$ eV (A)
	Au: $E_c - 0.04$ eV (A), $E_c - 0.2$ eV (A), $E_v + 0.15$ eV (A), $E_v + 0.05$ eV (D)
	Cu: $E_c - 0.26$ eV (A), $E_v + 0.32$ eV (A), $E_v + 0.045$ eV (A)
	B: $E_v + 0.010$ eV (A)
	Al: $E_v + 0.010$ eV (A)
	Ga: $E_v + 0.011$ eV (A)
GaAs	Si: $E_c - 0.0058$ eV (D), $E_v + 0.035$ eV (A)
	Ge: $E_c - 0.0061$ eV (D), $E_v + 0.0404$ eV (A)
	O: $E_c - 0.4$ eV (D), $E_v + 0.67$ eV (D)
	C: $E_v + 0.026$ eV (A)
	Be: $E_v + 0.028$ eV (A)
	Mg: $E_v + 0.028$ eV (A)
	Au: $E_v + 0.09$ eV (A)
	Mn: $E_v + 0.095$ eV (A)
	Cu: $E_v + 0.44$ eV (A), $E_v + 0.24$ eV (A), $E_v + 0.19$ eV (A), $E_v + 0.14$ eV (A), $E_v + 0.023$ eV (A)
	Fe: $E_v + 0.52$ eV (A), $E_v + 0.37$ eV (A)
InP	S, Si, Sn, Ge: $E_c - 0.0057$ eV (D)
	C: $E_v + 0.04$ eV (A)
	Hg: $E_v + 0.098$ eV (A)
	Zn: $E_v + 0.035$ eV (A)
	Si: $E_v + 0.03$ eV (A)
	Cu: $E_v + 0.06$ eV (A)
	Be: $E_v + 0.03$ eV (A)
	Mg: $E_v + 0.03$ eV (A)
	Ge: $E_v + 0.021$ eV (A)
	Mn: $E_v + 0.027$ eV (A)
GaN	Si: $E_c - (0.02 - 0.12)$ eV (D) (often instead of a single value, a range has been reported)
	Mg: $E_v + (0.14 - 0.21)$ eV (A) (often instead of a single value, a range has been reported)

(D) and (A) are indicators of energy level of donor and acceptor nature, respectively. The presented information for the first three semiconductors is more complete.

In the description of the *Fermi–Dirac* distribution function, a certain value of energy, known as *Fermi level* (i.e., E_f), serves as the reference value of energy in determination of the probability of occupation of all *energy states*. According to *Fermi–Dirac* statistics, an *energy state* with E_f as its energy value, at any given temperature, has a 50% chance of being occupied by an electron. *Energy states* above E_f have lower than 50% chance of occupation, and those below have higher than 50% chance of occupation. *Fermi–Dirac* statistics is presented by

$$f_\mathrm{D}(E) = \frac{1}{1+\exp((E-E_\mathrm{f})/kT)}. \tag{1.18}$$

In this equation, T is the *lattice temperature* in Kelvin and k is the *Boltzmann* constant.[31] At 0 K, $f_\mathrm{D}(E)$ reduces to a *Heaviside unitary step function*. Whereas at all temperatures $f_\mathrm{D}(E_\mathrm{f}) = 1/2$, an increase in temperature causes this step function to spread out further.

1.3.2 *Maxwell–Boltzmann* Statistics

The *Fermi–Dirac* distribution function when $E - E_\mathrm{f} \gg kT$ can be reduced to a simpler distribution function, which is referred to as *Maxwell–Boltzmann* distribution function,[32]

$$f_\mathrm{M}(E) = \exp\left(-\frac{E-E_\mathrm{f}}{kT}\right). \tag{1.19}$$

As shown in Appendix 1.A, the derivation of *Fermi–Dirac* statistics is bound by the constraints of *Pauli's exclusion principle*. Electrons and *holes*, as particles that are required to follow these constraints, are referred to as *Fermions*. Although in the case of *Maxwell–Boltzmann* statistics such a constraint is not present, simplifying *Fermi–Dirac* statistics by the distribution function of (1.19) does not impose any difficulty. The reason is rooted in the condition applied to this approximation (i.e., $E - E_\mathrm{f} \gg kT$). Under this condition, the chance of finding an electron at *energy state E* will be very small. As a result the chance of encountering two electrons at the same *state* and clashing with *Pauli's exclusion principle* will be equally small. The semiconductor for which *Maxwell–Boltzmann* statistics is applicable is referred to as a *nondegenerate*[33] semiconductor. Figure 1.16 presents the form of temperature variation of the

[31] $k = 8.62 \times 10^{-5}$ eV/K.

[32] Named after *James Maxwell* and *Ludwig Boltzmann*.

[33] In this case the definition of *degeneracy* is with reference to becoming *metallike*, which as hinted earlier happens when the *dopant* concentration is very high and as a result of overlapping the *impurity-band* and *conduction-band* (likewise *valence band* in case *holes*) *charge-carrier* concentration of *conduction-* or *valence-band* approaches that of the metal. This *degeneracy* is different from the one observed in our discussions of *E–k* diagram.

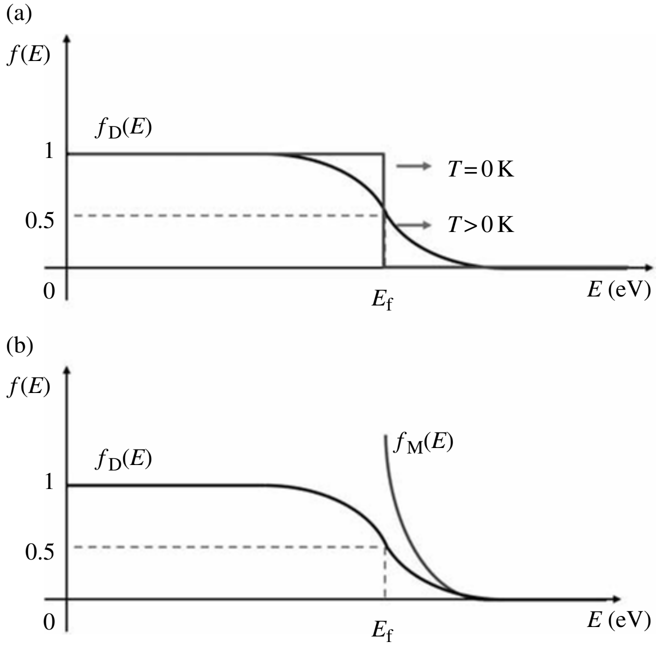


FIGURE 1.16 (a) Variation of *Fermi–Dirac* distribution function with temperature. (b) Comparison between the *Fermi–Dirac* and the *Maxwell–Boltzmann* distribution functions at a temperature above 0 K. Scales are linear.

Fermi–Dirac distribution function and its approximation with the *Maxwell–Boltzmann* distribution function.

For a *nondegenerate* semiconductor the distribution function of (1.18) can be simplified to a *Maxwell–Boltzmann* distribution function. Through adopting a semiclassical approach to the definition of energy such as

$$E = E_c + \frac{1}{2}m^* v^2, \tag{1.20}$$

one can rewrite (1.19) as

$$f_M \cong \exp\left(\frac{E_f - E_c}{kT}\right) \cdot \exp\left(-\frac{m^* v^2}{2kT}\right) = C \cdot \exp\left(-\frac{m^* v^2}{2kT}\right). \tag{1.21}$$

It can be proven that C for a given *doping* level is constant.[34] According to this equation, *charge-carrier* velocities of a *nondegenerate* semiconductor are distributed in the form of a *Gaussian* with an average velocity of 0 and a variance, which is a function of *lattice temperature*. In the three-dimensional setting,

[34] That is, using (1.23) and (1.32) or (1.33).

$$v^2 = v_x^2 + v_y^2 + v_z^2 \tag{1.22}$$

and as a result such a conclusion can also be drawn for *carrier* velocities along each of the axes of the *Cartesian* coordinate system.

1.3.3 Calculating Electron and Hole Concentration in Nondegenerate Semiconductors

As suggested earlier in this section, in calculating *thermal-equilibrium* electron concentration in the *conduction band* (i.e., n_0), one needs to multiply the *Fermi–Dirac* distribution function (or where appropriate the *Maxwell–Boltzmann* distribution function) with the *density of states* function of the *conduction band* and take an integral over the energy width of this *band*. Likewise, in calculating the *thermal-equilibrium hole* concentration in the *valence band* (i.e., p_0), one should multiply the *DOS* function of the *valence band* by the complementary of the *Fermi–Dirac* (or where appropriate the *Maxwell–Boltzmann*) distribution function and take an integral over the energy extent of the *band*. For cases in which *Maxwell–Boltzmann* statistics is applicable, such calculations result in the following closed-form expressions for n_0 and p_0:

$$n_0 = N_\mathrm{c} \exp\left(\frac{E_\mathrm{f} - E_\mathrm{c}}{kT}\right) \tag{1.23}$$

and

$$p_0 = N_\mathrm{v} \exp\left(\frac{E_\mathrm{v} - E_\mathrm{f}}{kT}\right) \tag{1.24}$$

where N_c and N_v are the *effective density of states* at the lower edge of the *conduction band* (i.e., E_c) and the higher edge of the *valence band* (i.e., E_v), respectively. These *effective* values are mathematical tools proposed to yield the above simple expressions.[35] While in reality in a *bulk* semiconductor, the *DOS* at the lower edge of *conduction band* and the higher edge of the *valence band* are either 0[36] or have a very small value (i.e., in case of formation of *band tail*), the values of N_c and N_v are quite large. This is because they *effectively* represent the total number of *states* distributed throughout the *bands* in the form of two *Dirac delta functions* defined at the edge of each *band*. These values, as expressed in (1.25) and (1.26), are dependent on *lattice temperature* and *effective mass* of electrons and *holes* (i.e., for N_c and N_v, respectively):

$$N_\mathrm{c} = 2\left(\frac{2\pi m_\mathrm{n}^* kT}{h^2}\right)^{3/2} \tag{1.25}$$

[35] In terms of these *effective density of states*, the aforementioned integral is calculated using the *gamma* function.

[36] See (1.14).

TABLE 1.4 A Number of Important Properties of Si, Ge, GaAs, and GaN at Room Temperature

	Semiconductor			
Property	Si	Ge	GaAs	GaN
Lattice type and constant a (Å)	Zinc blende 5.431	Zinc blende 5.658	Zinc blende 5.65325	Wurtzite 3.189 ($c = 5.186$)
Melting point T_m (°C)	1412	937	1237	2500
Bandgap E_g (eV)	1.12	0.67	1.42	3.39
Static relative permittivity ϵ_r	11.7	16	12.9	8.9
Specific heat C_p (J/g/K)	0.7	0.31	0.33	0.49
Thermal conductivity k (W/cm/K)	1.412	0.606	0.455	1.3
Electron mobility $\boldsymbol{\mu}_n$ (cm^2/V/s)	1417	3900	8800	1000 (approx.)
Hole mobility $\boldsymbol{\mu}_p$ (cm^2/V/s)	471	1900	400	200 (approx.)
Effective conduction-band density of states N_c (cm^{-3})	2.8×10^{19}	1.04×10^{19}	4.7×10^{17}	2.23×10^{18}
Effective valence-band density of states N_v (cm^{-3})	1.04×10^{19}	6×10^{18}	7×10^{18}	4.62×10^{19}
Intrinsic carrier concentration n_i (cm^{-3})	1.45×10^{10}	2.4×10^{13}	9×10^{6}	1.9×10^{-10}
Critical electric-field E_c (V/cm)	3×10^{5}	8×10^{4}	3.5×10^{5}	5×10^{6}
Effective transport electron mass $m_n{}^*$	$0.26m_0$	$0.12m_0$	$0.068m_0$	$0.2m_0$
Effective transport hole mass $m_P{}^*$	$0.386m_0$	$0.3m_0$	$0.5m_0$	$0.8m_0$
Optical phonon energy E_{op} (meV)	63	37	35	91.2

and

$$N_v = 2\left(\frac{2\pi m_p^* kT}{h^2}\right)^{3/2}. \tag{1.26}$$

Equations (1.23)–(1.26) imply that the *Fermi level* of an *intrinsic* semiconductor is very close the middle of the bandgap but not exactly there since $m_p^* > m_n^*$.

Table 1.4 presents a list of important properties, such as N_c and N_v, for a few well-known semiconductors.

Since in *nondegenerate doping* of a semiconductor, due to low levels of *dopant* concentration, structural properties of the semiconductor remain almost intact, under these conditions N_c and N_v will not change with *doping*. As a result, letting E_i

represent the *Fermi level* of the *intrinsic* semiconductor, (1.23) and (1.24) can be rewritten in terms of E_i and n_i (i.e., *intrinsic* electron and *hole* concentration):

$$n_0 = n_\mathrm{i} \exp\left(\frac{E_\mathrm{f} - E_\mathrm{i}}{kT}\right) \tag{1.27}$$

and

$$p_0 = n_\mathrm{i} \exp\left(\frac{E_\mathrm{i} - E_\mathrm{f}}{kT}\right). \tag{1.28}$$

1.3.4 Mass Action Law

Concentration profiles of electrons and *holes* in a semiconductor are not independent from one another. From (1.27) and (1.28) it is quite obvious that

$$p_0 n_0 = n_\mathrm{i}^2. \tag{1.29}$$

This relationship is referred to as the *mass action* law.

Also based on (1.23) and (1.24),

$$n_\mathrm{i} = \sqrt{N_\mathrm{c} N_\mathrm{v}} \exp\left(\frac{-E_\mathrm{g}}{2kT}\right). \tag{1.30}$$

On the basis of (1.30), for a semiconductor of smaller *bandgap*, at any given temperature the *intrinsic charge-carrier* concentration (i.e., n_i) is larger. This explains why a wider *bandgap* semiconductor remains *extrinsic* over a wider range of temperatures (Fig. 1.15).

Equation (1.23) shows that as a result of elevation of the *Fermi level* through n-type *doping*, E_f approaches E_c, and E–E_f at the bottom of the *band* will become smaller. When this difference becomes smaller than a few times kT, the approximation of *Fermi–Dirac* statistics by *Maxwell–Boltzmann* statistics is inadmissible. An identical situation will happen when E_f is lowered toward E_v. Under these situations, which are caused by heavy n- and p-type *doping*, approximate equations of (1.23)–(1.28) are no longer valid. A highly doped semiconductor like this is referred to as a *degenerate* semiconductor, since with the increase in *charge-carrier* concentration, the conduction properties of the semiconducting material have essentially *degenerated* into those of a metal.[37] In degenerate semiconductors, electron and *hole* concentrations should be calculated with the use of *Fermi–Dirac integrals*.

[37] We have previously come across this situation while speaking of the merger of the *impurity band* and *conduction* or *valence bands* in a highly *doped* semiconductor.

The *mass action* law is a *thermal-equilibrium* relationship. Whereas simply multiplying (1.27) and (1.28) proved the *mass action* law in a *nondegenerate* semiconductor, for all semiconductors under *thermal equilibrium*, such a relationship is readily extendable. In a general form, this can be easily proven through considering the counterbalancing processes of *generation* and *recombination* of EHP. While the process of *generation* is a thermally induced process (where its rate can be expressed in terms of a function of temperature: $f_1(T)$), the process of *recombination* is not only temperature dependent but also dependent on the population of electrons in the *conduction band* and *holes* in the *valence band*. The temperature-dependent part of the rate of the *recombination* process is expressed in terms of another function of temperature, which we will call $f_2(T)$. By equating the rates of the two counterbalancing processes, under *thermal equilibrium*,

$$G = R \Rightarrow f_1(T) = p_0 n_0 f_2(T) \Rightarrow \frac{f_1(T)}{f_2(T)} = p_0 n_0 = f_3(T). \tag{1.31}$$

Knowing that the function $f_3(T)$ is just a function of temperature and not the *doping* process, it will be evident that $f_3(T) = n_\text{i}^2$, which yields the *mass action* law (i.e., (1.29)).

On the basis of the *mass action* law and charge neutrality, one can prove that in a semiconductor *doped* with both *donors* (i.e., N_d in cm^{-3}) and *acceptors* (i.e., N_a in cm^{-3}), if all *dopants* were to be activated,

$$n_0 = \left(\frac{N_\text{d} - N_\text{a}}{2}\right) + \sqrt{\left(\frac{N_\text{d} - N_\text{a}}{2}\right)^2 + n_\text{i}^2} \quad \text{for} \quad N_\text{d} > N_\text{a}, \tag{1.32}$$

which yields $n_0 \cong N_\text{d} - N_\text{a}$ when $N_\text{d} - N_\text{a} \gg n_\text{i}$.

In addition,

$$p_0 = \left(\frac{N_\text{a} - N_\text{d}}{2}\right) + \sqrt{\left(\frac{N_\text{a} - N_\text{d}}{2}\right)^2 + n_\text{i}^2} \quad \text{for} \quad N_\text{a} > N_\text{d}, \tag{1.33}$$

which yields $p_0 \cong N_\text{a} - N_\text{d}$ when $N_\text{a} - N_\text{d} \gg n_\text{i}$.

Example

A *Ge* sample is uniformly *doped* with both *B* and *As* to the levels of 10^{16} and 10^{15} cm^{-3}, respectively. Determine the electron concentration in this sample at both 300 and 500 K. Assume that at both temperatures, all impurities are *activated*.

According to Table 1.4, the *bandgap* of *Ge* at room temperature is 0.67 eV, while $n_\text{i} = 2.4 \times 10^{13}$ cm^{-3}.

Based on Table 1.2 the size of the *bandgap* at 500 K is given by

$$E_\text{g}(T = 500\,\text{K}) = 0.742 - 4.8 \times 10^{-4} \frac{500^2}{735} \cong 0.579\ \text{eV}.$$

According to this value and the values of N_c and N_v at room temperature (provided in Table 1.4), (1.30) yields the *intrinsic* electron concentration at 500 K:

$$n_\mathrm{i}(T=500\,\mathrm{K}) = \sqrt{1.04\times10^{19}\times6\times10^{18}\left(\frac{500}{300}\right)^3}\cdot\exp\left(-\frac{0.579}{2\times0.0258\times\frac{500}{300}}\right)$$
$$\cong 2.03\times10^{16}\ \mathrm{cm}^{-3}$$

Based on these values, Equation (1.33), and the mass action law, we can calculate the electron concentration given $N_\mathrm{d}=10^{15}\ \mathrm{cm}^{-3}$ and $N_\mathrm{a}=10^{16}\ \mathrm{cm}^{-3}$.

For $T=300$ K,

$$p_0=\left(\frac{10^{16}-10^{15}}{2}\right)+\sqrt{\left(\frac{10^{16}-10^{15}}{2}\right)^2+\left(2.4\times10^{13}\right)^2}\cong 9\times10^{15}\ \mathrm{cm}^{-3}$$

and

$$n_0=\frac{n_\mathrm{i}^2}{p_0}=\frac{\left(2.4\times10^{13}\right)^2}{9\times10^{15}}\cong 6.4\times10^{10}\ \mathrm{cm}^{-3}.$$

For $T=500$ K,

$$p_0=\left(\frac{10^{16}-10^{15}}{2}\right)+\sqrt{\left(\frac{10^{16}-10^{15}}{2}\right)^2+\left(2.03\times10^{16}\right)^2}\cong 2.5\times10^{16}\ \mathrm{cm}^{-3}$$

and

$$n_0=\frac{n_\mathrm{i}^2}{p_0}=\frac{\left(2.03\times10^{16}\right)^2}{2.5\times10^{16}}\cong 1.6\times10^{16}\ \mathrm{cm}^{-3}.$$

1.3.5 Calculation of Electron and Hole Concentration in a Degenerate Semiconductor

According to the *density of states* functions presented Section 1.1.3 and the *Fermi–Dirac* distribution function of (1.18), the calculation of electron concentration in terms of *Fermi–Dirac integral* yields the following for *bulk* 3-D, 2-D, and 1-D semiconductors. The calculation of *hole* concentration follows the same principles.

In the case of the 3-D semiconductor, using (1.14),

$$n_0=\int_0^\infty f_\mathrm{D}(E)D_\mathrm{3D}(E)dE=N_\mathrm{3D}\mathrm{F}_{1/2}\left[\frac{E_\mathrm{f}-E_\mathrm{c}}{kT}\right]=N_\mathrm{3D}\mathrm{F}_{1/2}(\eta_\mathrm{F}) \qquad (1.34)$$

where $N_{3D} = 2\left(2\pi m_n^* kT/h^2\right)^{3/2}$ and $F_{1/2}$ is the order ½ *Fermi–Dirac integral*.[38]

As presented in Section 1.3.3, for a *nondegenerate* 3-D semiconductor, this simplifies to the familiar form of

$$n_0 = N_{3D} \exp\left(\frac{E_f - E_c}{kT}\right). \tag{1.35}$$

Similar to this, in the two-dimensional case,

$$n_s = N_{2D} F_0(\eta_F) \text{ cm}^{-3} \tag{1.36}$$

where

$$F_0(\eta_F) \equiv \frac{1}{\Gamma(1)} \int_0^\infty \frac{d\xi}{1 + \exp(\xi - \eta_F)} = \frac{1}{\Gamma(1)} \int_0^\infty \frac{\exp(\eta_F - \xi) d\xi}{1 + \exp(\eta_F - \xi)} = -\frac{1}{\Gamma(1)} \text{Ln}(1 + \exp(\eta_F - \xi))\big|_0^\infty. \tag{1.37}$$

As a result, in the case where the second *subband* is much above *Fermi level*,[39]

$$n_s = N_{2D} \text{Ln}\left[1 + \exp\left(\frac{E_f - E_1}{kT}\right)\right]. \tag{1.38}$$

E_1 is the bottom of the first *subband* identified in (1.8) and $N_{2D} = \left(m_n^* kT\right)/\pi\hbar^2$. Again, under *nondegenerate* conditions,

$$n_s = N_{2D} \exp\left(\frac{E_f - E_1}{kT}\right). \tag{1.39}$$

For the one-dimensional case,

$$n_L = N_{1D} F_{-1/2}(\eta_F) \tag{1.40}$$

where

$$N_{1D} = \frac{\sqrt{2m^{*kT}/_{\pi}}}{\hbar}. \tag{1.41}$$

[38] *Fermi–Dirac integral* of order j is defined in terms of the Γ-*function*: $F_j(\eta) \equiv 1/\Gamma(j+1) \int_0^\infty \left(\xi^j d\xi\right)/(1 + \exp(\xi - \eta))$.

[39] That is, at least 3–4kT above E_f. In this case, only the first *subband* will be having a chance of having electrons.

Under *nondegenerate* conditions, with the assumption of only the first *subband*,

$$n_{\mathrm{L}}(E_{\mathrm{f}}) = N_{\mathrm{1D}} \exp\left(\frac{E_{\mathrm{f}} - E_1}{kT}\right). \tag{1.42}$$

1.3.6 Quasi-Fermi Levels

It should be kept in mind that the definitions of *Fermi–Dirac* statistics and the *Fermi level* are only possible under *thermal equilibrium*. For a semiconductor removed from *thermal equilibrium*, through the application of external excitations (such as electrical, optical, magnetic, or nonuniform thermal excitations), the definition of the *Fermi level* is not possible. However, we tend to preserve the form of the two closed-form Equations (1.27)–(1.28) for calculating the electron and *hole* concentration. This is done through proposing the following equations[40]:

$$n = n_{\mathrm{i}} \exp\left(\frac{E_{\mathrm{fn}} - E_{\mathrm{i}}}{kT}\right) \tag{1.43}$$

and

$$p = n_{\mathrm{i}} \exp\left(\frac{E_{\mathrm{i}} - E_{\mathrm{fp}}}{kT}\right) \tag{1.44}$$

where n and p are the *nonthermal-equilibrium* concentration of electrons and *holes* in the *conduction* and *valence band*, respectively. These values can either be larger than the concentrations under *thermal equilibrium* (i.e., n_0 and p_0) or smaller. E_{fn} and E_{fp} are referred to as *quasi-Fermi levels* of electrons and *holes*, respectively. Unlike the *Fermi level*, there is no physical origin to the *quasi-Fermi* levels. Often *quasi-Fermi levels* are referred to as *imref* (i.e., *Fermi* spelled backward). Evidently, under *thermal equilibrium*, *quasi-Fermi levels* will overlap one another and the *Fermi level*. The degree of deviation from *thermal equilibrium* can be quantified in terms of the difference between the *quasi-Fermi levels* of electrons and *holes*.

1.3.7 Statistics of Dopant Activation Process

Whereas at high enough temperatures all *shallow dopants* are *ionized*, in the low-to-moderate temperature range the degree of *ionization* of *dopants* requires attention. In evaluating this degree of *ionization*, the occupation status of *donor* and *acceptor states* should be studied. In such an investigation *Fermi–Dirac* distribution functions, slightly different from the one presented in (1.18), are employed. The electron distribution functions determining *ionization* status of *donor* and *acceptor states* (i.e., E_{d} and E_{a}) are expressed as

[40] Subscript 0 assigned in the previous equations to electron and *hole* concentrations is reserved for *thermal-equilibrium* situations.

$$f_{\mathrm{d/a}} = \frac{1}{1+\beta_{\mathrm{d/a}}\exp\left(\frac{E_{\mathrm{d/a}}-E_{\mathrm{f}}}{kT}\right)}. \tag{1.45}$$

In here β_{d} and β_{a} are often approximated by 1/2 and 2,[41] respectively (Landsberg, 1969, p. 270). Spontaneously arriving at the *thermal-equilibrium* state, which is expressed in (1.45), normally requires a considerably long period of time. This time requirement is due to the fact that in this case the *donor* and *acceptor states* should reach systemic *thermal equilibrium* with *valence* and *conduction band*.

There are important differences in the application of *Fermi–Dirac* statistics to the occupancy of the *conduction* and *valence band* and those of the *donor* and *acceptor energy levels*. In the latter cases, attention should be paid to the fact that while an *energy state* in the *conduction* or *valence band* can accommodate two electrons (i.e., considering *spin degeneracy*[42]), a *donor state* (or an *acceptor state*) can accommodate only one electron (or one *hole*), which is to be donated to the *conduction band* (or *valence band*). This single electron or *hole*, however, can take any of the two *spin* values at its ground state and also excited states. This is the fact that results in the presence of $\beta_{\mathrm{d/a}}$ unequal to 1 in (1.45).

In *silicon* technology in the case of a *donor* atom, the *donor state* (i.e., E_{d}) is occupied if the fifth valence shell electron is occupying it. If *ionized*, this *donor state* will be unoccupied. What makes this process unlike the process of *EHP generation* is that this fifth electron has one of the two *spin-orientation* choices in occupying the *donor state* (i.e., one *spin-up* and one *spin-down*). However, all of the electrons in covalent bonds (i.e., in *valence band*) have such a choice. A *spin state* is associated with the presence of the electron. As a result, in the absence of an electron, the *donor state*, instead of two empty *states*, merely has one. The two *states* of opposite *spin* orientation have the same value of energy and as a result impose a twofold *degeneracy* in the *Fermi–Dirac* distribution function of electrons, which as we saw in (1.45) is represented by β_{d}^{-1}. In terms of this distribution function, the density of activated *donors* in a 3-D semiconductor is given by

$$n_{\mathrm{d}} = \frac{N_{\mathrm{d}}}{1+2\exp((E_{\mathrm{f}}-E_{\mathrm{d}})/kT)} \cong N_{\mathrm{d}} \quad \text{for} \quad E_{\mathrm{d}}-E_{\mathrm{f}} \gg kT \tag{1.46}$$

where N_{d} is the volume density of *donors*. Since for low-to-moderate *dopant* levels the *density of states* function of *dopant states* is representable by a *Dirac delta function*, integration over the product of the *density of states* and the distribution function of (1.45) was not required to be shown explicitly.

This discussion, with an important modification, can also be extended to the *acceptor states*. Since the *valence-band* electrons in many semiconductors are shown to have *wave functions* rooted in p-type orbitals (hence exhibiting a threefold *orbital degeneracy*), β_{a} is expected to go beyond the pure *spin degeneracy* (addressed by β_{d}).

[41] As we will see shortly, β_{a} in many situations is equal to 4.

[42] See the discussions in Section 1.1.2.

This threefold *degeneracy* that impacts the occupation function of *acceptor states* can be partially lifted by the *crystal field* and *spin–orbit* coupling. As a result, either a twofold *degeneracy* between the *heavy-* and *light-hole bands* or a pure *spin-degenerate valence-band edge* will evolve. For these two cases, β_a is equal to 4 or 2, respectively.[43]

As a result, without lifting the *degeneracy* between the *heavy-* and *light-hole bands*,[44] the density of *activated acceptors* is given by

$$p_\mathrm{a} = \frac{N_\mathrm{a}}{1 + 4\exp((E_\mathrm{a} - E_\mathrm{f})/kT)} \tag{1.47}$$

in which β_a is equal to 4.

In the development of this formalism, it has been assumed that the *doping* level has been low enough so that no serious perturbation on the *band* structure of the *host* crystal is induced. As a result, the definition of the *effective charge-carrier* mass is still possible in terms of the simple parabolic *band* approximation near the *band* edges (i.e., calculated in terms of (1.13)). As we took note earlier, with increasing the *doping* level, when the *dopants* are on average separated from one another only by about 10 nm, such a formalism cannot be used. In that case, since *dopants* are not acting independent of one another, electron–electron interactions develop. Caused by *Pauli's exclusion principle* and electron exchange of energy, electrons will spread in their momenta. This spreading is to avoid the overlapping of the *wave function* of individual electrons. As mentioned earlier, this is the cause of the evolution of the *donor* and *acceptor states* into *bands* and eventual formation of *band tails* and shrinkage of the *bandgap*. This issue causes the *optical bandgap* to become larger than the *electronic bandgap*, which is mathematically defined using the *mass action* law (i.e., Equation (1.30)).

1.4 CHARGE-CARRIER TRANSPORT IN SEMICONDUCTORS

On the basis of the *Bloch* theorem, as indicated in Section 1.1, quantum mechanics predicts that in a perfect semiconductor (i.e., devoid of any *lattice vibration* and crystal defects) *charge carriers* travel as *propagating waves* within the boundaries of the 3-D semiconductor. Nevertheless, the presence of nonidealities existing in real semiconductors imposes randomizing *scattering* processes on the trajectory of *charge carriers*, rendering them to be not as free as *Bloch* theorem alludes. Many of these *scattering* processes are *inelastic*.

In electronic devices of relatively large sizes, the randomizing effect of the *scattering* processes paves the way for the description of *charge transport*[45] in terms

[43] In the case of acceptors, the *degeneracy* factor appears as g in the complementary *Fermi–Dirac* function: $\frac{1}{1+\frac{1}{g}\exp((E_\mathrm{f}-E)/kT)}$, which is β_a in the *Fermi–Dirac* function of (1.45).

[44] Usually the *split-off band* due to its energy separation from the top of the *valence band* is not considered.

[45] From this point onward instead of speaking of *charge-carrier transport*, we will talk of *charge transport*.

of *averaging*. In order to be able to identify the length scale of the devices to which this language of *charge transport* is applicable, we will identify more quantifiable length scales in Section 1.4.2. The *transport* language used under these circumstances is *drift–diffusion transport*. In this *averaging* formalism, there are *charge carriers* that go through fewer *scattering* processes and obtain higher *carrier* velocities (and therefore energies), which are referred to as *lucky electrons* (or *holes*), and there are those that suffer from *scattering* with a higher frequency than average. As expected, over a shorter travel length, the chance of encountering *lucky electrons* grows, and as a result the *average-based drift–diffusion* formalism fails to account for the full behavior of *charge transport*. Although for modern *deep-submicron* and *nanoscale FETs* such a formalism, which is developed according to the *Drude*[46] model, is insufficient, we will still focus our treatment of *transport* in this chapter on this model. Later on in Chapter 6, we will discuss the *transport* formalism for these smaller-size devices.

As one of the reference scales of length, the concept of *mean free path* is adopted as the length over which an average electron goes through a *scattering* process, resulting in total loss of momentum. Under the condition of *charge transport* in a device of shorter length than the *mean free path*,[47] it is said that *charge carriers* travel *ballistically*. This situation is encountered in *deep-submicron FETs*. Later on in this chapter, other length scales, which are also important to the description of *charge transport*, are reviewed.

While asserting that under *thermal-equilibrium* spatial profiles of electron and *hole* concentration are time independent, it should not be imagined that *charge carriers* at temperatures above 0 K remain stationary. On the contrary, as invoked earlier, according to the laws of statistical mechanics at a temperature T (in Kelvin) per degree of freedom, an average particle possesses an amount of energy equal to $1/2kT$. As a result, in a three-dimensional semiconductor, the average semi-Newtonian charged particle (i.e., with *effective mass* of m^*) has a finite *thermal velocity* (i.e., v_{th}), which can be calculated through the following equation:

$$\frac{1}{2}m^* v_{\text{th}}^2 = \frac{3}{2}kT. \tag{1.48}$$

This random *thermal velocity*, however, results in a *zero net velocity* for the total population of *charge carriers*. It should be mentioned that the above equation, which is based on semiclassical quantum mechanics, is only an approximation and it suffers from improper averaging.

In the presence of an external source of energy (such as an electric field) or a *nonzero* gradient in *charge-carrier* concentration, this random *thermal velocity* can be slightly biased toward a certain trajectory to produce a *nonzero net velocity* of *charge carriers*. This results in electrical current. While (1.48) predicts a random *thermal velocity* in the order of 10^7 cm/s at room temperature, the *net velocity* would be much smaller than this. Therefore, only a semiclassical description of quantum mechanics

[46] Named after *Paul Drude*.

[47] In Section 1.4.2 we will present this condition on a more accurate ground.

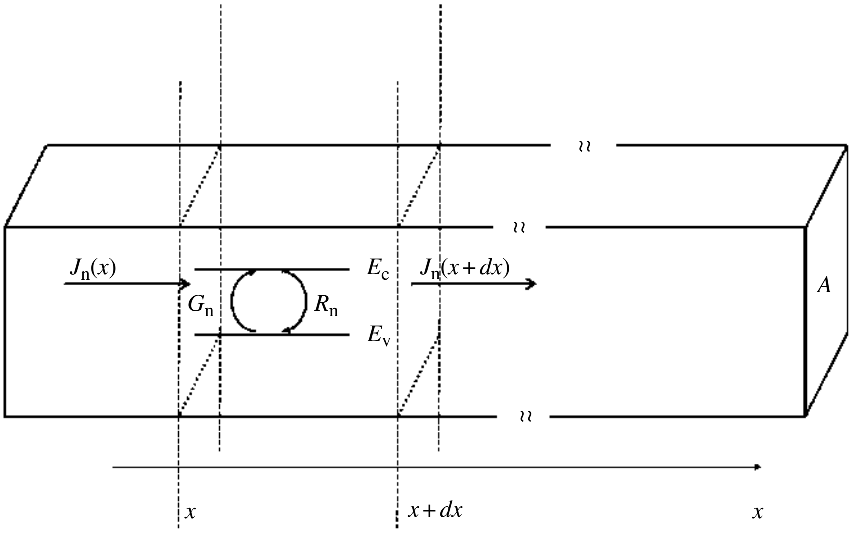


FIGURE 1.17 Schematic depiction of current continuity across a slab of semiconductor.

(certainly without invoking *relativistic mass*) needs to be considered in the formulation of *charge transport*.[48]

1.4.1 Current-Continuity Equation

A capable tool in the description of the *net* movement of *charge carriers* in a relatively large-size semiconductor is the *current-continuity equation*. According to Figure 1.17, the *charge-carrier* concentrations in a semiconductor's *conduction* and *valence bands* evolve with time as functions of *carrier transport* and *generation/recombination* rates of *carriers* (i.e., *G* and *R* in terms of the number of *carriers* per unit time, per unit volume). According to this one-dimensional depiction,

$$\frac{\partial n}{\partial t}A\cdot dx=[F_{\mathrm{n}}(x)-F_{\mathrm{n}}(x+dx)]A+[G_{\mathrm{n}}-R_{\mathrm{n}}]A\cdot dx \tag{1.49}$$

$$\frac{\partial p}{\partial t}A.dx=\left[F_{\mathrm{p}}(x)-F_{\mathrm{p}}(x+dx)\right]A+\left[G_{\mathrm{p}}-R_{\mathrm{p}}\right]A.dx. \tag{1.50}$$

In these equations, $F_{\mathrm{n/p}}(x)$ stand for the flow rates of *charge carriers* normal to the cross-sectional area *A*. Since electrons and *holes* are not always created (or annihilated) in pairs,[49] in the above equations (with indication of the appropriate

[48] In semiconductors, due to *Bragg diffraction* and evolution of negative mass, velocities much higher than 10^8 cm/s do not usually transpire.

[49] That is, through *direct band-to-band* processes.

subscripts[50]), the rates of *generation/recombination* of electrons and *holes* are differentiated from one another.

If the flow rates were to change almost linearly (i.e., expressed by the first-order *Taylor* series) between the two ends of the elemental volume *A.dx*, $F_{\mathrm{n/p}}(x) - F_{\mathrm{n/p}}(x+dx) \cong -\frac{\partial F_{\mathrm{n/p}}}{\partial x}dx$. As a result,

$$\frac{\partial n}{\partial t} \cong -\frac{\partial F_{\mathrm{n}}(x)}{\partial x} + [G_{\mathrm{n}} - R_{\mathrm{n}}] \tag{1.51}$$

$$\frac{\partial p}{\partial t} \cong -\frac{\partial F_{\mathrm{p}}(x)}{\partial x} + \left[G_{\mathrm{p}} - R_{\mathrm{p}}\right]. \tag{1.52}$$

While in terms of electron and *hole* current densities (i.e., J_{n} and J_{p}, in units of A/cm^2), $F_{\mathrm{n}} = J_{\mathrm{n}}/-q$ and $F_{\mathrm{p}} = J_{\mathrm{p}}/q$,[51] we have

$$\frac{\partial n}{\partial t} \cong \frac{1}{q}\frac{\partial J_{\mathrm{n}}(x)}{\partial x} + [G_{\mathrm{n}} - R_{\mathrm{n}}] \tag{1.53}$$

$$\frac{\partial p}{\partial t} \cong -\frac{1}{q}\frac{\partial J_{\mathrm{p}}(x)}{\partial x} + \left[G_{\mathrm{p}} - R_{\mathrm{p}}\right]. \tag{1.54}$$

According to the *continuity equation*, in the absence of a considerable *generation/recombination* rate, the time evolution of *charge-carrier* concentration will be purely explainable in terms of the spatial gradient of the electrical current density. The time constant determined in terms of the *charge-transport* processes for removing any amount of *excess charge* concentration in the semiconductor (i.e., beyond charge neutrality) is referred to as the *dielectric relaxation time*. *Dielectric relaxation time* is one of the important metrics in evaluation of *charge-transport* properties of a semiconductor.

1.4.2 Drift–Diffusion Formalism

We shall start our description of the *drift–diffusion* formalism with *drift transport* and also address the limitations of this *carrier-transport* formalism. This mode of *carrier transport*, due to its similarity with *electron transport* through conductors (which is normally expressed in terms of *Ohm*'s law), is relatively easier to understand for a reader not so familiar with semiconductors. According to *Ohm*'s law, for a rectangular slab of semiconductor with homogeneous concentration of *charge carriers*, *conductance* (*G* in *Siemens*) can be written in terms of its width (*W*), length (*L*), thickness (*T*), and the important material property known as *conductivity* (σ in 1/Ω cm):

$$G = \sigma\frac{W \cdot T}{L}. \tag{1.55}$$

[50] That is, n for *electrons* and p for *holes*.

[51] q is the absolute value of the charge of an electron.

As will be further explored in Chapter 6, with the reduction of the length scale of this slab on the order of a few *mean free paths* at normal temperatures of operation (i.e., into *deep submicrometers*), such an equation will not hold. One recent development in physics of semiconductor devices has taken place regarding determining these minimum dimensions. Physical realization of *deep-submicron* devices (i.e., of dimensions within the range or smaller than the *mean free path*) has led to significant progress in understanding the meaning of *resistance* at *nano-* and *molecular scales* and also the notion of *contact resistance* to these small-scale entities.[52]

With respect to this concept, a relatively new area has evolved in semiconductor electronics, which is referred to as *mesoscopic* electronics. In *mesoscopic* devices, dimensions are intermediate between the atomic scales and scales over which the *Ohmic* behavior of *drift transport* prevails. Although so far we have only talked about this length scale relative to *mean free path*, it is worth elaborating more on the matter. For a conductor to exhibit an *Ohmic* behavior, the dimensions ought to be much larger than each of the following three characteristic length scales (and not just one of them):

- The *de Broglie* wavelength of the electron
- The *mean free path*
- The *phase-relaxation length*

Whereas we are already acquainted with the first two, the last of these length scales indicates the distance that a *wave electron* travels before its initial *phase* information is lost.

These characteristic length scales are dependent on material, processing, temperature, and also external forces. As a result of the shortening of these characteristic length scales with increase in temperature and *bias*, a *mesoscopic* behavior is further encountered at lower temperatures and under lower *biases*. At room temperature, this characteristic is often washed out by the many existing *scattering* processes.

Caused by the tremendous push to realize electronic devices of ever-smaller dimensions, it is worth paying attention to the underlying physics of these length characteristics. This qualitative description paves the way to better understand three important (and often confused or overlooked) time constants in evaluation of the behavior of electronic devices: *momentum relaxation time*, *phase relaxation time*, and *energy relaxation time*.

Not all collisions and interactions that electrons go through are *inelastic*. As a result, *momentum relaxation time* (referred to as τ_m) in terms of collision time constant (i.e., τ_c) can be given by

$$\frac{1}{\tau_\mathrm{m}} = \frac{1}{\tau_\mathrm{c}}\alpha_\mathrm{m}, \tag{1.56}$$

[52] Measuring semiconductor properties requires creating a contact between the *macroscale* world and nano- or micron-scaled semiconductor devices. Starting from Chapter 2, through dealing with the concept of contact, we will address this issue more in-depth.

where α_{m} is a constant varying between 0 and 1, which indicates the effectiveness of the collision for destruction of momentum. For example, collisions resulting in small-angle *scattering* have very little impact on erasing the momentum information.

In order to get a more hands-on feeling about the length scales over which *mesoscopic transport* prevails, let's consider the case of low-temperature conduction through a 2-D populated channel, where the conductance is entirely determined by the electrons with energies close to the *Fermi level*,[53]

$$v_{\mathrm{f}} = \frac{\hbar k_{\mathrm{f}}}{m^*} = \frac{\hbar}{m^*}\sqrt{2\pi n}. \tag{1.57}$$

In this example, on the basis of (1.56), the *mean free path* is calculated in terms of *Fermi velocity* (i.e., v_{f}) as[54]

$$L_{\mathrm{m}} = v_{\mathrm{f}}.\tau_{\mathrm{m}}. \tag{1.58}$$

For an electron concentration on the order of 10^{11} cm^{-2}, *Fermi velocity* will be about 10^7 cm/s. As a result, for a *momentum relaxation time* of 100 ps, *mean free path* will amount to about 30 μm. Considering the size of modern semiconductor devices, this is extremely large.[55] As a result, as suggested earlier, application of equations such as (1.55) will not be extendible to devices operating under these regimes, even when the dimensions of the devices are still on micron scale.

In order to magnify the differences in nature of the three aforementioned length scales, it is worth paying attention to the special case of electron–electron *scattering*.[56] Among the many *scattering* events, electron–electron scattering is a *scattering* process that does not impact the *mean free path*. This is due to the fact that in such processes, no *net* loss in momentum of electrons occurs. In this *average* sense, a momentum lost from one electron is a momentum gained for the other. However, unlike the efficiency factor of this process (i.e., α_{m}), that of *phase relaxation time* is not equal to 0. *Phase relaxation* in general is the byproduct of *scattering* by fluctuating *scatterers*.

Whereas rigid sources of *scattering* such as impurities in general do not contribute to the *phase relaxation* process, magnetic impurities due to their internal degree of freedom (i.e., their time-fluctuating *spin*) serve as *phase randomizers* while *scattering* electrons.

[53] This is due to the fact that under low temperatures, the *Fermi–Dirac* function of (1.18) turns into a *step function*.

[54] Which is the velocity of electrons in *states* matched to *Fermi level*.

[55] In (1.57) electrons are assumed to be in a 2-D channel, which is almost like that of *MOSFETs* to be seen in Chapter 3. On the basis of assumptions of low temperature and occupation of only one *subband* in this channel, Equation (1.15) yields $n = \left(m^*/\pi\hbar^2\right)\left(E_{\mathrm{f}} - E_1\right)$. Using the semiclassical definition of energy as $E = \frac{1}{2m}(mv)^2 = \frac{1}{2m}(\hbar k)^2$, this yields the *Fermi wave number* $k_{\mathrm{f}} = \sqrt{2\pi n}$.

[56] In Section 1.7, we will deal with *scattering* mechanisms at a greater length.

In normal temperatures of operation, depending on the semiconductor's crystal structure and electron energy, an electron can propagate without perturbation over an *inelastic mean free path* in the range of 3×10^{-7} to 10^{-5} cm. Because these distances span many *lattice constants*,[57] electrons still see the crystal as a *phase-coherent* quantum mechanical entity. This explains why, between the *scattering* events, *perturbation theory* allows us to see the electrons in light of *Bloch states*, which are defined by the periodic crystal. However, after this unperturbed propagation, electrons, according to *Fermi golden rule*, *scatter* to other *Bloch states* and go through *dephasing*. As a result of this discussion, as indicated in the opening of Section 1.4, on geometric scales much larger than the *inelastic mean free path*, electrons are more or less seen as classical particles.

Evaluation of the aforementioned time constants and modes of *scattering* is extremely important in the formulation of the *Boltzmann transport equation* (*BTE*), of which the *drift–diffusion* formalism dealt with in this chapter is only a special case. For the *BTE* one needs to identify a distribution functional for electrons in terms of time, position, and momentum. Considering the constraint of the *uncertainty principle*, such a definition would be possible only if the geometrical flight length allows the application of the aforementioned semiclassical framework.

According to the earlier discussions and as a result of the *uncertainty principle*, the allowable value of *uncertainty* of the electron *wave vector* will be about 10^5 cm^{-1}, which is matched to the geometrical *uncertainty* in the range of 10^{-5} cm. This degree of accuracy in determining *wave vector* $\vec{k}$ is often deemed sufficient, since the *Brillouin zone*[58] boundary is about 10^8 cm^{-1}. In compliance with these conditions, a distribution functional $f(\vec{k}, \vec{r}, t)$ representing the probability of finding an electron having a *wave vector* between $\vec{k}$ and $\vec{k} + d\vec{k}$ and located in a space between $\vec{r}$ and $\vec{r} + d\vec{r}$ can be identified.

1.4.2.1 Drift Transport So far in our discussions, we have encountered three velocity terms:

- Random *thermal velocity* (which is of *zero net velocity*)
- *Drift velocity* or *net velocity* (which is a slight bias imposed on random *thermal velocity* through application of an external sources of energy)
- *Fermi velocity*

Our aim is now to see why in Equation (1.57) *Fermi velocity*, and not the other velocities, was used in determining the characteristic length scale L_{m}.

[57] The long-range order in a crystalline material indicates the presence of *unit cells*, which are repeated in the structure. As will be elaborated in Section 1.6, *lattice constant* is an indicatory dimension of these *unit cells*.

[58] Using *Schrödinger equation*, a *real-space* crystal structure can be transformed into *momentum space*. While in the *real space* periodicity of crystal presents itself through the notion of repetition of *unit cells*, in the *momentum space* also the information is repeated in the form of repetition of a building block, which is known as *Brillouin zone*. This is named after *Léon Brillouin*.

In a large homogeneous semiconductor, *carrier transport* is explained in terms of the *drift velocity* (i.e., v_{d}), which yields the electron *drift* current density as

$$J_{\mathrm{n}} = qnv_{\mathrm{dn}}. \tag{1.59}$$

The *drift hole* current density can also be expressed through an equation identical in form to (1.59). However, instead of electron concentration (i.e., n), *hole* concentration (i.e., p), and instead of *drift velocity* of electrons (i.e., v_{dn}), drift velocity of *holes* (i.e., v_{dp}) should be used. In the calculation of current under conditions very close to *thermal equilibrium*, we can still approximately rely on *Fermi–Dirac* statistics. However, for further deviations from *thermal equilibrium*, another distribution functional should be used in the current calculation.

In this equation, through incorporation of electron concentration n, it is implied that all *conduction-band* electrons are contributing to current conduction. However, in the case of conductors, or in *degenerate* semiconductors,[59] such a framework becomes misleading. In these cases, application of an external electric field does not induce a net velocity (i.e., v_{d}) in all conduction electrons. With the use of *energy-resolved* measurements, it has been shown that only those electrons with energy values within a few *kT* of the electron *quasi-Fermi level* are carriers of the *net* current. This occurs while the rest of the *conduction-band* electron population maintains at its *zero net velocity*. This finding results in a tremendous simplification in studying the behavior of these *solids* at very low temperatures, since only a very small concentration of conduction electrons are now required to be studied.

This fact highlights the difference between the impact of application of electric field on electrons as *individual* single particles or as an *ensemble* of electrons. From the point of view of an *ensemble* of electrons, the application of an electric field results in moving only a few electrons from *states* with negative *wave number* (i.e., $-k_{\mathrm{f}}$) to *states* with positive *wave number* (i.e., $+k_{\mathrm{f}}$). This is schematically illustrated in Figure 1.18. Accordingly, the *drift transport* equation in this case can be rephrased as

$$J_{\mathrm{n}} = q\left[n\frac{v_{\mathrm{dn}}}{v_{\mathrm{f}}}\right]v_{\mathrm{f}}, \tag{1.60}$$

where the quantity in the square brackets is representative of a small fraction of total concentration of electrons with energies within a few *kT* of *quasi-Fermi level*. These are the electrons that move with *Fermi velocity*. Although from a purely quantitative point of view this equation implies no change in the value of *drift* current density, it conceptually indicates why it is required to talk about *Fermi velocity* in metals as well as in low-temperature/low-dimensional *degenerate* semiconductor systems. As is shown in (1.60), in *degenerate* systems under low temperature, conduction takes place

[59] Which, as indicated earlier in this chapter, refers to the cases where *Fermi level* is positioned above the *conduction-band* edge for a three-dimensional *bulk* semiconductor or above the first *subband level* for the case of a lower-dimensional semiconductor $k_{\mathrm{B}}T \ll E_{\mathrm{f}} - E_1$, which materializes specially at lower temperatures. Many of the *FETs* have *degenerate* channels for current conduction.

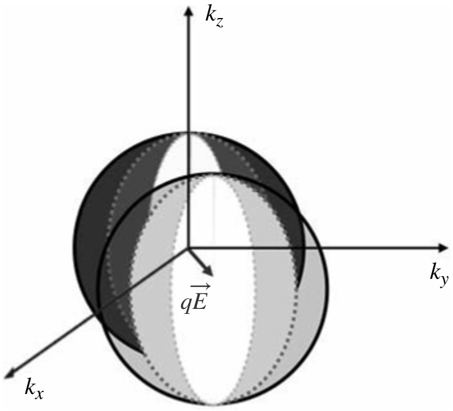


FIGURE 1.18 While prior to the application of an electric field all states within a sphere of radius k_{f} (i.e., Fermi wave number) are filled at low temperatures, upon application of an electric field $\vec{E}$, this sphere shifts according to the vector $q\,\vec{E}$. As a result, instead of symmetric filling of the band, states that carry the current along the electric-field vector are filled up to a higher energy. This elimination of symmetry between the states (k_x, k_y, k_z) and ($-k_x$, $-k_y$, $-k_z$) results in current conduction in the range of energies for which the two spheres are nonoverlapping.

through the movement of a small number of electrons at velocities much higher than indicated by the *drift velocity*.

However, aside from this *ballistic transport* in low-dimensional systems and low-temperature transport behavior of *degenerate* systems, transport under low electric fields can be easily described in terms of the *drift* (and to be added in Section 1.4.2.2 *diffusion*) formalism.

As expected, upon application of an electric field, the *net* movement of *charge carriers* is materialized either along the direction of electric field (for *holes*) or opposite of that (for electrons). In *macroscopic* electronic devices, an instantaneous relationship is proposed between the applied electric field and this *net velocity* (i.e., *drift velocity*: v_{dn} and v_{dp}):

$$v_{\mathrm{dn}} = -\mu_{\mathrm{n}} E \tag{1.61}$$

$$v_{\mathrm{dp}} = \mu_{\mathrm{p}} E. \tag{1.62}$$

The proportionality constants of these relationships are referred to as *low-field* electron and *hole mobility* (i.e., μ_{n} and μ_{p} in cm^2/V s, respectively). As will be shown in Section 1.7, these constants are dependent on the frequency of a large number of *scattering* processes that a *carrier* encounters when moving in a semiconductor. As a result, they depend on a long list of parameters including temperature, presence of impurities, presence of internal *polarization* in the semiconductor, etc.

Equations (1.61) and (1.62) certainly have their own limitations. Assumption of such instantaneous relationships even in portions of large-size semiconductor devices (i.e., in regions of rapidly changing electric field) has been proven to be misleading. In these relationships, it is assumed that an average semi-Newtonian electron,[60] while traveling through a piece of semiconductor across which an electric field (i.e., E) has been established, will at any given point experience a force equal to $-qE$. As a result of this force, assuming a constant electric field between any two *inelastic scattering* events (which are separated from one another by an amount of time typically referred to as *mean scattering time*), the electron will receive an amount of energy equal to $qE\lambda$ over the length of one *mean free path* (i.e., λ). If we assume that on average after a time period equal to the so-called mean scattering time (i.e., τ_{cn}, for electrons),[61] an electron undergoes an *inelastic scattering* process (and sees a drop in its field-contributed *drift velocity* to 0), we can write the following Newtonian equation for this particle:

$$qE\tau_{\text{cn}} = -m_{\text{n}}^{*}v_{\text{dn}} \tag{1.63}$$

in which m_{n}^{*} is the effective mass of electron.

As a result, based on (1.61),

$$\mu_{\text{n}} = \frac{q\tau_{\text{cn}}}{m_{\text{n}}^{*}}. \tag{1.64}$$

Likewise, we can calculate the *low-field hole mobility* as

$$\mu_{\text{p}} = \frac{q\tau_{\text{cp}}}{m_{\text{p}}^{*}}, \tag{1.65}$$

where τ_{cp} and m_{p}^{*} are the *mean scattering time* and *effective mass* of *holes*, respectively.

These approximations, in circumstances dealing with large-size electronic devices and when the applied electric field is small, hold well. Nevertheless, at higher electric fields since electrons (and *holes*) move away from the bottom of *conduction band* (and *valence band*), the *effective-mass* values calculated at the *band* boundaries will not be applicable. It is worth noting that in formalizing *drift transport*, if electric field were large, higher powers of E would be required in explaining the *transport*. This is because high electric field appreciably disturbs the equilibrium distribution of velocities of *carriers* (i.e., profile expressed in (1.21)). In addition, increase in energy will not only change the *scattering* rates, and as a result the value of *mean scattering time*, but will also trigger a number of new *scattering* processes. In Section 1.7.3, we will perform a fuller evaluation of a variety of *scattering* processes.

[60] While this discussion is being presented for electrons, in case of *holes*, evidently with the appropriate change of signs, appropriate equations are developed.

[61] It should be emphasized that τ_{cn} (also its counterpart in case of *holes*: τ_{cp}) is given by the *momentum relaxation time* and not the collision time constant in (1.56).

In a simple review of this problem, we see that we have thus far made a good number of assumptions, including:

- The electron, which is a *wave/particle dual*, is a Newtonian particle.
- *Transport* is one-dimensional.
- The electric field is constant between any two *inelastic scattering* events.
- An *inelastic scattering* process happens every τ_{cn} (i.e., *mean scattering time* for electrons) and τ_{cp} (i.e., *mean scattering time* for *holes*).

These assumptions prevent us from developing a thorough understanding of the behavior of the quantum *wave/particle* c*harge carrier*. The processes of receiving (or dissipating) energy of these *wave/particles* are not instantaneous. These processes are instead governed by a *relaxation time constant*, referred to as *energy relaxation time*.

Prompted by the difference in the energy of *phonons* and electrons, the *momentum relaxation time* is often shorter than the *energy relaxation time*. This is because the thermalization of electrons requires a number of inelastic *scattering* events, considering the large electron energy (i.e., electron–*phonon* energy exchange). However, in contrast, even with one *scattering* event, the momentum of the *charge carrier* can be randomized.

As mentioned before, the *low-field mobility* is dependent on the frequency of electron interaction with *scattering* processes. Temperature and presence of *ionized impurities* are two prime contributors to *scattering* processes, on which the *low-field mobility* depends. These contributions result in *ionized-impurity scattering* and *lattice-vibration scattering* (also known as *phonon scattering*). At low temperatures, due to insignificant *lattice vibrations*, *ionized-impurity scattering* is dominant. At higher temperatures, *phonon scattering* plays the dominant role.

As a result, the temperature dependence of the *low-field mobility* takes two completely different characteristics at the temperature extremes. At high temperatures, because of the strengthening of the dominant *scattering* process (i.e., *phonon scattering*), the frequency of *scattering* events increases, and as a result *carrier mobility* deteriorates. At lower temperatures, an increase in temperature results in a *mobility* improvement. This improvement in *mobility* is due to the increased *thermal velocity* of *carriers* and, as a result, reduction in the time spent by them in the vicinity of the dominant *scatterers* (i.e., *ionized impurities*). In semiconductors free of *ionized impurities* (i.e., in the absence of *ionized-impurity scattering*), at lower temperatures *mobility* becomes temperature independent. Whereas depending on the curvature of the *E–k* diagram and the value of mean *scattering time*, the values of electron and *hole mobility* of different semiconductors are different, at higher temperatures the *low-field mobility* becomes almost independent of the semiconductor.

With increase in *dopant* concentration and when *ionized-impurity scattering* is dominant, *carrier mobility* deteriorates. Figure 1.19 schematically illustrates these temperature- and *doping*-dependent variations of *carrier mobility*.

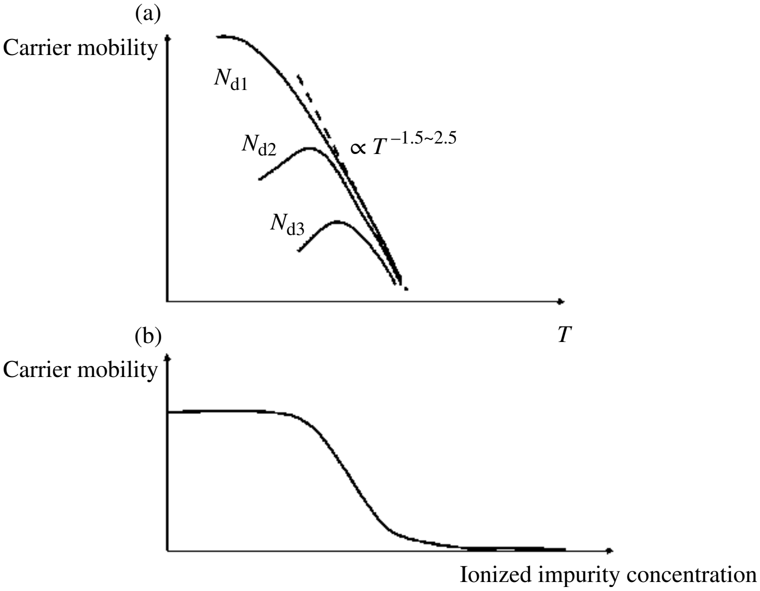


FIGURE 1.19 (a) Typical dependence of carrier mobility on lattice temperature for a semiconductor doped to three different levels of doping: N_{d1}, lightly doped; N_{d2}, moderately doped; N_{d3}, heavily doped. (b) Typical dependence of carrier mobility on the ionized-impurity concentration. Scales are linear.

1.4.2.2 Diffusion Transport Under the aforementioned circumstances, which were identified as suitable for the application of *drift-transport* formalism, the second mode of conduction is *diffusion*. *Diffusion* is a thermal process that, unless there is a gradient in the concentration of *carriers*, produces a *zero net* movement of *charge carriers*. This is because the average *carrier*, although moving with *thermal velocity*, has a random *zero-sum* movement. In the presence of a gradient in *charge-carrier* concentration, however, *carriers* will diffuse from regions of higher concentration to regions of lower concentration.

According to the one-dimensional situation depicted in Figure 1.20, with a few assumptions a simple equation can be produced for this *net* movement of charge in the presence of a *carrier* concentration gradient (while electric field is 0). In this one-dimensional situation, *carriers* at any given point have a 50% chance of traveling toward the right or the left side. Taking two points to the right and to the left of a given reference point, distanced from the reference point by the *mean free path*, one can assume that an average *carrier* starting at each of these two points and coming to the reference point will not go through any *inelastic scattering* event before reaching $x=0$. As a result, in the case of electrons, the net flux of *carriers* at $x=0$ will be equal to

$$F=\frac{1}{2}v_{\text{th}}n(x-\lambda)-\frac{1}{2}v_{\text{th}}n(x+\lambda). \tag{1.66}$$

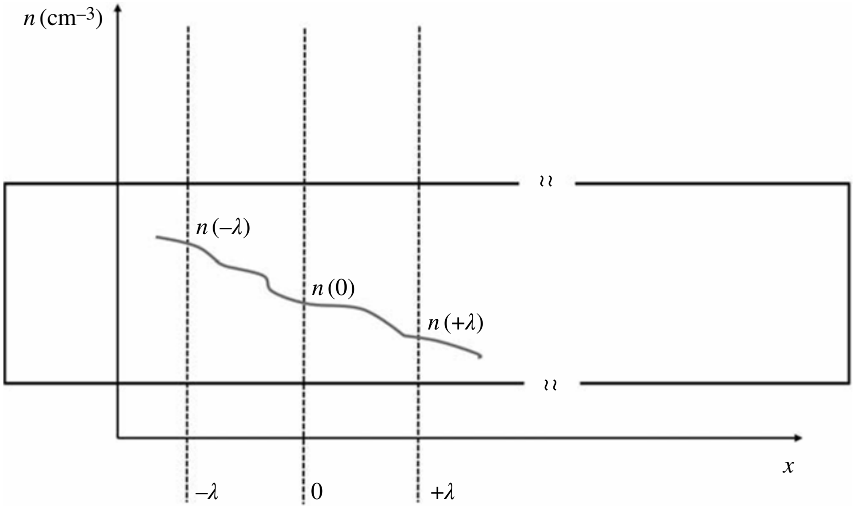


FIGURE 1.20 Schematic depiction of an arbitrary 1-D dependent carrier concentration profile (i.e., along the *x*-axis) on the cross section of a slab of semiconductor.

Assuming that the electron concentration gradient is small enough so that $n(x)$ can be approximated by the first-order *Taylor* series, we can write

$$F \cong -\lambda v_{\mathrm{th}} \frac{dn}{dx}. \tag{1.67}$$

Since for electrons *diffusion current* and flow rate of electrons can be related as $J_{\text{Diff-n}} = -qF$, the *diffusion*-current density can be written as

$$J_{\text{Diff-n}} \cong q\lambda v_{\mathrm{th}} \frac{dn}{dx}. \tag{1.68}$$

The factor λv_{th} is referred to as the *diffusion constant* of electrons (D_{n} in units of cm^2/s).

For the case of holes, with appropriate modifications, the following equation is developed:

$$J_{\text{Diff-p}} \cong -q\lambda v_{\mathrm{th}} \frac{dp}{dx}. \tag{1.69}$$

1.4.2.3 Einstein Relationship It is important to understand that the constants of *drift* and *diffusion* in terms of current are not independent of one another. At conditions close to *thermal equilibrium* (i.e., low *E*) and for *nondegenerate* semiconductors, a simple derivation will result in the famous *Einstein relationship*[62] between these constants.[63]

According to our earlier discussions,

$$D_{\mathrm{n}} = \lambda v_{\mathrm{th}} = v_{\mathrm{th}}^2 \tau_{\mathrm{cn}} = v_{\mathrm{th}}^2 \frac{\mu_{\mathrm{n}} m_{\mathrm{n}}^*}{q}. \tag{1.70}$$

For a one-dimensional semiconductor, with no electric field, according to the *equipartition* theorem,

$$\frac{1}{2} m_{\mathrm{n}}^* v_{\mathrm{th}}^2 = \frac{1}{2} kT. \tag{1.71}$$

As a result, based on (1.70) and (1.71),

$$D_{\mathrm{n}} = \mu_{\mathrm{n}} \frac{kT}{q}. \tag{1.72}$$

An identical relationship can also be established between the *diffusion constant* of *holes* and their *low-field mobility*. As a result, between the four transport factors of *drift–diffusion* formalism, we have

$$\frac{D_{\mathrm{n}}}{\mu_{\mathrm{n}}} = \frac{D_{\mathrm{p}}}{\mu_{\mathrm{p}}} = \frac{kT}{q}, \tag{1.73}$$

which, as indicated already, is referred to as *Einstein relationship*.

This version of *Einstein relationship* is only valid for a *nondegenerate* semiconductor (i.e., the *Maxwellian* semiconductor expressed in (1.21)). However, as proven in Appendix 1.B, for a *degenerate* semiconductor at low electric fields, in terms of *Dirac integrals*, we will have

$$\frac{D}{\mu} = \frac{kT}{q} \left[2F_{3/2}(\eta) / 3F_{1/2}(\eta) \right], \tag{1.74}$$

where

$$F_j(\eta) \equiv \int_0^{\infty} \frac{\xi^j d\xi}{1 + \exp(\xi - \eta)} \quad \text{and} \quad \eta = \frac{E_{\mathrm{f}} - E_{\mathrm{c}}}{kT}. \tag{1.75}$$

Notice that $\mathrm{F}_j(\eta)$, which we saw in Section 1.3.5, is the scaled version of $F_j(\eta)$ by $\Gamma(j+1)$.

[62] Named after *Albert Einstein*.

[63] In Appendix 1.B, we will show how a more sophisticated version of this relationship is also valid without these conditions.

According to the *drift–diffusion* formalism, the total electron and *hole* current density terms can be written as

$$J_{\mathrm{n}} = qn\mu_{\mathrm{n}}E + qD_{\mathrm{n}}\frac{dn}{dx} \tag{1.76}$$

$$J_{\mathrm{p}} = qp\mu_{\mathrm{p}}E - qD_{\mathrm{p}}\frac{dp}{dx}. \tag{1.77}$$

Through applying the definitions of *quasi-Fermi levels* to a *nondegenerate* semiconductor, we can combine the *drift*- and *diffusion*-current terms in one equation:

$$J_{\mathrm{n}} = n\mu_{\mathrm{n}}\frac{dE_{\mathrm{fn}}}{dx}. \tag{1.78}$$

Likewise we can write

$$J_{\mathrm{p}} = p\mu_{\mathrm{p}}\frac{dE_{\mathrm{fp}}}{dx}. \tag{1.79}$$

The *diffusion* process, in addition to being instigated by the presence of a finite gradient in *dopant* and *charge-carrier* concentration, can also be provoked by the existence of a temperature gradient. This can be clearly seen through the temperature dependence of the *quasi-Fermi levels* presented in Equations (1.43) and (1.44).

With explicit indication of transport due to a temperature gradient, *drift–diffusion* transport of electrons can be formulated as

$$J = nq\mu E + qD\nabla n + qS\nabla T, \tag{1.80}$$

where S is referred to as *Soret coefficient*.[64]

Although we have already provided the *drift–diffusion* formalism in a unified form, it is worth emphasizing again that, especially due to the restrictions of *Pauli's exclusion principle*, for *degenerate* semiconductors, one cannot assume that *drift* and *diffusion* are independent processes.

Example

Along the length of a slab of *n-doped germanium* with a peak *dopant* concentration of 10^{15} cm^{-3}, a constant electric field of 100 V/cm is applied. The length of the slab is 1 μm. Assuming that at room temperature along this sample a current of 10 A/cm^2 is flowing in the direction of the electric field, determine the spatial distribution profile of the *dopants*. Room-temperature values of electron and *hole mobility* everywhere along the length of this sample are taken equal to 1000 and 800 cm^2/V s, respectively.

[64] *Soret* coefficient is defined as $S_{\mathrm{n}} = \mu_{\mathrm{n}}(k/q)n$, and it is named after *Charles Soret*.

Since the sample is *n-doped*, we start with assuming that the current is being carried by electrons only. Taking the electric field to be in the +*x* direction, in terms of (1.76),

$$J_\mathrm{n} = qn\mu_\mathrm{n}E + qD_\mathrm{n}\frac{\partial n}{\partial x} = +10\ \mathrm{A/cm^2}.$$

Since $\mu_\mathrm{n} = 1000\ \mathrm{cm^2/V\ s}$, according to *mass action law*,

$$\frac{D_\mathrm{n}}{\mu_\mathrm{n}} = \frac{kT}{q} \Rightarrow D_\mathrm{n} \cong 25.8\ \mathrm{cm^2/s}.$$

Based on these values,

$$1.6\times10^{-19}\left[n(x)\times1000\times100+25.8\frac{\partial n}{\partial x}\right]=10,$$

which yields

$$n(x)=A.\exp\left(-3.88\times10^{3}x\right)+6.25\times10^{14}\ \mathrm{cm^{-3}}.$$

We assume first that $N_\mathrm{d}(x) \cong n(x)$ and then evaluate the validity of this assumption.

Since the peak of the *donor* concentration is $10^{15}\ \mathrm{cm^{-3}}$, $n(x=0)$ is set to this value. Accordingly,

$$N_\mathrm{d}(x)=3.75\times10^{14}.\exp\left(-3.88\times10^{3}x\right)+6.25\times10^{14}\ \mathrm{cm^{-3}}.$$

Based on this distribution profile, along the length of the sample stretching from $x=0$ to $x=10^{-4}$ cm, *dopant* concentration changes from 10^{15} to $8.79\times10^{14}\ \mathrm{cm^{-3}}$.

Within this range everywhere N_d is at least an order of magnitude greater than $n_\mathrm{i}(T=300\ \mathrm{K})=2.4\times10^{13}\ \mathrm{cm^{-3}}$, which validates our earlier assumption of $N_\mathrm{d}(x) \cong n(x)$.

We still have to validate another assumption (i.e., if electrons are the predominant charge carriers). Based on the electron concentration profile, the *hole* concentration profile is evaluated as

$$p(x)=\frac{\left(2.4\times10^{13}\right)^2}{3.75\times10^{14}.\exp\left(-3.88\times10^{3}x\right)+6.25\times10^{14}}.$$

According to which,

$$J_\mathrm{p} = qp\mu_\mathrm{p}E - qD_\mathrm{p}\frac{\partial p}{\partial x},$$

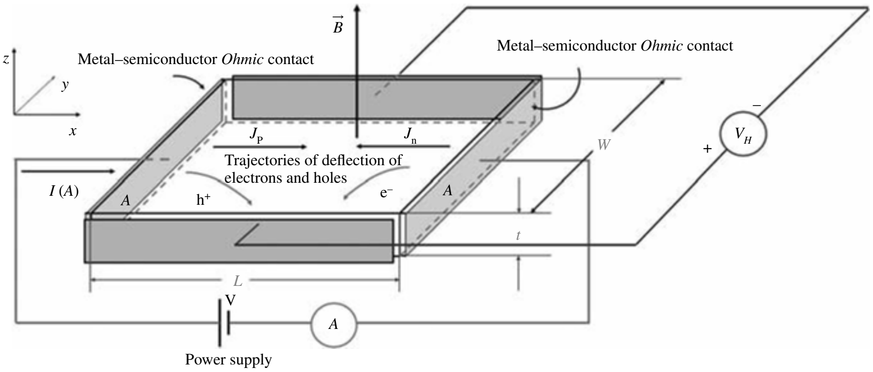
where $\mu_\mathrm{p} = 800\ \mathrm{cm^2/V\ s}$ and $D_\mathrm{p} = 800\times25.8\times10^{-3} \cong 20.64\ \mathrm{cm^2/s}$.

Based on these, since both $n \ll p$ and $(\partial p/\partial x) \ll (\partial n/\partial x)$, J_p would be much smaller than J_n, which sets electrons as the predominant charge carriers.

1.4.3 Characterization of Low Electric-Field Transport Parameters

Hall measurement is one of the most capable tools in characterizing semiconductors. This tool can be used to not only measure the *carrier mobility* in a semiconductor but also to reveal the *type* of its majority *carriers* and their concentration. Figure 1.21 illustrates a *Hall* measurement setup. In this setup in presence of both an electric field (which is induced due to the application of *bias V*) and the permanent magnetic field, while *charge carriers* are traversing along, or opposite to, the direction of the electric field (i.e., for the *holes* and electrons, respectively), they are deflected to the same transverse side of the semiconducting slab. In this deflection, charge neutrality will be imbalanced, which results in induction of a second electric field normal to the first. The balancing act between the *Lorentz* force imposed on the moving *charge carriers* by the magnetic field and the force exerted by this secondary electric field (which is referred to as *Hall* field) renders an equilibrium state. Through measuring the polarity and the magnitude of V_H, one can gain insight into the majority *carrier type*, *mobility*, and concentration. For example, for an n-type sample the balance of forces yields

$$E_\mathrm{H} = -\frac{J_\mathrm{n}}{qn}B. \tag{1.81}$$



$\vec{B}$
z
y
x
Metal–semiconductor *Ohmic* contact
Metal–semiconductor *Ohmic* contact
J_P
J_n
Trajectories of deflection of electrons and holes
I (A)
A
h⁺
e⁻
A
W
t
L
V
A
Power supply
V_H
+
−

FIGURE 1.21 Schematic depiction of a simple Hall measurement setup. A uniform magnetic field with flux density $\vec{B}$ is applied perpendicular to the *x*–*y* plane. According to the polarity of the connections to the power supply, a current in the positive *x*-direction is fed normal to the *y*–*z* plane of the slab with contact area indicated as $W \times t$. Based on the indicated directions of deflection for electrons and holes, in the case of p- and n-doped semiconductor, a positive or a negative Hall potential (i.e., V_H) will be observed on the voltmeter, respectively. Accordingly, the direction of the resulting Hall electric field E_H defined along the *y*-axis will depend on whether the semiconductor is n- or p-type doped.

Likewise, in a p-type sample we have

$$E_{\mathrm{H}} = \frac{J_{\mathrm{p}}}{qp} B. \tag{1.82}$$

Clearly V_{H}, which is equal to $E_{\mathrm{H}} \cdot W$, is positive in a p-type sample and negative in an n-type sample. E_{H} is often expressed as $R_{\mathrm{H}} \cdot J \cdot B$.

In the analysis of *Hall* measurement data, we have to pay attention to the inherent assumptions of *drift-current transport*. According to this first-order approximation expressed in (1.81) and (1.82), we have assumed that all *carriers* have the same *mean scattering time*. In reality, however, a proper averaging of *mean scattering time* must be adopted. Formal analysis by the *Boltzmann transport equation*, in the case of electrons, results in

$$\mu = \frac{q}{m^*} \cdot \frac{\langle v^2 \tau \rangle}{\langle v^2 \rangle} \tag{1.83}$$

$$R_{\mathrm{H}} = -\frac{1}{qn} \cdot \frac{\langle v^2 \tau^2 \rangle \langle v^2 \rangle}{\langle v^2 \tau \rangle^2}. \tag{1.84}$$

The symbol <> represents the averaging process over the *Boltzmann* distribution of *carriers*. As a result, the product of R_{H} and *conductivity* (i.e., σ in units of Ω^{-1} cm^{-1}) is different from the low-field *mobility* (i.e., *drift mobility*) that we have talked about in this section. This product is referred to as *Hall mobility*.

As a result of these relationships, the ratio of *Hall mobility* (i.e., μ_{H}) versus *drift mobility* (i.e., μ) can be written as

$$\frac{\mu_{\mathrm{H}}}{\mu} = \frac{R_{\mathrm{H}}.\sigma}{\mu} = \frac{\langle v^2 \tau^2 \rangle \langle v^2 \rangle}{\langle v^2 \tau \rangle^2}. \tag{1.85}$$

This ratio, although not equal to 1, is generally close to 1. Nonuniformity in the distribution of current, temperature nonuniformity across the sample, and lack of an ideal *Ohmic* contact[65] to the semiconductor sample are often the other sources of ambiguity in *Hall* measurement data.

1.4.4 High Electric-Field Drift Transport

As identified earlier, under high electric fields the semilinear relationship between the drift velocity and the electric field vanishes (i.e., (1.61) and (1.62)). Throughout

[65] As shown in Figure 1.21 and more clearly explored in Chapter 2, a metal–semiconductor contact between the outside world (in this case power supply) and the semiconductor is required. The main characters of such a contact are that it presents very little potential drop at the contact-site and behaviorally it follows *Ohm*'s law (hence the name Ohmic contact).

this section we have already named a few culprits, which we now intend to summarize.

The increase in the amount of *excess* kinetic energy gained by the average electron (or *hole*) from the electric field results in further drift of the *charge carrier* into the *conduction band* (or *valence band*). Consequently, the *effective mass* of the *charge carrier* and its *mobility* change. In addition, in some semiconductors, this can cause the migration of electrons from *lower valleys* to *higher valleys*[66] of *conduction band*. At the same time, the increase in energy of the *charge carriers* plays an important role in modifying the *scattering* rates and also in activating *scattering* processes that were not present in the case of lower *carrier* energies. This *excess* of energy under high electric-field conditions is often expressed in terms of a quantity referred to as *electron temperature* (i.e., T_e). This definition is in analogy with the statistical mechanics, notion of kT (where T is the *lattice temperature*) as a representation of average thermal energy of an electron. In this discussion, one should be careful to distinguish between the notions of *lattice temperature* and that of *electron temperature*.

1.4.4.1 Electron Temperature versus Lattice Temperature In order to understand the development of *electron temperature* beyond *lattice temperature* at high electric fields, we shall start with a review of our understanding of *thermal equilibrium*. Under *thermal equilibrium*, the equality of *charge carriers*' emission and absorption of energy results in a *zero-sum* gain of energy. This equality is maintained through equality of emission and absorption rates of *phonons*, which are quantum particles representative of *lattice vibrations*. Under this condition, the energy distribution is *Maxwellian*. However, applying an electric field across a semiconductor removes the system from the state of *thermal equilibrium* (as now *charge carriers* begin to gain additional energy from the electric field while still emitting some of it through the *phonon emission* process). As a result, not only for high electric fields but also for low-to-moderate values of electric field, electrons will become more energized than they were under *thermal equilibrium*. This *excess* of energy beyond *thermal equilibrium* is what is explained through the notion of *electron temperature*. In connection to this *excess* energy, it must be indicated that even under moderate electric-field conditions, the increase in energy is less than the amount predicted by the laws of electrostatics. This is because the emission rate of *phonons* is now also on the rise, which will eventually reach steady state for a given electric field by matching the absorption rate. This description partially explains the *bowing* in the *drift velocity* versus electric-field characteristics (also known as v_d–E characteristics) at medium values of electric field (Fig. 1.22).

Figure 1.22, in addition to demonstrating the bowing of v_d–E characteristics, shows two other important phenomena: (i) saturation of *drift velocity* at high electric fields and (ii) peaking of the v_d–E and evolution of a *negative differential mobility*[67] region

[66] Also referred to as *satellite valley*.

[67] *Differential mobility* is defined unlike the *linear mobility* and is calculated around a given value of electric field as $\Delta v/\Delta E$.

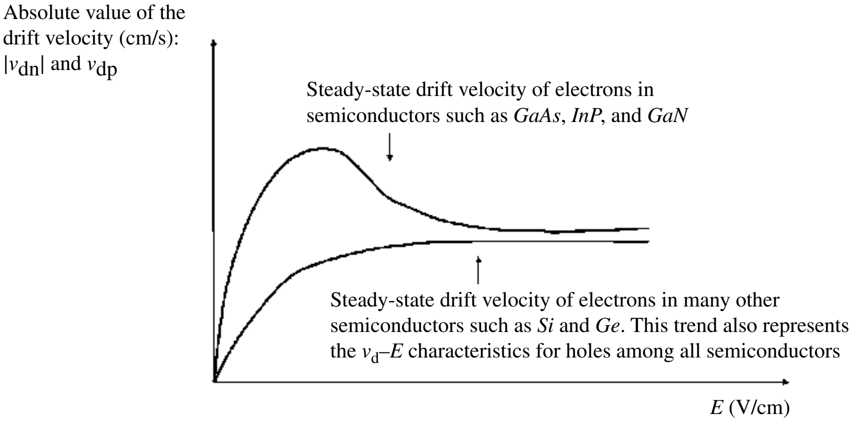


FIGURE 1.22 A behavioral depiction of v_d–E characteristic for electrons and holes in different semiconductors. Scales are linear.

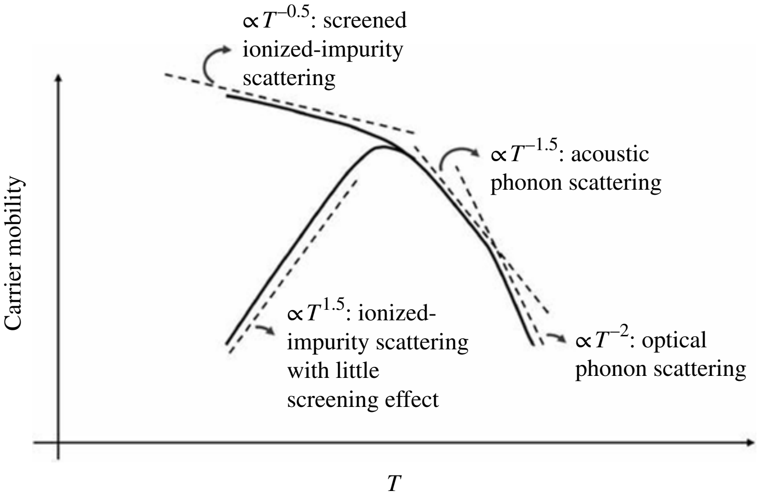


FIGURE 1.23 Correlation between the typically observed temperature dependences of low-field carrier mobility in semiconductors and the dominance of scattering processes. Scales are linear.

in the intermediate range of electric field only for electrons (and not *holes*) in some semiconductors.

While for small electric fields the *low-field carrier mobility* remains descriptive of the achievable *carrier drift velocity*, Figure 1.23 behaviorally presents the dependence of this parameter on temperature under the dominance of a number of *scattering* processes (which will be dealt with especially in Section 1.7).

For moderate values of electric field, which do not result in transfer of electrons from one *valley* to the next, we can equate the rates of emission and absorption of energy at a given electric field in order to quantify the *electron temperature*. For the relatively simpler cases of *Ge* and *Si*, which are semiconductors with no *intervalley* electron transfer, balancing the rate equations results in the following definition for the ratio of *electron temperature* to *lattice temperature*:

$$\frac{T_{\mathrm{e}}}{T}=\frac{1}{2}\left[1+\sqrt{1+\frac{3\pi}{8}\left(\frac{\mu_{\mathrm{n}}E}{c_{\mathrm{s}}}\right)^{2}}\right]. \tag{1.86}$$

In terms of which,

$$v_{\mathrm{dn}}=-\mu_{\mathrm{n}}E\sqrt{\frac{T}{T_{\mathrm{e}}}}. \tag{1.87}$$

In these equations, μ_{n} is the low-field electron *mobility* and c_{s} represents the velocity of sound. As shown in (1.86) and (1.87), for moderate fields (i.e., when the linearly calculated *drift velocity* is comparable to the sound velocity), the *electron temperature* increases beyond the *lattice temperature*. As a result, *drift velocity* will become smaller than the linearly predicted value of $\mu_{\mathrm{n}}E$.

Other more general frameworks have also been developed for determining the *electron temperature*. With certain approximations, it can still be proven that the distribution function of electrons under conditions far from *thermal equilibrium* follows the form of *Maxwell–Boltzmann* statistics. The major difference is that the *lattice temperature* (i.e., T) should be replaced by the electron temperature (i.e., T_{e}), where approximately

$$T_{\mathrm{e}}\approx T\left(1+\frac{E^{2}}{E_{\mathrm{C}}^{2}}\right). \tag{1.88}$$

In (1.88) E is the electric field and E_{C} is called the *critical* electric field (which for the case of electrons in *silicon* is about 10^{4} V/cm). At this *critical* electric field, T_{e} is approximately equal to $2T$. $3/2kT_{\mathrm{e}}$ is a measure of average energy of electron. This can result from the *Boltzmann transport equation*, assumption of *E–k* diagram, and *scattering* in terms of the *Fermi golden rule*.

In the aforementioned derivation of the *Einstein relationship* in Section 1.4.2.3, it has been assumed that the first-order *Taylor* series representation of electron concentration is sufficient. However, in certain situations, such as under high electric fields, this picture is not necessarily acceptable. Obviously, the aforementioned derivation was performed from a *thermal-equilibrium* perspective in which it was assumed that kinetic energy is equal to thermal energy. Such an assumption, which ignores *drift energy*, is not permissible under high electric fields. However, under high electric fields, one can establish a relationship close to the *Einstein relationship* by simply replacing the *lattice temperature* with the *electron temperature* (i.e., $(D/\mu) = (kT_e/q)$). This is valid only for *nondegenerate* (i.e., *Maxwellian*) semiconductors.

The electric fields encountered in a transistor can be very well in excess of E_C. As a result, *electron temperature* can be much higher than *lattice temperature*. However, we should avoid confusing the *electron temperature* with the actual device temperature. A transistor with thousands of degrees of *electron temperature* is still cool to the touch. For electrons to leave the crystal and have their temperature felt, they have to overcome a *work function*. This *work function* itself is typically several *electron volts*, which interestingly enough corresponds to thousands of degrees *Kelvin* (i.e., just like the T_e). It should also be mentioned that the earlier formalism loses validity at extremely high fields (i.e., $E \sim 10^6$ V/cm). At such high electric fields (and as a result high energies), the *density of states* function becomes extremely nonlinear. However, in the derivation of the above, a *DOS* function such as (1.14) has been employed. For those high electric-field cases, numerical techniques such as the *Monte Carlo* method should be used to calculate the *electron temperature*.

As a result of the discussions on *electron temperature*, it should be indicated that in the definition of *quasi-Fermi levels*, these *energy levels* are functions of T_e and not the *lattice temperature*.

1.4.4.2 Steady-State Velocity Overshoot and Saturation At higher electric fields, increase in the *carrier* energy (and *carrier temperature*) triggers the interaction of *carriers* with other sources of *scattering* known as *optical phonons*.[68] This is an interaction that is, however, not incorporated into the *balance* equations employed in the derivation of (1.86) and (1.87). The *optical-phonon emission* process, being a very efficient process of removing *excess* energy (beyond that of *optical phonons*) from *carriers*, results in a *drift velocity* quite independent of electric field (i.e., saturation of v_d–E). For all semiconductors, such a field-independent saturation of *drift velocity* is observed for both electrons and *holes* at high electric fields (Fig. 1.22). The onset of this saturating behavior is determined by the *optical-phonon emission* energy of the given semiconductor.

Saturation velocity, being determined by *optical-phonon* emission *scattering*, is essentially independent of *doping* concentration. It is also quite independent of the semiconductor itself. The reason for the similarity of saturation velocity of different semiconductors is that at high energies, corresponding to the high electric fields, the *DOS* functions of all semiconductors are quite similar. In spite of this, the saturation velocity decreases with increase in temperature, since under these conditions at lower electric fields, electrons acquire a sufficient amount of energy for triggering the *optical-phonon* emission process.

As pointed out on Figure 1.22, for the case of electrons in only *direct-bandgap* semiconductors, a more interesting characteristic emerges before *velocity saturation* takes over. In these semiconductors, due to the migration of the energetic[69] electron

[68] We will pay a more in-depth attention to these *scatterers*, among others, in Section 1.7.3.7.

[69] Called also *hot* electrons, in connection to the notion of *electron temperature*.

from lower *valleys* of lower *effective mass*[70] to higher *valleys* of higher *effective mass*, a region of *negative differential mobility* develops. This is because while such a transfer is taking place, electron energy is still insufficient for interaction with *optical phonons*. To maintain the value of kinetic energy as a consequence of this migration, since electrons are feeling heavier, the *drift velocity* will become less as the electric field is increasing. This region of *negative differential mobility* provides fascinating opportunities for designing microwave sources and oscillators in forms such as *Gunn diodes*. Due to the absence of multiple *valley*s in the *valence band*, *negative differential mobility* is not observed on *drift-transport* characteristics of *holes*.

The saturation velocity for the case of *Si* and *Ge*, which have a simple saturating v_d–E characteristic, is

$$v_s = \sqrt{\frac{8}{3\pi}\frac{E_p}{m_0}} \cong 10^7 \text{ cm/s}, \tag{1.89}$$

where E_p is the *optical-phonon* energy (which as indicated in Table 1.4 is 63 meV in the case of *Si*).

Oftentimes, approximate closed-form analytical models are used to explain the variation of *drift velocity* versus electric field in the regions of small, medium, and high electric fields. One such an empirical relationship used for *Si* is

$$v_d = \frac{\mu E}{\left[1 + (\mu E/v_s)^{C_2}\right]^{1/C_2}}. \tag{1.90}$$

The constant C_2 is a temperature-dependent fitting parameter, which is in the range of two for electrons and one in the case of *holes*. Evidently, (1.90) represents the *drift velocity* as an absolute value.

For the case of the *direct-bandgap* semiconductor of *GaAs*, which is the most investigated *III–V* semiconductor,[71] due to the presence of *negative differential mobility*, the v_d–E characteristic is more complicated (i.e., as suggested in Fig. 1.22). As indicated already, in order to explain the hump in this characteristic, careful knowledge of the *E–k* diagram and also *optical-phonon* energy is required. While in the lower *valley*, which is located at the center of *Brillouin zone* (i.e., Γ-*valley*), electron *mobilities* as high of 8000 cm^2/V s can be achieved at room temperature, the satellite *valley* (located along the <*111*> axes) offers an electron *mobility* in the range of 100 cm^2/V s. This satellite *valley* is located 0.3 eV above the *lower valley*. The electron *effective mass* in the lower and the satellite *valleys* are equal to 0.068m_0 and 0.55m_0, respectively.

[70] That is, the central valley of the *Brillouin zone*: Γ valley.

[71] Unlike *silicon*, which is a monoatomic semiconductor, *GaAs* is a *compoundsemiconductor* composed of atoms from groups *III* and *V* of the *periodic table* (hence the name *III–V*).

1.4.4.3 Nonsteady-State Velocity Over- and Undershoot With only one exception mentioned in the opening of Section 1.4, all of our discussions are thus far concerned with steady-state *drift velocity*. In order to arrive at steady state, *charge carriers* need to go through a sufficient number of *scattering* events. For devices of shorter length scale than the *mean free path*, however, such a steady-state condition will not be achieved. Under such circumstances, *drift velocity* is observed to reach values larger than the steady-state values. This is referred to as *velocity overshoot*. This condition is instigated by what was called earlier in this section *ballistic transport*. Under this condition, in analogy to our discussion on *low-field mobility*, we can say that velocity increases with time and as a result with distance according to $\approx (qEt/m^*)$. However, it should be emphasized that these values, which are higher than the steady-state values, are only attained momentarily, within a limited span of space and time. Besides the phenomenon of *velocity overshoot*, an *undershoot* in the nonsteady-state *drift velocity* is also possible.

Generally speaking, these nonsteady-state phenomena are not just encountered in *nanoscale* devices. Instead they are encountered when *carriers* are suddenly exposed to a large variation of electric field. For a Γ-*valley* electron suddenly exposed to a high electric field, electron energy can very well exceed the *intervalley* separation, while due to the short time of travel, an electron does not get a chance to *scatter* to a higher *valley*. As a result, *drift velocity* exceeds the value it could obtain under steady state and in the higher *valley* (i.e., of higher *effective mass*). This is the case of the velocity *overshoot*. However, if an electron originally occupying the higher *valley* (of higher *effective mass*) is suddenly introduced into a region in which electric field is suddenly reduced, electron velocity remains much lower than the steady-state value that it could obtain from that electric-field strength. This happens because the electron has not received a chance to *scatter* to the lower *valley*. This is the case of the velocity *undershoot*.

1.4.4.4 Summary of Observations on the Choice of Carrier-Transport Formalism As observed through the so far presented discussions in Section 1.4, *charge-carrier transport* is a sophisticated problem. In *carrier-transport* problem, interactions of *charge carriers* with *scattering* potential of a variety of sources (e.g., *ionized impurities*, *acoustic phonons*, *optical phonons*, *polar optical phonons*, *piezoelectric polarization*, etc.),[72] internally induced and externally applied electric fields are investigated. While in large-size electronic devices, oftentimes phenomenological description of *carrier transport* in terms of the *drift–diffusion* formalism is capable of providing sufficient insight into the problem, in smaller-size devices (and also where the electric-field variation is strong), effects such as *nonsteady-state velocity overshoot* (or *undershoot*) and *ballistic transport* are required to be considered.[73]

[72] Only some of which are named prior to this in the text. Section 1.7 will deal with a larger number of these processes at a deeper level.

[73] Later on, where appropriate, we will add more items to this list, such as *real-space transfer*.

The ongoing trend in the design of *transistors* of scaled dimensions has rendered the *drift–diffusion* formalism more and more insufficient. As an intermediate step in incorporation of some of these *carrier-transport* effects, steady-state models of *drift* characteristics have been used to incorporate effects such as *transferred electron* and *velocity saturation* (which as noticed earlier are due to take over at intermediate to high electric fields). In these models of *carrier transport*, *mobility* and *diffusion constant* are defined as functions of local electric field. Accordingly, even beyond low electric fields, the *drift velocity* in terms of the instantaneous relationship to electric field is identified using (1.61) and (1.62). On the basis of this *high-field* linear *mobility*, *Einstein relationship* is then employed to produce the high electric-field *diffusion constant*.

However, as it was mentioned earlier, reduction of the size of the *channel* along which *carriers* are to fly (compared to *mean free path*) renders even these intermediate models insufficient. This is because under these circumstances, the steady-state assumption vanishes. At the same token *carrier transit time* becomes comparable to *energy relaxation* and *momentum relaxation time constants*. Under these conditions, *carrier* distribution functions determined by the local value of electric field are no longer valid, and *carrier transport* both in time and in space is *nonlocal*. As a result, the device's current characteristics cannot be explained by the continuity equations (i.e., Equations (1.49) and (1.50)).

Now that we have presented some of the limitations of different formalisms of *carrier transport* in semiconductor devices, it is time to point out that in design and analysis of electronic devices, the first step is the selection of the appropriate *transport* formalism. This process of selection is done in terms of comparison of device sizes to those characteristic length scales identified in Section 1.4.2 and also the *carrier* flight time with the relevant timescales. In the case of *silicon* as the semiconductor medium of *carrier transport*, this is summarized in Figure 1.24. In this regard, for a rapidly changing potential on scale of the *lattice constant*, only a full quantum mechanical treatment, and not a semiclassical *effective-mass* approach, will be sufficient. In this quantum mechanical treatment, the *wave equation* with incorporation of all potentials (i.e., including *scattering*, externally applied and internally induced) should be solved.

So far in our discussion, we have invoked a few properties of *Si* and *GaAs*. Table 1.5 provides an essential summary of a few of their important *carrier-transport* properties.

1.4.5 Thermionic and Field Emission

In macroscale devices, not all components of *carrier transport* are explained through the *drift–diffusion* formalism. There are two other frequently encountered modes of current conduction in semiconductor devices: *thermionic* and *field emission*. *Thermionic* and *field emission* modes evolve in association with the presence of a potential barrier.

In the case of the *thermionic emission*, existence of a potential barrier restricts the movement of majority *charge carriers* normal to the barrier, especially if the *carrier*

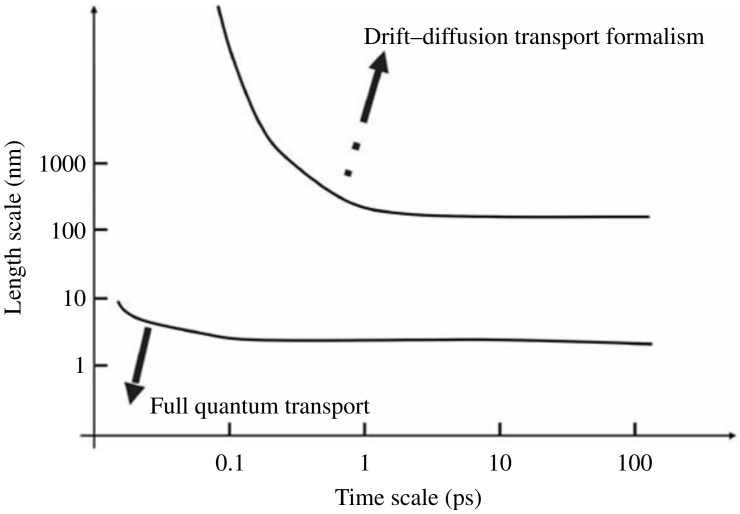


FIGURE 1.24 Approximate boundaries for the applicability of drift–diffusion carrier transport and the necessity for dealing with the full quantum mechanical description of transport in a silicon medium, as a function of time and length scale. Although, more or less such trends are applicable to other semiconductor media, the indicated values are dependent on the medium. Adapted from Lundstrom (2000, p. 347). Copyright 2000, Cambridge University Press. Reprinted with the permission of the Cambridge University Press.

TABLE 1.5 A Summary of Important Transport Properties of Si and GaAs

	Semiconductor	
Property	Si	GaAs
Longitudinal acoustic velocity (cm/s)	9.04×10^5	5.24×10^5
Transverse acoustic velocity (cm/s)	5.34×10^5	3.0×10^5
Electron acoustic deformation potential (eV)	9.5	7.01 (Γ-*valley*), 9.2 (*L-valley*), and 9.0 (*X-valley*)
Electron optical deformation potential (eV/cm)	—	3.0×10^8 (*L-valley*)
Hole acoustic deformation potential (eV)	5.0	3.5
Hole optical deformation potential (eV/cm)	6.0×10^8	6.48×10^8

temperature is low (and as a result the kinetic energy of the *charge carrier* is less than the barrier height). Nevertheless, energized (i.e., *hot*) *charge carriers* can rise over the barrier and move across it. It is due to this temperature dependence that the mechanism of *carrier transport* is referred to as *thermionic emission*. In this mode of *transport*, the critical parameter is the height of the potential barrier and not its shape. An example is provided in Figure 1.25.

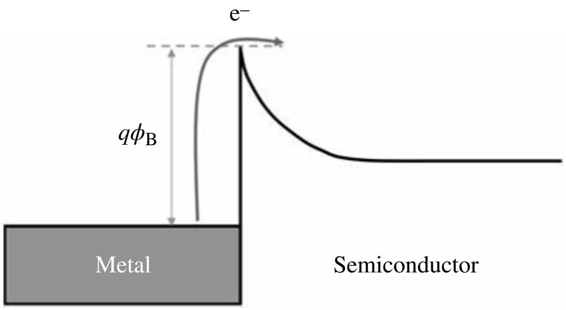


FIGURE 1.25 An example of thermionic emission across a potential barrier of height $q\phi_B$ formed between a metal and a semiconductor. We will talk about these contacts in details in Section 2.2.

Thermionic emission is observed in a variety of semiconductor devices including *Schottky barrier diodes*.[74] Ideally, for this mechanism to be the controlling mode of *transport*, the *carrier transport* in the barrier is supposed to be without *scattering*. In other words, *carriers* injected into the barrier should go through this material *ballistically* (i.e., the barrier width is to be smaller than the *mean free path*), instead of following the *drift–diffusion transport*. In addition to this, injected *carriers* should be moved out of the barrier on the opposite interface (which is often an interface with a metal), without *scattering*.

In order to quantify the *thermionic* current, we should remind ourselves of the fact that according to *Fermi–Dirac* statistics, while the number of *carriers* with energy values deeper into the *conduction* and *valence band* decreases with the separation from E_c and E_v, these *carrier* concentrations are not equal to 0. In addition, the *carrier* concentration in the *states* deeper above E_c, or below E_v, increases as the *carrier temperature* increases. As a result, the integrated number of *carriers* with energy values above the barrier height increases with this *temperature*. These *carriers*, which are not confined by the barrier, can now participate in current conduction through the *thermionic emission* process. The electron current density over a barrier of height $q\phi_B$ can be quantified as

$$J_n = A^* \cdot T^2 \cdot \exp\left(\frac{-q\phi_B}{kT}\right), \tag{1.91}$$

where A^*, which is often referred to as *effective Richardson constant*,[75] is defined as

$$A^* \equiv \frac{4\pi q m_n^* k^2}{h^3}. \tag{1.92}$$

[74] Which will be explored in Chapter 2. These *diodes* are named after *Walter Schottky*.

[75] Named after *Owen Richardson*.

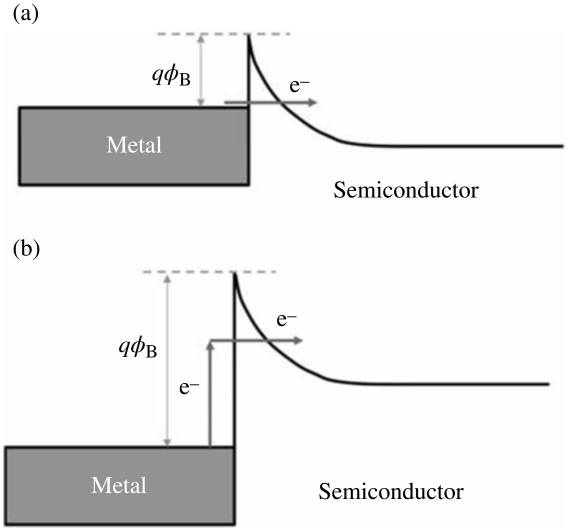


FIGURE 1.26 (a) An example of field emission across a potential barrier formed between a metal and a semiconductor. (b) An example of field-assisted thermionic emission across a potential barrier of height $q\phi_B$.

As suggested earlier, the *thermionic emission* process is not the only mode of *transport* through a potential barrier. Transport through such a barrier can be assisted, or even overwhelmed, by another mode of *carrier transport*, which is rooted in the quantum mechanical *tunneling*. These modes, which are known as *field emission* and *field-assisted thermionic emission*, are illustrated in Figure 1.26. In those cases, A^* will be modified by processes such as *reflection* and *tunneling* of the *electron wave*.

Quantum *tunneling* is rooted in the *wave* nature of electrons. According to quantum mechanics, even *charge carriers* with less energy than the barrier height exhibit a finite probability of *tunneling* through the barrier. According to the shape of the barrier, at each energy value, the effective thickness of the barrier that *carriers* need to *tunnel* through is different. Hence, unlike *thermionic emission* this mode of *transport* is quite dependent on the shape of the barrier. The smaller the effective width of the barrier at a certain energy value, the larger is the *tunneling* probability for *charge carriers* with that much energy. This is an exponential dependence.

Although the potential barriers often encountered in semiconductor devices do not have a rectangular form, for the sake of simplicity, it is worth considering the simple case of a rectangular barrier of height V_0 and of thickness W to describe the quantum *tunneling* process. This situation is illustrated in Figure 1.27. According to quantum mechanics, in this one-dimensional scenario, electrons on either side of the potential barrier are expressed in terms of *propagating waves*. However, at positions within the thickness of the barrier, the time-independent portion of the *wave function* follows a decaying exponential represented by $\exp(-|kx|)$, where *wave number* $k=\sqrt{(2m^*(V_0-E))/\hbar^2}$. In this case, the energy of the electron (i.e., E) is less than

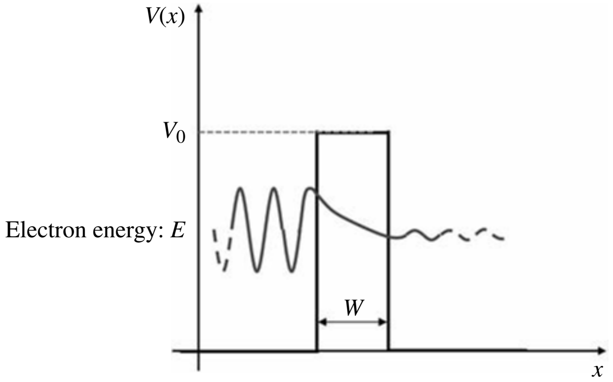


FIGURE 1.27 Quantum tunneling across a rectangular potential barrier of height V_0.

the barrier height V_0. Establishing the boundary conditions (i.e., in terms of continuity, single valuedness, and finiteness of the *wave function* and its first derivative) results in *tunneling* probability of

$$T_\mathrm{t} = \left[1 + \frac{V_0^2 \cdot \sinh(|k|W)}{4E(V_0 - E)}\right]^{-1} \approx \frac{16E(V_0 - E)}{V_0^2} \cdot \exp\left(-2\sqrt{\frac{2m^*(V_0 - E)}{\hbar^2}}W\right). \quad (1.93)$$

In the case of more complicated barrier shapes, *WKB*[76] approximation is employed as a simplification over solution of the *Schrödinger* equation. This approximation is applicable where potential *V*(*x*) is not varying rapidly with position. In terms of this approximation, according to the schematics of Figure 1.28, the *tunneling* probability can be calculated by

$$T_\mathrm{t} \approx \exp\left\{-2\int_{x_1}^{x_2} |k(x)| dx\right\} = \exp\left\{-2\int_{x_1}^{x_2} \sqrt{\frac{2m^*}{\hbar^2}[V(x) - E]} dx\right\}. \quad (1.94)$$

Knowing the *tunneling* probability one can calculate the *tunneling* current density J_t by integration versus energy over the product of three components: the *tunneling* probability, the number of *carriers* at each value of energy at the originating side of the barrier, and the number of empty *states* at that value of energy on the receiving side of the barrier:

$$J_\mathrm{t} = \frac{qm^*}{2\pi^2\hbar^3}\int F_\mathrm{E} \cdot N_\mathrm{E} \cdot T_\mathrm{t} \cdot (1 - F_\mathrm{R}) \cdot N_\mathrm{R} dE. \quad (1.95)$$

[76] Named after *Gregor Wentzel*, *Hendrik Anthony Kramers*, and *Léon Brillouin*.

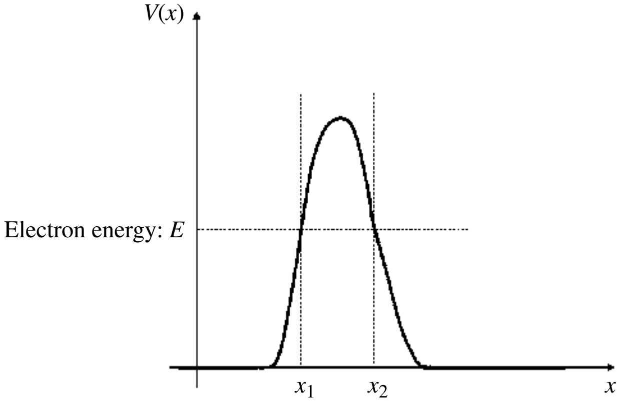


FIGURE 1.28 Quantum tunneling of an electron of energy E across an arbitrary potential barrier.

F_E, F_R, N_E, and N_R stand for the *Fermi–Dirac* distributions and *densities of states* in the emitting and receiving sides of the barrier, respectively. This framework describes the mode of *transport*, which is purely explicable in terms of quantum *tunneling* (i.e., *field emission*).

Transport through a potential barrier can also take place through a mixture of *thermionic-* and *field emission* processes. In this fashion, whereas *carriers* are *thermally* raised behind the barrier, depending on the shape of the barrier, they see an effective reduction in the barrier width and as a result an improvement in *tunneling* probability. As suggested earlier, this mode of *carrier transport* is known as *thermionic field emission* or *field-assisted thermionic emission*.

1.5 BREAKDOWN IN SEMICONDUCTORS

In discussing the high electric-field effects on electrons, the other mechanism that is of interest is *impact ionization*. *Intervalley scattering* is not the only outcome of electrons becoming *hot* (i.e., rise in T_e). As electrons become *hot*, they also become capable of breaking covalent bonds and producing *EHP*. This is shown in Figure 1.29. This happens in the form of an *avalanche*. One *hot charge carrier* produces two more *charge carriers*, and if the high electric-field region is large enough, each of these *charge carriers* will develop a chance for producing an extra *EHP*. As a result, *carrier* concentration will multiply itself. Due to this, the other name for this process of *EHP generation* is *avalanche breakdown*. The minimum amount of energy needed for a *hot carrier* to instigate this process should be larger than the *bandgap* of the semiconductor. Assuming the same *effective mass* for electrons and *holes*, this energy can be easily proven to be 50% in excess of the *bandgap* (i.e., $1.5 \times E_g$).

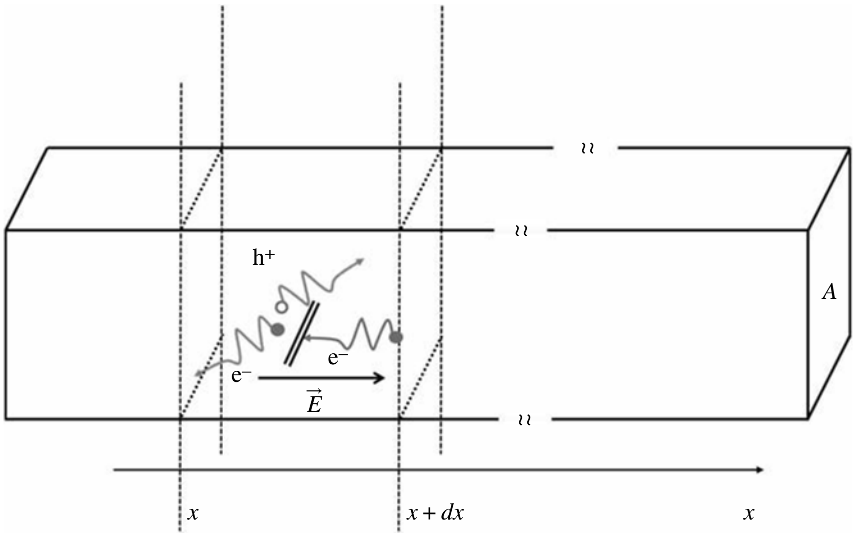


FIGURE 1.29 Generation of an electron–hole pair upon collision of a hot electron generated within the high electric-field section of a semiconductor slab with a covalent bond.

The *avalanche* multiplication process is characterized by an *ionization rate*, which is defined by the number of *EHP* generated while the electron is traversing a unit of distance. Assuming the electron velocity to be v_n, this rate can be written as

$$\alpha_n = \frac{1}{n}\frac{dn}{d(tv_n)} = \frac{1}{nv_n}\frac{dn}{dt}. \tag{1.96}$$

In terms of this *ionization rate* and a similarly defined rate for *holes*, the time variation of electron and *hole* concentrations can be written as

$$\frac{dn}{dt} = \frac{dp}{dt} = \alpha_n n v_n + \alpha_p p v_p = \frac{\alpha_n J_n}{q} + \frac{\alpha_p J_p}{q}. \tag{1.97}$$

According to the *continuity equation*, we can also have

$$\frac{dJ_n}{dx} = \alpha_n J_n + \alpha_p J_p \tag{1.98}$$

and

$$\frac{dJ_p}{dx} = -\alpha_n J_n - \alpha_p J_p. \tag{1.99}$$

Based on these, the spatial derivative of $J_n + J_p$ remains 0, resulting in

$$\frac{dJ_n}{dx} = -\frac{dJ_p}{dx}. \tag{1.100}$$

As expected, the *ionization rates* have a strong dependence on electric field. The following physical expression represents this dependence:

$$\alpha(E) = \frac{qE}{\varepsilon_I}\exp\left\{-\frac{E_I}{E[1+(E/E_P)]+E_T}\right\}. \tag{1.101}$$

In this equation ε_I is the *high-field effective ionization threshold energy*. E_T, E_P, and E_I stand for the electric-field strength needed by *carriers* to overcome the decelerating effects of *thermal*, *optical phonon*, and *ionized-impurity scattering*, respectively. In the case of *silicon*, ε_I for electrons and *holes* is equal to 3.6 and 5 eV, respectively. For a limited range of electric field, the above equation can be replaced by the following simpler equations:

$$\alpha(E) = \frac{qE}{\varepsilon_I}\exp\left(-\frac{E_I}{E}\right), \quad \text{if} \quad E_T < E < E_P \tag{1.102}$$

or

$$\alpha(E) = \frac{qE}{\varepsilon_I}\exp\left(-\frac{E_I \cdot E_P}{E^2}\right), \quad \text{if} \quad E > E_P \text{ and } E > \sqrt{E_P \cdot E_T}. \tag{1.103}$$

Figure 1.30 depicts the experimentally observed *ionization rates* for *Ge*, *Si*, *SiC*, and *GaN*.

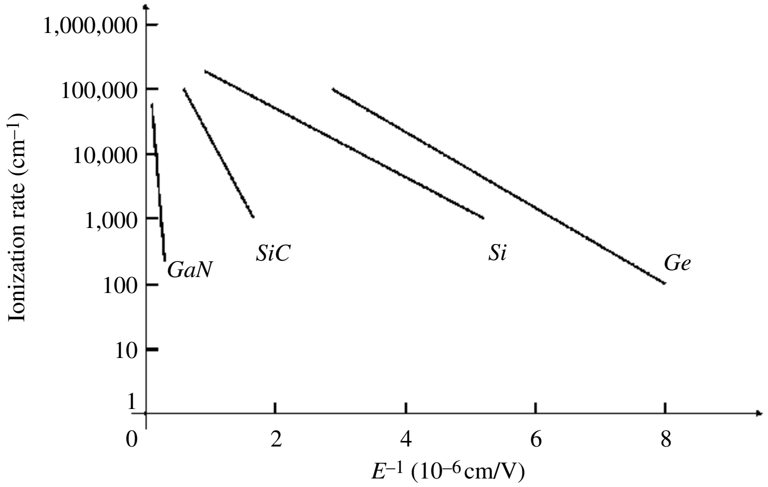


FIGURE 1.30 Approximate variation of ionization rate of electrons versus inverse of the electric field for *Ge*, *Si*, *SiC*, and *GaN* at room temperature. Adapted from Sze and Ng (2006, p. 41). Copyright 2006, John Wiley and Sons. Reprinted with the permission of the John Wiley and Sons.

In electronic devices, in connection to the process of *avalanching*, we often speak of *avalanche breakdown voltage*. This refers to the voltage difference applied to the terminals of a device that causes the induction of electric fields beyond the onset of an *avalanche* process. Production of large concentrations of *charge carriers* induces heating and either permanent structural damage to the electronic device or temporary malfunction.

As expected from (1.101), for larger *bandgap* materials the *ionization rates* are smaller. As a result, wide *bandgap* semiconductors such as *SiC*, *GaN*, *AlN*, and *ZnO* are deemed suitable for high electric-field/high-voltage applications. In addition to *bandgap*, *optical-phonon* energy is also an important factor in determining the *breakdown voltage*, which as suggested earlier is the minimum voltage needed to create a rush of current. The higher the *optical-phonon energy*, the higher the chance of *charge carriers* becoming *hot*, and as a result, the smaller would be the applied voltage causing *breakdown*. *GaN*, which is a wide *bandgap* material, possesses a large value of *optical-phonon energy*. These two factors promote adversary agenda with regard to the *breakdown voltage*.

With respect to *ionization rates*, it is also important to mention that at a constant electric-field *ionization energy* is a declining function of temperature. This is due to the fact that at high temperatures, *phonon scattering* deteriorates the chance of *carriers* becoming *hot*.

The other mechanism instigating *breakdown* at high electric fields is *Zener*[77] *breakdown*. In this mechanism, the high electric field directly breaks covalent bonds and produces a large number of *EHP*s needed to create a rush in current.

The *impact ionization* process, unlike the *Zener breakdown* mechanism, is an avalanching process, which is why it requires *carriers* to travel a distance multiple times longer than the *mean free path*. As a result of this requirement, *avalanche breakdown* can only happen in devices that have a relatively long high electric-field region.

1.6 CRYSTALLINITY AND SEMICONDUCTOR MATERIALS

At the beginning of this chapter, it was pointed out that while from the point of view of crystallinity semiconductors can be found in *amorphous*, *polycrystalline*, or *monocrystalline* forms, *monocrystalline* semiconductors are often preferred in electronic applications. An interesting question to address in dealing with semiconductor devices, which is at the root of semiconductor electronics, is: why should we prefer *monocrystalline* semiconductors over *polycrystalline* or *amorphous*?

The reason for this choice is the short wavelength of electrons. As a result of this short wavelength, *charge-carrier transport* is critically dependent on the atomic arrangement in the *solid*. The higher the crystalline order, the better the *transport* properties. As invoked earlier, according to *Bloch* theorem, for a perfect *monocrystalline*

[77] Named after *Clarence Zener*.

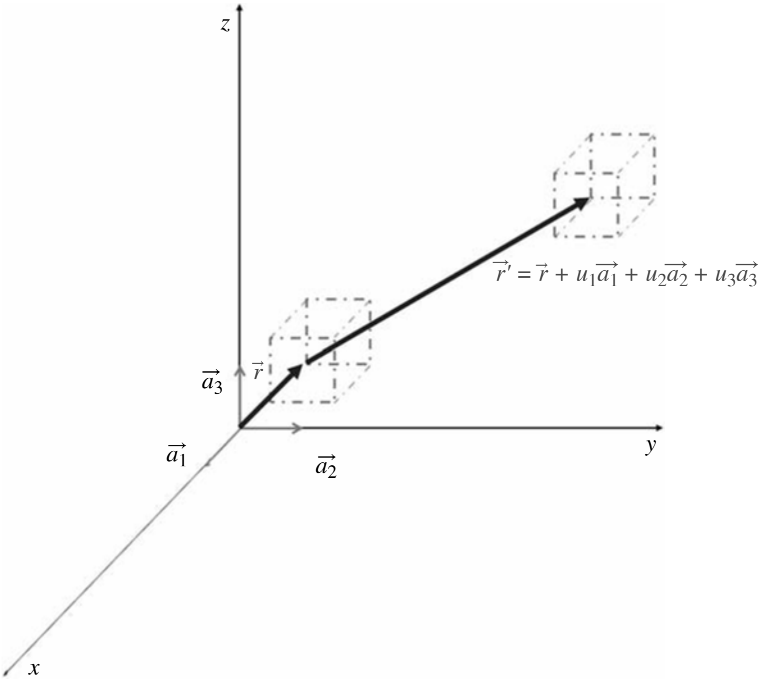
solid electrons travel as *propagating waves*, and the atoms in the crystal essentially pose no cause for *scattering* of the *electron wave*. However, in contrast to electrons, *photons* have a longer wavelength and are not impacted by the mere *short-range* order of the *host* material. As a result, noncrystalline materials such as glasses are suitable for optical applications while not so for electronic applications.

An important line of research in the area of semiconductor electronics focuses on crystalline growth of semiconductor materials. Interaction of these crystals with *X-rays* in the form of *diffraction* is one of the capable tools for studying the degree of crystallinity of a *solid*. *X-rays* are suitable for this purpose because their wavelengths are very close to the *lattice constants* of the semiconductor crystals.

The structure of an ideal crystal can be explained in terms of *copying* and *pasting* of a building block or a so-called unit cell according to a certain *translation vector*,

$$\vec{r'} = \vec{r} + u_1\,\vec{a_1} + u_2\,\vec{a_2} + u_3\,\vec{a_3}, \tag{1.104}$$

where u_1, u_2, and u_3 are the three arbitrary integers, $\vec{r}$ is the position of an arbitrary point in the building block, and $\vec{a_1}$, $\vec{a_2}$, and $\vec{a_3}$ are referred to as the *translation vectors*. This is schematically presented in Figure 1.31.



z
$\vec{r'} = \vec{r} + u_1\vec{a_1} + u_2\vec{a_2} + u_3\vec{a_3}$
$\vec{a_3}$
$\vec{r}$
$\vec{a_1}$
$\vec{a_2}$
y
x

FIGURE 1.31 Translation of a cubic building block of a lattice in space according to the translational vector $u_1\,\vec{a_1} + u_2\,\vec{a_2} + u_3\,\vec{a_3}$.

1.6.1 Bravais Lattices

In the representation of (1.104), considering all possible integer values taken by u_1, u_2, and u_3, an infinite three-dimensional space has to be completely filled (i.e., with no voids). This requirement imposes certain restrictions on the *degree of symmetry* of the *unit cells* of a crystal. For example, while one, two, three, four, or sixfold rotational symmetry is allowed, five and seven are inadmissible (see Fig. 1.32). The degrees of symmetry for a given semiconducting crystal have considerable importance in its *carrier-transport* properties.

These building blocks are often explained in terms of a *lattice* and a *basis*. While *lattice* is the abstract arrangement of points in space representing a block, *basis* refers to the atoms assigned to these points. An infinite number of building blocks can be proposed for constructing a crystal. This is shown in Figure 1.33. The smallest of all these building blocks (or *unit cells*) is referred to as a *primitive cell*. While the number of atoms assigned to a *primitive cell* and its volume (which is identified by the translation vectors as $\vec{a_1} \cdot \vec{a_2} \times \vec{a_3}$) are unique, the vectors representing the *primitive cell* are by no means unique. An easy way to envision a *primitive cell* is through following the *Wigner–Seitz*[78] procedure.

In the *Wigner–Seitz* procedure, the boundaries of the *primitive cell* are determined by the intersection of imagined planes normal to the lines connecting a given *lattice* point to all its neighboring points. In the case of a two-dimensional *lattice*, this is illustrated in Figure 1.34.

Generally speaking, different *lattice types*, which are explained in terms of *lattice* and *basis*, are referred to as *Bravais*[79] *lattices*. In the case of three-dimensional

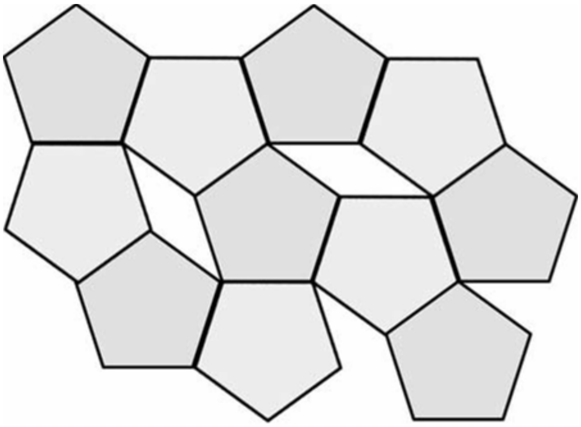


FIGURE 1.32 Schematic demonstration of inability of fivefold symmetry in filling of an infinite space.

[78] The *primitive cells* produced according to this procedure, after *Eugene Wigner* and *Frederick Seitz*, are referred to as *Wigner–Seitz cells*.

[79] Named after *Auguste Bravais*.

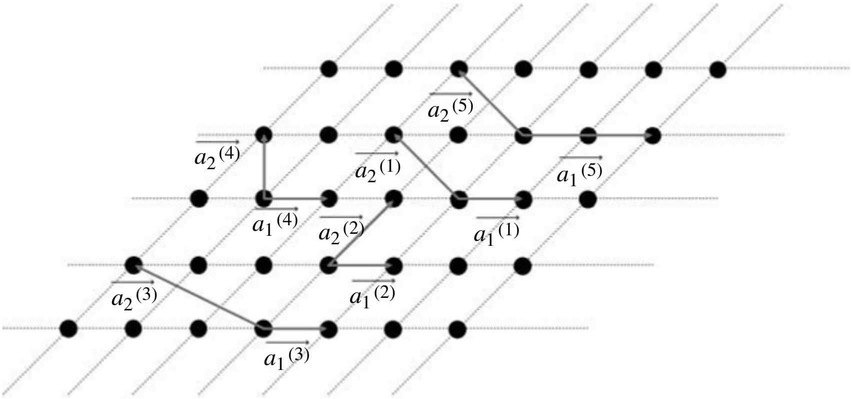

$\overrightarrow{a_2}^{(5)}$
$\overrightarrow{a_2}^{(4)}$
$\overrightarrow{a_2}^{(1)}$
$\overrightarrow{a_1}^{(5)}$
$\overrightarrow{a_1}^{(4)}$
$\overrightarrow{a_2}^{(2)}$
$\overrightarrow{a_1}^{(1)}$
$\overrightarrow{a_2}^{(3)}$
$\overrightarrow{a_1}^{(2)}$
$\overrightarrow{a_1}^{(3)}$

FIGURE 1.33 A schematic depiction of lattice and basis on a two-dimensional *Bravais lattice*. Atoms are represented by the dark circles. While as examples the parallelograms represented by a pair of vector $\vec{a_1}^n$ and $\vec{a_2}^n$ for $n = 1–4$ represent various primitive cells for this 2-D crystal, the parallelogram represented by $\vec{a_1}^5$ and $\vec{a_2}^5$ owing to its twice as large the area does not represent a primitive cell.

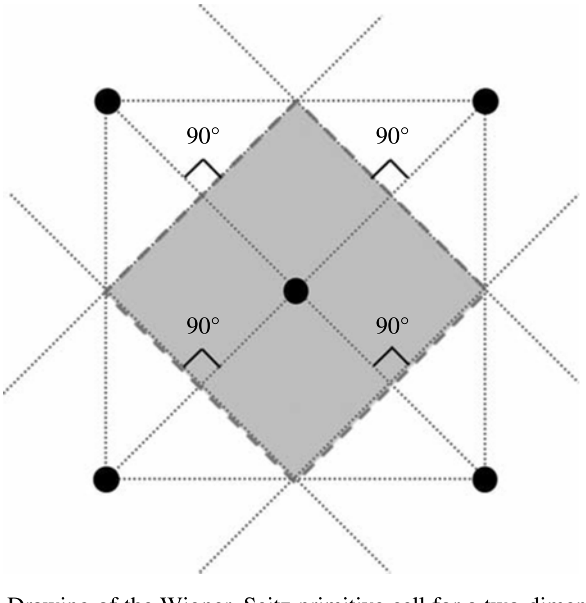


FIGURE 1.34 Drawing of the Wigner–Seitz primitive cell for a two-dimensional lattice. In this case, instead of envisioning a volume defined by the intersection of normal planes passing through the middle of the lines connecting neighboring atoms of the crystal, a primitive cell is formed by the intersection of normal lines passing through these points on the plane of the 2-D crystal. This plane is marked in gray, where atoms are presented by the dark circles.

TABLE 1.6 The Fourteen 3-D Lattice Types

Crystalline System	Number of Admissible Lattices	Relationships between the Sizes of the Axes of the Cell (a_1, a_2, a_3) and the Angles Defined Sequentially between These Axes (α, β, and γ)
Triclinic	1	$a_1 \neq a_2 \neq a_3$ $\alpha \neq \beta \neq \gamma$
Monoclinic	2	$a_1 \neq a_2 \neq a_3$ $\alpha = \gamma = 90° \neq \beta$
Orthorhombic	4	$a_1 \neq a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$
Tetragonal	2	$a_1 = a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$
Cubic	3	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma = 90°$
Trigonal	1	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma < 120°, \neq 90°$
Hexagonal	1	$a_1 = a_2 \neq a_3$ $\alpha = \beta = 90°$ $\gamma = 120°$

lattices, the number of *lattice types* is restricted to 14. This is due to the limitations posed by the *point symmetry groups*. These *lattice types* are indicated in Table 1.6. Out of these 14 *lattice types*, only *two* (which belong to *cubic* and *hexagonal* systems) are of interest in solid-state electronics and optoelectronics.

1.6.1.1 Hexagonal Crystals In the *hexagonal* system, the *unit cell* is in the form of a right prism. Figure 1.35a illustrates this right hexagonal prism through indication of 14 atoms, 7 arranged on each of the two basal hexagons. The separation between the two basal hexagons is referred to as c (which clearly denotes the height of the hexagonal prism). In order to clearly imagine this structure, envision *seven* identical atoms on a plane, *six* of which are centered at the *six* corners of a hexagon and *one* sitting right in the middle. If we imagine the atoms as hard spheres, in the so-called *hexagonal close-packed* (or *hcp*) structure, these spheres are supposed to tangentially touch each other. Now looking from above normal to this plane, we see cavities among the spheres. Obviously in the right prism, from this angle the *seven* atoms of the top basal hexagon will also have their centers coincide with those of the bottom basal hexagon.

The unit cells of many important semiconductors are formed through interpenetration of two of these hexagonal prisms according to a certain *translation vector*. As a result of this interpenetration, a second layer of atoms will also appear between the two basal hexagons of the first prism. As suggested in Figure 1.35b, the centers of these atoms coincide with the identified set of cavities. In terms of the vectors identified in Figure 1.35a, the second hexagonal prism is displaced by $2\vec{a_1}/3 + \vec{a_2}/3 + \vec{a_3}/2$. As shown in this figure, $\vec{a_3}$ has a magnitude equal to c and is identified normal to the basal hexagon. $\vec{a_1}$ and $\vec{a_2}$ defined in the basal plane have an

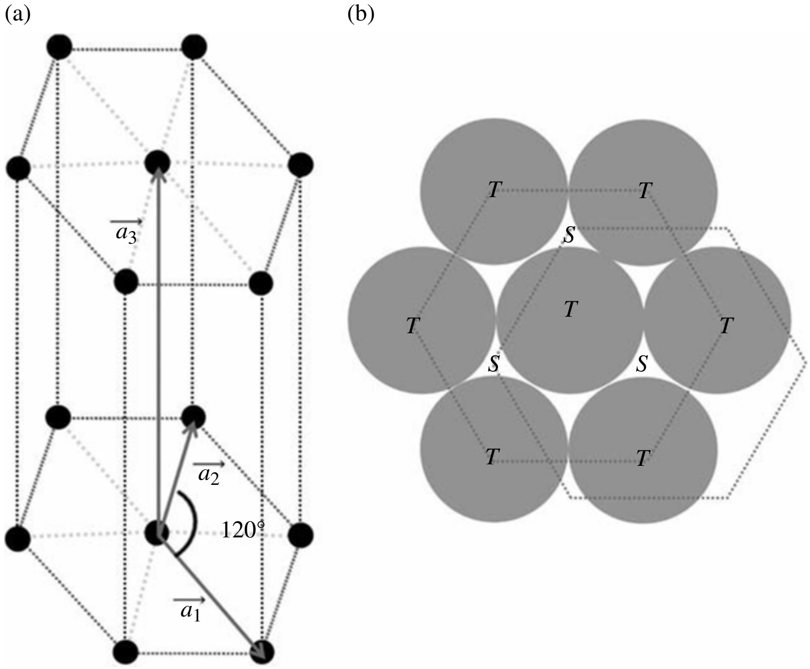


FIGURE 1.35 (a) A right hexagonal prism with the indication of placement of atoms as dark circles. (b) Position of the center of the atoms on the first, second, and third layer of a wurtzite crystal. While the gray circles represent the atoms of the first layer, letters *S* and *T* stand for the position of the center of the atoms on the second and third layer from a top perspective, respectively.

equal magnitude, which we denote as *a*, defined at 120° from one another. This crystal structure is known as *wurtzite*.[80] In order to guarantee *maximum sphere packing*, *c*/*a* should be equal to $\sqrt{8/3}$. This condition identifies an ideal *hcp* crystal.

1.6.1.2 Cubic Crystals The *cubic* system can be in the forms of *simple cubic*, *body-centered cubic* (or *BCC*), and *face-centered cubic* (or *FCC*). Shared between these three *lattice types* is a cube with edge length *a* (which is known as the *lattice constant*). While in a *unit cell* of *simple cubic* crystals only eight atoms are present (each centered at one of the eight corners of the cube), in *BCC* an extra atom is added to the middle of the cube. In *FCC*, six extra atoms are centered in the middle of the six faces of the cube. The atomic arrangements of these *cubic lattice types* are illustrated in Figure 1.36. A number of important characteristics of *cubic lattices* are summarized in Table 1.7.

[80] Named after *Charles-Adolphe Wurtz*.

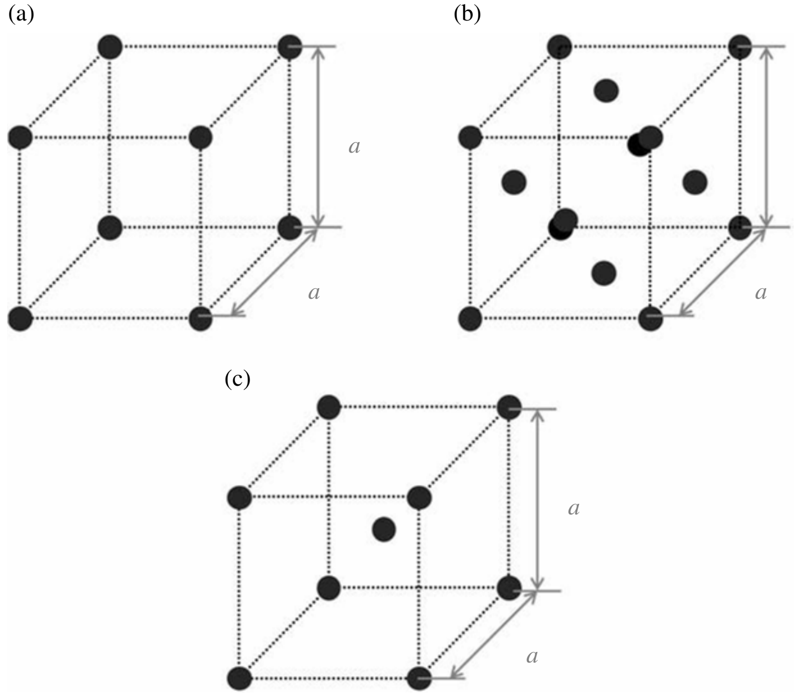


FIGURE 1.36 Atomic arrangements in the unit cells of (a) a simple cubic, (b) a face-centered cubic, and (c) a body-centered cubic crystal. Atoms are marked by full circles.

TABLE 1.7 Summary of the Characteristics of Cubic Lattice Types

	Lattice Type		
Characteristic	Simple Cubic	BCC	FCC
Volume of the conventional cell in terms of the lattice constant a	a^3	a^3	a^3
Volume of the primitive cell	a^3	$0.5a^3$	$0.25a^3$
Number of nearest neighbors	6	8	12
Nearest neighbor distance	a	$0.866a$	$0.707a$
Maximum packing ratio	0.524	0.680	0.740

Among the *cubic* crystals, *FCC Bravais lattice* is the *lattice type* that explains the largest number of semiconductors of interest in optoelectronics and electronics (e.g., *Si*, *Ge*, *GaAs*, *AlAs*, *InP*, etc.). However, as suggested already, not all semiconductors of interest are explained by the *FCC Bravais lattices*. Many semiconductors, and also metals, crystallize in form of *hcp* structures (e.g., *BN*, *AlN*, *GaN*, *SiC*, etc.).[81]

[81] Some semiconductors such as *GaN* can be grown in both *cubic* and *hexagonal* forms.

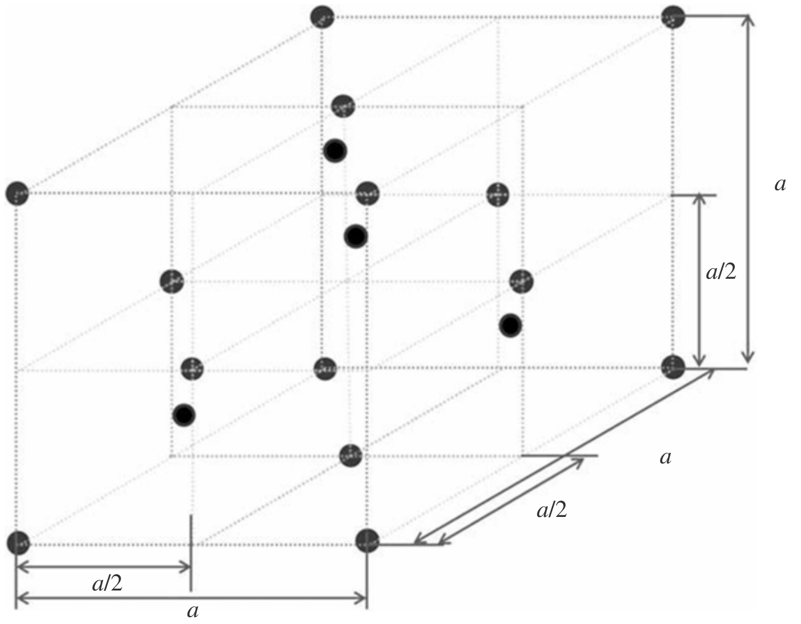


FIGURE 1.37 Placement of atoms on a diamond crystal unit cell. The gray full circles are sitting at the *FCC* positions. The other four atoms indicated by the darker circles are placed in the middle of the four inner cubes of edge length *a*/2.

Silicon, which is the most popular semiconductor, has a *lattice type* that results from interpenetration of *two* identical *FCC lattices*. Each of these *FCC lattices* is referred to as a *sublattice*. The *translation vector* between the two *FCC sublattices*, in terms of the *lattice constant a*, is (*a*/4, *a*/4, *a*/4). The crystal structure of *silicon*, which is constructed in this fashion, is referred to as *diamond* (since this is also the crystal structure of diamond). This is illustrated in Figure 1.37. *Germanium* also crystallizes in this form, although with a different *lattice constant*.

Whereas other important semiconductors such as *GaAs* and *InP* are also formed through interpenetration of *two FCC sublattices*, in their cases the *two sublattices* are not identical. In these cases, one *sublattice* has metallic atoms from group *III* of the *periodic table* (e.g., *Ga* and *In*) as its *bases*, while the other one has group *V* atoms (hence the name *III–V* semiconductors). These crystals are referred to as *zinc blende*.[82]

In contrast to *silicon* and *germanium*, which are known as *elemental* semiconductors, those semiconductors in which more than one atom is present in their *undoped* structures are referred to *compound* semiconductors. There are *compound* semiconductors that are *cubic* and those that are *hexagonal*. There are those such as *GaAs* that are *III–V* and those such as *ZnO* that are *II–VI*. There are those such as *InP* that are

[82] Named after the mineral *zinc blende* (*sphalerite*).

binaries and those such as Ga_xIn_{1-x}P (i.e., x part *Ga*, $1 - x$ part *In*) that are *ternaries*.[83]

Example

Calculate the volume density of the atoms in a diamond crystal of *lattice constant a*.

Since according to Figure 1.37 in a *diamond* crystal out of the 18 atoms associated with the *cubic unit cell* of volume a^3 only *four* completely reside within this volume while eight atoms share only 1/8th of their volume with the cube and the other six share half of their volume, the volume density is expressed as $\frac{(4+(8/8)+(6/2))=8}{a^3}$.

In this evaluation we have taken the atoms as hard spheres.

1.6.1.3 Miller Indexing System Due to differences in atomic arrangement along different directions of a *lattice*, *carrier transport* will be dependent on the alignment of the designed *channel* of the device with the crystalline axes. For that reason, it is important to use an indexing system to distinguish between different directions in a crystal. This system is known as the *Miller index*.[84] This indexing system is used to denote different planes and directions in a crystal.

To identify the *Miller indices* for a plane in a *cubic* crystal, the following procedure is used:

- Choose a *Cartesian* coordinate.
- Find the intersection of the given plane with the three axes of the coordinate.
- Construct a vector composed of the inverses of the intersections with *x*-, *y*-, and *z*-axis.
- Multiply this vector by the smallest common denominator.
- The resulting vector, which is often referred to as (*hkl*), refers to the aforementioned plane.

In a *cubic* crystal direction normal to the plane (*hkl*) is denoted by [*hkl*]. This is not generally extendable to all crystalline types. This procedure is shown schematically in Figure 1.38a.

In this procedure, the choice of the *Cartesian* coordinate was an arbitrary one. Due to this and the existing degrees of *symmetry* among different crystals, there are a number of equivalent planes and directions in a crystal. The existence of these equivalencies is a source for *degeneracy* in crystals, with which we have already made an acquaintance in Section 1.2.1. As a result of this *degeneracy*, {*hkl*} refers to a group of identical planes, and <*hkl*> denotes a group of identical directions.

Since many semiconductors crystallize in the hexagonal form, it is also of interest to indicate how the planes and directions in the *hexagonal unit cells* are indexed. The adopted indexing system in these cases is represented in terms of four digit indices known as *Miller–Bravais* indices. In dealing with *hexagonal* crystals, instead of orthogonal

[83] We can equally well have more than three atoms in a *compound*.

[84] Named after *William Miller*.

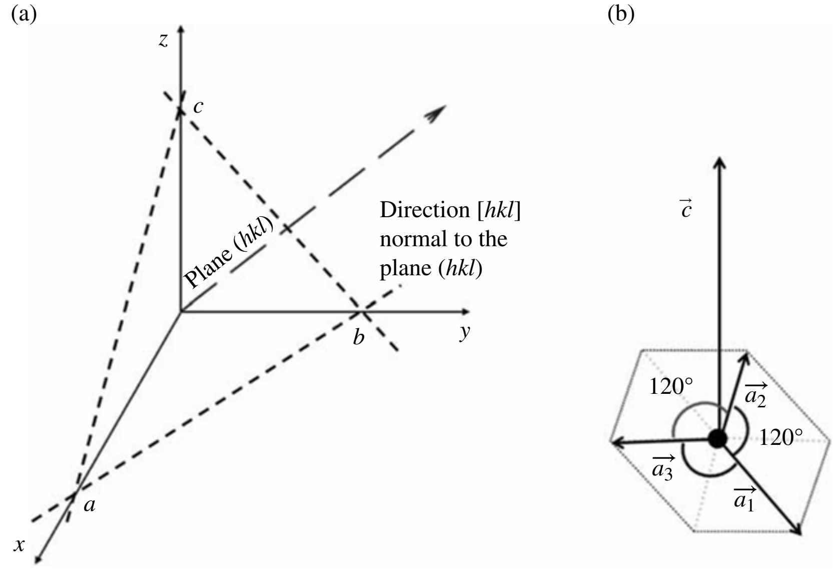


FIGURE 1.38 (a) Definition of *Miller indices* in a cubic system. Vector (*h,k,l*) is created by multiplying the vector (1/*a*, 1/*b*, 1/*c*) by the smallest common denominator between the three elements of the vector. (b) Coordinate system used in identifying *Miller–Bravais* indices in a hexagonal system.

coordinates, three nonorthogonal *basis vectors* (i.e., $\vec{a_1}$, $\vec{a_2}$, and $\vec{a_3}$) in the basal plane of the *hexagonal unit cell* and one *height vector* (i.e., $\vec{c}$) along the height of this hexagon are chosen. Directions of these vectors are identified in Figure 1.38b. In terms of the previously mentioned procedure identified for evaluation of *Miller indices* in cubic crystals, *Miller–Bravais* indices of *hexagonal* crystals are identified by finding out the intersection of a plane with these four vectors. Following that procedure, designations for a plane and the direction perpendicular to that plane are identified by (*hkil*) and [*hkil*], respectively. In evaluating these, care should be taken that the amplitude of unit vector $\vec{c}$ is different from that of the three other vectors. As an example, the plane parallel to the base of the unit cell has *Miller–Bravais* indices of (0001), while a face plane on the side that intersects $\vec{a_1}$, $\vec{a_2}$, $\vec{a_3}$, and $\vec{c}$, at 1, ∞, −1, and ∞ (which are scaled with the amplitudes of these vectors), is called $(10\bar{1}0)$. Orientations identified in Figure 1.38b for $\vec{a_1}$, $\vec{a_2}$, and $\vec{a_3}$ indicate that $h + k + i$ is always equal to 0.

1.6.2 Strain and Techniques of Epitaxy

Originally, *epitaxial* crystal growth techniques such as *MOCVD*[85] and *MBE*[86] were developed for growing planar crystals on a substrate. More recently, there has been

[85] Which, as identified earlier, stands for *metal organic chemical vapor deposition*.

[86] Also indicated earlier to stand for *molecular beam epitaxy*.

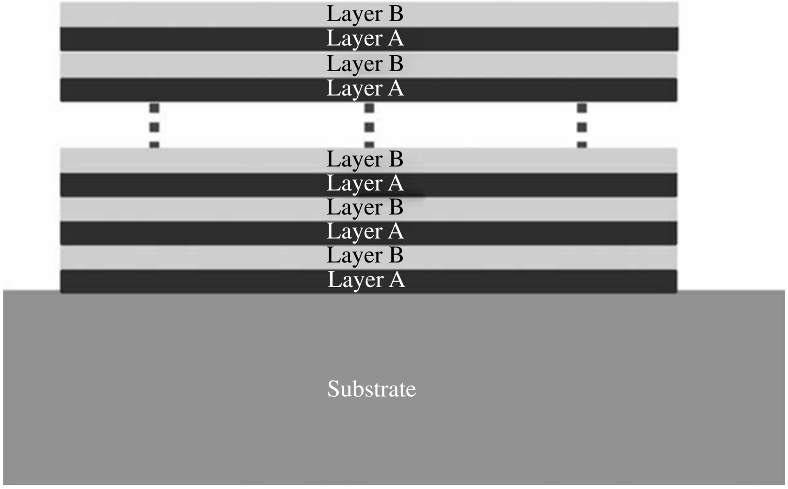


FIGURE 1.39 A superlattice is formed by growing a stack of a large number of alternating thin layers of at least two different semiconductors on a substrate.

an increasing interest in selective *overgrowth* (i.e., growth over the windows opened through a *masking* material). For local *overgrowth* applications, an advantage of *epitaxial* crystal growth techniques relying on chemical reactions (e.g., *vapor deposition epitaxial* process of *MOCVD*) over *MBE* is that they can take advantage of lateral temperature control for realizing local area growth. Based on this principle, for some materials laser-assisted local area growth has been proven possible.

Over the past three decades, techniques of *epitaxy*, either through realizing local area or planar crystal growth, have contributed significantly to realization of new opportunities in the *FET* technologies.

1.6.2.1 Ordered and Random Mixing of Semiconductors Whereas fundamentals such as free-energy minimization dictate the natural choice of the *growth mode* and *lattice constant* of crystals, the advances made with crystal growth techniques have made it possible to grow artificial crystals with structures other than those that can be found based on natural laws. Such a possibility is the outcome of the availability of techniques that enable atomic placement with exact precision.

With this degree of precision, the so-called *superlattice* structures have been grown. Realization of *superlattices* takes place in the form of *epitaxial* growth of alternating layers of different crystalline materials. While in each of these layers the crystalline structure is defined according to the laws of nature, control over the *pitch* and the *period* of these alternating layers helps in realizing an arbitrary degree of periodicity along the direction normal to the interface. This is schematically presented in Figure 1.39.

As the periodicity of the crystal impacts the *carrier-transport* properties, realization of *superlattices* provides a new dimension in engineering the *carrier transport*. These structures have been used in a variety of applications such as realization of *drain* and *source Ohmic* contacts in a few *FET* technologies such as *AlGaN/GaN HFET* technology[87] or production of high-frequency *signals*.

One of the important contributions of crystal growth techniques is that they provide the possibility of randomly mixing semiconductors to create *semiconductor alloys*. Whereas ordered structuring of a number of semiconductors can be of interest in forms such as *superlattices* (i.e., in realizing artificial crystals), random alloying of semiconductors provides the opportunity for tuning parameters such as *lattice constant*, *bandgap*, and *effective carrier mass* between the elements of the *alloyed* semiconductor. The *lattice constant* of such a *semiconductor alloy* is often provided in terms of a linear weighted average between those of the parent semiconductors. This approximation is referred to as *Vegard's law*. We will deal with this law in further detail shortly.

1.6.2.2 Coherent and Incoherent Growth of Heterojunctions In *epitaxy*, crystal growth is performed over a crystalline template. An *epitaxial* junction formed between two different semiconductors is referred to as a *heterojunction*. This is in contrast to *homojunctions*, which are formed between two pieces of one semiconductor that are, for example, merely *doped* differently. The terms *heterostructure*, *heterointerface*, and *heteroepitaxy* are also used in this context. Depending on the matching of the *lateral lattice constant* of the template (which is often the substrate) and the *free-standing*[88] *lateral lattice constant* of the *overgrowing* crystal, one can either have a *lattice matched* or a *lattice mismatched* mode of crystal growth. Figure 1.40 offers a schematic representation of these two modes of *epitaxial* growth. *Lattice-matched* (also known as *coherent*) growth, which is the result of the similarity of the two *lattice constants*, produces the highest-quality *heterojunction* between the two crystals. On the other hand, the presence of *lattice mismatch* results in formation of a large number of *dangling bonds* at the *heterointerface* and also crystal faults in the *overgrown* layer.

While only a few of the known semiconductors are *lattice matched* to one another, there is a strong demand for realization of *heterojunctions* between a variety of *lattice-mismatched* semiconductors. In order to avoid the unwanted effects of *lattice-mismatched* growth, techniques of *strained epitaxy* have been developed in a large number of semiconductor families. Through engineering the *lateral lattice constant* of the overgrowing crystals to match that of the template (i.e., often substrate), these techniques realize the so-called *pseudomorphic* mode of *epitaxy*.

1.6.2.2.1 Strain Calculation Depending on the sign of the algebraic difference between the *lateral lattice constant* of the substrate and that of the *freestanding* overgrown material, a built-in strain in either *tensile* or *compressive* form is induced at the *pseudomorphically* grown *heterointerface*. The presence of this strain, if *tensile*, limits the maximum thickness of the overgrown material. This is because the increasing of the thickness of the overlayer builds up strain energy, eventually surpassing the bonding energy of this film. Consequently, if not instantaneously, overtime cracks will appear in the film. In contrast, *compressive* strain poses no limit of this kind

[87] To be dealt with in detail in Chapter 5.

[88] That is, when grown as bulk and not in contact with another crystal.

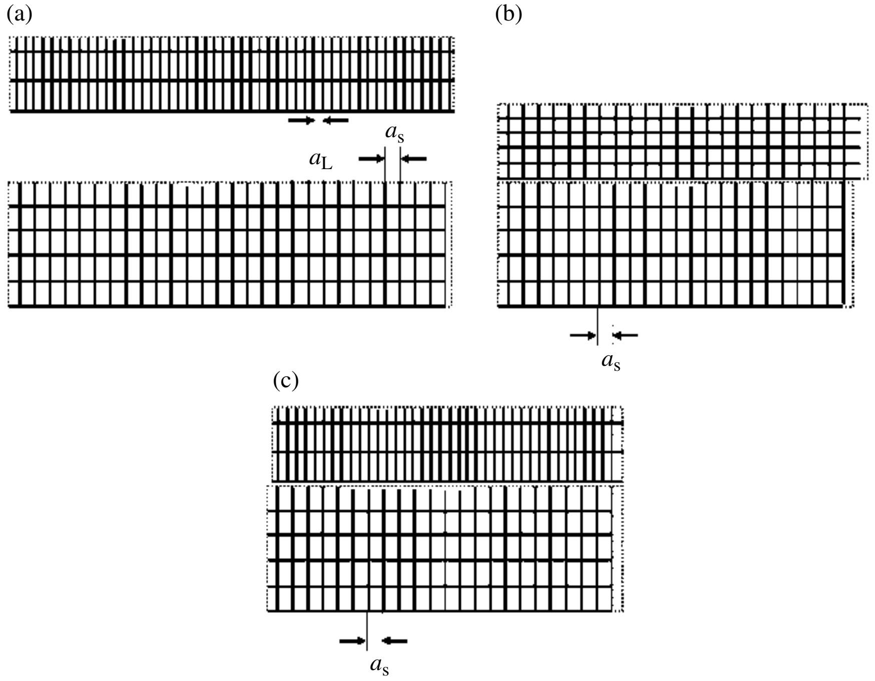


FIGURE 1.40 (a) Two lattice-mismatched 2-D lattices. (b) Strained epitaxial growth. Under this mode of tensile-strained epitaxy, the lateral lattice constant of the overlayer is expanded to match that of the substrate (i.e., a_S). (c) Lattice-mismatched epitaxial growth.

on film thickness. A film grown under *compressive* strain, however, eventually relaxes to its own *freestanding lattice constant* as its thickness grows beyond a critical value. Figure 1.41 provides schematics for these growth modes.

One of the recent developments in the area of semiconductor devices has taken shape in the form of engineering this *heteroepitaxial* strain for realizing different levels of charge concentration in the vicinity of the *pseudomorphically* grown *heterojunctions*.[89] This property has been used especially in polar *III-Nitride* technology, which enjoys large *piezoelectric coefficients* in all its *binaries* (i.e., *AlN*, *GaN*, and *InN*) and their *alloys*.

Another recent development in this area is the realization of *compliant* substrates. In these substrates, it is not just the overgrown material whose *lattice constant* is being modified. Researches on development of *compliant* substrates, which contribute to decoupling the strain induced at the *wafer* substrate interface from the *heterojunctions*, and also development of substrates for *heterostructures* such as *AlGaN/GaN*, which traditionally lack a viable *freestanding lattice-matched* substrate, are some of the major activities in this area. Such modes of *epitaxy* are also seen as major hopes

[89] We will provide an in-depth analysis of this situation in Chapter 5.

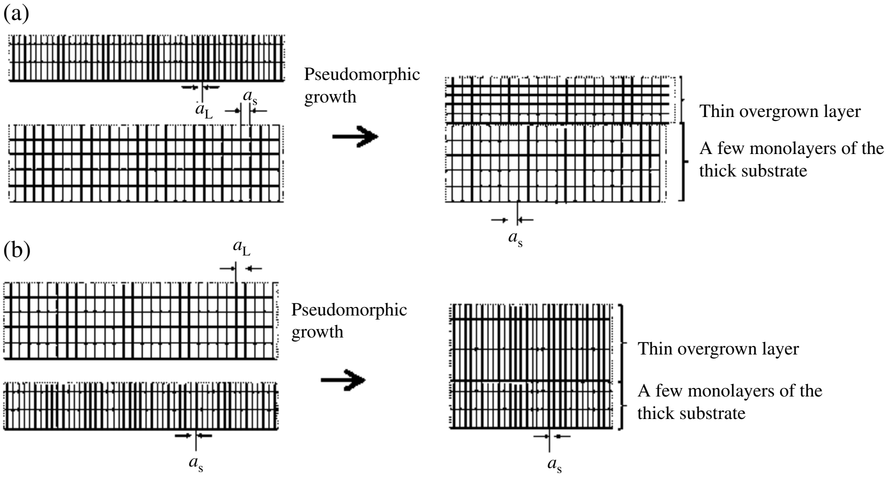


FIGURE 1.41 Two-dimensional schematic depictions of pseudomorphic growth under (a) tensile and (b) compressive strained epitaxy.

for integration of devices and circuits of different semiconductor technologies on the same *chip*.

In order to explain the outcomes of *lattice-mismatched epitaxy*, it is worth acquainting ourselves with its language of the *strain tensor*.[90] In a planar *epitaxial* growth, the overgrown layer is *biaxially strained* in the plane of substrate (indicated by $\epsilon_{\|}$) while *uniaxially strained* in a direction normal to the substrate (indicated by $\epsilon_{\perp}$). For a thick *noncompliant* substrate, the amount of strain in terms of *lateral lattice constant* of the substrate (i.e., a_S) and that of the *freestanding* overgrowing crystal (i.e., a_L) is calculated as

$$\epsilon_{\|} = \frac{a_S}{a_L} - 1 = \epsilon. \tag{1.105}$$

Since there is only *in-plane* stress, the amount of perpendicular strain $\epsilon_{\perp}$ will be calculated in terms of *in-plane* strain and *Poisson's ratio*[91] (i.e., σ),

$$\epsilon_{\perp} = \frac{-\epsilon_{\|}}{\sigma}. \tag{1.106}$$

According to this situation of *zero* stress along the growth direction, for a *strained growth* over a (001) substrate of *FCC* type, it can be shown that

[90] Within a more general context, the notion of *strain tensor* is presented in Appendix 1.C.

[91] Named after *Siméon Poisson*.

$$
\begin{aligned}
\sigma &= \frac{c_{11}}{2c_{12}} \\
\epsilon_{xx} &= \epsilon_{\|} \\
\epsilon_{yy} &= \epsilon_{\|} \\
\epsilon_{zz} &= -\frac{2c_{12}}{c_{11}}\epsilon_{\|} \,. \\
\epsilon_{xy} &= 0 \\
\epsilon_{yz} &= 0 \\
\epsilon_{zx} &= 0
\end{aligned}
\tag{1.107}
$$

c_{ij}'s present in (1.107) and the rest of this discussion are the *elastic constants*.

Such relationships are obviously developed on the basis of the knowledge of the plane over which growth is taking place. As another example in case of growth over a (*111*) *FCC* substrate, we have

$$
\begin{aligned}
\sigma &= \frac{c_{11}+2c_{12}+4c_{44}}{2c_{11}+4c_{12}-4c_{44}} \\
\epsilon_{xx} &= \left[\frac{2}{3}-\frac{1}{3}\left(\frac{2c_{11}+4c_{12}-4c_{44}}{c_{11}+2c_{12}+4c_{44}}\right)\right]\epsilon_{\|} \\
\epsilon_{yy} &= \epsilon_{xx} \\
\epsilon_{zz} &= \epsilon_{xx} \\
\epsilon_{xy} &= \left[-\frac{1}{3}-\frac{1}{3}\left(\frac{2c_{11}+4c_{12}-4c_{44}}{c_{11}+2c_{12}+4c_{44}}\right)\right]\epsilon_{\|} \\
\epsilon_{yz} &= \epsilon_{xy} \\
\epsilon_{zx} &= \epsilon_{yz}.
\end{aligned}
\tag{1.108}
$$

It can be observed from (1.107) and (1.108) that while in (*001*) growth the *strain tensor* is diagonal, in (*111*) growth (among a few other directions) the *strain tensor* has nondiagonal terms. The distortion caused by *strained epitaxy* to the *cubic* lattice, depending on the *growth orientation*, can produce a reduced degree of *crystal symmetry*. The nondiagonal terms introduced into the *strain tensor* by the reduced *crystal symmetry* are often used in production of built-in *polarization* fields in the *heterostructure*. Since we have already made a connection between the *degree of symmetry* and *transport*-related properties such as the presence of *degeneracies* at the *band* edges, we can expect the *strained epitaxy* to be capable of extending opportunities for engineering these properties. We will deal with these aspects in further details later in Chapter 4.

As we have already mentioned, due to the possession of large *piezoelectric coefficients*, *strained growth* has even more important impacts on electronic devices realized in *hcp pseudomorphic heterostructures* of *III-Nitride* semiconductors. These *heterostructures* are often realized in the form of Al_xGa_{1-x}N or In_xGa_{1-x}N *alloyed ternaries*[92] on a thick *relaxed GaN* layer, grown along *c-axis* of the *hexagonal* prism. In this case, the *strain tensor* is given by

[92] As an example, *alloyed ternary* of Al_xGa_{1-x}N is created by random mixing of *AlN* and *GaN binaries* in proportions identified by $x(AlN) + (1 - x)GaN$.

$$\begin{aligned} \epsilon_{xx} &= \epsilon_{yy} = \frac{a_{\mathrm{S}}}{a_{\mathrm{L}}} - 1 \\ \epsilon_{zz} &= -\frac{2c_{13}}{c_{33}}\epsilon_{xx} \end{aligned} . \tag{1.109}$$

In Chapter 5, we will establish a connection between this component of strain and problem of charge induction into the *heterostructure*'*s quantum well* formed at the *heterointerface*. Recently, a growing degree of interest is also observed on *a-axis* growth in this family. This endeavor has been developing with the goal of reducing the amount of *polarization*-induced charge.

1.6.2.2.2 Band Diagram Engineering The *band* structure of a semiconductor, which is naturally defined through its chemical composition and crystalline structure, can be engineered through a number of different ways including:

- Alloying two or more semiconductors
- Implementing quantum confinement in *heterostructures*
- Implementing built-in strain in *pseudomorphically* grown *heterojunctions*

BAND ENGINEERING THROUGH ALLOYING In the discussion of *alloyed epitaxy*, we have already indicated that a weighted averaging law can produce a first-order approximation for the properties of the semiconductor *alloy*. In regard to the application of this law (i.e., *Vegard's law*), it should be emphasized that this law of weighted averaging is applicable only when the *alloy* is random and the components in the *alloy* have the same crystalline structures. This law is not extendable to *phase-separated alloys*, which are *alloys* in which components of the *alloy* are separated into regions. Obviously, this need not be a *superlattice* for which case we have already hinted that *Vegard's law* is not applicable.

Most of the *alloys* used in semiconductor electronics are *random alloys*. In these *alloys*, in the absence of periodicity in the background crystal potential, the description of the *electron wave* in terms of a *traveling wave* (i.e., defined by *Bloch* theorem) is not possible. Instead, the *wave function* and the *probability density function* are position dependent. As will be discussed later in this chapter, this is a cause for *scattering* and degradation of *carrier mobility* in *alloyed* crystals.

In the application of *Vegard's law*, which is motivated by weighted averaging of the *virtual crystal approximation*, it should be appreciated that in most *alloys* bowing effects (i.e., in excess of linear averaging) arise from the increasing disorder due to *alloying*. This bowing is usually modeled through adding a parabolic term to *Vegard*'s linear weighted average. In the *virtual crystal approximation*, the *effective carrier mass* is defined in terms of the *effective carrier masses* in the parent semiconductors (i.e., A and B) of the *alloy* A_xB_{1-x}[93]:

[93] That is, x part A and $1-x$ part B.

$$\frac{1}{m^*_{\text{alloy}}} = \frac{x}{m^*_{\text{A}}} + \frac{1-x}{m^*_{\text{B}}}. \tag{1.110}$$

This is because in terms of the *Vegard's law*,

$$E_{\text{alloy}}(k) = \frac{\hbar^2 k^2}{2m^*_{\text{alloy}}} = x\frac{\hbar^2 k^2}{2m^*_{\text{A}}} + (1-x)\frac{\hbar^2 k^2}{2m^*_{\text{B}}}. \tag{1.111}$$

Based on (1.111), the law of weighted averaging used to determine the *electron affinity* and *bandgap* of a *direct-bandgap* alloy is formulated as

$$q\chi_{\text{alloy}} = xq\chi_{\text{A}} + (1-x)q\chi_{\text{B}} \tag{1.112}$$

$$E_{\text{g-alloy}} = xE_{\text{g-A}} + (1-x)E_{\text{g-B}}. \tag{1.113}$$

The *lattice constant* also approximately follows the same relationship.

As mentioned, Equation (1.113) can only be used when parent semiconductors are both *direct*. Otherwise, (1.111) should be used to evaluate the bottom of the *conduction band* for all values of *k* and then to identify the *bandgap* as the smallest energy gap between the *conduction* and *valence band*.

BAND ENGINEERING THROUGH QUANTUM CONFINEMENT AND STRAIN As indicated earlier in this section, the implementation of *carrier* confinement is another way for altering the *density of states* functions and the semiconductor *band* structure. Nowadays, with an increased intensity, a combination of crystal growth (i.e., *MOCVD* and *MBE*) and processing techniques are being explored for realizing 2-D, 1-D, and 0-D confined *carriers* in semiconductor *heterostructures*. Implementation of *heterostructures* in different material systems has created a variety of new possibilities for the design of electronic and optoelectronic devices. Differences between the size of the *bandgap* and the *electron affinity* of different semiconductors provide three different forms of *band lineup* between the crystals from which a *heterostructure* can be created (i.e., nested *bandgap*: *Type-I*, staggered *bandgap*: *Type-II*, and offset *bandgap*: *Type-III*). The three types of *band lineup* are depicted in Figure 1.42. The various possibilities for *band lineup* are among the determining factors for the choice of materials in designs involving *carrier* confinement.

Among the types of *band lineup*, *Type-I* is the one that is most studied for electronic applications. Especially in Chapter 5, an in-depth analysis of *Type-I band lineup* with respect to *FETs* is provided.

Type-II band lineup, because of its small *effective bandgap*,[94] is of considerable interest in the design of long wavelength optoelectronic devices. An important observation of *Type-II band lineup* is the small spatial separation of electrons and *holes*.

[94] Which is essentially formed between the *conduction-band* edge of one semiconductor and the *valence-band* edge of the other.

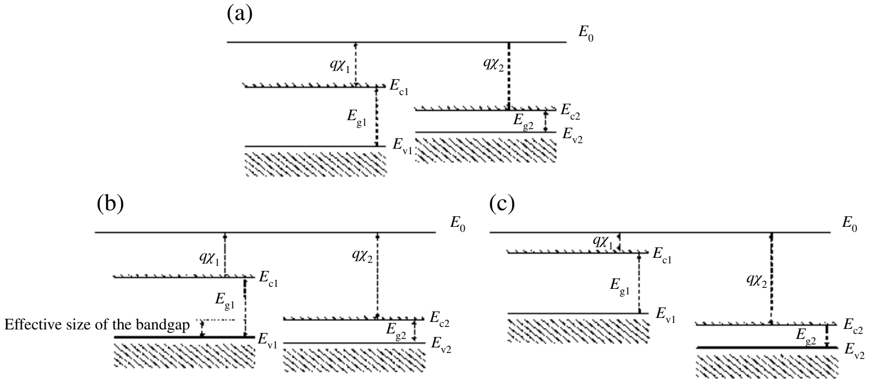


FIGURE 1.42 Types of band lineup: (a) *Type-I*, (b) *Type-II*, and (c) *Type-III*. Occupation of the bands by electrons is indicated by the hash marks.

In this case as demonstrated in Figure 1.42b, electrons and *holes* are accumulated on the opposite sides of the *heterojunction*. This can be very important in the design of *infrared detectors* when it comes to the consideration of instantaneous *EHP recombination*.

One should also be aware that, while oftentimes the definition of *electron affinity* is employed in envisioning the type of *band lineup* in a *heterostructure*, this definition is often not so accurate when two different semiconductors form a *heterostructure*. This is due to the presence of *charge sharing* across the *heterointerface* atoms. Although a few theoretical techniques have been developed to decide the type of *band lineup*, due to the complexity of these techniques, experiments are often used to shed light on the type of *band lineup*.

According to our discussions in Section 1.1.3 in terms of *subband* energy *levels*, the *density of states* functions in a *quantum well* is formulated as

$$N(E) = \sum\nolimits_i \frac{m^*}{\pi\hbar^2} U(E - E_i) \text{ for conduction band} \tag{1.114}$$

$$N(E) = \sum\nolimits_i \sum\nolimits_{j=1}^{2} \frac{m_j^*}{\pi\hbar^2} U\left(E_{ij} - E\right) \text{ for valence band,} \tag{1.115}$$

where U is the *Heaviside* step function and E_i is the *subband* energy *level*. The presence of double subscripts in Equation (1.115) points to the lift in *degeneracy* of *heavy-* and *light-hole subbands* even in the absence of strain.

Often in simple descriptions of *quantum well*s, the *conduction-band states* are seen as pure s-type *states* and a simple *effective-mass* theory like the one suggested in (1.13) is employed. For a more accurate description, a full *band* structure, which is referred to as an *eight-band model*, is employed. As a result of this more sophisticated calculation, it is observed that while the *conduction-band states* in *unstrained heterostructures* are not affected by the incorporation of this fuller mode, in *highly strained*

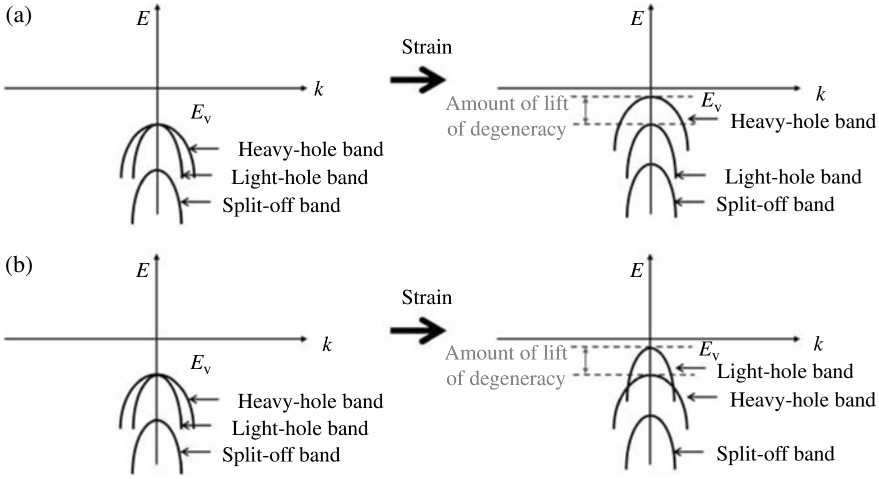


FIGURE 1.43 Lifting the valence-band degeneracy of heavy- and light-hole bands through the application of (a) compressive and (b) tensile biaxial strain, depicted on the *E*–*k* diagram.

heterostructures the more complicated calculation is definitely required. An example of this would be the case of so-called self-assembled quantum dots. These are 3-D *quantum well*s that are created as the result of large *lattice mismatch* between the substrate and the *freestanding lattice constant* of the overgrown material. Excessive *tensile* strain induces a 3-D mode of growth instead of a planar growth. The density of these islands of 3-D growth is determined by the material properties and *lattice mismatch*. These *self-assembled quantum dots* have been realized between *AlAs* and a *GaAs* substrate.

The description provided in (1.115) of *subbands* for *valence-band states* is only approximately valid. While *heavy-hole* and *light-hole states* are pure *states* at $k = 0$, they strongly mix away from $k = 0$.

Built-in strain has been successfully used to lift *degeneracies* in the *band* edges, change the character of *band*-edge *wave functions*, and engineer the *density of states* at the *band* edges.

In *direct-bandgap* semiconductors, with regard to the *conduction band*, strain only moves the position of the *band edge* and has a very limited impact on the *carrier mass*. Because the bottom of *conduction band* is not *degenerate*, this shifting of the *band edge* does not result in lifting any *degeneracy* or change in the *DOS* and *effective mass*. However, while the *valence-band* edge of semiconductors are *degenerate*, strain causes lifting of that *degeneracy*. The amount of lift in *degeneracy* caused by quantum confinement is usually about 10–15 meV and is less than the amount of shift due to strain. *Compressive biaxial* strain raises the *band edge* and lifts the *degeneracy* on the order of 100 meV. In this case, the *HH band* is lifted above *LH band*. Under *tensile biaxial* strain this order is reversed. This is depicted in Figure 1.43.

In *indirect-bandgap* semiconductors, due to the presence of *degeneracy* in the *conduction-band* edge, the role of strain in modifying the *conduction-band* characteristics is much different. Whereas the bottom of *conduction band* of a *direct* semiconductor is in the Γ-*valley* (which has no *degeneracy*), for the case of the *indirect* semiconductor *Si*, this is located close to the *X-valley*. According to the form of the *unit cell* of the *diamond* crystal of *Si* and *Ge*, there are six equivalent faces (or, consistently, six equivalent *X-valleys* in their *Brillouin zone* or *E–k* diagram).

As a result, in contrast to the *direct-bandgap* semiconductors, the *conduction band* of *indirect alloy* of Si_xGe_{1-x} is significantly altered by the strain imposed through *pseudomorphic* growth on *Si* substrate. In the case of (*001*) growth, strain lifts in the *degeneracy* between the six equivalent *valleys* of the *conduction band*. This *biaxial compressive* strain causes breaking of the sixfold *degeneracy* into a fourfold (*in-plane*) and a twofold (*out of plane*) *degeneracy*, where the fourfold *degeneracy* is lower in energy. The resulting reduction of *DOS* at the bottom of the *conduction band* helps with reducing the *effective electron mass*. An additional result of this lift in *degeneracy* is a rapid reduction of the size of the *bandgap* of *SiGe alloy* with the *Ge mole fraction* (i.e., $1-x$). In the context of improving the *carrier-transport* properties of the *channel* of the modern *FET* technologies, this concept will be heavily dealt with in Chapter 4.

Despite the differences in the *conduction-band* structures, due to the similarity of the *valence bands* of *direct* and *indirect* semiconductors, the aforementioned discussion on the impact of strain on the *valence band* of *direct* semiconductors is readily extendible to *indirect* semiconductors.

In the *valence band*, lift in *degeneracy* is also accompanied by large changes in the *band* curvature. In the edge *states* of this *band*, strain can cause the *DOS effective mass* to be scaled down by as large as a factor of *three*.

We have already talked about the consequences of the formation of the bottom of the *conduction band* of *Si* near the *X-valley*. In this case, constant-energy surfaces for electrons form *six* ellipsoids along the <*100*> directions. The ellipsoidal form is expressed in terms of two values of effective mass. For an *x–y* surface, whereas the *two* ellipsoids in the *z*-direction have an *effective mass* (referred to as *longitudinal*), which is equal to $m_l = 0.98m_0$, the other *four* ellipsoids' electron *effective mass* is the *transverse effective mass* (given by $m_t = 0.19m_0$). Among these, in determining the *subband* energies, the *effective mass* in the direction of the confining potential is employed.

In calculating the *DOS* function, it is easy to incorporate ellipsoidal shape of the constant-energy surfaces. In the case of *silicon*, m^* must be replaced by $\left(m_l^* m_t^{*2}\right)^{1/3}$. Multiplying this by 6 takes care of the sixfold *degeneracy* inside the *Brillouin zone* of *silicon*.[95] In the case of *germanium*, because the *conduction band* is defined at the *L-valley*, an eightfold *degeneracy* is present. Hence, rather than *six degenerate* ellipsoids, *eight* are present. However, since only *one-half* of each ellipsoid falls inside the *Brillouin zone*, the overall *degeneracy* will be only fourfold. Instead of multiplying the mass by 6 (as is the case with *silicon*), it should only be multiplied by 4.

[95] In the calculation of *conductivity* in *silicon*, however, $\frac{1}{m_n^*} = \frac{1}{3}\left(\frac{1}{m_l^*} + \frac{2}{m_t^*}\right)$.

1.7 QUANTUM TRANSPORT PHENOMENA AND SCATTERING MECHANISMS IN SEMICONDUCTORS

Whereas *scattering* processes are of interest to physicists in studying the electron–electron and electron–matter interactions at a fundamental level, electrical engineers need to know how these processes contribute to the problem of *charge-carrier transport* in a semiconductor. In evaluation of the *conductivity* of semiconductors, accurate understanding of the motion of electrons in the *solid* (i.e., including the *scattering* events) is required. It is only in the presence of *scattering* events (i.e., imperfections) that the known notion of *conductivity* in terms of the *Ohm*'s law is extendible to semiconductors. Without these imperfections, as mentioned earlier, *ballistic transport* and a number of other important modes of *persistent current* or *oscillations* will prevail.

In order to establish the connection between *conductivity* and *scattering*, the *velocity–field* relationships of different semiconductors are studied. These relationships, as encountered in Section 1.4, are often illustrated in terms of the *drift* process of *charge carriers* under steady-state conditions. According to those discussions, in the description of the steady-state *drift-transport* characteristics, three regions are identified:

1. When the electric field is low and *drift velocity* changes as a linear function of the electric field: $\vec{v_{\mathrm{d}}} = \mu \vec{E}$. Under this condition, the *Boltzmann transport equation* yields an analytical form.
2. When the electric field takes on moderate values (usually larger than 1 kV/cm). Under this regime, $v_{\mathrm{d}} - E$ does not follow a linear trend. Numerical methods are often used in evaluation of *carrier transport* under this regime. Interpolation of an analytical relationship between the *drift velocity* and electric field is often used to produce an analytical basis for evaluation of *carrier transport*.
3. When the electric field exceeds the breakdown field of the semiconductor (usually larger than 10^5 kV/cm). Under this regime, either due to *impact ionization* or electron *tunneling* from *band* to *band* (i.e., *Zener* breakdown), the semiconductor breaks down.

As mentioned in Section 1.4, in order to reach steady state, an electron needs to undergo several tens of collisions. Considering the collision times on the order of *picoseconds* and assuming electrons to travel at velocities as high as 10^7 cm/s, the traveling distance to reach steady state will amount to *micron*-size distances. As a result, in small submicron devices the steady-state *transport* formalism becomes less and less applicable.

It has been already indicated in this chapter that electrons can be *scattered* through a variety of *elastic* and *inelastic scattering* processes. Two major categories of these processes are *ionized-impurity scattering* and lattice-vibration (or *phonon*) *scattering*. It should also be mentioned that *lattice vibrations* themselves (i.e., *phonons*) are also *scattered* by *elastic scatterers* such as *ionized* and *neutral impurities*, the presence of different atomic isotopes, and surfaces. This is the cause for the slow propagation of heat in a semiconductor, which is identified in terms of *heat resistivity*. While ideally in a *solid* heat should propagate with the *velocity of sound*, these *scattering* mechanisms cause the *heat transfer* to become much slower.

1.7.1 Quantum Phenomena in Carrier Transport: A Snapshot

We already know that in the presence of *scattering* events, such as *lattice vibrations* (which result in *inelastic scattering*), a few of the important predictions of quantum mechanics are not realizable. Among these is the *Bloch oscillation*. This mode of oscillation originates from reflection of *electron waves* at the boundaries of the *Brillouin zone*. According to *Bloch theorem*, electrons in a perfectly periodic crystal follow the *bands* to the edges of the *Brillouin zone* and then return. This situation for an arbitrary *E–k* diagram is depicted in Figure 1.44. Considering the small size of the *Brillouin zone*, these oscillators are expected to operate at very high frequencies.

However, in the presence of *phonon scattering*, due to the *inelastic* nature of the *scattering* process, electrons do not get a chance to reach the boundaries of the *Brillouin zone*. As a result all attempts to realize a *Bloch oscillator* have remained unsuccessful.

Another mode of oscillation envisioned for *perfect* semiconductors is referred to as *Esaki–Tsu oscillation*. Although these oscillations are also never realized in practice, the theoretical drive behind them prompted major advances in *epitaxial* growth of *superlattices*. As shown later, quantum mechanics predicts that the application of a constant *DC* electric field to a periodic structure (i.e., with periodic *E–k* diagram) results in very high-frequency oscillations. The frequency of these oscillations is determined proportional to the *lattice constant* of the periodic structure. These oscillators in theory are capable of generating very high-frequency tunable *signals*.

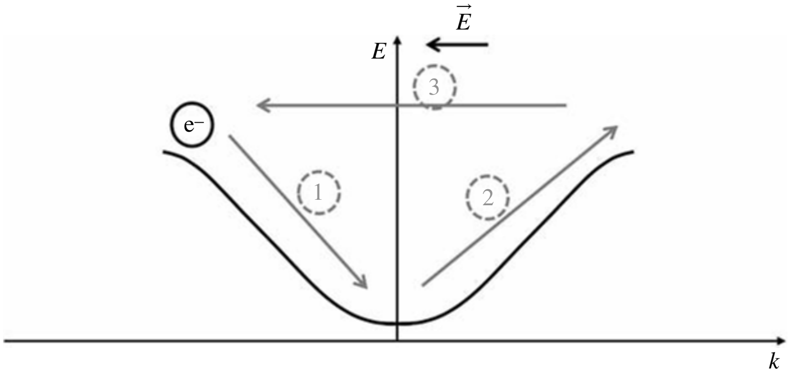


FIGURE 1.44 Trajectory of movement of an electron in an arbitrary band. In a perfect crystal, electron after climbing up the band tail, reflects upon reaching the zone edge. This reflection looks like scattering by a reciprocal lattice vector. The numbers are supposed to provide a snapshot of the position of electron in sequential instances of time.

In order to gain a first-order mathematical insight into the *Esaki–Tsu oscillators*, we will consider the *band* structure of a one-dimensional crystal. This *E–k* diagram similarly applies to a simple cubic three-dimensional crystal,

$$E(k) = E_0(1 - \cos ka). \tag{1.116}$$

As suggested in (1.12), for electrons in the *conduction band*, we can calculate their *group velocity* as

$$v = \frac{1}{\hbar}\nabla_k E(k) = \frac{1}{\hbar}\frac{dE}{dk} = \frac{E_0 a}{\hbar}\sin(ka). \tag{1.117}$$

In addition, in terms of the definition of *momentum* in quantum mechanics, we know that $\hbar(dk/dt) = -q\mathrm{E}$. This equality results in

$$k = -q\mathrm{E}\frac{t}{\hbar} + k_0. \tag{1.118}$$

Assuming that the electron starts at *zero* momentum at $t = 0$, we will have

$$v = -\frac{E_0 a}{\hbar}\sin\left(q\mathrm{E}a\frac{t}{\hbar}\right). \tag{1.119}$$

According to (1.119), the current density created by the constant electric field E is a sinusoidal *signal* of angular frequency: $(q\mathrm{E}a)/\hbar$. Clearly this angular frequency can be tuned by changing the period of the structure of the crystal and its *lattice constant*: *a*. As a result, realization of a *superlattice* with a very large period is expected to result in materialization of a very high-frequency oscillation.

1.7.2 Drude's Model: A Close-UP

Due to the presence of a large number of interfering factors (including dependence of *scattering* on the momentum of electrons and *phonons*, presence of many electrons and as a result importance of consideration of *many-body interactions*, requirement for statistical evaluation of electron propagation, and the role played by the *band* structure in *transport*), the theory of *carrier transport* is quite complicated. However, as long as the electric field is weak to moderate, instead of a detailed description of all these intricacies, simpler models are usually adopted.

1.7.2.1 Boltzmann Transport Theory The *Boltzmann transport equation* (i.e., *BTE*) is the foundation of these simpler formalisms, through which the *mean scattering time constant* needed in calculation of low-field *mobility* is evaluated (see (1.64) and (1.65)). According to this equation, upon application of an external perturbation to a system under *thermal equilibrium*, the distribution functions of *charge carriers* can still be represented in terms of the *Fermi–Dirac* distribution function (i.e., *thermal-equilibrium* distribution). As we took note in Section 1.3.1, such a distribution function shows the spread of electrons in *energy space* (and as result *momentum space*). According to *BTE*, this distribution function is then used to evaluate all of the *transport* properties of *charge carriers*.

Assuming $f_{\vec{k}}\left(\vec{r}\right)$ as the local occupation function of electron in *momentum state* $\vec{k}$ and in position $\vec{r}$, the first step in *Boltzmann transport theory* is to study the time evolution of this function. This time evolution is prompted by the *thermal motion* of the electrons, the *drift* of electrons due to an external force, or due to *scattering* between different *momentum states*. For these three different causes, we can explain the time evolution of $f_{\vec{k}}\left(\vec{r}\right)$ in terms of the following equations:

1. *Thermal motion* of the electrons

$$\left.\frac{\partial f_{\mathrm{k}}}{\partial t}\right|_{\mathrm{diff}} = -\frac{\partial f_{\mathrm{k}}}{\partial \vec{r}} \cdot \vec{v}_{\mathrm{k}}, \tag{1.120}$$

where $\vec{v}_{\mathrm{k}}$ is the velocity of a *carrier* in the *state* $\vec{k}$.

2. *Drift* of electrons due to an external electromagnetic force

$$\left.\frac{\partial f_{\mathrm{k}}}{\partial t}\right|_{\mathrm{external\ force}} = -\frac{q}{\hbar}\left[\vec{E} + \vec{v} \times \vec{B}\right] \cdot \frac{\partial f_{\mathrm{k}}}{\partial \vec{k}}. \tag{1.121}$$

3. *Scattering* between different *momentum states*

$$\left.\frac{\partial f_{\mathrm{k}}}{\partial t}\right|_{\mathrm{scattering}} = \int \left[f_{\mathrm{k}'}(1-f_{\mathrm{k}})W(k',k) - f_{\mathrm{k}}(1-f_{\mathrm{k}'})W(k,k')\right]\frac{d^3k'}{(2\pi)^3}, \tag{1.122}$$

where $W(k,k')$ represents the rate of *scattering* from *momentum state* k to k'. According to the process of *microscopic reversibility*, the *scattering* rates between the before and after *momentum states*, as long as *scattering* is elastic, are equal:

$$W(k,k') = W(k',k). \tag{1.123}$$

The inclusion of the *scattering* rate in *BTE* is rooted in the wave nature of electrons.

The role of time constant τ in *Boltzmann transport theory* is to model the time constant for relaxation of the aforementioned perturbations. This is based on the so-called relaxation time approximation. As suggested earlier, this time constant can be calculated according to the *scattering* rate of different *scattering* processes. Under steady-state conditions, the time evolutions created by the aforementioned sources cancel one another out.

In formulation of the *scattering* rate, attention should be given to the nature of the collision processes, which cause the *scattering*. For example, *alloy scattering* (which is present in semiconductor *alloys*, due to their deficient periodicity) and *impurity scattering* are results of collisions in which electron's energy remains unchanged (i.e., *elastic* collisions). However, many *scattering* processes are *inelastic*. For example, *phonon scattering* due to the change caused in the electron energy is an *inelastic scattering* process. Instigated by the domination of different *scattering* processes in *carrier transport* and a number of other complications listed in Section 1.4.3, *Hall mobility* can be quite different from *drift mobility* (see (1.85)).

In the case of *elastic scattering* processes in parabolic *bands*, the calculation of *relaxation time* is quite trivial. However, many of the *scattering* processes are not so lenient.

As long as the energy gained from the electric field is smaller than the thermal energy, *carrier transport* stays under the linear regime expressed in terms of the low-field *mobility*. However, with further increase of energy, the simple approximations used in solving the *Boltzmann transport equation* will become insufficient. As mentioned earlier in this section, under this regime complex numerical techniques such as the *balance equation* and *Monte Carlo* method are often used. The continuity equations, (1.49) and (1.50), are two of these so-called balance equations. *Monte Carlo* method treats electrons as particles whose *scattering* events between *Bloch states* are described probabilistically through the so-called Fermi golden rule. In this numerical technique, *carrier transport* is seen as periods of free flight and instantaneous *scattering* events, which are accurately described in terms of the probability of the involved *scattering* processes.

Monte Carlo has been proven to be a versatile technique in addressing a variety of *transport*-related problems, such as evaluation of steady-state *drift-transport* characteristic, *electron temperature*, *valley* occupation, distribution in *k-space*, noise, *ballistic transport*, transit time, *carrier* injection/*thermalization* process,[96] and *impact ionization*.

1.7.2.2 Drude's Model *Drude*'s model of conduction, which is devised on the basis of the *Boltzmann transport equation*, is the model that we have so far adopted in this volume (i.e., in the definition of low-field *diffusion constant* and *mobility*). Now that we are further acquainted with the foundations of this model, before getting too involved in a rigorous discussion of *scattering* rates of different *scattering* processes, it seems quite instructive to pay a closer visit to this model. As implicitly suggested in Section 1.4, *Drude*'s model is built on the following assumptions:

1. All electrons move with the same velocity v, which in terms of applied force F_0 is identified as $\hbar k = mv = F_0 t$.
2. Instead of taking the whole *band* structure into account, the *effective-mass* notion (i.e., m^*) has been introduced to take the position of the mass of an electron.

[96] *Thermalization* of electrons refers to the process of electrons losing energy and coming back to the bottom of *conduction band* through emitting *phonons*.

3. *Scattering* processes are envisioned as a friction force in the above equation of motion (i.e., F_{f}), which is defined in terms of the relaxation time constant of the friction force: $F_{\mathrm{f}} \cong (mv)/\tau$. In terms of the *Boltzmann transport equation*, the presence of this friction force is presented through the modified electron distribution function: $f_{\vec{k}}(\vec{r}) = f_{\vec{k}+\frac{q\tau\vec{E}}{\hbar}}{}^{0}(\vec{r})$.

As a result of these assumptions, the total equation of motion is given by[97]

$$m^* \frac{dv}{dt} = F_0 - F_{\mathrm{f}} = F_0 - \frac{m^* v}{\tau}, \tag{1.124}$$

where in the absence of an applied force (i.e., F_0),

$$m^* \frac{dv}{dt} = -\frac{m^* v}{\tau} \tag{1.125}$$

and

$$\vec{v} \propto \exp\left(-\frac{t}{\tau}\right). \tag{1.126}$$

As a result, τ will also be the time constant in decaying $\vec{v}$. The current density is then evaluated by multiplying this velocity by the charge of an electron and the electron concentration.

In this framework, in the presence of a number of independent *scattering* processes, the total *scattering* rate can be calculated in terms of *mean scattering* times of individual *scattering* processes

$$\frac{1}{\tau_{\mathrm{tot}}} = \sum\nolimits_i \frac{1}{\tau_i}. \tag{1.127}$$

Only when the various *scattering* rates follow the same energy dependence (i.e., the same *effective mass*), the above equation results in an overall low-field *mobility* that follows *Matthiessen's rule*[98]:

$$\frac{1}{\mu_{\mathrm{tot}}} = \sum\nolimits_i \frac{1}{\mu_i} \tag{1.128}$$

However, the above condition is usually not satisfied. In spite of this, *Matthiessen*'s rule has been widely used and found to be reasonably accurate.

[97] As suggested in (1.121), magnetic force can be easily incorporated into this equation (i.e., $m^* \frac{dv}{dt} = -q\left(E + \vec{V} \times \vec{B}\right) - \frac{m^* v}{\tau}$).

[98] Named after *Augustus Matthiessen*.

As examples for the application of *Drude's theory*, in the presence of only an electric field, it is instructive to apply (1.124) to the following two special cases:

1. When the electric field is finite but time independent.

   Due to lack of time dependence in the external source of energy, this case refers to a steady-state situation where $\frac{dv}{dt}=0$ and $m^*\vec{v}=-q\vec{E}\tau$. This relationship results in a *DC* current of

   $$J=-qnv=\frac{q^2\tau n}{m^*}E \tag{1.129}$$

   with no chance of *persistent current* and *oscillation* (since in *Drude*'s model the full *E–k* diagram has been overlooked by the notion of the *effective mass*). This special case refers to the case that we have investigated earlier in our discussions in Section 1.4. In this case, the *conductivity* σ is observed to be a scalar value given by $\sigma=qn\mu$, where $\mu=q\tau/m^*$.
2. When the electric field is a low-frequency field.

   The definition of the "low frequency" is given in relation to the *energy relaxation time*. In this case, $E\equiv E_{\mathrm{ac}}\exp(j\omega t)+E_{\mathrm{ac}}\exp(-j\omega t)$. In solving for $\vec{v}$, the velocity vector is written as $\vec{v}=v_0\exp(j\omega t)+v_0\exp(-j\omega t)$, which results in

   $$m^*j\omega v_0e^{j\omega t}=-qE_{\mathrm{ac}}\exp(j\omega t)-m^*v_0\frac{\exp(j\omega t)}{\tau}. \tag{1.130}$$

   Accordingly,

   $$-v_0=\frac{qE_{\mathrm{ac}}}{jm^*\omega+(m^*/\tau)}=\frac{\mu E_{\mathrm{ac}}}{1+j\tau\omega}. \tag{1.131}$$

   This result indicates that the *drift velocity* is equal to the *DC* velocity divided by $1+j\tau\omega$. As a result, at frequencies comparable to $1/\tau$ (usually greater than 10 GHz), the semiconductor instead of a purely resistive behavior expresses a resistive–inductive behavior. The imaginary part can be seen as a contribution of free electrons to the *dielectric constant* of the semiconductor. The other interesting point to mention is that the real part is inversely proportional to the relaxation time at very high frequencies, meaning that for this frequency range, the *conductivity* is 0 even in the absence of *scattering*.

1.7.3 Major Scattering Processes

According to *Boltzmann transport theory*, *scattering* rates between different *states* of *momentum space* are to be evaluated for a range of important *scattering* processes. Before embarking any deeper into this discussion, it is worth pointing out that *momentum space* (which is also referred to as *k-space*) is really a *Fourier space*.[99] In this analogy, it should be emphasized that in treating a *scattering* problem, a *scattering* matrix should be established between each initial *state* and a range of final *states* of an electron. These initial and final *states* are *plane wave states* that can be calculated by taking the *Fourier transform* of the potential (i.e., solving the *Schrödinger equation* for *carrier wave function*). Analogous to the role of *Fourier series*, the transformation from almost periodic *real space* to *k-space* sometimes greatly simplifies the equations. As a simple example, the *Poisson equation* according to this transformation will be simplified as

$$\nabla^2 \phi = -\frac{\rho}{\epsilon} \rightarrow \phi = \frac{\rho}{\epsilon k^2}. \tag{1.132}$$

At this point in our discussion, without getting into details, we present a few important facts such as *scattering* rates to explain different *scattering* processes that are encountered by electrons in semiconductors. A few of these processes, such as *ionized-impurity scattering* and *phonon scattering*, have been introduced in Section 1.4. In addressing *ionized-impurity scattering* and the *Coulombic* potential imposed by these impurities on *charge carriers*, one needs to pay attention to the issue of *charge screening*. As an example, we can consider the case of a positive charge placed in an *electron gas*. Movement of electrons around this charge can essentially *screen off* its potential. As will be seen in this section, in the case of *ionized-impurity scattering* and also *carrier–carrier scattering*, *screening* can be dealt with by replacing the *Coulombic* potential of a *point charge* by a *screened Coulombic* potential.

1.7.3.1 Ionized-Impurity Scattering In this case, the *scattering* rate given by the *Fermi golden rule* in terms of *Dirac's delta function* is represented by

$$W(k,k') = \frac{2\pi}{\hbar}\left(\frac{Zq^2}{V\epsilon}\right)\frac{\delta(E_{\mathrm{k}} - E_{\mathrm{k}'})}{\left(4k^2 \sin^2(\theta/2) + \lambda^2\right)^2}, \tag{1.133}$$

where Zq, V, and θ represent the charge of the impurity, volume, and polar *scattering* angle, respectively. ϵ is the semiconductor's permittivity. In this equation, λ is defined as

$$\lambda = \sqrt{\frac{n_0 q^2}{\epsilon k T}}, \tag{1.134}$$

where n_0 is the mean background *carrier* concentration.

In the presence of a large *free carrier* concentration, a *screened Coulombic potential* has been used in this framework:

$$\phi(r) = \frac{q}{4\pi\epsilon r}\exp(-\lambda r). \tag{1.135}$$

[99] Named after *Joseph Fourier*.

For the situations without *screening* (i.e., low *free carrier* concentration), the rate equation is evaluated when λ tends toward 0, for which case

$$W(k,k') \propto \frac{1}{16k^4 \sin^4(\theta/2)}. \tag{1.136}$$

However, under strong *screening* $\lambda \to \infty$ and

$$W(k,k') \propto \frac{1}{\lambda^4}. \tag{1.137}$$

On this basis, one can see that while forward *scattering* is dominant in the case of weak *screening*, for strong *screening* angular dependence is not present. As a result, because forward *scattering* does not imply *carrier mobility* reduction, the *ionized-impurity scattering* in presence of weak *screening* is less important.

In the formalism presented in (1.133), *ionized impurities* are treated independent of one another (i.e., the average distance >10 nm). This assumption is not extendable to heavily *doped* semiconductors (i.e., >10^{18} cm^{-3}) for which case the *ionized-impurity scattering* is more complex. For a *degenerate* semiconductor, the *screening parameter* λ should be changed to

$$\lambda = \sqrt{\frac{3n_0 q^2}{2\epsilon E_\mathrm{F}}}, \tag{1.138}$$

where E_F refers to the *Fermi energy level* measured from the *band* edge.

The Equation (1.138) results in a faster drop in *mobility* with the concentration of *ionized impurities*. This is due to the effect of multi-impurity *scattering*.

In *ionized-impurity scattering*,

$$\mu \propto \left(Z^2 N_\mathrm{i}\right)^{-1} \tag{1.139}$$

where N_i is the concentration of *ionized impurities* and Z, as indicated earlier, is the charge of the *donor*. The presence of Z in this relationship explains one of the reasons why impurities that can offer more than *one* electron from each atom are not used as proper *dopants*. In order to explain this further, we can imagine *dopant* atoms that offer *two* electrons instead of *one*. In this case, to achieve the same electron concentration, Z^2 will become *four* times larger, whereas N_i will be only divided by a factor of 2. As a result, according to this change of *dopants*, *mobility* will be reduced by 50%.

As indicated in Section 1.4.2.1, *carrier mobility* as a function of temperature follows an improving characteristic (which is of the form $\mu \propto T^{3/2}$), when *ionized-impurity scattering* is dominant. This is a distinguishing feature of this process of *scattering*.

1.7.3.2 Alloy Scattering We have already suggested that the root of this *scattering* mechanism is the presence of disorder in the crystal's potential. According to the *hard sphere model*, the *scattering* potential is represented by

$$\Delta V(r) = \begin{vmatrix} V_0 & \text{for} & |r| \le r_0 \\ 0 & \text{for} & |r| > r_0 \end{vmatrix} \tag{1.140}$$

in which r_0 and V_0 refer to interatomic distance and the maximum potential difference between any two points in the *solid*, respectively. Using the *Fermi golden rule*, the *scattering* rate of an *alloy scattering* process is represented by

$$W(k) = \frac{2\pi}{\hbar}\sum\nolimits_{k'} |M_{\mathrm{kk'}}|^2 \delta(E_{\mathrm{k}} - E_{\mathrm{k'}}) \tag{1.141}$$

where matrix elements are given by

$$M_{kk'} = \int \exp\left[j\left(\vec{k} - \vec{k'}\right) \cdot \vec{r}\right] \Delta V(r) d^3 r. \tag{1.142}$$

Using the fact that *scattering* potential extends only to a *unit cell*, the exponential term in the above equation tends toward one. As a result, in *alloy scattering* there is no angular dependence because there is no k and k' dependence on the matrix elements. Hence, for an isotropic *density of states* function, there will be no angular dependence of *scattering* rate. After performing a proper ensemble averaging, *relaxation time* of *alloy scattering* results in

$$\frac{1}{\langle\langle \tau \rangle\rangle} = \frac{3\pi^3}{8\hbar} V_{\mathrm{unit}} V_0^2 x(1-x) \frac{m^{*3/2}(kT)^{1/2}}{\sqrt{2}\pi^2\hbar^3} \frac{1}{0.75} \tag{1.143}$$

where V_{unit} is the volume of the unit cell.

This equation shows that the *carrier mobility* in terms of *alloy scattering* degrades with temperature as

$$\mu \propto T^{-1/2}. \tag{1.144}$$

In Equation (1.143), V_0 is on the order of 0.5 eV. The assumptions behind these equations are that there are no clusters formed in the *alloy* and that the smallest region over which a disorder is present is a *unit cell*.

1.7.3.3 Neutral Impurity Scattering The presence of neutral impurities and defects is another reason for *carrier scattering* in semiconductors. This presence can be caused by *substitutional* impurities and *dopants* that are not *activated*. *Scattering* by neutral impurities can be seen in the same light as *alloy scattering* (i.e., in the form of a disturbance to the periodicity of the crystal potential seen by *Bloch states*). The resulting *scattering* rate, not unlike the case of *alloy scattering*, is given by

$$W(k) = \frac{2\pi}{\hbar}\left(\frac{4\pi}{3} r_0^3 V_0\right)^2 N(E_k) \tag{1.145}$$

in which V_0 denotes the *scattering* potential and r_0 stands for the radius of the hard sphere representing the defect. Accordingly, the *scattering* time constant is given by

$$\frac{1}{\langle\langle\tau\rangle\rangle}=N_{\mathrm{imp}}\frac{2\pi}{\hbar}\left(\frac{4\pi}{3}r_0^3V_0\right)^2\frac{m^{*3/2}(kT)^{1/2}}{\sqrt{2}\pi^2\hbar^3}\frac{1}{0.75} \tag{1.146}$$

where N_{imp} stands for the concentration of neutral impurities.

Only in the presence of a very large concentration of these types of impurities ($>10^{18}\,\mathrm{cm}^{-3}$) is the *neutral impurity scattering* mechanism worthy of consideration.

1.7.3.4 Interface Roughness Scattering The *channel* of the majority of *FETs* is formed in the vicinity of an interface (i.e., of either two semiconductors, the interface of a semiconductor and a metal or the interface of a semiconductor and an insulator). As a result, a *scattering* process often encountered in the *channel* of these *transistors* is the *interface-roughness scattering*. The degree of roughness at these interfaces is dependent on fabrication technology. Roughness at these interfaces imposes an element of disturbance on the potential felt by *Bloch states*. This is in addition to the role played by the *surface states* (as either neutral or charged *states*), which were discussed in Section 1.1.6. The *scattering* rate caused by these interface potential bumps in an *inversion MOSFET channel* to be explored in Chapter 3 is formulated as

$$W(k)=\frac{1}{A}\frac{2\pi}{\hbar}\frac{1}{4\pi^2}\int_0^{2\pi}d\theta\int_0^{\infty}qdq|M(k,k')|^2\delta(E_{\mathrm{k}}-E_{\mathrm{k}'}) \tag{1.147}$$

where A is the area and $q=2k\sin(\theta/2)$.

In this equation, it is assumed that half of the electric field drops across the insulator.

1.7.3.5 Carrier–Carrier Scattering Whereas the *scattering* sources so far discussed in this section are fixed in time and in space, not all *scattering* sources operate in this way. *Carrier–carrier scattering* is one of the *scattering* processes that belongs to this latter group. *Scattering* processes that are fixed in time and in space result in *elastic* scattering (i.e., *zero* change in *carrier* energy), which is due to the large mass of *scatterers* in those processes. This is not the case in *carrier–carrier scattering*. *Carrier–carrier scattering* can be in the form of *electron–hole scattering* or *electron–electron scattering* (which is analogous to *hole–hole scattering* in the *valence band*). Due to the *Fermionic* nature of electrons, *scattering* of two identical *carriers* is a more complex process. *Carrier–carrier scattering* gains significant importance only when *carrier* concentration exceeds $10^{18}\,\mathrm{cm}^{-3}$.

As an approximation, assuming that the *effective mass* of a *hole* is much larger than that of an electron, we can envision the role of *electron–hole scattering* on electrons in a p-type semiconductor (*minority-carrier scattering*) by multiplying the *scattering* rate due to *ionized acceptor impurities* by a factor of 2.

1.7.3.6 Auger Processes and Impact-Ionization Scattering An important characteristic of these two *scattering* processes is that they result in a change in *carrier* concentration in the *conduction* and *valence band*. While in *Auger*[100] process, *scattering* results in *recombination* of an electron and a *hole* (and as a result reduction of *carrier* concentration in both *conduction* and *valence band*), *impact ionization* results in an increase in *carrier* concentration in both *bands* (i.e., through *EHP generation*). We have already visited the process of *impact ionization* while addressing the *breakdown* processes in Section 1.5.

Auger is the inverse process of *impact ionization*. This is a *nonradiative recombination* process and is detested in optoelectronic *photon* generators. According to this process, even in *direct-bandgap* semiconductors, an electron and a *hole* can recombine without *generation* of a *photon*. This is possible since in this interaction the energy is being transferred to another *carrier* or to a *phonon*. These processes can be assisted by *Coulombic* interactions (i.e., *electron–electron scattering*), by *phonons*, or by *trap states*. Only in high-purity *direct-bandgap* semiconductors is the *photon generation* dominant.

In *impact-ionization* process, as noted in Section 1.5, *carriers* gain energies in excess of the size of the *bandgap* (from sources such as the electric field). This can cause a high-energy electron in the *conduction band* to *scatter* an electron from the *valence band* (i.e., to break a covalent *bond* and produce an *EHP*). Each of these three *carriers* can then go through the same process and as a result instigate an avalanche in the number of *carriers* in both *bands*.

As a result of the requirement for the conservation of momentum and energy, rates of *Auger* and *impact-ionization* process are obviously dependent both on the size of the *bandgap* and temperature. Reduction of the size of the *bandgap* among different semiconductors results in an exponential increase in the rate of both processes. The rate of *Auger* process increases with the number of *carriers* and likewise with temperature. The *impact ionization* rate, however, reduces with increasing temperature. This is because of the increase in the rates of other *inelastic scattering* processes and the subsequent increased difficulty for *carriers* to gain energy from the electric field. In semiconductors with *bandgaps* larger than 1.5 eV, *Auger* process has a limited presence.

Complete calculation of the rate of *impact-ionization* process requires knowledge from the entire *band* structure. In the case of *Auger* process, the threshold value of energy can be approximated by

$$E_1 = \frac{1+2\mu}{1+\mu} E_\mathrm{g}, \text{ where } \mu = \frac{m_\mathrm{n}^*}{m_\mathrm{p}^*}. \tag{1.148}$$

However, due to the anisotropic nature of the *bands*, at higher energies this threshold has a strong angular dependence.

[100] Named after *Pierre-Victor Auger*.

In the case of parabolic *bands*, *Ridley* has successfully formulated the *impact-ionization* rate in terms of a simple closed-form expression:

$$W_{\text{impurity}} = 4.139 \times 10^{16} \times \left\{ \frac{\sqrt[4]{m_{\text{n}}^{*} m_{\text{p}}^{*}}}{m_0} \left(\frac{m_{\text{n}}^{*}}{m_0} + \mu \right) \left(\frac{\epsilon_0}{\epsilon} \right)^2 \left[\frac{E_1}{E_{\text{g}}} - (1 + \mu) \right] \right\}. \qquad (1.149)$$

1.7.3.7 Lattice Vibration or Phonon Scattering Vibration of the crystal *lattice* can be seen as a hindrance on the periodicity of the *lattice* and, as a result, a source of *scattering* between the *Bloch states*.

As indicated before, crystallization in a particular crystalline form is the result of the tendency of a system to achieve minimum energy. Due to this tendency, as the atoms start to move about their *lattice* points, for example, due to exertion of thermal energy, a restoring force will be developed. This combination results in *lattice* vibrations. Depending on the similarity, or lack thereof, of atoms in *sublattices* of the crystal, their charge status, and the direction of the movement of these *sublattices*, *lattice* vibrations can have a few different modes (e.g., *acoustic*, *optical*, *polar*, *piezoelectric*). As indicated earlier in this chapter, lattice vibrations are expressed in terms of quantum particles known as *phonons*.

1.7.3.7.1 Optical and Acoustic Vibrations As a result of *lattice* vibrations, the position of atoms within the crystal can then be expressed as *traveling waves*. Solving the force equations for these *traveling waves* yields certain *dispersion* characteristics (i.e., ω or E vs. k dependence, where k is the *wave number*). An example is depicted in Figure 1.45. In this figure, a is the equilibrium distance between identical atoms along the direction of vibration.

In Figure 1.45, *two* branches of vibration can be identified. In the lower-frequency branch, known as the *acoustic phonon branch*, the frequency of vibration tends toward 0 as the *wave number* reduces to 0. This is the result of the movement of the *two sublattices* of the crystal along the same direction with equal velocity at any given time. However, if keeping the condition on velocity we reverse the direction of the movement of one *sublattice* versus the other, a high-frequency oscillation results (i.e., at k equal to 0). The branch identifying with this high-frequency mode is known as the *optical-phonon branch* (due to its higher frequency). This *branch* even for k equal to 0 is not producing 0 in its *dispersion* characteristic.

The *acoustical branch* represents the propagation of *sound waves* in a crystal. Using the *dispersion* characteristic of this branch, *sound velocity* is given by

$$v_{\text{s}} = \frac{d\omega}{dk} = \sqrt{\frac{C}{M_{\text{av}}}} a \qquad (1.150)$$

where M_{av}, a, and C stand for the average mass of the two atoms, minimum separation between identical planes in the crystal, and *spring constant* of the vibratory system, respectively.

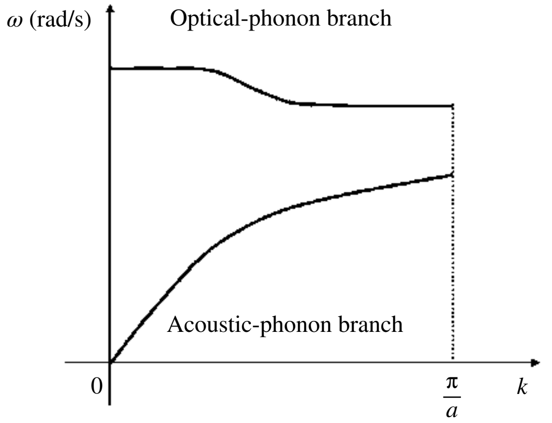


FIGURE 1.45 Schematic depiction of the dispersion diagram of a 1-D diatomic lattice, expressing the optical- and acoustic-phonon branches.

Whereas in Figure 1.45 we envisioned the vibrations in a one-dimensional *lattice*, in the case of three-dimensional *lattices*, a few other branches also appear on the *dispersion* characteristic. In reality, for a three-dimensional crystal, for each *wave vector one longitudinal* and *two transverse modes* of vibration will be present. This is true in the case of both *optical* and *acoustical* branches. Due to the difference in arrangement of atoms along different directions of most crystals, frequencies of vibration differ between the *longitudinal* and *transverse* branches.

DEFORMATION POTENTIAL AND BANDGAP VARIATION The distorting effect of *phonons* on the crystal, among a number of other appearances, takes the form of inducing a *deformation* potential. This *deformation* potential can be envisioned through the variation of the semiconductor *bandgap* (as the *lattice deformation* is causing the *interatomic* distance to change). According to this description, the displacement of the *lattice* by an amount *u* results in changing the energy of the *conduction* or *valence bands* in the form

$$\Delta E_{\mathrm{c,v}} = E_{\mathrm{c,v}}(a) - E_{\mathrm{c,v}}\left(a + \frac{du}{dx}a\right). \tag{1.151}$$

Here *a* is the *lattice constant*, which in the description of (1.151) has been assumed to be smaller than the wavelength of the *phonon* (i.e., *phonon* wavelength spans over many *lattice constants*, so that the displacement of the *lattice* can be taken for the expansion and contraction of the whole crystal). Through *Taylor* series expansion, (1.151) results in

$$|\Delta E_{\mathrm{c,v}}| = \frac{dE_{\mathrm{c,v}}}{da} \cdot \frac{du}{dx} a. \tag{1.152}$$

WAVE-FUNCTION SYMMETRY AND CRYSTAL VIBRATIONS In a 3-D *lattice* the volumetric changes induced by *acoustic* and *optical phonons* are fundamentally different. Therefore, *optical-phonon scattering* has been shown to be very sensitive to the symmetry of the *band* structure (i.e., in the range relevant to *carrier scattering*). This has very important implications on *carrier mobility*. As a result of this sensitivity, if an electron is *scattered* close to the Γ-*valley* minimum (e.g., in *GaAs*) or near the *X-valley* minima and has a spherically symmetric *wave function*, *optical deformation potential scattering* is forbidden. The existence of this symmetry at the *conduction band* minimum of *GaAs* and lack thereof in *Si* result in superior electron *mobility* in *GaAs*.

However, while for both *GaAs* and *Si* the top of the *valence band* happens at the Γ-*point*, for these cases the *hole wave function* does not have spherical symmetry (i.e., they possess *sp* rather than *s*-orbital nature). Therefore, the *optical deformation potential* is present for *holes* in both semiconductors, which results in low *hole mobility* for both cases.

POLAR OPTICAL AND PIEZOELECTRIC VIBRATIONS In *ionic crystals*, in which the atoms of the *two sublattices* are different (e.g., in the *zinc blende* crystal of *GaAs*), *optical* vibrations also induce *vibrating polarization fields*. This is due to the vibration of *cation* and *anion sublattices* in opposite directions to one another. Presence of these vibrating fields is important in *longitudinal* vibrations but not in *transverse* vibrations. Hence, in the case of *longitudinal* vibration, there is an additional restoring force due to the *long-range polarization*. As an example, Figure 1.46 shows the dispersion relationship of *GaAs*, in which the frequency of the *longitudinal optical* mode is higher than the *transverse optical* mode.

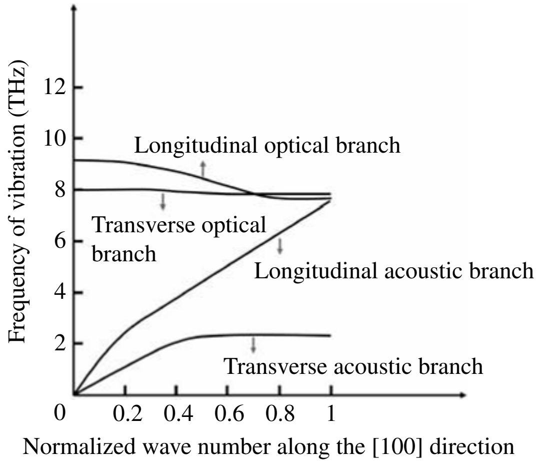


FIGURE 1.46 Approximate representation of phonon dispersion characteristics in *GaAs*. Adapted from Singh (2003, p. 228). Copyright 2003, Cambridge University Press. Reprinted with the permission of the Cambridge University Press.

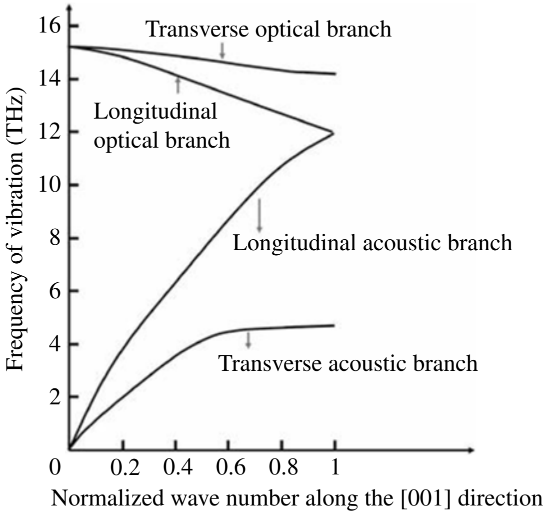


FIGURE 1.47 Approximate representation of phonon dispersion characteristics in *Si*. Adapted from Singh (2003, p. 228). Copyright 2003, Cambridge University Press. Reprinted with the permission of the Cambridge University Press.

These *vibrating polarization fields* are indicated by the presence of *polar optical phonons*. While *polar optical phonons* are not present in group *IV* semiconductors, in *III–V* semiconductors they play a very important role. This is why in group *IV* semiconductors there is no split between the *transverse* and *longitudinal optical* branches of the *dispersion* characteristic. Figure 1.47 illustrates this situation.

Optical phonons (especially *transverse optical phonons*) are much less dispersive than *acoustic phonons*. This is especially true at low *k* values. As shown in Figure 1.47, *longitudinal optical* and *longitudinal acoustic dispersions* tend toward the same value of angular frequency as *k* increases. However, this is not the case with transverse modes.

So far, we have only spoken of the role of *optical phonons* in inducing *polar* vibrations. Acoustic vibrations also cause *polar* vibrations, which are referred to as *piezoelectric effects*. *Piezoelectric scattering*, while much weaker than *polar optical-phonon scattering*, only becomes important in very high-purity samples at low temperatures.

1.7.3.7.2 Phonon Distribution Function Although we have already mentioned that *phonons* are quantum particles, at this point in our discussion, it is worthwhile to explain the quantum nature of these particles a little further. According to quantum mechanics, oscillation energy of crystal vibrations is not continuous. The *quantum oscillator* has a minimum energy of $\hbar\omega/2$, while the energy changes only in $\hbar\omega$ quanta, referred to as a *phonon*. With regard to this *quanta*, quantum number *n* refers to the occupation number of *phonons* in the system. Just like electrons, in order to evaluate the number of *phonons* in a given mode ω_k and at a given temperature *T*, a distribution function is required. However, because *phonons* do not obey *Pauli's exclusion principle*, they do not follow *Fermi–Dirac* statistics. Instead, the proper distribution function for them is defined in terms of *Bose–Einstein* statistics. Hence, *phonons*, rather than being *Fermions*, are referred to as *Bosons*. Because *phonons* are *Bosons*, their *occupation number* at *thermal equilibrium*, which is denoted as

$$\langle n_\omega \rangle = \frac{1}{\exp\frac{\hbar\omega}{kT} - 1}, \tag{1.153}$$

is not a probability of occupation unlike the case of *Fermions*.

As expressed by this distribution function, the number of *phonons* increases with the temperature, since *vibration* also becomes stronger at higher temperatures.

Obviously, at low temperatures the occupancy of the *optical phonons* will be very small. This is due to the fact that for any value of *k*, the energy of an *optical phonon* is large unlike that of the *acoustic phonon*. As a result, *acoustic-phonon scattering* is present even at low temperatures. For *optical-phonon scattering*, temperature or energy of the *carriers* must be beyond a certain value. This threshold of energy is determined by the *band* structure of the crystal.[101]

1.7.3.7.3 Quantum Mechanical Foundations of Phonon Scattering Electrons and *phonons* are treated similarly by solving the *Schrödinger equation* in a periodic potential. However, there is a qualitative difference between the two cases, which results from the difference in their *de Broglie wavelength*. In the case of electrons, only when the dimensions of the *quantum well* approach the *de Broglie wavelength* of the electron (i.e., ~10 nm) the *band offsets* at semiconductor *heterostructures* render the *heterostructure* effects important. However, the equivalent length scale in the case of *phonons* is about a few monolayers. This difference results in the development of *phonon* modes, associated with *interfaces* and *superlattices* of small periods.

As an example of the similarity between electrons and *phonons*, we can take a look at interface between *AlAs* and *GaAs*. This structure has a *Type-I band lineup*, which restricts the movement of *carriers* normal to the *heterointerface*. It also restricts the movement of the *optical phonons* from one material to the next. This is because the *optical* branches of the *dispersion* characteristics of *GaAs* and *AlAs* do not overlap. This is a fact that we will appreciate more after covering the materials in Chapter 2, where we get ourselves further acquainted with *junctions* and *interfaces*.

Phonons, *lattice* vibrations, and *band* structure are, to a first-order approximation, separated from one another. This treatment overlooks the obvious variations of the *band* diagram depicted in Figure 1.48, which are expressed in equations such as (1.151) and (1.152). As a result of this approximation, *scattering* by *phonons* is treated merely as a *perturbation*. Frequencies of lattice vibrations (i.e., *phonons*) are in the range of *terahertz*. Yet, in dealing with this electronic system, even in presence of *phonons*, it has been proven possible to imagine a *band* structure defined by a

[101] Previously in Section 1.4.4.2, we talked about *optical phonons* in relation to saturation of *drift velocity*.

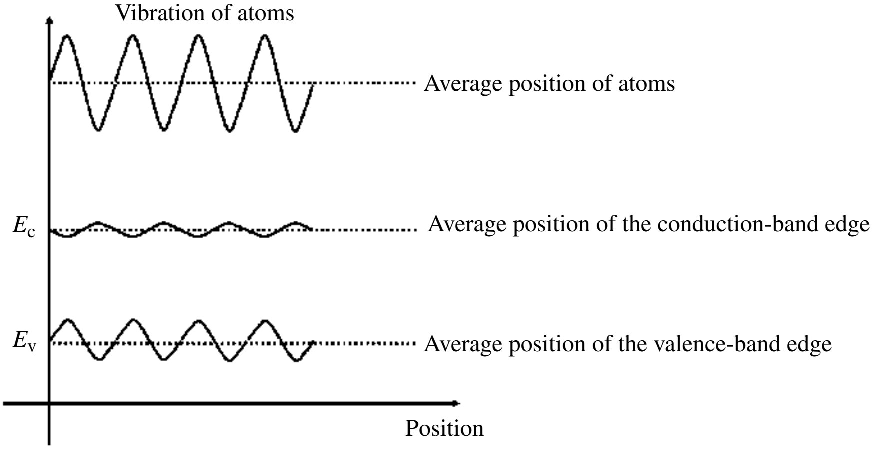

Vibration of atoms
Average position of atoms
E_c
Average position of the conduction-band edge
E_v
Average position of the valence-band edge
Position

FIGURE 1.48 Schematic depiction of position dependence of band edges due to lattice vibrations. Adapted from Singh (2003, p. 228). Copyright 2003, Cambridge University Press. Reprinted with the permission of the Cambridge University Press.

time-independent periodic potential. Due to the much smaller mass of electrons than atoms, which gives them a *de Broglie wavelength* on the order of 10 nm, this is a valid approximation.

As a result of the small *wavelength*, the electron frequency will be in the order of 10^{16} Hz (i.e., $\frac{c}{\lambda} = \frac{3 \times 10^8}{10^{-8}} = 3 \times 10^{16}$ Hz). This is about four orders of magnitude higher than the *phonon* frequency, which renders the *Hamiltonian* describing the system almost stationary. This approximation is referred to as *adiabatic approximation*.

In the *adiabatic approximation*, electrons see the impact of *phonon*-caused energy fluctuations in terms of *scattering* between existing *states*. These interactions are seen as processes of *emission* and *absorption* of *phonons*, depending on whether a loss or a gain in electron energy results. The *scattering* rate is deduced based on the *Fermi golden rule*.

Through *phonon scattering*, both the energy and momentum of *charge carriers* are changed. While the *wave vectors* of electrons and *phonons* are similar, the energy of an electron is much larger (i.e., due to larger frequency). These are very consequential points in establishing the energy and momentum conservation.

As a result of the small energy of *phonons*, under low electric fields, *interband scattering* of *holes* (and not electrons) is probable. This will be in the form of *scattering* between the *two degenerate heavy-* and *light-hole bands*. However, due to the small amount of energy of *phonons*, even *interband scattering* to the split-off band, which is only separated by several hundred meV, is not possible. For *interband scattering* to occur to the *split-off band*, the electric field should be much stronger. However, because of the much smaller difference between the *degenerate bands* and *split-off band* at the top of the *valence band* of *silicon* (i.e., 44 meV), this semiconductor is an exception. This is one of the contributing factors to the deterioration of *hole transport* in *silicon*.

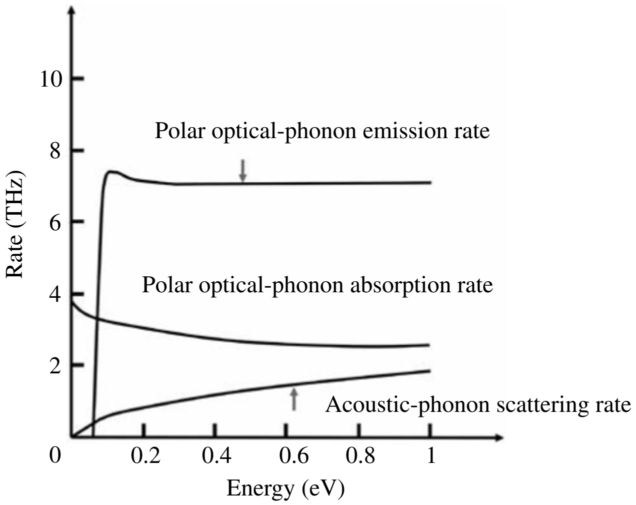


FIGURE 1.49 Comparison of absorption and emission rates of polar optical phonons with acoustic-phonon scattering rate of *GaAs* at room temperature. Adapted from Singh (2003, p. 228). Copyright 2003, Cambridge University Press. Reprinted with the permission of the Cambridge University Press.

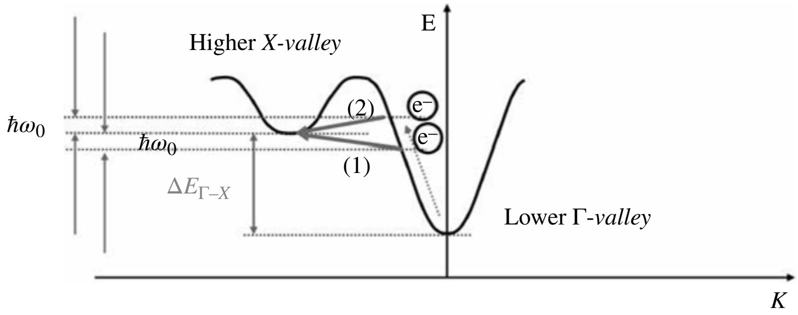


FIGURE 1.50 Schematic depiction of the possibility of an energetic electron scattering from the central valley to the *X-valley* through (1) absorption or (2) emission of a phonon of energy $\hbar\omega_0$.

In *III–V* semiconductors, while at low temperatures and low electric fields *polar optical-phonon scattering* is not important, at room temperature and also in the presence of high electric fields (even at low temperatures), the process of *emission* of *polar optical phonons* becomes the dominant *scattering* mechanism. As an example see Figure 1.49 for an approximate depiction of the rates of *acoustic* and *polar optical-phonon scattering* processes in *GaAs* at room temperature.

The processes of *phonon emission* and *absorption* in *intervalley scattering* are illustrated in Figure 1.50. As mentioned in Section 1.4.2.2, the *intervalley scattering* process plays a very important role in manifesting *negative differential mobility* in some semiconductors.

Now that we have visited a number of important *scattering* mechanisms, it is time to point out that there are important differences between *ionized-impurity scattering* and *phonon scattering*. The first important difference is that in regard to *phonon scattering*, we have both *absorption* and *emission* processes, which result in variation of the *wave functions* of both the *phonon* and the electron. The second important difference between the *phonon scattering* and *ionized-impurity scattering* is rooted in the formation of a *scattering* potential. As we have seen already in this section, the distorting effect of *phonons* on the crystal *lattice* can take a number of different forms (i.e., *deformation* potential, *piezoelectric* potential, and *polar optical* potential). Due to these differences, dealing with *ionized-impurity scattering* is much less complicated.

1.7.3.8 Carrier Scattering in Lower-Dimensional Systems So far in our discussions, we have maintained a focus on *carrier transport* in three-dimensional systems. In lower-dimensional systems, in addition to the aforementioned processes of *scattering*, due to the differences in *DOS* functions, a few additional *scattering* mechanisms become important.

As a special case, we have already identified *interface-roughness scattering* as a *scattering* process important only to two-dimensional *carrier transport*. Such a *scattering* mechanism is also sometimes present when *carrier transport* is confined to a one-dimensional *channel* (i.e., a *quantum wire*). In such lower-dimensional systems, due to the development of energy *subbands* (i.e., which are expressed by the modification in the 2-D *DOS* presented in (1.114) and (1.115) and also the so-called minibands in smaller lower-dimensional systems), both *intra*- and *intersubband scattering* are present. In a 2-D system, the larger the energy difference between the *subbands*, the less probable would be the *intersubband scattering*.

The *DOS* in 1-D systems (such as *quantum wires*) is qualitatively different from 3-D and 2-D systems, which results in very important implications on *carrier transport*. As illustrated in Figure 1.51, in the case of a 1-D system, *intrasubband scattering* is very restricted, while *elastic scattering* is considered. As depicted in this figure, in *elastic scattering* electron can *scatter* only to a *state* with the same or opposite momentum (i.e., k and $-k$). While the former results in no change on *transport*, the latter event requires a very short-ranged potential to conserve momentum. This requirement severely limits the *scattering*. Consequently, *mobilities* as high as 10^7 cm^2/V s are predicted for *quantum wires*. However, due to fabrication and crystal growth difficulties, this has not been experimentally observed yet.

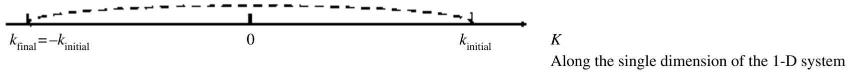

$k_{\text{final}}=-k_{\text{initial}}$
0
k_{initial}
K
Along the single dimension of the 1-D system

FIGURE 1.51 In the event of elastic scattering in a 1-D semiconductor, the equal energy surface has only two viable k-states. In the case of 2-D or a 3-D system, this equal energy surface evolves into the surface of a circle or a sphere, respectively, of radius k_{initial} containing many more k-states available to scattering. Adapted from Singh (2003, p. 228). Copyright 2003, Cambridge University Press. Reprinted with the permission of the Cambridge University Press.

FURTHER READING

Following references have been used in preparation of the materials presented in this chapter. They are also suggested as sources for further reading.

SOLID-STATE THEORY

S. Datta, *Quantum Phenomena*, Addison-Wesley, Reading, MA, 1989.

S. Datta, *Electronic Transport in Mesoscopic Systems*, Cambridge University Press, Cambridge, UK, 1995.

K. Hess, *Advanced Theory of Solid State Devices*, IEEE Press, Piscataway, NJ, 2000.

C. Kittle, *Introduction to Solid State Physics*, John Wiley & Sons, Inc., Hoboken, NJ, 2005.

P. T. Landsberg (Ed), *Solid State Theory: Methods and Applications*, John Wiley & Sons, Inc., Hoboken, NJ, 1969.

M. Lundstrom, *Fundamentals of Carrier Transport*, Cambridge University Press, Cambridge, UK, 2000.

J. Singh, *Electronic and Optoelectronic Properties of Semiconductor Structures*, Cambridge University Press, Cambridge, UK, 2003.

PHYSICS OF SEMICONDUCTOR DEVICES

R. S. Muller, T. I. Kamins, and M. Chan, *Device Electronics for Integrated Circuits*, John Wiley & Sons, Inc., Hoboken, NJ, 2003.

M. Lundstrom and J. Guo, *Nanoscale Transistors: Device Physics, Modeling and Simulation*, Springer, New York, 2006.

W. Shockley, *Electrons and Holes in Semiconductors with Applications to Transistor Electronics*, Van Nostrand, New York, 1950.

S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, John Wiley & Sons, Inc., Hoboken, NJ, 2007.

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

SEMICONDUCTOR MATERIALS AND HETEROSTRUCTURES

S. Adachi, *Properties of Semiconductor Alloys: Group IV, III-V, and II-VI Semiconductors*, John Wiley & Sons, Inc., Hoboken, NJ, 2009.

Ioffe Physico-Technical Institute, Electronic Archive of New Semiconductor Materials, Characteristics and Properties: http://www.ioffe.rssi.ru/SVA/NSM/. Accessed September 18, 2015.

M. Levinshtein, S. Rumyantsev, and M. Shur, *Properties of Advanced Semiconductor Materials: GaN, AlN, InN, BN, SiC, SiGe*, John Wiley & Sons, Inc., Hoboken, NJ, 2001.

Y. Sun, S. Thompson, and T. Nishida, *Strain Effect in Semiconductors: Theory and Device Applications*, Springer, New York, 2009.

PROBLEMS

1.1 Using the time-independent 3-D Schrödinger equation, prove that for an electron in a 3-D infinite potential well described by $U(x,y,z) = \begin{cases} 0 & \text{for } 0 \le x \le W,\ 0 \le y \le W, 0 \le z \le W \\ \infty & \text{elsewhere} \end{cases}$, the allowed energy levels are expressed by $E = \frac{\hbar^2 \pi^2}{2m^* W^2}\left(n^2 + m^2 + p^2\right)$ where n, m, and p are integers.

Wave function and its first spatial derivative are continuous, single-valued functions.

1.2 Demonstrate that the Fermi level of an intrinsic semiconductor instead of residing in the middle of the bandgap has a slight offset from this position, which is determined by the following expression: $\frac{3}{4}kTLn\left(\frac{m_{\mathrm{p}}^*}{m_{\mathrm{n}}^*}\right)$.

1.3 According to (1.21) calculate the average thermal energy of Maxwellian particles having a single degree of freedom. What is the thermal energy if the particles were allowed to move in the 2-D or the 3-D space?

1.4 Calculate the effective density of states at the lower edge of the conduction-band N_{c}, for silicon at room temperature. Notice the difference in the definitions of transport and *DOS* effective masses and take the transverse and longitudinal effective mass to be equal to $0.19m_0$ and $0.98m_0$, respectively.

1.5 Prove Equation (1.32).

1.6 For a silicon sample doped with arsenic, only a quarter of the impurities are ionized at room temperature. Elaborate on the cause(s) of the observation and describe how the electron concentration is calculated in this case.

1.7 Prove Equations (1.78) and (1.79).

1.8 Formulate the Hall effect when the semiconductor is behaving semi-intrinsically. Presentation of the Section 1.4.3 was provided with the assumption of a significant extrinsic characteristic.

1.9 Perform a literature survey on the values of the magnetic flux density often used in Hall effect measurements and indicate why Hall measurements are not performed under high magnetic fields.

1.10 Employing *WKB* approximation, calculate the tunneling probability of an electron of energy 1 eV across a rectangular barrier of height 5 eV and width 1 nm. Repeat this problem for a triangular barrier of the same height and also the same width at the energy level of the electron.

1.11 Prove that assuming the same effective mass for electrons and holes, an electron requires 50% more kinetic energy than the size of the bandgap to instigate impact ionization.

1.12 Prove that to guarantee maximum sphere packing, the *c*/*a* ratio of the hexagonal prism of a wurtzite crystal should be equal to $\sqrt{8/3}$.

1.13 Calculate the entries of Table 1.7.

1.14 Indicate the plane $(\bar{1}100)$ on a hexagonal unit cell.

1.15 Perform a literature survey to determine the type of the band lineup between the following pairs:

a. *InP*/*InGaAs*.

b. *AlGaN*/*GaN*.

c. *GaSb*/*InAs*.

1.16 Perform a literature survey to determine which material system is more suitable for solar cell applications:

a. AlInGaAs

b. AlInGaN

Determine the energy coverage of the bandgap of each compound and compare to solar spectrum.

1.17 In strained epitaxy we often simplistically define the critical thickness of the overlayer causing the generation of dislocations within this film as $\frac{a_S}{2|\epsilon|}$. Considering the pseudomorphic growth of $In_{0.2}Ga_{0.8}As$ over a thick relaxed layer of *GaAs*, determine the critical thickness of *InGaAs*. Perform the calculations at room temperature for which $a_{InAs} = 6.058$ Å and $a_{GaAs} = 5.653$ Å.

1.18 Through producing schematic depictions of a 1-D lattice, demonstrate why the dispersion diagram of the acoustic-phonon branch tends toward 0 at small values of wave number while for the optical-phonon branch the energy remains sizable (see Fig. 1.45).

APPENDIX 1.A DERIVATION OF *FERMI–DIRAC* STATISTICS

The key point in the derivation of *Fermi–Dirac* statistics is to acknowledge that *Fermions* (electrons included) are indistinguishable particles obeying *Pauli's exclusion principle*. According to this definition, the total number of ways that N_i *Fermions* can be arranged among g_i *states* corresponding to *i*th energy *level* (which are determined in terms of *density of states* function) is given by

$$P_i = \frac{g_i!}{N_i!(g_i - N_i)!}. \tag{1A.1}$$

According to *Pauli's exclusion principle*, g_i is greater than or equal to N_i.

Since the *density of states* function is a double-density function (i.e., representing the number of *states* per unit energy per unit volume), in studying the *thermal-equilibrium* distribution of *Fermions* among these *states*, we should not only consider the *states* corresponding to the *i*th energy *level* but also all those other *states* assigned to other values of energy. Accordingly, considering a total of N electrons, where N_i of which take the *i*th *energy level*, we will have

$$N = \sum_{i=1}^{n} N_i \tag{1A.2}$$

where n is the number of energy-wise distinguishable *levels*.

In terms of this distribution, the total energy is given by

$$E = \sum_{i=1}^{n} N_i E_i \tag{1A.3}$$

where E_i is the energy of the *i*th *level*.

As a result of (1A.2), the total number of ways of arranging $N_1, N_2, \ldots N_n$ indistinguishable particles among n energy *levels* would be

$$P = \prod_{i=1}^{n} P_i = \prod_{i=1}^{n} \frac{g_i!}{N_i!(g_i - N_i)!}. \tag{1A.4}$$

On the basis of (1A.2) and (1A.3), in order to identify the distribution function presenting the least amount of systemic energy (i.e., representative of *thermal equilibrium*), the most likely distribution function should be selected. Toward that end N_i's should be arranged so that P is maximized (i.e., representing the most plausible distribution) for a fixed N.

According to the discussions of Section 1.3.5, N_i can be replaced by $g_i f_i$. This substitution results in

$$P = \prod_{i=1}^{n} \frac{g_i!}{(g_i f_i)!(g_i - g_i f_i)!}. \tag{1A.5}$$

Using *Stirling*'s approximation of *factorials* and assuming that n is greater than 20,

$$n! \cong \sqrt{2\pi n} n^n \exp\left[-n + \frac{1}{12n}\right], \tag{1A.6}$$

which gives

$$Ln(n!) \cong nLn(n) - n + \frac{1}{2}Ln(2\pi n). \tag{1A.7}$$

Neglecting the third term on the right side of (1A.7), (1A.5) can be rewritten as

$$Ln(P) \cong \sum_{i=1}^{n} Ln(P_i) = \sum_{i=1}^{n} [g_i Ln(g_i) - g_i f_i Ln(g_i f_i) - g_i(1 - f_i)Ln(g_i - g_i f_i)]. \tag{1A.8}$$

This has been proven to be a valid approximation.

To maximize *P*, according to the *Lagrange* method of *undetermined multipliers* and Equations (1A.2) and (1A.3), the following function should be maximized by the appropriate choice of α and β to get *P* or *Ln*(*P*) maximized

$$f(g_i f_i) = Ln(P) - \alpha \sum\nolimits_i g_i f_i - \beta \sum\nolimits_i E_i g_i f_i. \tag{1A.9}$$

As a result, the derivative of function $f(g_i f_i)$ versus $g_i f_i$ is set to 0 in order to calculate α and β:

$$\frac{\partial}{\partial(f_i g_i)} \left[Ln(P) - \alpha \sum\nolimits_i g_i f_i - \beta \sum\nolimits_i E_i g_i f_i \right] = 0. \tag{1A.10}$$

Assuming fixed g_i (which is determined by the *density of states* function of the semiconductor) and through the following mathematical manipulation, f_i results in

$$\begin{aligned} & -Ln(g_i f_i) + Ln(g_i - f_i g_i) - \alpha - E_i \beta = 0 \\ & \Rightarrow Ln\left(\frac{g_i - g_i f_i}{g_i f_i}\right) = \alpha + \beta E_i \Rightarrow f_i = \frac{1}{1 + \exp(\alpha + \beta E_i)}. \end{aligned} \tag{1A.11}$$

As expressed below, the function f_i is a *Fermi–Dirac* distribution function with $E_{\mathrm{f}} = -\alpha/\beta$:

$$f_{\mathrm{D}}(E_i) = \frac{1}{1 + \exp\left(\dfrac{E_i - E_{\mathrm{f}}}{1/\beta}\right)}. \tag{1A.12}$$

Knowing that $E = \sum\nolimits_{i=1}^{n} E_i f_i g_i$,

$$dE = \sum\nolimits_{i=1}^{n} E_i d(f_i g_i) + \sum\nolimits_{i=1}^{n} f_i g_i d(E_i). \tag{1A.13}$$

Besides from (1A.10),

$$\frac{1}{\beta} \partial(LnP) = \sum\nolimits_{i=1}^{n} E_i d(f_i g_i) + \frac{\alpha}{\beta} \sum\nolimits_{i=1}^{n} d(f_i g_i) \tag{1A.14}$$

where $E_{\mathrm{f}} = -\alpha/\beta$ and $\sum\nolimits_{i=1}^{n} d(f_i g_i) = dN$.

As a result, (1A.14) can be rewritten as

$$dE = \frac{1}{\beta} \partial Ln(P) + E_{\mathrm{f}} dN + \sum\nolimits_{i=1}^{n} f_i g_i d(E_i). \tag{1A.15}$$

Through incorporating dV as a volumetric variation,

$$dE = \frac{1}{\beta}\partial Ln(P) + E_{\mathrm{f}}dN + \sum\nolimits_{i=1}^{n} f_i g_i \frac{dE_i}{dV}dV. \tag{1A.16}$$

This relationship is similar to thermodynamic identity of $dE = TdS - PdV + \mu dN$. According to this analogy, $\beta = 1/kT$, $S = kLnP$, and μ (the energy of the *Fermions*) is represented by E_{f}.

As a result (1A.12) evolves into the familiar form of

$$f_{\mathrm{D}}(E) = \frac{1}{1 + \exp\left(\frac{E - E_{\mathrm{f}}}{kT}\right)} \tag{1A.17}$$

FURTHER READING

G. Fournet, English edition edited by S. Chomet, *Solid State Electronics (first published in French: Physique Électronique des Solids)*, Iliffe Books, London, 1968.

APPENDIX 1.B DERIVATION OF EINSTEIN RELATIONSHIP IN DEGENERATE SEMICONDUCTORS

In the case of one-dimensional *carrier transport* and according to the definition of *thermal equilibrium*,

$$J_n = qn\mu_n E + qD_n \frac{dn}{dx} = 0 \tag{1B.1}$$

where $E = \frac{1}{q}\frac{dE_i}{dx}$ or equivalently $E = \frac{1}{q}\frac{dE_{\mathrm{c}}}{dx}$.

In addition, based on the discussions in Section 1.3.5, $n = N_{\mathrm{c}}\mathrm{F}_{1/2}(\eta_{\mathrm{F}})$ where $\eta_{\mathrm{F}} = \frac{E_{\mathrm{f}} - E_{\mathrm{c}}}{kT}$.

Resulting from the definition of electric field E and n,

$$\frac{dn}{dx} = -\frac{1}{kT}\frac{dn}{d\eta_{\mathrm{F}}}\frac{dE_i}{dx} = -\frac{q}{kT}\frac{dn}{d\eta_{\mathrm{F}}}E. \tag{1B.2}$$

Based on (1B.1) and (1B.2), the following general form of the *Einstein* relationship results:

$$\frac{D_{\mathrm{n}}}{\mu_{\mathrm{n}}} = \frac{kT}{q}\frac{n}{(dn/d\eta_{\mathrm{F}})} \tag{1B.3}$$

in which $n = N_{\mathrm{c}}\mathrm{F}_{1/2}(\eta_{\mathrm{F}})$ and $\mathrm{F}_{1/2}(\eta) = (2/\sqrt{\pi})F_{1/2}(\eta)$ where $F_{1/2}(\eta) = \int_0^\infty \frac{\xi^{1/2}d\xi}{(1 + e^{\xi - \eta})}$.

A relationship identical in form to (1B.3) can also be developed for *holes*.

In general,

$$\mathrm{F}_j(\eta) \equiv \frac{1}{\Gamma(j+1)}\int_0^{\infty} \frac{\xi^j d\xi}{(1+e^{\xi-\eta})} \tag{1B.4}$$

$$\mathrm{F}_j(\eta) \rightarrow e^{\eta} \text{ as } \eta \rightarrow -\infty \tag{1B.5}$$

$$\frac{d}{d\eta}\mathrm{F}_j = \mathrm{F}_{j-1}(\eta) \tag{1B.6}$$

$$\mathrm{F}_{1/2}(\eta) \cong [e^{-\eta} + \xi(\eta)]^{-1} \tag{1B.7}$$

where

$$\xi(\eta) = 3\sqrt{\frac{\pi}{2}}\left[(\eta+2.13) + \left(|\eta-2.13|^{2.4}+9.6\right)^{5/12}\right]^{-3/2} \tag{1B.8}$$

with a maximum error of ~ ±0.5%,

$$\eta \cong \frac{Ln(U)}{1-U^2} + \frac{(3\sqrt{\pi}U/4)^{2/3}}{1+\left[0.24+1.08(3\sqrt{\pi}U/4)^{2/3}\right]^{-2}} \tag{1B.9}$$

where $U \equiv \mathrm{F}_{1/2}(\eta)$ with a maximum error of ~ ±0.5%.

A convenient approximation for $(n)/(dn/d\eta_\mathrm{F})$ is presented in Nilsson (1978), in the form of $\mathrm{F}_{1/2}(\eta_\mathrm{F})/\mathrm{F}_{-1/2}(\eta_\mathrm{F})$.

In its *nondegenerate* limit, n becomes equal to $N_\mathrm{c}\exp(\eta_\mathrm{F})$ and $(n)/(dn/d\eta_\mathrm{F})$ tends toward *one*. As a result (1B.3) takes on to its familiar form,

$$\frac{D_n}{\mu_n} = \frac{kT}{q}. \tag{1B.10}$$

While this derivation relies on *thermal-equilibrium* conditions, it can be shown that the *Einstein* relationship is also valid under *nonthermal-equilibrium* conditions.

FURTHER READING

N. G. Nilsson, "Empirical approximation applied to generalized Einstein relation for degenerate semiconductors," Phy Stat Sol (a), vol. **50**, 1978b, K43–K45.

R. F. Pierret, *Advanced Semiconductor Fundamentals, vol. VI Modular Series in Solid State Devices*, Prentice Hall, Upper Saddle River, NJ, 2003.

APPENDIX 1.C STRAIN TENSOR

Deformations in the crystal structure result in induction of strain. Strain is defined in terms of the relative change in the *lattice constants* with regard to the freestanding *lattice constants* of a crystal. Figure 1.C.1 illustrates a simple 2-D *lattice* under strain. The outcomes of this pictorial insight can be mathematically represented in the 3-D form with the use of the *strain tensor*.

As illustrated in Figure 1.C.1a, in the 2-D case we can use two *unit vectors* $\hat{x}$ and $\hat{y}$ of the *Cartesian* coordinate system to represent the *unstrained lattice*. In a simple *lattice*, these vectors correspond to the *basis vectors* of the *lattice*. A small uniform deformation of the *lattice* results in distortion of the *unit vectors* both in magnitude and orientation (see Fig. 1.C.1b). These distortions result in a new set of vectors identified by

$$\hat{x'} = (1+\epsilon_{xx})\hat{x} + \epsilon_{xy}\hat{y} + \epsilon_{xz}\hat{z} \tag{1C.1}$$

$$\hat{y'} = \epsilon_{yx}\hat{x} + \left(1+\epsilon_{yy}\right)\hat{y} + \epsilon_{yz}\hat{z}. \tag{1C.2}$$

In the 3-D case, we also have a distortion with regard to the z-axis of the coordinate system,

$$\hat{z'} = \epsilon_{zx}\hat{x} + \epsilon_{zy}\hat{y} + (1+\epsilon_{zz})\hat{z}. \tag{1C.3}$$

Elements ϵ_{ij} are referred to as *strain coefficients*. These coefficients are dimensionless. The 3×3 matrix composed of these elements is referred to as *strain tensor*:

$$\bar{\bar{\varepsilon}} = \begin{bmatrix} \epsilon_{xx} & \epsilon_{xy} & \epsilon_{xz} \\ \epsilon_{yx} & \epsilon_{yy} & \epsilon_{yz} \\ \epsilon_{zx} & \epsilon_{zy} & \epsilon_{zz} \end{bmatrix}. \tag{1C.4}$$

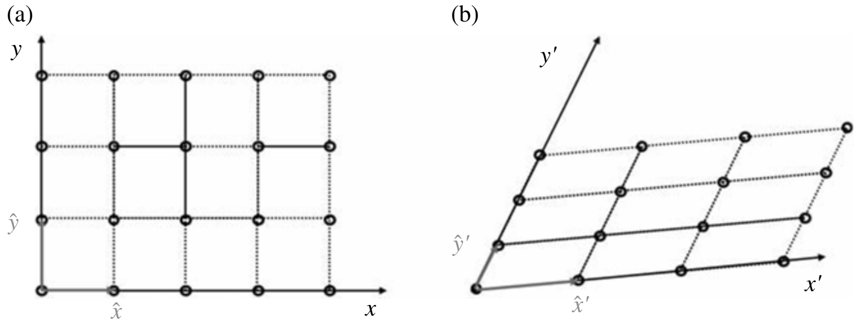


FIGURE 1.C.1 (a) Schematic depiction of a 2-D *Bravais* lattice. (b) Arbitrary deformation of the lattice drawn in (a).

Generally speaking, *tensor* is a mathematical notion used to describe a linear relationship between two physical quantities. Depending on the degrees of freedom, a *tensor* can be a *scalar* quantity (i.e., *zero rank*), a vector (i.e., *first rank*), or as in this case a matrix (i.e., *second rank*). As an example, we can look into the strain-caused deformation with regard to a *lattice point* represented by the vector $\vec{r} = x\hat{x} + y\hat{y} + z\hat{z}$. Considering a uniform deformation to evolve this point to $\vec{r'} = x\hat{x'} + y\hat{y'} + z\hat{z'}$ for a general varying strain, the *strain tensor* is formulated as

$$\epsilon_{ij} = \frac{\partial u_i}{\partial x_j} \tag{1C.5}$$

where $u_i = u_x, u_y, u_z$, $x_j = x, y, z$.

In this definition, u_i refers to the displacement of the *lattice* point (e.g., $\vec{r}$) along x_i.[102] According to this definition, it is obvious that without rotation, the *strain tensor* is symmetric and

$$\epsilon_{ij} = \epsilon_{ji} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{1C.6}$$

Oftentimes, instead of the aforementioned set of *strain tensor* components, the following are used:

$$\begin{cases} e_{xx} = \epsilon_{xx}; e_{yy} = \epsilon_{yy}; e_{zz} = \epsilon_{zz} \\ e_{xy} = \hat{x'} \cdot \hat{y'} = \epsilon_{xy} + \epsilon_{yx} \\ e_{yz} = \hat{y'} \cdot \hat{z'} = \epsilon_{yz} + \epsilon_{zy} \\ e_{zx} = \hat{z'} \cdot \hat{x'} = \epsilon_{zx} + \epsilon_{xz} \end{cases}. \tag{1C.7}$$

As presented in (1C.7), the *strain components* e_{xy}, e_{yz}, and e_{zx} are defined with respect to changes of angle between the *basis* vectors. This definition is provided with neglecting the terms of order ϵ_{ij}^2 in the small-strain approximation.

The above six *strain coefficients*, shown by an array $e = \{e_{xx}, e_{yy}, e_{zz}, e_{yz}, e_{zx}, e_{xy}\}$, offer a complete definition of strain. This set provides a more convenient way for describing the relationship between the strain and the strain-related physical quantities. The form of presentation expressed in (1C.4), however, gets complicated very quickly. This is because the relationship between two *second-rank tensors* (i.e., one for the *strain tensor* and one for the strain-related physical quantity) is representable through a *fourth-rank tensor*. However, describing each of the two *second-rank tensors* by a vector only requires dealing with a *second-rank tensor* for evaluation of the interactions.

[102] According to (1C.1)–(1C.3), $\vec{r'} = [x(1+\epsilon_{xx}) + y\epsilon_{yx} + z\epsilon_{zx}]\hat{x} + [x\epsilon_{xy} + y(1+\epsilon_{yy}) + z\epsilon_{zy}]\hat{y} + [x\epsilon_{xz} + y\epsilon_{yz} + z(1+\epsilon_{zz})]\hat{z}$. Hence, as an example, $u_x = x(1+\epsilon_{xx}) + y\epsilon_{yx} + z\epsilon_{zx} - x$, and as a result $\partial u_x/\partial x$ would be equal to ϵ_{xx}.

In terms of the *first-rank tensor* description of (1C.7), the *dilation* of the *unit cell* of the crystal can be simply evaluated through calculating the volume of the *unit cell* by

$$V' = \hat{x}'.\hat{y}' \times \hat{z}' = 1 + e_{xx} + e_{yy} + e_{zz}. \tag{1C.8}$$

Equation (1C.8) shows the *dilation* δ as

$$\delta = \frac{\delta V}{V} = e_{xx} + e_{yy} + e_{zz}. \tag{1C.9}$$

Interestingly enough, this is the *trace* of the *strain tensor*. This *dilation* is the same as the negative of the *hydrostatic pressure*.

It is worthwhile indicating that under a *hydrostatic pressure P*, the *shear* stress is 0 and the stress along any principal direction is equal to $-P$,

$$\tau = \begin{bmatrix} -P & 0 & 0 \\ 0 & -P & 0 \\ 0 & 0 & -P \end{bmatrix}. \tag{1C.10}$$

According to this sign convention, the *tensile* stress is indicated with a positive sign, while the *compressive* stress is indicated in terms of negative values.

In the case of the uniaxial stress T along the [*001*] direction, all stress components but τ_{zz} are 0, and $\tau_{zz} = T$. In other words,

$$\tau = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & T \end{bmatrix}. \tag{1C.11}$$