1

Semiconductor Physics

1.1 Introduction

The definition of semiconductors is somewhat broad. For instance, semiconductors may be defined as materials possessing a bandgap (a fundamental concept to be explained later, signifying the minimum electronic excitation energy) ranging between 0 and 4 eV. Materials characterized by a zero bandgap are classified as metals or semimetals, whereas those featuring energy gaps exceeding 4 eV are termed insulators. Alternatively, semiconductors can be defined by their electrical conductivity, which lies between that of conductors and insulators, typically within the range of approximately 10^{-2}–10^{9} Ω cm. Unlike metals, the conductivity of semiconductors generally increases with elevated temperature. However, these definitions are not absolute. The conductivity of semiconductors, for example, can be drastically modified via doping—a technique involving the deliberate introduction of specific foreign atoms into the semiconductor lattice to modulate its conductive properties. Consequently, crystalline diamond, exhibiting a bandgap near 6 eV and typically considered an insulator, can be engineered into a functional semiconductor suitable for potential power electronic devices.

Common examples of semiconductors include silicon, germanium, and gallium arsenide, among others. Undoubtedly, silicon remains the most extensively studied and widely recognized semiconductor. It serves as the primary functional material in numerous electronic applications, including solar cells, analog circuits, and large-scale integrated circuits. The prominence of silicon stems from several key factors: its natural abundance, moderate energy bandgap, established doping techniques, and the exceptional stability of both silicon and its native oxide, silicon dioxide. Semiconductor properties critically depend on their chemical compositions and crystal structures, which fundamentally govern respective band structures, doping characteristics, and carrier transport mechanisms. As silicon-based electronic devices approach material performance limits while semiconductor markets demand enhanced device performance, research focus increasingly shifts toward alternative materials such as wide-bandgap semiconductors and oxide semiconductors. The fundamental physics of these semiconductors is comprehensively addressed within this chapter.

1.2 Properties of Semiconductors

The importance of semiconductors in electronics stems from their unique physical properties, determined by their atomic structure and governed by the physics of semiconductor materials. An intrinsic semiconductor exhibits very low conductivity. However, introducing a relatively small concentration of impurity atoms, known as doping, enables modulation of the material's conductivity over a large dynamic range. Crucially, depending on the type of dopant employed—such as phosphorus or boron in silicon—semiconductors are categorized as *n*-type or *p*-type, dictating whether electrical current conduction is mediated primarily by electrons or holes, respectively. Moreover, semiconductor conductivity can be significantly modulated by applied electric fields. Collectively, these properties render semiconductor materials exceptionally valuable for constructing electronic devices, enabling the realization of critical functionalities including unidirectional current flow, variable resistance, variable capacitance, photosensitivity, voltage-controlled conductivity, signal amplification, switching operations, and energy conversion.

1.2.1 Atomic Bonding of Semiconductors

The properties of semiconductors are predominantly governed by their chemical compositions and structures, which, in accordance with semiconductor physics, subsequently dictate their overall characteristics. Generally, atomic bonds can be classified into ionic, covalent, metallic, van der Waals, and hydrogen bonding, as illustrated in Figure 1.1, contingent upon electron transfers and the charge states of the constituent atoms. Certain atoms exhibit a strong proclivity to attain

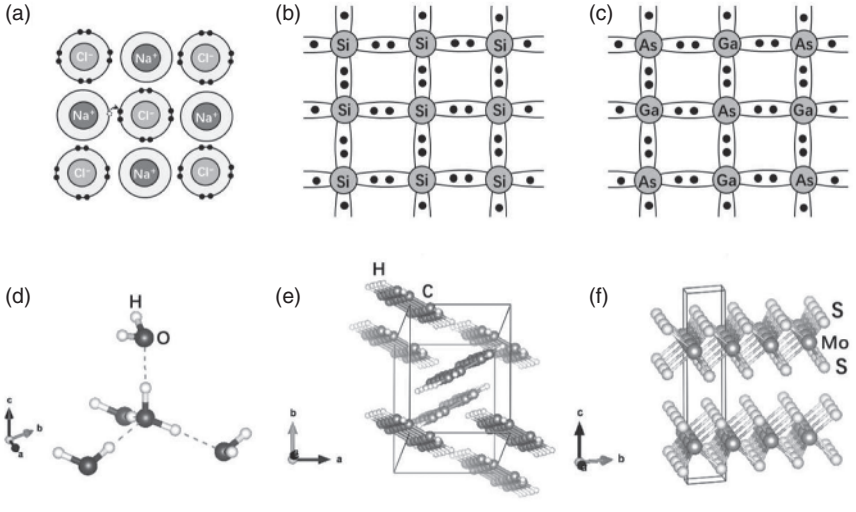

Figure 1.1 (a) Schematic of NaCl crystal with ionic bonding. (b) Schematic of Si crystal with covalent bonding. (c) Schematic of GaAs crystal with covalent bonding. (d) Hydrogen bonding between water molecules. (e) Pentacene and (f) MoS_2 with van der Waals bonding.

a completed outer electron shell by either gaining or losing a limited number of valence electrons, thereby forming charged ions. Ionic bonding—exemplified by the NaCl crystal structure shown in Figure 1.1a—arises from Coulombic attraction between positively and negatively charged ions. Ionic bonding also plays a significant role within semiconductors. For instance, in perovskite $CsPbBr_3$, ionic bonding occurs between Cs^+ ions and the surrounding $PbBr_3^-$ network, while it has been established that the bonding character between Pb and Br is metavalent, intermediate between covalent and metallic [1].

Covalent bonding is formed through the sharing of electrons between atoms, rather than through the gain or loss of electrons. The shared electrons create a bond that holds the atoms together. This bonding mechanism is essential for many important elemental and compound semiconductors (Figure 1.1b,c). Examples of elemental semiconductors involving covalent bonding include silicon (Si), germanium (Ge), and semiconducting single-walled carbon nanotubes (SWCNTs). In compound semiconductors, covalent bonding produces numerous critical materials. For instance, III–V semiconductors result from covalent bonding between Group III and Group V elements, such as gallium nitride (GaN), gallium phosphide (GaP), gallium arsenide (GaAs), and indium phosphide (InP). Similarly, II–VI semiconductors are formed by covalent bonding between Group II and Group VI elements, including zinc oxide (ZnO), zinc sulfide (ZnS), mercury telluride (HgTe), and cadmium telluride (CdTe). The bonding in these semiconductors typically exhibits a coordination number of 4, meaning each atom has four nearest neighbors, forming a tetrahedral structure.

When all atoms release their valence electrons, metallic bonding is formed, with the free electron fluid acting as the bonding agent. Although typically characteristic of metals, metallic bonding can also contribute to the lattice mechanical strength of semiconductors with a very high density of free carriers. Hydrogen bonding resembles ionic bonding and results from the attraction between a positively charged proton involved in a covalent bond and a nearby negatively charged atom (Figure 1.1d). This bonding type is highly localized due to the small size of the proton.

For atoms or molecules characterized by saturated covalent bonds, bonding can still arise via van der Waals interactions, including permanent dipole-dipole attractions and induced dipole interactions (Figure 1.1e,f). Many organic semiconductors, such as pentacene and polyacetylene, whose constituent building blocks are organic molecules, are primarily held together by van der Waals forces [2]. van der Waals forces are also responsible for the stacking of two-dimensional materials, such as transition metal dichalcogenides [3].

1.2.2 External Modulation Effects

Semiconductors possess a medium-sized bandgap, distinguishing them from metals and insulators. A precise understanding of the bandgap concept necessitates quantum physics, which describes electron motion within a crystal lattice. Broadly defined, the bandgap represents the minimum energy required for electronic excitation. Due to the presence of this bandgap, the intrinsic conductivity of a

pure semiconductor is markedly low, as the excitation of free carriers necessitates overcoming the bandgap energy. This energy is typically substantial relative to thermal fluctuation energy at room temperature. Nevertheless, the conductivity of semiconductors can be profoundly modulated through the controlled introduction of foreign impurity atoms, known as dopants. Dopants significantly enhance the concentration of free charge carriers within the crystal lattice. Semiconductors accommodate two types of charge carriers, namely electrons and holes; doping processes associated with generating these carriers are designated as *n*-type and *p*-type, respectively. This versatility in controlling carrier type and concentration enables the fabrication of diverse semiconductor junctions. These junctions form the foundation for realizing electronic devices exhibiting specific functionalities, such as the rectifying properties characteristic of a PN junction.

Besides doping, the properties of semiconductors, particularly their electrical conductivity, can be modulated by various external forces, including temperature, applied voltage, magnetic field, light, mechanical stress, and chemical adsorption. These factors enhance the capability of semiconductors to serve as functional devices, as demonstrated in Figure 1.2. For example, numerous semiconductor properties can be significantly influenced by temperature. Elevated temperatures lead to lattice expansion, typically reducing the bandgap. Furthermore, high temperatures facilitate intrinsic carrier excitation and dopant activation, while diminishing the effect of Coulomb scattering. Consequently, semiconductors generally exhibit increased conductivity at heightened temperatures. However, excessively high temperatures induce more pronounced lattice oscillations, manifesting as heightened phonon excitation; this scatters carriers, reducing carrier mobility and thus

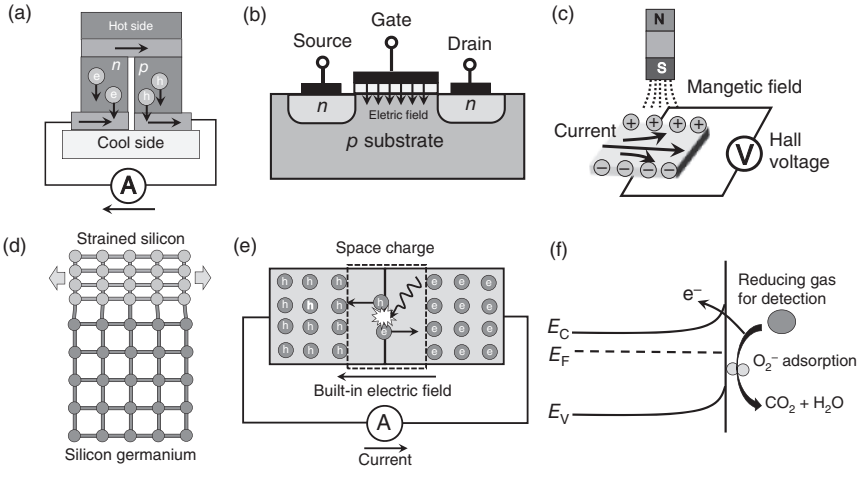

Figure 1.2 (a) Electricity generation via the thermoelectric effect (Seebeck effect) of semiconductors. (b) Metal-oxide-semiconductor field-effect transistor utilizing the electric field modulation effect of semiconductors. (c) Semiconductor characterization or magnetic field detection using the Hall effect. (d) Modulation of silicon by tensile strain introduced by the silicon–germanium substrate. (e) Photon detection based on the *p*–*i*–*n* diode. (f) Reducing gas sensing via the absorption on semiconductor and the subsequent modulation of Schottky barrier height.

decreasing conductivity. Other thermal effects include the thermoelectric phenomenon in semiconductors, arising from temperature gradients. Carriers migrate from the high-temperature region toward the low-temperature region, generating electric current, which can be exploited for electricity generation (Figure 1.2a), temperature measurement, or producing temperature differences via the reverse effect—the Peltier effect.

Owing to the relatively low carrier concentration within semiconductors, the carriers can be effectively modulated by an applied electric field. This field can be applied using various structures, such as a semiconductor junction or a metal-oxide-semiconductor capacitor. The electric field alters the carrier type and concentration, thereby influencing current conduction through mechanisms including drift, diffusion, or injection. Numerous devices exploit this modulation effect, such as the field-effect transistor (FET) and the bipolar junction transistor (BJT). The metal-oxide-semiconductor field-effect transistor (MOSFET), illustrated in Figure 1.2b, exhibits extremely high input impedance and can be manufactured with minimal dimensions. Consequently, it has become the foundation for modern digital integrated circuits. Conversely, the BJT operates as a current-controlled device, regulating a larger output current via a small input control current. Owing to its high operating speed, high transconductance, and low output resistance, the BJT remains a competitive device for amplification and switching applications, particularly in demanding analog circuits such as radio-frequency circuits for wireless systems.

The charge carrier transport within semiconductors is influenced by magnetic fields. When a magnetic field is applied perpendicular to the current direction in a semiconductor, a transverse voltage difference, known as the Hall voltage, develops across the electrical conductor (see Figure 1.2c). This phenomenon, termed the Hall effect, was first observed by Edwin Hall in 1879. The Hall voltage is fundamentally related to the carrier type and concentration, exhibiting opposite polarity for electron and hole conduction. Consequently, the Hall effect in semiconductors serves as a powerful characterization technique for determining carrier type, concentration, and mobility. Furthermore, this phenomenon is utilized in sensor applications for measuring magnetic fields or current, offering the advantages of non-contact measurement and high immunity to environmental contamination.

The properties of semiconductors can be modulated by strain. A typical example is the strain effect in silicon (Figure 1.2d), which finds important applications in the semiconductor industry to enhance further the performance of silicon-based devices. Strain influences silicon properties through the alteration of its interatomic distances. Silicon strain can be introduced through several techniques. Tensile strain can be created by depositing silicon over a silicon–germanium substrate, owing to the larger atomic size of germanium. This tensile strain increases silicon atom distances, resulting in higher electron mobility and thus lower energy dissipation in the fabricated transistor. Strain can similarly be created by doping the source and drain regions of a transistor with larger or smaller atoms. For instance, carbon doping in the source and drain regions of an *n*-type MOSFET induces tensile strain in the channel, enhancing electron mobility. Conversely, germanium doping induces

compressive strain, thereby increasing hole mobility. Strain may also be introduced by depositing a strained dielectric, such as a stressed silicon nitride layer, onto the device channel.

Other factors, such as photon excitation or chemical adsorption, can also influence semiconductor conductivity, enabling the development of sensors like photodetectors and chemical gas sensors for gas detection. Photon excitation modifies semiconductor properties through distinct mechanisms, including the photoelectric effect and photothermal effects. For instance, a *p*–*n* junction acts as a photodetector via the photoelectric effect; following photon absorption, electron-hole pairs generated within the junction are separated by the electric field in the depletion region, inducing an increase in electrical current, as demonstrated in Figure 1.2e. Similarly, the chemical adsorption of molecules onto a semiconductor surface alters its conductivity. This phenomenon underpins chemical sensors, exemplified by the application shown in Figure 1.2f. To achieve enhanced sensitivity, semiconductor sensing materials require a substantial surface area exposed for gas adsorption. Consequently, diverse semiconductor materials, including metal oxides and polymers—particularly low-dimensional structures such as nanowires and nanotubes—have been investigated for this purpose. For example, tin oxide (SnO_2), a widely studied metal oxide, exhibits notable advantages for gas sensing due to its facile fabrication, low cost, high sensitivity, and stability.

1.3 Crystal Structure and Types of Semiconductors

1.3.1 Lattice and Crystal Systems

Although the chemical compositions of semiconductors primarily define key properties—for instance, silicon can be used to fabricate solar cells in amorphous, polycrystalline, or single-crystalline forms—their crystallinity significantly influences bandgap, optical absorption, carrier transport, carrier recombination, and related phenomena. The fundamental distinction among amorphous, polycrystalline, and single-crystalline semiconductors lies in the spatial scale of their long-range order. Single crystals exhibit extensive long-range order, enabling their properties to be modeled by considering infinite periodic structures wherein atomic arrangements remain invariant under finite translations along specific symmetry directions. In contrast, polycrystalline solids comprise numerous small single-crystalline domains exhibiting short-range order and localized periodicity. Amorphous solids lack long-range atomic order entirely; nevertheless, their atomic configurations deviate from true randomness and retain short-range similarities to crystalline counterparts, explaining their partial property resemblance to single crystals. The periodicity and long-range order in single crystals facilitate structural description via finite repeating unit cells and enable precise calculation and analysis of properties using quantum mechanical principles. Conversely, non-crystalline materials necessitate complex modeling of many-particle systems involving astronomically large numbers of atoms.

The description and understanding of semiconductor crystal structures form the foundation for understanding semiconductor properties. The most essential characteristic of semiconductor crystals is their periodicity, which imparts translational symmetry. In a bulk crystal, this translational symmetry extends along three orthogonal dimensions. Consequently, the entire crystal structure composes itself by the periodic replication of a finite unit cell arrangement in these three directions. While the dimensionality of repetition reduces in two-dimensional materials like graphene or one-dimensional systems such as carbon nanotubes, the descriptive formalism remains analogous; this discussion will therefore concentrate on the bulk crystal case. These three-dimensional repetitions generate a spatial framework termed the crystal lattice. The crystal lattice can be mathematically defined by equation,

$$\vec{\boldsymbol{R}} = n_1\vec{\boldsymbol{a}} + n_2\vec{\boldsymbol{b}} + n_3\vec{\boldsymbol{c}} \tag{1.1}$$

where $\vec{\boldsymbol{a}}$, $\vec{\boldsymbol{b}}$, and $\vec{\boldsymbol{c}}$ are the three primitive vectors defined by the repetition lengths and directions (Figure 1.3a), and n_1, n_2, and n_3 are the corresponding repetition numbers along those directions. The atomic structure is constructed by repeating a finite set of atoms over the crystal lattice (Figure 1.3b, Figure 1.3c, and Figure 1.3d). The smallest block of atoms replicated in this manner is termed the basis (Figure 1.3c), while the parallelepiped containing the atoms (Figure 1.3b) constituting the basis and defining the spatial periodicity is called the unit cell of the crystal.

For a given crystal structure, both the basis and the unit cell are non-unique. Using a two-dimensional crystal as an example illustrates the variety of possible unit cell definitions. The smallest possible unit cell, termed the primitive unit cell, itself lacks uniqueness. Nevertheless, conventional definitions of the primitive unit cell exist to better reflect inherent crystal symmetries. For certain crystals, symmetries remain obscured when viewed within a primitive cell. In these instances, a conventional unit cell—distinct from a primitive unit cell and employing a larger basis—is often adopted to facilitate symmetry recognition. Consider the body-centered cubic (bcc) lattice, routinely depicted as a cubic unit cell containing an additional lattice point at its center, as shown in Figure 1.4a. Its actual primitive unit cell, occupying half the volume, is defined by selecting the center point as the origin, with primitive vectors extending to three non-adjacent corner vertices. Analogously, the face-centered

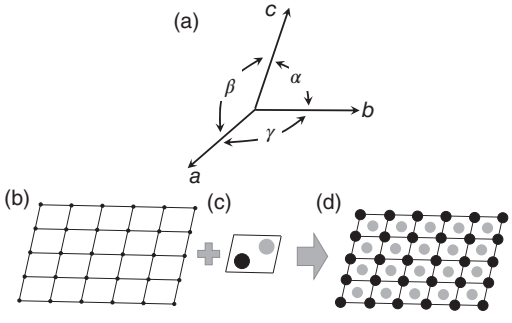

Figure 1.3 (a) Primitive lattice vectors and angles. (b) Bravais lattice grids, (c) the basis, and (d) the constructed atom structure.

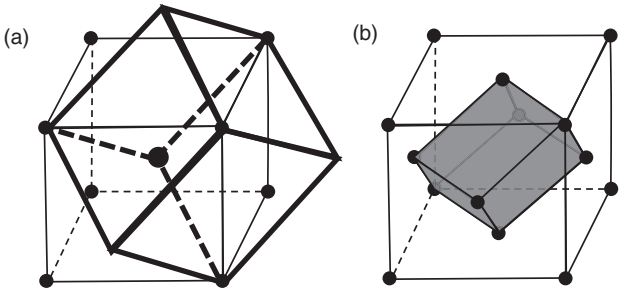

Figure 1.4 Cubic unit cell and the primitive unit cell of (a) the body-centered cubic lattice and (b) the face-centered cubic lattice.

cubic (fcc) lattice (Figure 1.4b) is typically represented as a cubic cell featuring additional lattice points at each face center. Here, the primitive vectors connect the origin (similarly defined) to three adjacent face centers. Symmetries for primitive unit cells in both bcc and fcc crystals derive from the rhombohedral lattice. Due to the ease with which cubic representations reveal higher symmetrical properties, they are preferred over rhombohedral descriptions.

Grids formed by translating the unit cell are termed Bravais lattices. These lattices can be classified based on their symmetry properties; lattices sharing the same symmetry belong to the same type. In three dimensions, only 14 distinct Bravais lattice types exist. These are further grouped into seven lattice systems, distinguished by the relationships among the lengths and angles of the conventional lattice vectors: triclinic, monoclinic, orthorhombic, tetragonal, rhombohedral (also known as trigonal), hexagonal, and cubic. Within the unit cells of these seven lattice systems, there may be more than one equivalent lattice point. The additional equivalent points can be located in different positions within the unit cell basis, leading to specific naming conventions: primitive (P) denotes lattices with points only at the unit cell corners; base-centered (A, B, or C, depending on the centered pair of parallel faces) has one additional point at the face center; body-centered (I) possesses one additional point at the body center; face-centered (F) features one additional point at the center of each face (Figure 1.5a–d). Accounting for these centered lattice points expands the seven basic lattice systems to the complete set of 14 three-dimensional Bravais lattice types (Table 1.1).

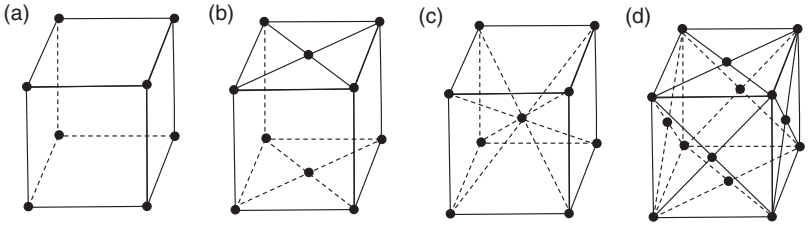

Figure 1.5 The four different Bravais lattices of the orthorhombic lattice system. (a) The primitive (P), (b) base-centered (S), (c) body-centered (I), (d) face-centered (F) Bravais lattices.

Table 1.1 Lattice systems, space groups, point groups, and crystal systems.

Lattice system	Lattice vector length relation	Lattice vector angle relation	Bravais lattice types	Space group types	Point group types	Crystal system
Triclinic			1	2	2	Triclinic
Monoclinic		$\alpha = \gamma = 90°$	2	13	3	Monoclinic
Orthorhombic		$\alpha = \beta = \gamma = 90°$	4	59	3	Orthorhombic
Tetragonal	$a = b$	$\alpha = \beta = \gamma = 90°$	2	68	7	Tetragonal
Rhombohedral	$a = b = c$	$\alpha = \beta = \gamma$	1	7	5	Trigonal
Hexagonal	$a = b$	$\alpha = \beta = 90°, \gamma = 120°$	1	18		
				27	7	Hexagonal
Cubic	$a = b = c$	$\alpha = \beta = \gamma = 90°$	3	36	5	Cubic
Total			14	230	32	7

The atomic structure of a crystal is constructed by repeating the basis throughout the Bravais lattice. The number of atoms within the basis may exceed one, thereby altering the lattice symmetry. The combination of basis atoms and lattice symmetry necessitates the classification of crystals into 230 space groups. These space groups encompass symmetry operations about points (e.g., rotation, reflection, and inversion) and incorporate translational symmetry, including operations combining translation with rotation. Specifically, the latter category comprises screw axes and glide planes: a screw axis represents a rotation combined with a translation, whereas a glide plane constitutes a reflection combined with a translation.

Symmetry operations centered on points—such as rotation, reflection, and inversion—constitute point groups. Only 32 distinct point groups exist among the 230 space groups. The 14 Bravais lattices represent a specific subset of space groups, exhibiting seven distinct point groups and categorized into seven crystal systems: triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic. Point groups are commonly designated using Schönflies notation, employing C, D, T, and O to denote cyclic, dihedral, tetrahedral, and octahedral symmetries, respectively, supplemented by subscripts indicating rotational symmetry order. For example, C_{3h} signifies a threefold cyclic rotational symmetry coupled with a mirror plane perpendicular to the rotational axis. Space groups are typically described using Hermann–Mauguin notation.

1.3.2 Reciprocal Lattice and Miller Indices

A direction connecting two distinct lattice points is uniquely specified by a corresponding parallel lattice vector,

$$\vec{\boldsymbol{R}} = n_1\vec{\boldsymbol{a}}_1 + n_2\vec{\boldsymbol{a}}_2 + n_3\vec{\boldsymbol{a}}_3 \tag{1.2}$$

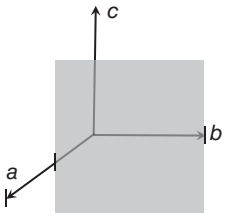

Figure 1.6 Identifying the (210) plane index by reading the interceptions of the plane on the lattice vectors.

which is generated by the lattice vectors $\vec{a}_1$, $\vec{a}_2$ and $\vec{a}_3$. For simplicity, the crystallographic direction can be denoted by the conventional notation $[n_1 n_2 n_3]$. For example, the [100] direction in a simple cubic lattice indicates the direction parallel to the positive x-axis. Due to lattice symmetry, the directions [100], [−100], [010], [0–10], [001], and [00–1] are crystallographically equivalent. The complete set of these equivalent directions corresponding to $[n_1 n_2 n_3]$ is conventionally indexed using angle brackets as $<n_1 n_2 n_3>$.

Certain lattices lie within a specific plane. Miller indices represent a symbolic notation for designating lattice planes. The lattice plane positioned closest to the origin intercepts the three lattice vectors at positions $\vec{a}_1/h$, $\vec{a}_2/k$ and $\vec{a}_3/l$. This plane is indexed as (hkl) using round brackets. For instance, the plane intercepting the a and b axes a $\vec{a}/2$ and $\vec{b}$, respectively, while being parallel to the c-axis (thus having an infinite intercept equivalent to $\vec{c}$), is designated as (210), as illustrated in Figure 1.6. The set of all crystallographically equivalent planes to (hkl) is denoted by{hkl} within curly brackets. The concept of Miller indices is more readily understood through the definition of a complementary set of vectors called the reciprocal lattice vectors, $\vec{b}_1$, $\vec{b}_2$ and $\vec{b}_3$. These reciprocal lattice vectors are defined by the relationship,

$$\vec{a}_i \cdot \vec{b}_j = 2\pi\delta_{ij} \tag{1.3}$$

where δ_{ij} is the Kronecker delta. δ_{ij} equals 1 only for $i = j$. Reciprocal lattice vector $\vec{b}_1$ can be directly constructed by

$$\vec{b}_1 = 2\pi \frac{\vec{a}_2 \times \vec{a}_3}{\vec{a}_1 \cdot \vec{a}_2 \times \vec{a}_3} \tag{1.4}$$

while $\vec{b}_2$ and $\vec{b}_3$ can be derived similarly. Reciprocal lattices serve as important tools for understanding the electronic band structure and the diffraction patterns of high-energy particles, such as X-rays, electrons, or neutrons. The reciprocal lattice of a simple cubic lattice is itself simple cubic. However, the reciprocal lattice of a body-centered cubic lattice is face-centered cubic, whereas the reciprocal lattice of a face-centered cubic lattice is body-centered cubic.

The orthogonality relationship between lattice and reciprocal lattice vectors renders the calculation of the inner product of a lattice direction vector and a reciprocal lattice direction vector particularly efficient. For example, for $\vec{R} = n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3$ and $\vec{G} = h\vec{b}_1 + k\vec{b}_2 + l\vec{b}_3$, their inner product will be

$$\vec{R} \cdot \vec{G} = 2\pi(n_1 h + n_2 k + n_3 l) \tag{1.5}$$

Utilizing this relationship, it follows that any vector $\vec{v}$ in the lattice plane (hkl), defined using the axial intercept method, satisfies $\vec{v} \cdot \vec{G} = 0$. Therefore, the lattice plane (hkl) is perpendicular to the reciprocal lattice vector $\vec{G} = h\vec{b}_1 + k\vec{b}_2 + l\vec{b}_3$, which can consequently be indexed as (hkl). The distance between adjacent (hkl) crystal planes can be calculated by

$$d = \frac{2\pi}{|h\vec{b}_1 + k\vec{b}_2 + l\vec{b}_3|} \tag{1.6}$$

For cubic crystal with lattice constant of a, this equation can be simplified as

$$d = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \tag{1.7}$$

from which we can clearly see that crystal planes with higher Miller indices have smaller plane distance, usually making the corresponding crystal plane more difficult to cleavage due to stronger plane interaction.

1.3.3 Common Semiconductor Crystal Structures

The symmetry of the crystal structure significantly influences the properties of semiconductors. This symmetry is characterized by the space group. There exist 230 distinct space groups, categorized into seven lattice systems based on Bravais lattice symmetry or seven crystal systems based on point group symmetry. Lattice systems and crystal systems generally correspond. However, an exception occurs for crystals possessing a hexagonal Bravais lattice that crystallize within the trigonal system due to reduced symmetry resulting from inclusion of the basis. For instance, α-quartz exhibits this configuration (Figure 1.7 and Table 1.1).

Most significant semiconductors have cubic or hexagonal Bravais lattices and belong to the cubic, hexagonal, or trigonal crystal systems. For instance, the single crystals of elemental semiconductors, such as diamond, Si, and Ge, possess cubic Bravais lattices (Figure 1.8). Specifically, these materials exhibit face-centered cubic Bravais lattices. However, their basis consists of two atoms, meaning the primitive unit cell contains two atoms. Their atomic structures can be understood as the combination of two face-centered cubic lattices, displaced by one-quarter of the cubic body diagonal relative to each other. Structures of this type are referred to as diamond structures. Each atom is tetrahedrally coordinated by four nearest neighbors, resulting in a coordination number of 4, and these four atoms form a tetrahedron.

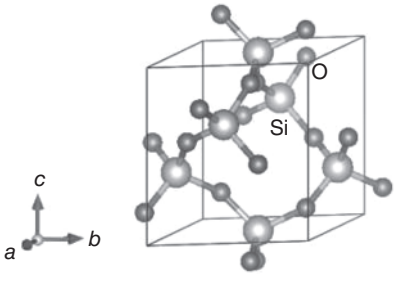

Figure 1.7 Atomic structure of α-quartz, which has hexagonal Bravais lattice but belongs to the trigonal crystal system.

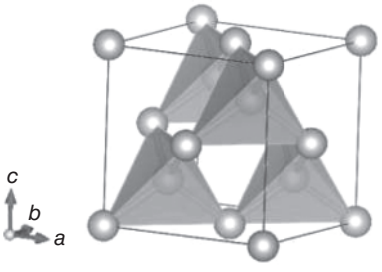

Figure 1.8 Diamond crystal structure of C, Si or Ge crystal with FCC Bravais lattice.

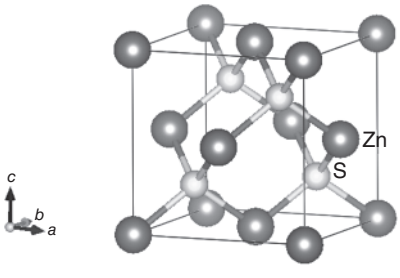

Figure 1.9 Zincblende crystal structure of ZnS.

In addition to elemental semiconductors from group IV, certain compound semiconductors exhibit crystal structures closely resembling the diamond structure. For example, binary AB compounds such as III–V or II–VI semiconductors can adopt the zincblende crystal structure. Similar to diamond, the zincblende structure possesses an fcc Bravais lattice and a basis of two atoms (Figure 1.9). However, in the zincblende basis, the two atoms are chemically distinct. Each central atom is tetrahedrally coordinated by four nearest-neighbor atoms of the opposite type. Semiconductors with zincblende structures include II–VI compounds (ZnS, ZnO, CdTe, HgTe) and III–V compounds (GaN, GaAs, InP). Furthermore, SiC, which is an IV-IV binary compound composed of elements from group IV, can also crystallize in the zincblende structure.

Besides the zincblende lattice, II–VI or III–V compounds can also adopt a hexagonal Bravais lattice. The simplest hexagonal structure is known as the wurtzite crystal structure (Figure 1.10). Both zincblende and wurtzite structures exhibit tetrahedral coordination with a coordination number of four. These structures represent different polytypes of the same compounds. The key distinction lies in the packing:

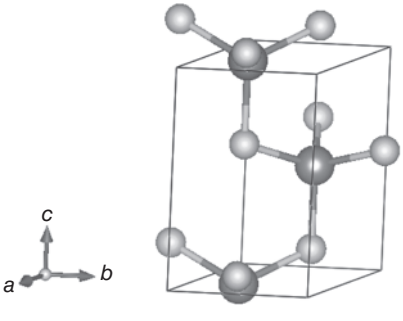

Figure 1.10 Wurtzite crystal structure of ZnS.

zincblende structures feature cubic close-packed (CCP) atoms in a face-centered cubic lattice, whereas wurtzite structures possess hexagonal close-packed (HCP) atoms in a hexagonal lattice. Both crystal types can be described using a hexagonal unit cell. Three different layer-by-layer stacking sequences exist for tetrahedrally coordinated structures. Zincblende crystals follow an ABCABCABC stacking pattern and possess a primitive unit cell with two atoms. Conversely, wurtzite crystals exhibit an ABABAB stacking pattern and contain four atoms per primitive unit cell. Zincblende is generally the low-temperature stable phase, while wurtzite becomes more stable at higher temperatures. Rapid cooling following heating can freeze-in and form the wurtzite structure at room temperature. Beyond zincblende and wurtzite, numerous other stacking sequences are possible for II–VI, III–V, or IV–IV compounds, resulting in additional polymorphs. For instance, SiC exhibits several polytypes such as 2H, 3C, 4H, and 6H [4]. Understanding these fundamental structures facilitates the construction and understanding of more complex semiconductor crystals. $Al_xGa_{1-x}N$ serves as a specific example; its structure is based on the wurtzite lattice of GaN. Al and Ga, being homologous group III metals, are randomly distributed in $Al_xGa_{1-x}N$, with Ga sites occupied by a mixture of Al and Ga atoms consistent with their respective concentrations [5].

Another important crystal structure for semiconductors is the perovskite structure. Perovskite has a chemical formula of ABX_3, in which A and B are two cations (positively charged ions), and X is a negatively charged ion (i.e., an anion). A and B are typically of significantly different sizes, with A ions larger than B ions. A representative example is $SrTiO_3$. ABX_3 adopts a cubic structure where the B cation is surrounded by an octahedron of X anions in sixfold coordination, and the A cation is surrounded by 12 X anions in cuboctahedral coordination. Perovskite structures occur in an extensive range of materials. All-inorganic cesium lead halide perovskites ($CsPbX_3$, where X = I, Br, or Cl) exhibit crystal structures analogous to $SrTiO_3$. Considering $CsPbBr_3$ as an example (Figure 1.11), it adopts the simple

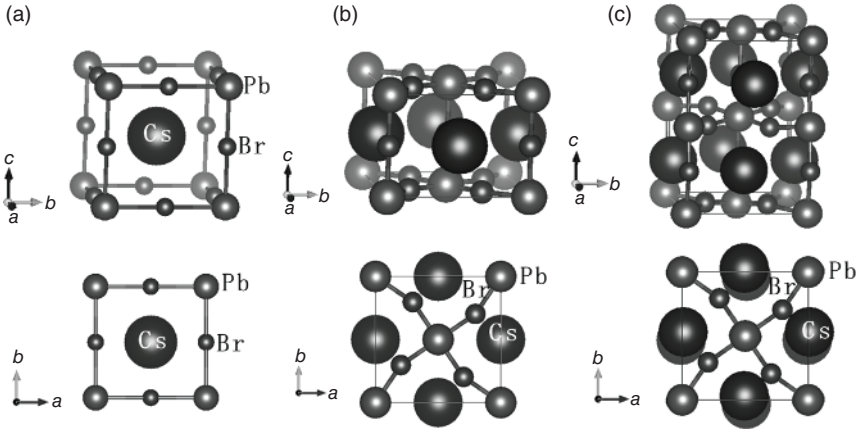

Figure 1.11 Side and top views of $CsPbBr_3$ in (a) cubic phase, (b) Pnma phase and (c) P4/mbm phase.

cubic Bravais lattice with $Pm\bar{3}m$ space group symmetry solely at temperatures exceeding 130 °C. $CsPbBr_3$ in the $Pm\bar{3}m$ phase contains 5 atoms per unit cell. Between 88 and 130 °C, the stable phase of $CsPbBr_3$ is tetragonal, belonging to the P4/mbm space group with 10 atoms in the unit cell; this phase can be derived from a $\sqrt{2} \times \sqrt{2} \times 1$ supercell of the cubic structure with distortion. At temperatures below 88 °C, the tetragonal phase transforms to the orthorhombic phase with Pnma space group symmetry [6]. The orthorhombic phase can be constructed based on a $\sqrt{2} \times \sqrt{2} \times 2$ supercell of cubic $CsPbBr_3$ with distortion. The cubic $Pm\bar{3}m$ phase possesses the highest symmetry, characterized by 48 symmetry operations; the tetragonal P4/mbm phase exhibits the second highest symmetry with 16 symmetry operations; and the orthorhombic Pnma phase has the lowest symmetry with 8 symmetry operations. $CsPbBr_3$ with the orthorhombic Pnma phase, stable at room temperature, achieves high carrier mobility and excellent optoelectronic properties, rendering it suitable for applications in various optoelectronic devices, such as light-emitting diodes, solar cells, photodetectors, lasers, among others.

1.4 Carrier Transport

Semiconductors contain two types of charge carriers: electrons and holes. An accurate description of electrons and holes requires the framework of electronic band structure, which will be presented in subsequent sections. Broadly, if an intrinsic semiconductor possesses additional electrons, these electrons occupy the conduction band and transport similarly to free electrons, albeit with a distinct effective mass. Conversely, if an intrinsic semiconductor loses electrons, holes exist within the valence band. These holes transport as positively charged particles with properties analogous to electrons but with a positive charge and a different effective mass. Electrons and holes coexist within the semiconductor simultaneously. However, they can recombine upon encountering each other, leading to the annihilation of both carriers. This recombination releases energy in the form of light emission, lattice excitations, or excitations of other particles. Conversely, an electron-hole pair can be generated through sufficient energy input, for example, via light absorption, thermal activation, or impact ionization. Due to the variability in carrier types and underlying physical mechanisms within semiconductors compared to metallic conductors, carrier transport exhibits more complex behaviors, including drift, diffusion, recombination, generation, thermionic emission, and tunneling.

1.4.1 Carrier Drift and Diffusion

An electric field exerts a force on both electrons and holes, as they are charged carriers, thereby accelerating them and inducing drift transport. In the absence of scattering mechanisms, these carriers would accelerate continuously, achieving arbitrarily high velocities. However, scattering events arising from disturbances

such as lattice vibrations (phonons) and crystal defects decelerate the carriers. This subsequent deceleration dissipates the acquired kinetic energy of the carriers, converting it into heat. Consequently, an average drift velocity is attained, which is described by the relationship given below,

$$v_e = \mu_e E \tag{1.8}$$

$$v_h = \mu_h E \tag{1.9}$$

where v_e and v_h are the average velocities of electrons and holes, respectively, E is the electric field strength, and μ_e and μ_h are coefficients termed carrier mobility. The velocity has units of cm/s, while the electric field strength has units of V/cm; therefore, the unit of carrier mobility is $\mathrm{cm}^2/(\mathrm{V}\cdot\mathrm{s})$. Mobility characterizes the rate at which a carrier moves when subjected to an electric field. It is noted that electrons and holes possess opposite charges; consequently, they experience opposing forces and transport in opposite directions within the semiconductor. However, despite their opposing transport directions, the opposite charges of electrons and holes therefore result in their electric currents having the same direction. If the carrier densities for electrons and holes are n_e and n_h, respectively, the total electric current density caused by carrier drift is

$$J_{\text{drift}} = en_e v_e + en_h v_h = en_e \mu_e E + en_h \mu_h E \tag{1.10}$$

The drift current can be rewritten as

$$J_{\text{drift}} = \sigma_e E + \sigma_h E = \sigma E \tag{1.11}$$

$$\sigma_e = en_e \mu_e \tag{1.12}$$

$$\sigma_h = en_h \mu_h \tag{1.13}$$

where σ_e and σ_h are the electrical conductivity of electrons and holes, respectively, and e is the fundamental electron charge. The electrical conductivity is related to resistivity by $\sigma = 1/\rho$. Carrier scattering imposes a limit on their drift velocity, which gives rise to Joule heating. The volumetric Joule heating power density generated by an electric field is given by

$$P = J_{\text{drift}} E = \sigma E^2 \tag{1.14}$$

The carrier mobility is fundamentally governed by the lifetime and mean-free-path of carriers during transport and scattering events. According to the Drude model, the carrier's momentum is accelerated by the applied electric field and dissipated through momentum relaxation during scattering events. Assuming a carrier scattering lifetime of τ, the equation of motion accelerated by the electric field is expressed as:

$$m\frac{\mathrm{d}v}{\mathrm{d}t} = eE - m\frac{v}{\tau} \tag{1.15}$$

where m is the carrier mass and e is the carrier charge, depending on the electron or hole carrier type. At steady state, $\mathrm{d}v/\mathrm{d}t = 0$, we can get

$$v = \frac{e\tau}{m}E \tag{1.16}$$

In associate with previous equation, the carrier mobility can be written as

$$\mu = \frac{e\tau}{m} \tag{1.17}$$

and the electric conductivity will be

$$\sigma = \frac{e^2\tau}{m} n \tag{1.18}$$

Multiple scattering mechanisms exist, exhibiting distinct dependences on temperature and thereby influencing the scattering lifetime. For acoustic phonon scattering via the deformation potential, mobility varies proportionally to $T^{-3/2}$. In contrast, for acoustic phonon scattering via the piezoelectric mechanism, mobility varies proportionally to $T^{-1/2}$. Ionized impurity scattering results in a temperature dependence of mobility scaling with $T^{3/2}$. At low electric fields, typically below 10^3 V/cm in silicon and gallium arsenide, the drift velocity increases linearly in accordance with Ohm's law. However, at higher fields, the carrier velocity exhibits a sublinear increase with field strength and eventually saturates at approximately 10^7 cm/s, a phenomenon termed velocity saturation. Certain semiconductors, such as GaAs, InP, and InGaAs, exhibit more complex behavior characterized by a velocity maximum preceding saturation as illustrated for fields between 3×10^3 V/cm and 2×10^5 V/cm. This maximum constitutes velocity overshoot, leading to a decrease in current with increasing electric field strength, known as negative differential resistance.

Besides the drift motions induced by the electric field, carrier transport also arises due to the gradient of carrier density, known as diffusion current. The diffusion current is proportional to the magnitude of the carrier density gradient, expressed by the following equation,

$$J_{\text{diff}} = eD_e \frac{\mathrm{d}n_e}{\mathrm{d}x} - eD_h \frac{\mathrm{d}n_h}{\mathrm{d}x} \tag{1.19}$$

which has considered the diffusion of both electrons and holes. D_n and D_p represent the diffusion coefficients of electrons and holes, respectively. The differing signs in the equation arise from the opposite charge polarity between electrons and holes. From a microscopic perspective, the diffusion coefficient can be determined using

$$D = \frac{v_{\text{rms}}^2}{3}\tau \tag{1.20}$$

where v_{rms} is the root-mean-square velocity of either electrons or holes, and τ is the same scattering lifetime as mentioned in the mobility. Remember that $mv_{\text{rms}}^2/2 = 3kT/2$, we can get the Einstein relationship between the diffusion coefficient and the carrier mobility,

$$D = \mu \frac{kT}{e} \tag{1.21}$$

The Einstein relation characterizes the diffusion of charge carriers, specifically electrons and holes governed by Boltzmann statistics. This relation can also be derived from the equilibrium condition occurring in a semiconductor possessing a built-in electric field, E. Under such conditions, the net current attributable to any particular carrier type, electrons or holes, must vanish to maintain equilibrium.

Using electron transport as an example, the combined contributions of the drift current and the diffusion current imply that the following equation must hold

$$\mu_e n_e E = -D_e \frac{\mathrm{d}n_e}{\mathrm{d}x} \tag{1.22}$$

While at equilibrium the carrier concentration, because of the electric field induced energy difference, according to the Boltzmann distribution, the spatial electron density distribution should follow,

$$n_e(x) = n_e(x_0) e^{-\frac{eE(x-x_0)}{kT}} \tag{1.23}$$

$$\frac{\mathrm{d}n_e}{\mathrm{d}x} = -n_e \frac{eE}{kT} \tag{1.24}$$

where x_0 is a reference coordinate with electron density of $n_e(x_0)$. Similar equations hold for holes as well. Combining these equations, we can get the Einstein relation.

The total current inside a semiconductor will be the combination of drift current and diffusion current, as indicated below,

$$J_{\mathrm{tot}} = J_{\mathrm{drift}} + J_{\mathrm{diff}} = en_e\mu_e E + eD_e \frac{\mathrm{d}n_e}{\mathrm{d}x} + en_h\mu_h E - eD_h \frac{\mathrm{d}n_h}{\mathrm{d}x} \tag{1.25}$$

which can be separated into electron and hole contributions respectively as below (Figure 1.12),

$$J_e = en_e\mu_e E + eD_e \frac{\mathrm{d}n_e}{\mathrm{d}x} \tag{1.26}$$

$$J_h = en_h\mu_h E - eD_h \frac{\mathrm{d}n_h}{\mathrm{d}x} \tag{1.27}$$

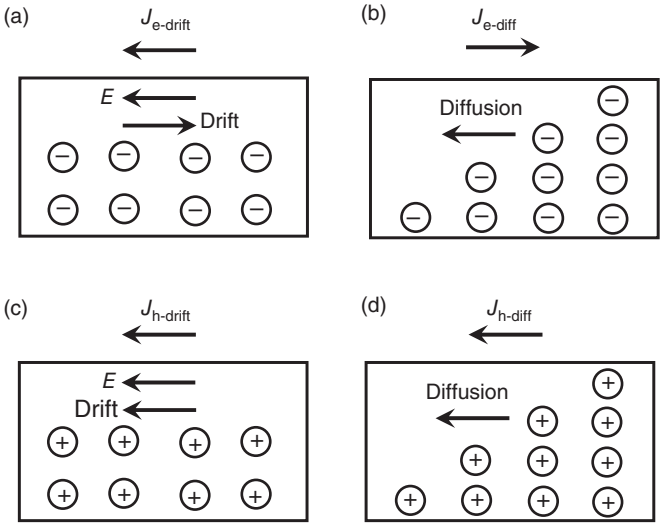

Figure 1.12 (a) Drift current caused by electric field and (b) diffusion current caused by concentration gradient for electrons. (c) Drift current caused by electric field and (d) diffusion current caused by concentration gradient for holes.

1.4.2 Quasi-Fermi Levels and Electrochemical Potentials

After excitations such as optical illumination or electron/hole injection, electron and hole densities deviate from their thermodynamic equilibrium values. Nonetheless, quasi-Fermi levels can still describe these densities; for instance, the Boltzmann distribution applies to nondegenerate cases. Assuming the electrostatic potential induced by the electric field is $\varphi(x)$, the electric field strength E is given by $E = -\mathrm{d}\varphi/\mathrm{d}x$. Considering the quasi-Fermi levels E_{Fe} and E_{Fh} for electrons and holes, respectively, the electron and hole densities can be expressed according to the Boltzmann distribution as

$$n_e(x) = n_e(x_0)e^{\frac{E_{\mathrm{Fe}}(x)+e\varphi(x)-E_{\mathrm{Fe}}(x_0)-e\varphi(x_0)}{kT}} \tag{1.28}$$

$$n_h(x) = n_h(x_0)e^{\frac{-E_{\mathrm{Fh}}(x)-e\varphi(x)+E_{\mathrm{Fh}}(x_0)+e\varphi(x_0)}{kT}} \tag{1.29}$$

where x_0 is a reference coordinate with an electron density of $n_e(x_0)$ and a hole density of $n_h(x_0)$. Since the electrostatic potential $-e\varphi(x)$ influences both the conduction and valence band positions by $-e\varphi(x)$, the relative position between the quasi-Fermi level and the conduction or valence band edge is given by $E_{\mathrm{Fe}}(x) + e\varphi(x)$ or $E_{\mathrm{Fh}}(x) + e\varphi(x)$. For electrons, a higher value of $E_{\mathrm{Fe}}(x) + e\varphi(x)$ indicates that the quasi-Fermi level is closer to the conduction band minimum, resulting in a higher electron density. In contrast, for holes, a higher value of $E_{\mathrm{Fe}}(x) + e\varphi(x)$ indicates that the quasi-Fermi level is more distant from the valence band maximum, resulting in a lower hole density. More specifically, we can get

$$\frac{\mathrm{d}n_e}{\mathrm{d}x} = n_e\left(\frac{\mathrm{d}E_{\mathrm{Fe}}}{\mathrm{d}x} + e\frac{\mathrm{d}\varphi}{\mathrm{d}x}\right) \tag{1.30}$$

$$\frac{\mathrm{d}n_h}{\mathrm{d}x} = n_h\left(-\frac{\mathrm{d}E_{\mathrm{Fh}}}{\mathrm{d}x} - e\frac{\mathrm{d}\varphi}{\mathrm{d}x}\right) \tag{1.31}$$

Inserting these two equations in the electron and hole current equations respectively, we can get the following equations for the total current after considering $E = -\mathrm{d}\varphi/\mathrm{d}x$ and Einstein relation $D = \mu kT/e$,

$$J_e = \frac{eD_e n_e}{kT}\frac{\mathrm{d}E_{\mathrm{Fe}}}{\mathrm{d}x} = n_e\mu_e\frac{\mathrm{d}E_{\mathrm{Fe}}}{\mathrm{d}x} \tag{1.32}$$

$$J_h = \frac{eD_h n_h}{kT}\frac{\mathrm{d}E_{\mathrm{Fh}}}{\mathrm{d}x} = n_h\mu_h\frac{\mathrm{d}E_{\mathrm{Fh}}}{\mathrm{d}x} \tag{1.33}$$

Similar to the electrostatic potential, the electrochemical potentials for electrons and holes are defined by

$$E_{\mathrm{Fe}} = e\varphi_{\mathrm{Fe}} \tag{1.34}$$

$$E_{\mathrm{Fh}} = e\varphi_{\mathrm{Fh}} \tag{1.35}$$

Consequently, the total electron and hole current densities become,

$$J_e = en_e\mu_e\frac{\mathrm{d}\varphi_{\mathrm{Fe}}}{\mathrm{d}x} = \sigma_e\frac{\mathrm{d}\varphi_{\mathrm{Fe}}}{\mathrm{d}x} \tag{1.36}$$

$$J_h = en_h\mu_h\frac{\mathrm{d}\varphi_{\mathrm{Fh}}}{\mathrm{d}x} = \sigma_h\frac{\mathrm{d}\varphi_{\mathrm{Fh}}}{\mathrm{d}x} \tag{1.37}$$

This formalism is analogous to the description of conduction current in a conductor or drift current in a semiconductor. The crucial modifications involve replacing the electrostatic potential with the quasi-Fermi levels for electrons and holes, respectively. From the preceding equations, it is evident that the gradient of the electrochemical potential governs carrier flow. In semiconductors under thermal equilibrium conditions, no net current exists, and the quasi-Fermi levels for electrons and holes are equal. Therefore, $J_e + J_h = 0$ and $\varphi_{\mathrm{Fe}} = \varphi_{\mathrm{Fh}}$, implying that the electrochemical potential remains constant throughout the semiconductor material.

1.4.3 Carrier Generation and Recombination

At the thermal-equilibrium condition, the electron concentration p and hole concentration n satisfy the following equation,

$$pn = n_i^2 \tag{1.38}$$

where n_i is the intrinsic electron and hole concentration. During carrier transport, electron and hole concentrations may deviate from thermal equilibrium conditions. Electron-hole recombination dominates when $pn > n_i^2$, whereas thermal generation dominates when $pn < n_i^2$. During electron-hole recombination, energy is conserved through photon emission, excitation of another free carrier (Auger recombination), or phonon emission. Conversely, the energy required for the thermal generation of electron-hole pairs originates from the lattice thermal energy. Recombination and generation mechanisms are classified as either band-to-band transitions or trap-mediated transitions (Figure 1.13). Band-to-band transitions primarily occur in direct-bandgap semiconductors, such as GaAs and other III–V compounds. For indirect-bandgap semiconductors like Si and Ge, however, trap-mediated transitions predominate.

For the band-to-band transitions, the net recombination or generation rate can be calculated by

$$U = R\left(pn - n_i^2\right) \tag{1.39}$$

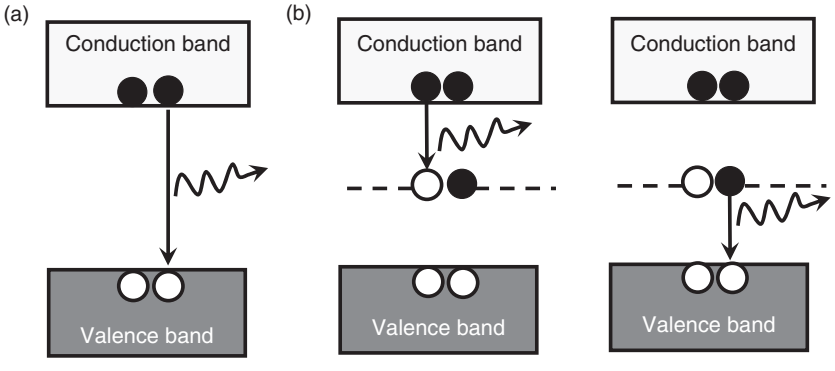

Figure 1.13 Schematics of (a) band-to-band recombination process and (b) Shockley-Read-Hall recombination process.

where R is the recombination coefficient, net recombination of electron-hole pairs occurs when $pn > n_i^2$; otherwise, net generation occurs. Under low-level injection conditions—for instance, in an n-type semiconductor with $n \approx N_{\mathrm{D}}$ and injected hole density $p = \Delta p + p_0$, where the equilibrium concentrations satisfy $np_0 = n_i^2$—the net recombination rate arising from band-to-band transitions is

$$U \approx \frac{\Delta p}{\tau_p} \tag{1.40}$$

$$\tau_p = \frac{1}{RN_{\mathrm{D}}} \tag{1.41}$$

Similarly, for a low-level injection of n-type semiconductor, the net recombination rate due to band-to-band transitions is

$$U \approx \frac{\Delta n}{\tau_n} \tag{1.42}$$

$$\tau_n = \frac{1}{RN_{\mathrm{A}}} \tag{1.43}$$

where $p \approx N_{\mathrm{A}}$ is the hole and acceptor doping density.

For recombination through trap levels, the capture of electrons and holes by the trap levels occurs sequentially. The capture probabilities for electrons and holes can be characterized by the capture cross sections, carrier thermal velocities, and trap density, respectively. The electron capture rate is determined

$$\frac{n}{\tau_{n,\mathrm{cap}}} = \sigma_n \upsilon_{\mathrm{th}} n(N_{t,\mathrm{tot}} - N_{t,\mathrm{occ}}) \tag{1.44}$$

where $N_{t,\mathrm{tot}}$ and $N_{t,\mathrm{occ}}$ are the total and occupied trap densities, respectively, and τ_n is the electron lifetime due to capture by unoccupied traps, characterized by the density $N_{t,\mathrm{tot}} - N_{t,\mathrm{occ}}$. Here, σ_n and υ_{th} are the electron capture cross-section and average thermal velocity, respectively. Similarly, the hole capture probability is proportional to,

$$\frac{p}{\tau_{p,\mathrm{cap}}} = \sigma_p \upsilon_{\mathrm{th}} p N_{t,\mathrm{occ}} \tag{1.45}$$

where $N_{t,\mathrm{occ}}$ is the occupied trap density. Besides electron and hole capture processes, charges can also undergo thermal excitation to the conduction and valence bands, respectively. The corresponding generation rates can be derived from the detailed balance principle.

$$\frac{N_{t,\mathrm{occ}}}{\tau_{n,\mathrm{gen}}} = \sigma_n \upsilon_{\mathrm{th}} n_i N_{t,\mathrm{occ}} e^{\frac{E_t - E_i}{kT}} \tag{1.46}$$

$$\frac{N_{t,\mathrm{tot}} - N_{t,\mathrm{occ}}}{\tau_{p,\mathrm{gen}}} = \sigma_p \upsilon_{\mathrm{th}} n_i (N_{t,\mathrm{tot}} - N_{t,\mathrm{occ}}) e^{\frac{E_i - E_t}{kT}} \tag{1.47}$$

Considering all these factors, the net recombination rate can be described using Shockley-Read-Hall statistics

$$U = \frac{\sigma_n \sigma_p \upsilon_{\mathrm{th}} N_t \left(pn - n_i^2\right)}{\sigma_n \left(n + n_i e^{\frac{E_t - E_i}{kT}}\right) + \sigma_p \left(p + n_i e^{\frac{E_i - E_t}{kT}}\right)} \tag{1.48}$$

Recombination due to surface trap states can be described by analogous/similar equations. However, the surface trap density N_t has units of cm^{-2}; consequently, the net recombination rate U has unit of $cm^{-2}{\cdot}s^{-1}$ rather than $cm^{-3}{\cdot}s^{-1}$. The equation above demonstrates that a net recombination or generation rate exists only if $pn \neq n_i^2$, and this rate is proportional to $pn - n_i^2$. Furthermore, the trap-assisted recombination rate is maximized when the trap level E_t lies close to the mid-bandgap, i.e., $E_t \approx E_i$. Thus, the net Shockley-Read-Hall recombination rate is given by:

$$U = \frac{\sigma_n \sigma_p \upsilon_{th} N_t \left(pn - n_i^2\right)}{\sigma_n (n + n_i) + \sigma_p (p + n_i)} \tag{1.49}$$

Under low-level injection conditions, such as in an n-type semiconductor with doping concentration $n \approx N_D$ and an injected hole density $p = \Delta p + p_0$ satisfying $np_0 = n_i^2$, the net Shockley-Read-Hall recombination rate is expressed as

$$U \approx \frac{\Delta p}{\tau_p} \tag{1.50}$$

$$\tau_p = \frac{1}{\sigma_p \upsilon_{\text{th}} N_t} \tag{1.51}$$

Similarly, for low-level injection in an n-type semiconductor, the net Shockley-Read-Hall recombination rate is

$$U \approx \frac{\Delta n}{\tau_n} \tag{1.52}$$

$$\tau_n = \frac{1}{\sigma_n \upsilon_{\text{th}} N_t} \tag{1.53}$$

For high-level injections $\Delta n = \Delta p > n_i$, the recombination lifetimes due to band-to-band recombination and trap-assisted recombination, respectively, are derived as

$$\tau_n = \tau_p = \frac{1}{R \Delta n} \tag{1.54}$$

$$\tau_n = \tau_p = \frac{\sigma_n + \sigma_p}{\sigma_n \sigma_p \upsilon_{\text{th}} N_t} \tag{1.55}$$

For trap-assisted recombination. These equations indicate that the band-to-band recombination lifetime decreases with injection level, while the trap-assisted recombination lifetime increases by approximately a factor of two when $\sigma_n = \sigma_p$. The increase in trap-assisted recombination lifetime may be explained by the half-occupation of trap states under high-level injection. Impurities or lattice defects typically introduce deep-level defect states within the bandgap, which shorten the minority-carrier lifetime. For instance, gold impurities in silicon introduce trap levels that can reduce the lifetime from several microseconds to several nanoseconds. This lifetime reduction can be advantageous for high-speed devices.

Besides band-to-band recombination and Shockley-Read-Hall recombination, Auger recombination represents another crucial recombination process. This nonradiative mechanism is significant in semiconductors, occurring when an electron and hole recombine by transferring their energy to a third charge

carrier—either another electron or hole—instead of emitting a photon. The net Auger recombination rate is described by the following equation

$$U = (C_n n + C_p p)\left(pn - n_i^2\right) \tag{1.56}$$

where C_n is the Auger coefficient for electron-dominated processes (cm^6/s), and C_p is the Auger coefficient for hole-dominated processes (cm^6/s). Auger recombination becomes increasingly prominent at high carrier densities as it competes with radiative recombination. This competition significantly reduces the efficiency of optoelectronic devices, including light-emitting diodes (LEDs), solar cells, and lasers.

1.5 Doping Process

Intrinsic semiconductors exhibit very low carrier densities, resulting in negligible conductivity. The reason semiconductors are immensely valuable in electronic devices is that their electronic properties can be precisely modified by introducing a relatively small concentration of defects, such as impurity dopants, typically less than one part per million. For example, silicon can be doped to form an *n*-type or *p*-type semiconductor by adding P atoms or B atoms, respectively, enabling conduction through electron or hole transport. More complex structures can be fabricated by spatially varying the doping profile within semiconductors, leading to significant applications and realizing functionalities unattainable with conventional devices. Typical examples include the diode and the bipolar junction transistor. A diode achieves unidirectional conduction; its fundamental structure is the PN junction, which conducts current under forward bias and blocks current under reverse bias. The BJT, based on PNP or NPN junctions, functions as a current amplifier. These unique properties of doped semiconductor devices underpin information processing in crucial circuits.

1.5.1 Point Defects and Energy Levels

Understanding defects and achieving controllable doping are of fundamental importance for utilizing semiconductors in electronic device applications. As previously emphasized, the required density of targeted defects is typically very low. Moreover, not every defect type contributes beneficially to device performance. Consequently, the creation and study of specific defects rely on the ability to first synthesize a semiconductor crystal of exceptionally high purity exhibiting negligible intrinsic defect density. Upon achieving this pristine base material, controlled introduction and subsequent investigation of the desired defect types and quantities become feasible. Based on spatial extent, defects are categorized as point defects, defect complexes, or line defects. Point defects involve only a few isolated atoms localized within the host crystal lattice. Defect complexes consist of several aggregated point defects. Line defects, primarily dislocations, involve one-dimensional rows of atoms. The functional point defects critical for the doping process will be the focus of this section.

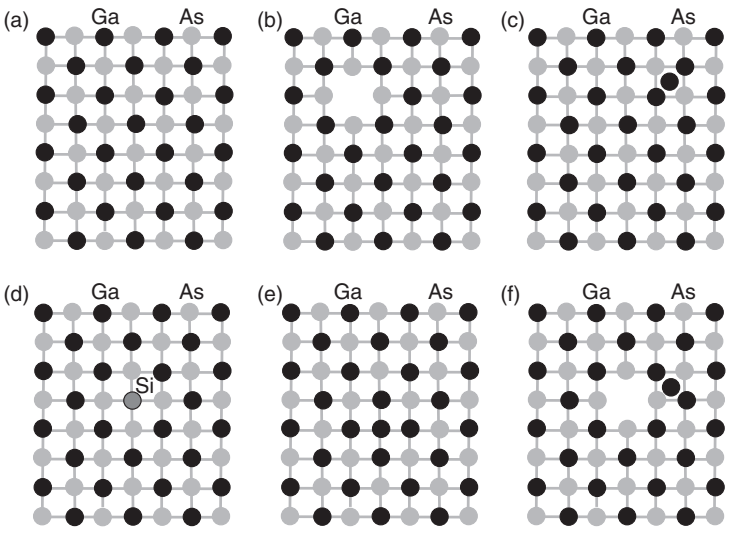

Figure 1.14 Schematic representation of various defects in GaAs in a 2D plane. (a) Pristine GaAs, (b) Ga vacancy defect (V_{Ga}), (c) Ga interstitial defect (I_{Ga}), (d) Ga substitution defect by a Si atom (Si_{Ga}), (e) Ga antisite defect (Ga_{As}), and (f) a Frenkel complex defect with Ga atom displaced from the lattice site but occupying a nearby interstitial site (V_{Ga}–I_{Ga}).

Point defects are further categorized into vacancy, interstitial, substitutional, antisite, and Frenkel pair. We exemplify these defects using GaAs (Figure 1.14). Figure 1.14a is the pristine GaAs. A vacancy in GaAs occurs due to a missing host atom, such as a missing Ga atom denoted V_{Ga} (Figure 1.14b). An interstitial defect arises when a host atom occupies an interstitial site, as exemplified by a Ga atom in an interstitial position, denoted I_{Ga} (Figure 1.14c). Substitutional defects result from the replacement of a host atom by an impurity atom, as when a Ga atom is substituted by a Si atom, denoted Si_{Ga} (Figure 1.14d). An antisite defect represents a specific substitutional type where one host atom replaces another host atom; for example, a Ga atom occupies an As site, denoted GaAs (Figure 1.14e). A Frenkel pair is a complex defect comprising a host atom displaced from its lattice site into a nearby interstitial position, illustrated by V_{Ga}–I_{Ga}, which involves a displaced Ga atom (Figure 1.14f). Defects involving impurity atoms are termed extrinsic defects. In contrast, defects such as vacancies, interstitials, and antisites, which lack impurity atoms, are referred to as intrinsic or native defects.

Electrically active defects, which can generate free carriers following thermal excitation, are categorized as donors or acceptors. Donors provide free electrons, whereas acceptors generate holes. Illustrating with silicon, substitutional defects involving Group V atoms (e.g., P, As, Sb) or interstitial defects involving monovalent atoms (e.g., Li, Na) act as donors (Figure 1.15a,b). Conversely, substitutional defects involving Group III atoms (e.g., B, Al, Ga, In) serve as acceptors (Figure 1.15c,d). Group V atoms possess one more valence electron than Group IV atoms, explaining their donor behavior, while Group III atoms lack one electron, accounting for their acceptor properties. Similarly, substitutional Group VI atoms (e.g., S, Se, Te)

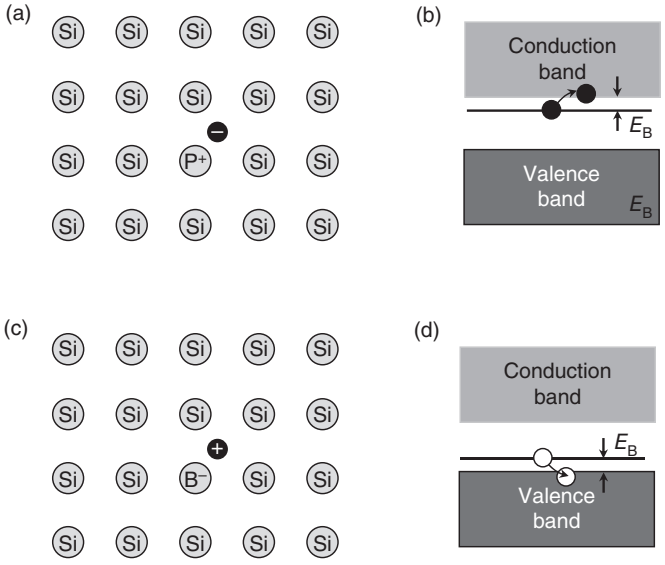

Figure 1.15 (a) Schematic representation of substitutional P doping of Si. (b) The donor level and free electron excitation in the conduction band of substitutional P defect. (c) Schematic representation of substitutional B doping of Si. (d) The acceptor level and free hole excitation in the valence band of substitutional B defect.

function as double donors, and substitutional Group II atoms (e.g., Be) function as double acceptors in silicon, releasing two free electrons or holes, respectively, upon excitation. Substitution by a Group IV foreign atom, such as C_{Si}, creates an isovalent center characterized by the same number of valence electrons as the host silicon lattice. Generally, isovalent centers remain electrically inactive; however, under some conditions, they can function as donors or acceptors.

The doping mechanism can be further elucidated through the effective mass approximation. For phosphorus (P) substitutional defects in silicon (Si), the P nucleus possesses one extra positive charge compared with the Si nucleus. However, the P atom also contributes an additional valence electron, rendering the defect charge-neutral. A Coulomb attraction exists between this extra positive charge and the additional electron. This interaction can be modeled using the effective mass approximation, wherein the extra electron is treated as a quasi-free electron in the conduction band, characterized by an effective mass (m^*) determined by the conduction band dispersion discussed later. This electron becomes bound to the positively charged P ion, analogous to the electron in a hydrogen atom. Crucially, the presence of the Si host crystal introduces a dielectric screening effect, distinct from the vacuum Coulomb potential in hydrogen. Incorporating this dielectric screening and the effective mass approximation, the binding energy for the electron bound to the positively charged donor is expressed as

$$E_B(n) = \frac{R_\infty}{n^2}\frac{m^*}{m_e}\frac{\varepsilon_0^2}{\varepsilon_s^2} \tag{1.57}$$

where $R_\infty = 13.6$ eV is the Rydberg constant, n is the principal quantum number for the bound states, m_e is the free electron mass, m^* is the conduction band effective mass, ε_0 is the vacuum permittivity, and ε_s is the static permittivity of the semiconductor. For the ground state ($n = 1$), the binding energy is maximal, and the corresponding Bohr radius of the electron probability density is

$$a^* = a_B \frac{m_e}{m^*} \frac{\varepsilon_s}{\varepsilon_0} \tag{1.58}$$

where $a_B = 0.529$ Å is the hydrogen Bohr radius. Given the anisotropy of silicon effective masses, a density-of-states average effective mass ($m^* = 0.36m_e$) and a static permittivity of $\varepsilon_s = 11.7\varepsilon_0$ yield a calculated binding energy of 36 meV. This value is reasonably close to the experimentally determined donor ionization energy of 45 meV. The calculated Bohr radius is 17.2 $\mathring{A}$, significantly larger than the silicon lattice constant (5.4 $\mathring{A}$), validating the application of the effective mass approximation. At room temperature, the thermal energy $kT = 25$ meV is sufficient to ionize this weakly bound electron into the conduction band, rendering P in Si an efficient donor. A parallel analysis applies to boron (B) doping in Si; a key distinction is that the B acceptor is negatively charged and binds a positively charged hole, analogous to an antihydrogen atom model.

Some defects, such as phosphorus-doped silicon (P_{Si}) or boron-doped silicon (B_{Si}), can be interpreted using the effective mass approximation as binding free electrons or free holes from the conduction band and valence band, respectively. These defects are commonly termed shallow impurities or shallow defects. However, certain defects possess properties that cannot be correctly explained by this effective mass approximation, and these are referred to as deep centers. Substitutional impurities resembling the host atoms, like P_{Si} in silicon or germanium-doped gallium arsenide (Ge_{Ga}) in GaAs, typically behave as shallow donors or acceptors. In contrast, if the substituting atoms differ significantly from the host atoms and induce a strongly localized potential, such as a substantial strain field, the resulting impurity introduces a deep-level center.

1.5.2 Doping Techniques

Doping can be implemented either during or after growth processes. For example, during bulk crystal growth or the epitaxial deposition of thin films, impurity dopants may be introduced concurrently with the source materials. However, such in situ doping inherently results in bulk or large-area doping distributions, which offer insufficient design flexibility for advanced device fabrication requirements. To fully unlock the potential of semiconductors, post-growth controllable doping techniques have been developed. Diffusion and ion implantation constitute the most extensively utilized and studied doping methodologies within the semiconductor industry, which will be detailed in the remainder of this section.

1.5.2.1 Thermal Diffusion

Diffusion is a doping technique that utilizes a concentration gradient to introduce dopant atoms from a high-concentration source into the semiconductor material.

Although the thermal motion of dopant atoms is inherently random, their net diffusivity depends exponentially on temperature and varies with specific dopant and semiconductor types. A concentration gradient induces net dopant transport from regions of high concentration to regions of low concentration. Conventional diffusion typically produces deep doping profiles. However, using doped oxides (e.g., borosilicate glass, BSG, for boron) or nitrides as diffusion sources combined with rapid thermal annealing (RTA) enables the realization of relatively shallow doping profiles. Selective doping, such as that required for the source and drain regions of MOSFETs, is essential for semiconductor device fabrication. Selective doping within a defined window region can be achieved via diffusion by patterning a mask layer (e.g., thick silicon dioxide or silicon nitride) to form the diffusion window.

Original sources containing dopants at high concentrations may exist in gaseous, liquid, or solid phases, as illustrated in Figure 1.16. For boron (B) doping of silicon, gaseous trimethylboron $(CH_3)_3B$ or diborane B_2H_6, liquid boron tribromide BBr_3, or a solid boron nitride (BN) disc coated with B_2O_3 may serve as the source. However, the actual source material facilitating boron diffusion is a layer of B_2O_3 deposited on the silicon surface. This intermediary B_2O_3 layer is introduced by oxidizing the gaseous, liquid, or solid precursors. The B_2O_3 subsequently reacts with elemental silicon, generating boron atoms according to the following reaction:

$$2B_2O_3 + 3Si \rightarrow 3SiO_2 + 4B \tag{1.59}$$

Similarly, gaseous phosphine PH_3, liquid phosphorus oxychloride $POCl_3$, or solid phosphorus pentoxide P_2O_5 functions as sources for phosphorus (P) doping of silicon. Consequently, the effective diffusion source is P_2O_5 deposited on the silicon surface. In all cases, the reaction between the precursor source material and silicon results in the simultaneous growth of an oxide layer.

The diffusion of dopants relies on defects for assistance (Figure 1.17). Diffusion mechanisms can be categorized into vacancy-assisted and interstitial-assisted diffusion. In vacancy-assisted diffusion, the impurity migrates via transition into

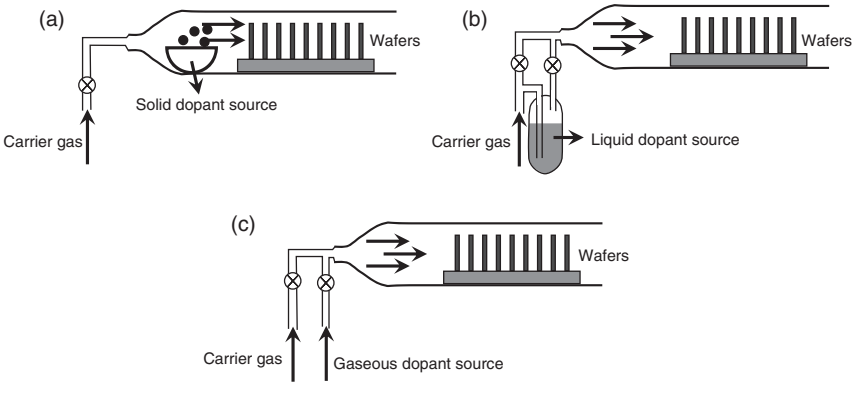

Figure 1.16 Schematic representations of diffusion systems with (a) solid, (b) liquid, and (c) gaseous dopants.

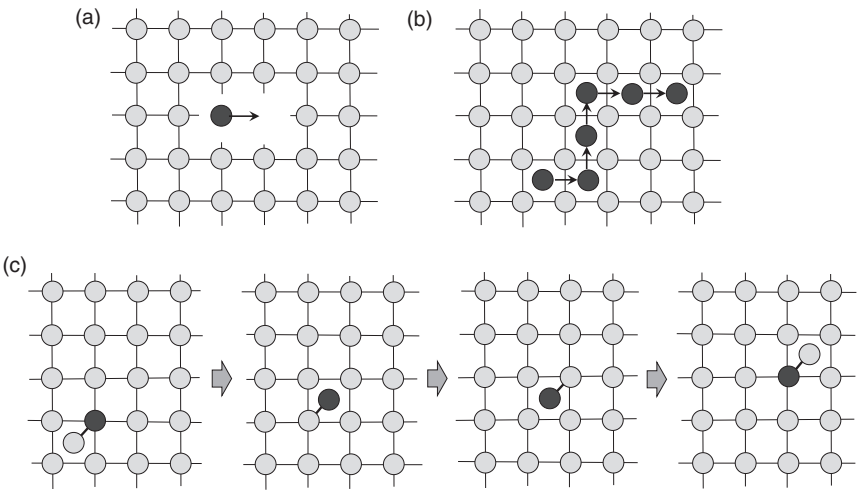

Figure 1.17 Diffusions of impurity atoms by (a) impurity jumping into a vacancy site, (b) the interstitial sites, and (c) impurity-defect pair (interstitialcy).

a vacant lattice site. Conversely, interstitial-assisted diffusion entails an impurity migrating from one interstitial position to another. Small-sized elements—such as hydrogen, helium, argon, lithium, sodium, and potassium (Group I and II)—exhibit rapid diffusion kinetics in silicon crystals. This phenomenon occurs consequently to interstitial-assisted diffusion. The interstitial mechanism also prevails for diffusion of boron, phosphorus, aluminum, or gallium in silicon. Conversely, the diffusion of arsenic and antimony is dominated by vacancy-assisted diffusion.

Dopant atoms are introduced and redistribute within the semiconductor wafer inside a furnace at elevated temperatures. The resultant doping profiles can be controlled by the source concentration, diffusion temperature, and diffusion time. Increasing the temperature enhances the diffusivity of the dopant species. A higher source concentration, higher diffusion temperature, or longer diffusion time generally increases the junction depth. To quantitatively model the diffusion process, Fick's laws of diffusion are applied. According to the diffusion flux equation derived from this theory, which remains valid under dilute impurity concentration conditions ($<10^{18}/\text{cm}^3$), the flux J, defined as the net rate of particle flow through a unit area perpendicular to the diffusion direction, is given by

$$J = -D\frac{\partial n}{\partial x} \tag{1.60}$$

where D is the diffusion coefficient (diffusivity) in units of cm^2/s, and $\partial n/\partial x$ is the impurity concentration gradient. The negative sign indicates that diffusion occurs down the concentration gradient, from regions of high concentration to regions of low concentration. The magnitude of the diffusivity governs the speed of the diffusion process and depends on the specific impurity type and the host crystal structure.

The diffusivity exhibits a strong temperature dependence described by the following Arrhenius relationship:

$$D = D_0 e^{-\frac{E_A}{kT}} \tag{1.61}$$

where D_0 is an impurity-specific constant and E_A is a measure of the diffusion energy barrier. For B and P diffusion in silicon, the D_0 values are 0.76 and 3.85 cm^2/s, respectively, and the E_A values are 3.46 and 3.66 eV, respectively. The diffusion will change the local impurity concentration in the semiconductor by the following continuity equation.

$$\frac{\partial n}{\partial t} = -\frac{\partial J}{\partial x} = D\frac{\partial^2 n}{\partial x^2} \tag{1.62}$$

This equation, obtained by substituting Fick's first law into the continuity condition, states that the rate of concentration change is determined by the spatial gradient of the flux. Subject to appropriate initial and boundary conditions, this partial differential equation can be solved to determine the evolving dopant distribution profile.

Under specific conditions, analytical solutions exist for the diffusion profile. For diffusion driven by a dopant source maintaining a constant surface concentration (as depicted in Figure 1.18a), the time-evolving concentration profile is given by

$$n(x,t) = n_0 \text{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \tag{1.63}$$

where n_0 represents the constant dopant concentration at the surface, and erfc(x) is the complementary error function. In practice, dopant introduction during furnace diffusion initially forms a predeposition layer near the semiconductor surface. The surface dopant concentration is typically established by either the dopant partial pressure in the gaseous source or the solid solubility limit, thus validating the assumption of constant interface concentration used in the solution. In Eq. (1.63), the diffusion length is defined as $L = \sqrt{Dt}$. As time proceeds, L increases, signifying greater penetration depth of the impurities into the semiconductor. Integrating the

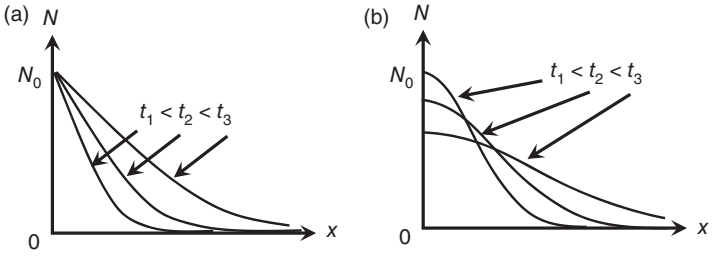

Figure 1.18 Impurity profiles with increasing time for (a) constant source diffusion and (b) constant dose diffusion.

impurity concentration yields the total dopant amount per unit area, Q, within the semiconductor after a specific diffusion time.

$$Q(t) = \frac{2n_0\sqrt{Dt}}{\sqrt{\pi}} \tag{1.64}$$

Analytical solutions for diffusion can also be derived when the dopant source is located at the semiconductor surface with a fixed amount of impurity concentration per unit area (Figure 1.18b). These surface dopants are subsequently driven deeper into the semiconductor during the diffusion process. In accordance with Fick's diffusion laws, the diffused dopant concentration conforms to a Gaussian distribution function.

$$n(x,t) = \frac{Q}{\sqrt{\pi Dt}} e^{-\frac{x^2}{4Dt}} \tag{1.65}$$

1.5.2.2 Ion Implantation

Besides diffusion, ion implantation serves as another fundamental process for doping semiconductors. In contrast to diffusion, implantation is a low-temperature technique that introduces dopants by accelerating dopant ions to high energies and bombarding the semiconductor. The penetration depth of the ions depends on both the dopant species and the incident ion energy. Ion implantation currently plays a central role in semiconductor circuit fabrication.

An implanter is utilized for ion implantation, which consists of an ion source, an accelerating column, a mass analyzer, a scanning system, and a wafer chamber (Figure 1.19). The final implantation performance is influenced by the ion type, kinetic energy, and dose. To generate energetic ions, source materials in gaseous, liquid, or solid phases are first vaporized or sputtered and introduced as gases into the source chamber. These gases are subsequently ionized within a plasma system generated via a hot filament or microwave discharge. Desired positive ions within the plasma are then extracted and accelerated to high energies (typically ranging from 10 to 100 keV) by the potential difference applied between the source chamber and the extraction electrode. Singly charged ions, such as B^+, BF_2^+, P^+,

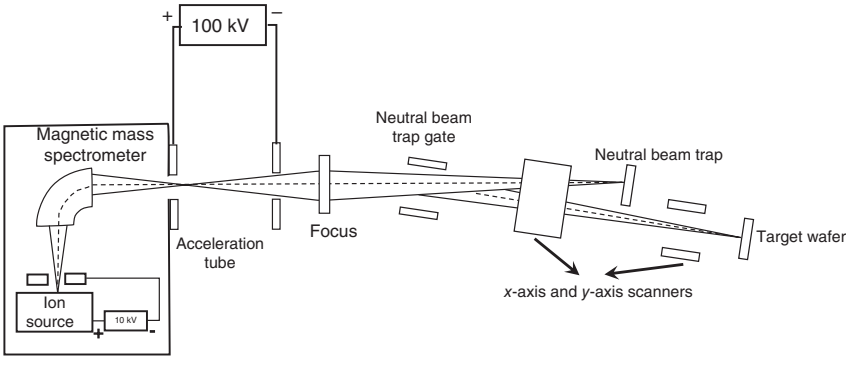

Figure 1.19 Schematic structure of an implantation system.

and Sb^+, constitute the most abundant species in the plasma, while doubly and triply charged ions are significantly less prevalent. For a given applied acceleration voltage, multiply charged ions attain higher kinetic energies than their singly charged counterparts. This facilitates achieving deeper doping profiles, albeit at the cost of a reduced implantation flux. For instance, an ion energy of 200 keV can be attained by accelerating doubly charged ions using a 100 keV accelerator.

The desired ions are selected by the mass analyzer. A magnetic field is applied perpendicular to the direction of ion transport, thereby deflecting the ions according to their mass-to-charge ratio. The deflection radius R is calculated by

$$R = \sqrt{\frac{2mV}{qB^2}} \tag{1.66}$$

where V is the accelerating voltage, B is the magnetic field strength, and m/q is the mass-to-charge ratio of the ions.

The deflected ions are selected by a slot aperture, which usually has the desired ion species, and can be further focused and accelerated to higher energies before being directed to the target substrate for implantation. To avoid Coulombic repulsion forces during transport, trapped electrons are often injected into the ion beam. Additionally, a low-energy electron shower is applied to the target wafer to maintain charge-neutrality during implantation, minimizing beam disturbance. Since the ion beam possesses a relatively narrow diameter, homogeneous implantation across the semiconductor wafer requires uniform sweeping of the beam over the wafer surface within the target chamber. Common sweeping techniques include the use of electrostatic deflection plates, mechanical motion of the wafer, or a combination of both scanning methods.

Ion implantation offers several advantages. Firstly, it enables doping profiles inaccessible via diffusion. Diffusion, driven by concentration gradients, inherently produces higher dopant concentrations near the semiconductor surface. In contrast, implantation provides superior profile control, allowing the realization of designed impurity distributions unattainable through simple diffusion. It can introduce dopants deep within the substrate, forming buried layers where the highest concentration resides internally or creating retrograde profiles through implants at single or multiple energies. Following ion implantation, the dopants are typically electrically inactive, necessitating high-temperature annealing for activation. However, the annealing duration for implantation is substantially shorter, resulting in a significantly lower thermal budget. RTA is commonly employed, capable of achieving temperature ramping rates of ~150 °C/s to ~1000 °C, compared with furnace processing rates of ~1 °C/s. Consequently, ion implantation is the predominant technique for doping gallium arsenide (GaAs), as the compound semiconductor decomposes substantially at the elevated temperatures required for diffusion. Furthermore, the lower temperatures involved permit the use of temperature-sensitive materials, such as photoresist, as implantation masks for selective doping within window regions.

Ion implantation does exhibit certain disadvantages. Primarily, high-energy ions induce crystalline radiation damage. These defects enhance the diffusivity

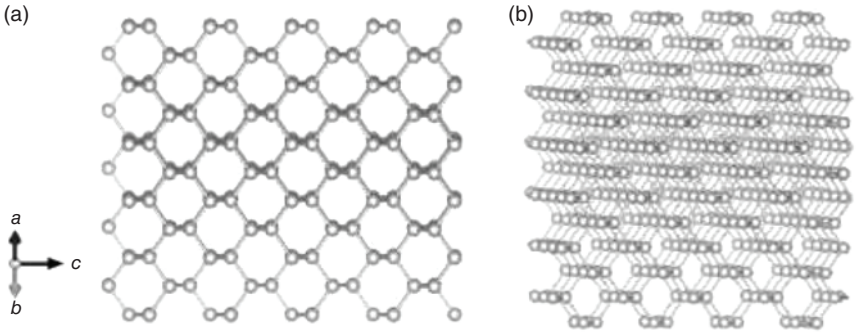

Figure 1.20 (a) Silicon crystal structure viewed from <110> direction with strong channeling effect during implantation. (b) Silicon structure tilted 10° from <110> direction.

of specific species, rendering the fabrication of ultra-shallow doping profiles challenging, even at comparatively low annealing temperatures. Additionally, residual radiation defects inevitably remain within the crystal lattice, increasing junction leakage currents in fabricated devices. Furthermore, energetic ions cause sputtering of metal surfaces, potentially contaminating the wafer. Unlike implantation in amorphous materials, high-energy ions penetrating crystalline semiconductors along specific crystallographic directions undergo the channeling effect, leading to deviations from the intended doping concentration. Controlling the incident ion beam angle or forming an amorphous surface layer mitigates this effect. For <110> wafers, perpendicular incidence induces significant channeling with substantial ion penetration depths; however, tilting the wafer 10° off-normal yields an approximately random ion distribution (Figure 1.20). The substantial kinetic energy carried by implanted ions thermally dissipates upon striking mask materials, necessitating restricted implantation fluxes to prevent overheating thermally sensitive components such as photoresists, whose operational temperatures typically must not exceed ~100 °C. Finally, the inherently low ion flux compromises throughput compared with furnace diffusion techniques.

1.6 Energy Band Theory

1.6.1 Quantum Mechanical Equation and Bloch Wavefunction

The investigation of electron behaviors within semiconductors constitutes the most fundamental approach to understanding their electronic properties. However, semiconductor crystals consist of atoms arranged periodically along specific dimensions, resulting in electron densities typically on the order of 10^{23} electrons/cm^3. Solving problems involving such a vast number of interacting electrons and nuclear ions proves formidable. Consequently, approximations become necessary to simplify this complexity.

Atomic ions are much heavier than electrons. For example, the proton in a hydrogen atom is approximately 1836 times heavier than an electron. Since the

kinetic energy E of a particle depends on momentum p according to $E = p^2/2m$ and the velocity v is given by $v = p/m$, both kinetic energy and velocity are inversely proportional to the particle mass m. Consequently, for identical momentum, both the kinetic energy and velocity of an electron are significantly larger than those of the heavier proton.

Owing to the considerably smaller energies and slower response times of ions, electrons respond instantaneously to ionic displacements, whereas ions cannot track the rapid electron motion and instead experience a time-averaged electron potential. For investigations of electronic properties, ions may therefore be regarded as stationary. This approximation, which is fundamental in quantum mechanics, is termed the Born–Oppenheimer or adiabatic approximation.

Accurate understanding of the electronic properties of a semiconductor necessitates the application of quantum mechanics to solve the kinetics and dynamics of electrons within the crystal lattice. Utilizing the Born–Oppenheimer approximation, the electronic Hamiltonian can be expressed as

$$\hat{H}_e = \sum_i \frac{\hat{p}_i^2}{2m_e} + \frac{1}{2}\sum_{i,i'} \frac{e^2}{4\pi\varepsilon_0 |\hat{r}_i - \hat{r}_{i'}|} - \sum_{i,j} \frac{Z_j e^2}{4\pi\varepsilon_0 |\hat{r}_i - R_{j0}|} \tag{1.67}$$

where $\hat{p}_i$ and $\hat{r}_i$ represent the momentum and position vectors of electron i, ε_0 denotes the vacuum permittivity, and Z_j corresponds to the atomic number of the positively charged ion at equilibrium position R_{j0}. In this equation, the first term signifies the kinetic energy of the electrons, the second term describes the Coulomb interaction energy between electrons, and the third term accounts for the Coulomb interaction energy between electrons and ions. The ion–ion interactions can be treated as essentially stationary due to the Born–Oppenheimer approximation (Figure 1.21a). The state of the electrons is characterized by

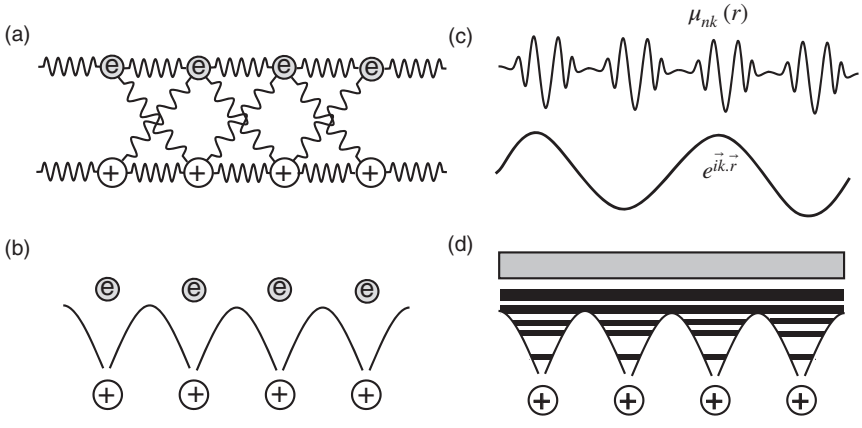

Figure 1.21 (a) Electron–ion, electron–electron, and ion–ion interactions in a crystal. (b) Mean-field potential experienced by electrons as independent particles with fixed ions in the Born–Oppenheimer approximation. (c) The periodic part and plane wave part of the Bloch wavefunction. (d) Filling of the energy levels from low energy to high energy according to Pauli principle.

the many-particle wavefunction $\Phi(r_1, r_2, \ldots, r_N)$ for an N-electron system. This wavefunction is determined by solving the eigenvalue problem derived from the Schrödinger equation.

$$\hat{H}_e\Phi(r_1, r_2, \ldots, r_N) = E\Phi(r_1, r_2, \ldots, r_N) \tag{1.68}$$

where E represents the total energy of the electron system.

Solving this many-body problem in quantum mechanics remains intractable. Consequently, another approximation, known as the mean-field approximation, simplifies this many-body problem to that of a single particle. Within this approximation, the complex electron–electron interactions in the Hamiltonian are represented by the interactions of independent electrons with a mean field generated by the electron ensemble. Therefore, all electrons become independent under the influence of the same average potential $V(r)$, which combines both the electron-derived potential (e.g., the Hartree potential) and the ionic potential. The Schrödinger equation describing the motion of each electron is thus expressed as

$$\hat{H}\Phi_n(r) = \left(\frac{\hat{p}^2}{2m_e} + V(r)\right)\Phi_n(r) = E_n\Phi_n(r) \tag{1.69}$$

where $\Phi_n(r)$ denotes the wavefunction for the n_{th} electron and E_n represents its corresponding energy eigenvalue. The wavefunctions and eigenvalues are solved independently using this mean-field approach.

The one-electron potential $V(r)$ is determined by the atomic species, atomic positions, and the electron density. The atomic nucleus generates a Coulombic potential, which can lead to complex effects such as rapid oscillations in the wavefunction near the nucleus, particularly for atoms with high atomic numbers and wavefunctions characterized by large quantum numbers. To circumvent these difficulties, a smoothly varying pseudopotential is typically employed as an approximation. This substitution enables the description of the wavefunction using fewer Fourier components, thereby lowering the cutoff energy required for plane wave basis sets and facilitating numerical convergence with computationally feasible resources. The pseudopotential yields a pseudo-wavefunction that accurately reproduces the true wavefunction outside the core region. After defining the potential of the atomic nuclei, the electron density must also be determined to account for the mean-field electron–electron interaction. This allows the final one-electron potential $V(r)$ to be obtained for solving the one-electron wavefunctions (Figure 1.21b). Because the electron density is itself determined by the occupied one-electron wavefunctions, this calculation must be performed self-consistently, typically through iterative cycles updating the wavefunctions and the electron density.

After getting the one-electron potential $V(r)$, solving the Schrödinger equation for a crystal with so many electrons is still a complicated task. Fortunately, if the crystal is large enough, it can be treated as a system with an infinite size and a translation symmetry, which can greatly simplify the problem. The translation symmetry means that the system is unchanged after translation along the Bravais lattice. We can introduce the translation operator T_R, which transform a function $f(x)$ by

$$\hat{T}_R f(x) = f(x + R) \tag{1.70}$$

where $R = n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3$ is a Bravais lattice vector. The translational symmetry of the crystal system implies that the operators $\hat{T}_R$ and $\hat{H}$ commute, denoted by the operator equation.

$$\hat{H}\hat{T}_R\Phi_n(r) = \hat{T}_R\hat{H}\Phi_n(r) = E_n\hat{T}_R\Phi_n(r) \tag{1.71}$$

Therefore, we recognize that $\hat{T}_R\Phi_n(r)$ is also an eigenfunction of the Hamiltonian operator $\hat{H}$ with the identical eigenvalue E_n. Since the translation operator $\hat{T}_R$ and $\hat{H}$ commute, eigenfunctions $\Phi_n(r)$ can always be constructed to be simultaneous eigenfunctions of both $\hat{T}_R$ and $\hat{H}$. Consequently,

$$\hat{T}_R\Phi_n(r) = \Phi_n(r+R) = e^{i\vec{k}\cdot\vec{R}}\Phi_n(r) \tag{1.72}$$

where the wave vector $\vec{k}$ arises from the phase difference induced by the lattice translation vector $\vec{k}$. Consequently, in $\Phi_{nk}(r)$, n and k serve as quantum indices characterizing the eigenfunctions $\Phi_{nk}(r)$ and eigenvalues E_{nk}. The eigenfunction $\Phi_{nk}(r)$ can be expressed in the form

$$\Phi_{nk}(r) = e^{i\vec{k}\cdot\vec{r}}\mu_{nk}(r) \tag{1.73}$$

where $\mu_{nk}(r)$ is a periodic lattice function satisfying $\mu_{nk}(r) = \mu_{nk}(r + R)$. This expression is known as a Bloch function (Figure 1.21c), indicating that the electronic wavefunction within a crystal, while similar to a plane wave characteristic of free electrons, is spatially modulated by a periodic function. We note that electrons are fermions governed by the Pauli exclusion principle; therefore, each distinct wavefunction can accommodate a maximum of two electrons possessing opposite spin orientations. In the ground state of the electron ensemble, electronic states are populated sequentially starting from the lowest available energy levels (Figure 1.21d).

1.6.2 Band Structures and Carrier Transport

For a reciprocal lattice vector $\vec{G} = m_1\vec{b}_1 + m_2\vec{b}_2 + m_3\vec{b}_3$, we know that $\vec{G}\cdot\vec{R} = 2\pi(m_1n_1 + m_2n_2 + m_3n_3)$, which results in no net phase shift. Consequently, both the wave vectors $\vec{k}$ and $\vec{k} + \vec{G}$ yield identical phase shifts upon lattice translation and can serve as quantum numbers to index the wavefunction. Conventionally, only the equivalent wave vector $\vec{k}$ with the smallest magnitude is employed. The region encompassing all such possible $\vec{k}$ vectors is termed the first Brillouin zone. A plot of the eigenvalue E_{nk} versus $\vec{k}$ within the first Brillouin zone gives the electronic band structure, which serves as a fundamental tool for characterizing and understanding the electronic properties of crystals. The first Brillouin zone corresponds to the Wigner–Seitz cell of the reciprocal lattice. It is defined as the set of points in reciprocal space closer to the origin than to any other reciprocal lattice point. This region, containing the origin point, is constructed by bisecting all reciprocal lattice vectors (Figure 1.22). Examples of first Brillouin zones for typical lattices are depicted in Figure 1.23. The reciprocal lattices inherit the symmetries of their corresponding real-space lattices. Consequently, electronic band diagrams are

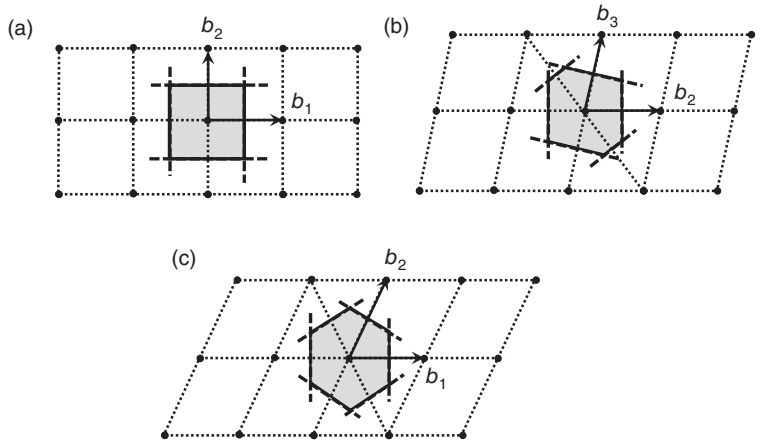

Figure 1.22 Construction of first Brillouin zones of (a) simple cubic lattice, (b) monoclinic lattice, and (c) hexagonal lattice by bisecting the reciprocal lattice vectors.

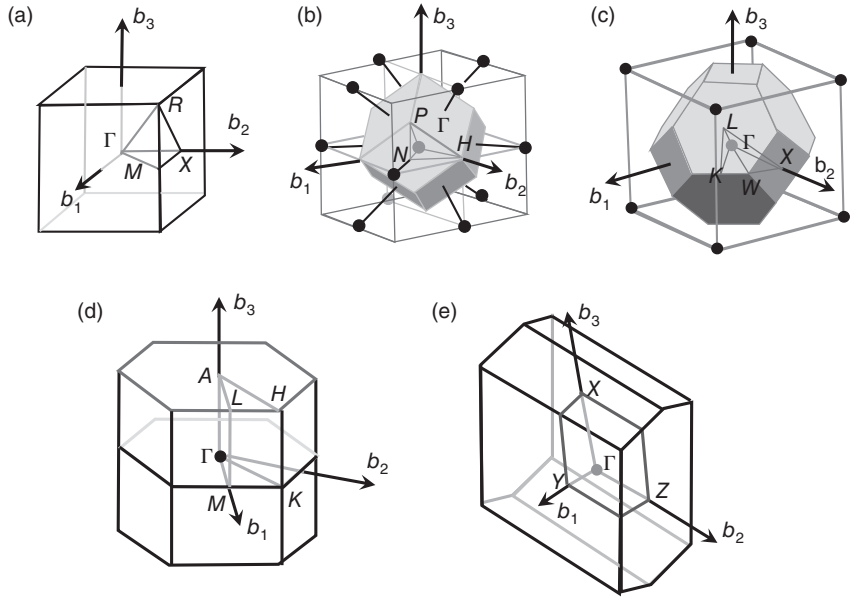

Figure 1.23 First Brillouin zones and high-symmetry point notations of (a) simple cubic lattice, (b) body-centered cubic lattice, (c) face-centered cubic lattice, (d) hexagonal lattice, and (e) monoclinic lattice.

typically plotted along $\vec{k}$-space paths connecting high-symmetry points within the Brillouin zone. Conventionally, high-symmetry points and lines inside the Brillouin zone are denoted by Greek letters, while points on the zone surfaces are denoted by Roman letters. Γ invariably designates the zone origin. For cubic lattices, X commonly represents the intersection point of the Brillouin zone surface with any of the main axes.

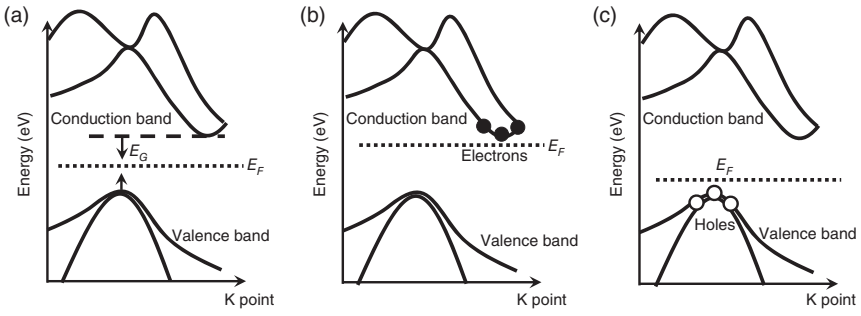

Figure 1.24 Schematic band diagrams and carrier types in (a) the intrinsic, (b) *n*-type, and (c) *n*-type semiconductors.

Electrons occupy the electronic band structure from low to high energies. Each energy band accommodates up to two electrons in the primitive cell, accounting for the spin degeneracy of electrons. Depending on the band filling, crystals are classified as metals, semiconductors, or insulators. Metals possess partially filled bands, with the Fermi level traversing one or more bands. In contrast, semiconductors and insulators have fully occupied valence bands and unoccupied conduction bands, separated by a bandgap, as shown in Figure 1.24a. Semiconductors are distinguished from insulators by their smaller bandgap. Semimetals, such as graphene, constitute an exception with an absence of a bandgap. The presence or absence of a bandgap profoundly influences the electron transport properties of a crystal. Electrons in completely filled bands cannot sustain net current flow, explaining the low conductivity of insulators and intrinsic semiconductors at cryogenic temperatures.

However, after doping with impurities, electrostatic modulation, or more directly, optical or thermal excitation, the formerly completely occupied or unoccupied bands can become partially occupied, enabling electrical conduction. The previously unoccupied bands are termed conduction bands, while the previously occupied bands are termed valence bands. Carrier transport in the conduction or valence bands can be understood by treating the carriers analogously to classical particles, referred to as electrons in the conduction bands or holes in the valence bands, respectively (Figure 1.24b,c). This analogy arises because electrons and holes are predominantly located at the conduction band minimum or valence band maximum, which exhibit nearly parabolic energy dispersion akin to classical particles, albeit with distinct effective masses. Specifically, the effective mass near the conduction band minimum can be derived from the electronic band structure using the relation:

$$m^* = \frac{\hbar^2}{\frac{\partial^2 E}{\partial k^2}} \tag{1.74}$$

where an isotropic band dispersion along all wave vector directions is assumed for simplicity. In contrast, anisotropic band dispersions necessitate a description using anisotropic effective masses.

Electrons with wave vector $\vec{k}$ in the conduction band can be treated as a wave packet with momentum $\vec{p} = \hbar\vec{k}$, where $\hbar$ is the Planck constant. If an external

force $\vec{F}$, such as that produced by an electric field $\vec{E}$ or a magnetic field $\vec{B}$, acts on the electron, the wave vector evolves according to the following equation:

$$\hbar \frac{d\vec{k}}{dt} = \vec{F} = q\vec{E} + q\vec{v} \times \vec{B} \tag{1.75}$$

where q is the elementary charge and $\vec{v}$ is the group velocity. For a band dispersion $E(\vec{k})$, the group velocity can be calculated by

$$\vec{v} = \frac{1}{\hbar} \nabla_k E(\vec{k}) \tag{1.76}$$

As an example, for one-dimensional band dispersion along the x axis, the group velocity is given by $v_x = \partial E/\hbar\, \partial k_x$. For an electron with an effective mass m^*, the band dispersion near the conduction band minimum can be approximated as $E = \hbar^2 k^2/2m^*$, yielding a group velocity $v = \hbar k/m^*$, analogous to a classical particle. Holes in the valence band can similarly be described as classical particles. For this interpretation to hold, the charge, momentum, and energy assigned to a hole must exhibit signs opposite to those of the corresponding electron states within the band structure. For instance, a hole possesses a charge of the same magnitude as an electron, albeit opposite in sign (positive).

1.7 PN Junctions

A PN junction is formed by bringing p-type and n-type semiconductors into contact (Figure 1.25a). Studying the current conduction characteristics of PN junctions is

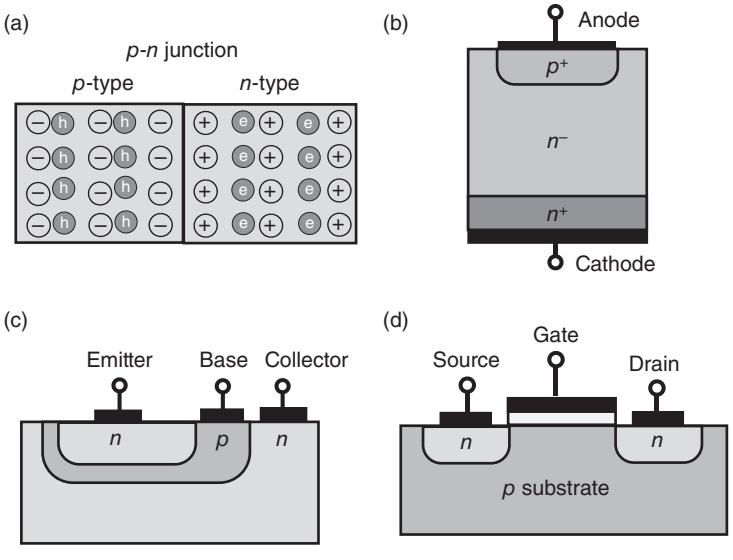

Figure 1.25 (a) Schematic of a PN junction formed by bringing n-type and n-type semiconductors into contact. Cross-sectional structures of (b) a $p^+ - n^- - n^+$ junction diode, (c) NPN bipolar junction transistor, and (d) an n-type metal–oxide–semiconductor field-effect transistor. The PN junctions can be found in these devices.

crucial for implementing and understanding semiconductor devices. An essential and fundamental device based on the PN junction is the PN diode (Figure 1.25b). This two-terminal device exhibits nonlinear current–voltage characteristics, conducting with low resistance in one direction, while presenting high resistance in the reverse direction. Consequently, diodes are widely used as rectifiers to convert alternating current (AC) to direct current (DC). Additionally, specific diodes are engineered to emit light efficiently for use as illumination sources. Furthermore, PN junctions constitute the fundamental building blocks of BJTs and MOSFETs (Figure 1.25c,d).

1.7.1 Depletion Region and Built-in Potential

To understand the properties of the PN junction, it is necessary to first consider the depletion region formed due to charge transfer. The abrupt junction serves as an illustrative model for this phenomenon. An abrupt junction consists of semiconductors undergoing an abrupt transition from acceptor doping to donor doping (Figure 1.26). Owing to the concentration gradients of electrons and holes, electrons diffuse from the *n*-type region into the *p*-type region, while holes diffuse from the *p*-type region into the *n*-type region. Since electrons are negatively charged and holes are positively charged, this diffusion process establishes space-charge regions. The resulting space charge generates an electric field, which induces a drift current of carriers in the opposite direction to the diffusion currents (Figure 1.26a). Consequently, the carrier dynamics within the PN junction ultimately reach thermal equilibrium.

At thermal equilibrium without an applied voltage, no net current flows. From the perspective of carrier transport, the electron and hole currents are directly related to the gradients of their respective quasi-Fermi levels, as expressed by

$$J_e = n_e \mu_e \frac{\mathrm{d}E_{\mathrm{Fe}}}{\mathrm{d}x} \tag{1.77}$$

$$J_h = n_h \mu_h \frac{\mathrm{d}E_{\mathrm{Fh}}}{\mathrm{d}x} \tag{1.78}$$

from which we know at equilibrium the quasi-Fermi level will be equal and constant in the whole sample, i.e., $E_{\mathrm{F}} = E_{\mathrm{Fe}} = E_{\mathrm{Fh}}$. The alignment of Fermi level is achieved by

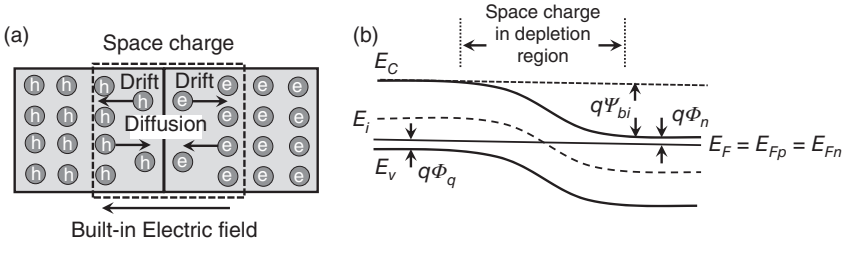

Figure 1.26 (a) The static equilibrium formed due to the balance of diffusion and drift currents in a PN junction. (b) Energy band diagram of a PN junction at equilibrium with aligned quasi-Fermi levels.

the built-in potential in the depletion region. According to the notation in the band diagram of Figure 1.26b, the built-in potential strength will simply be

From these equations, it follows that at equilibrium, the quasi-Fermi levels are equal and constant throughout the entire sample, denoted by $E_F = E_{Fe} = E_{Fh}$. This alignment of the Fermi level is facilitated by the built-in potential within the depletion region. Referring to the notation in the band diagram of Figure 1.26b, the magnitude of the built-in potential is given by

$$\psi_{bi} = E_G - \phi_n - \phi_p \tag{1.79}$$

For nondegenerate semiconductors, this equation can be further expressed as

$$\psi_{bi} = \frac{kT}{q} \ln\left(\frac{n_{n0}}{n_i}\right) + \frac{kT}{q} \ln\left(\frac{p_{p0}}{n_i}\right) \tag{1.80}$$

where n_{n0} and p_{p0} are the electron concentration in the n-region and the hole concentration in the p-region, respectively, and n_i is the electron or hole concentration of the intrinsic semiconductor. Since the relationships $n_{n0}p_{n0} = n_i^2$ and $p_{p0}n_{p0} = n_i^2$ hold for carriers in the *n*- and *p*-regions, the expression can be equivalently represented as

$$\psi_{bi} = \frac{kT}{q} \ln\left(\frac{n_{n0}}{n_{p0}}\right) = \frac{kT}{q} \ln\left(\frac{p_{p0}}{p_{n0}}\right) \tag{1.81}$$

where n_{p0} and p_{n0} are the minority-carrier concentrations in the *n*-type and *p*-type semiconductor, respectively. The aforementioned derivations rely on the Boltzmann statistics for nondegenerate semiconductors. For degenerate semiconductors, Fermi–Dirac statistics must be applied.

The built-in potential arises from the space charge within the depletion region. To facilitate analysis, it is commonly assumed that the space charge adopts a box profile in the depletion region, wherein all donors and acceptors are fully depleted (Figure 1.27a). No space charge exists outside this region. Given the donor and acceptor concentrations N_D and N_A, respectively, the depletion width W_D and W_A in the *n*-type and *p*-type semiconductors satisfy $N_D W_D = N_A W_A$ due to charge conservation. The electric potential is governed by the Poisson equations:

$$\frac{d^2\varphi}{dx^2} = \frac{qN_A}{\varepsilon_S} \quad \text{for } -W_A < x < 0 \tag{1.82}$$

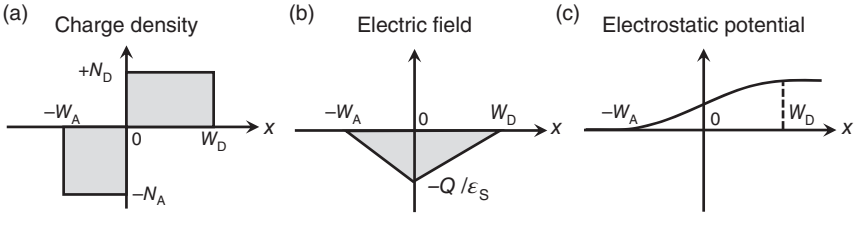

Figure 1.27 (a) Space-charge distribution in the depletion region of PN junction according to the box-profile approximation. (b) Electric field distribution generated by the space charge. (c) Electrostatic potential obtained by integration of the electric field.

$$\frac{d^2\varphi}{dx^2} = -\frac{qN_D}{\varepsilon_S} \quad \text{for } 0 < x < W_D \tag{1.83}$$

Integrating these equations yields the electric field distribution (Figure 1.27b), which exhibits its maximum magnitude at $x = 0$.

$$E_{max} = -\frac{d\varphi}{dx}(x = 0) = -\frac{qN_A W_A}{\varepsilon_S} = -\frac{qN_D W_D}{\varepsilon_S} \tag{1.84}$$

Further integration of the electric field (Figure 1.27c) provides the total potential difference across the PN junction as

$$\psi_{bi} = \frac{qN_A {W_A}^2}{2\varepsilon_S} + \frac{qN_D {W_D}^2}{2\varepsilon_S} \tag{1.85}$$

From this expression, combined with $N_D W_D = N_A W_A$, the depletion widths are derived as

$$W_D = \sqrt{\frac{2\varepsilon_S \psi_{bi} N_A}{qN_D(N_D + N_A)}} \tag{1.86}$$

$$W_A = \sqrt{\frac{2\varepsilon_S \psi_{bi} N_D}{qN_A(N_D + N_A)}} \tag{1.87}$$

A special case is the one-sided abrupt junction, featuring a heavily doped *p*-type region or *n*-type region; the depletion width equations then simplify for p^+-*n* and n^+-*p* junctions, respectively.

$$W_D = \sqrt{\frac{2\varepsilon_S \psi_{bi}}{qN_D}} \tag{1.88}$$

$$W_A = \sqrt{\frac{2\varepsilon_S \psi_{bi}}{qN_A}} \tag{1.89}$$

1.7.2 Current-Voltage Characteristics

When a voltage is applied across the terminals of a PN junction, the static equilibrium condition is disrupted, yielding carrier transport and net current flow. The current flow behavior in a PN junction under ideal conditions is described by the Shockley equation. The derivation of the Shockley equation relies upon several key assumptions. First, the depletion region is assumed to remain abrupt even under applied bias, with the semiconductor remaining neutral outside this region. Second, the semiconductor is treated as nondegenerate so that Boltzmann statistics are applicable. Third, the magnitude of minority-carrier injection induced by the applied voltage is significantly smaller than the majority-carrier concentration on either side. Finally, owing to the relatively short depletion width compared to the diffusion length, carrier recombination and generation within the depletion region are negligible.

Even though the carrier statistics depart from static equilibrium conditions, quasi-Fermi levels remain valid for describing the nonequilibrium state. Analogous to thermal equilibrium conditions, the electron and hole concentrations corresponding to quasi-Fermi levels E_{Fe} and E_{Fh} are expressed by

$$n = n_i e^{\frac{E_{\mathrm{Fe}} - E_i}{kT}} \tag{1.90}$$

$$p = n_i e^{\frac{E_i - E_{\mathrm{Fh}}}{kT}} \tag{1.91}$$

The quasi-Fermi levels across a PN junction under forward and reverse bias can be qualitatively represented (Figure 1.28a,b). At forward bias, $pn = {n_i}^2 e^{(E_{\mathrm{Fe}} - E_{\mathrm{Fh}})/kT}$ exceeds n_i^2, indicating that injected minority carriers undergo recombination. In contrast, under reverse bias, pn falls below n_i^2, signifying that electrons and holes undergo generation within the space-charge region.

Under the low-injection assumption, within the charge-neutral region outside the space-charge region, there is no drift current; thus, only diffusion current and carrier recombination require consideration. For the minority-carrier holes in the n-type region, at steady state, the hole density n_p satisfies the following equation:

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(-D_n \frac{\mathrm{d}n_p}{\mathrm{d}x}\right) = \frac{n_p - n_{p0}}{\tau_n} \tag{1.92}$$

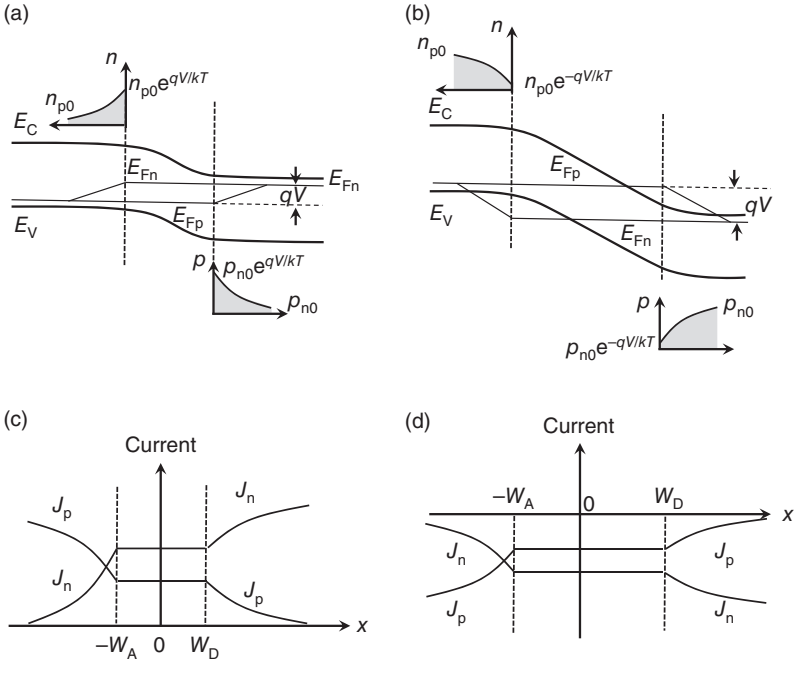

Figure 1.28 (a) Energy band diagram, quasi-Fermi level positions, and minority-carrier density at forward and reverse bias. (b) Energy band diagram, quasi-Fermi level positions, and minority-carrier density at reverse bias. Current contributions at (c) forward and (d) reverse biases.

where D_n is the hole diffusivity and τ_n is the hole recombination lifetime. Depending on the recombination mechanism—whether band-to-band or Shockley–Read–Hall recombination—the hole recombination lifetime is inversely proportional to the acceptor concentration or the trap density, respectively. This equation expresses the balance between diffusion current and recombination current. The boundary conditions are at the space-charge region edge ($x = 0$) $n_p = n_{p0}e^{qV/kT}$ and far from the space-charge region ($x = -\infty$) $n_p = n_{p0}$. The solution to the equation is

$$n_p = n_{p0}\left(e^{\frac{qV}{kT}} - 1\right)e^{\frac{x}{\sqrt{D_n\tau_n}}} + n_{p0} \tag{1.93}$$

where we define $L_n = \sqrt{D_n\tau_n}$ as the hole diffusion length in the n-type semiconductor. At $x = 0$, the hole diffusion current density is

$$J_n = qD_n\frac{\mathrm{d}n_p}{\mathrm{d}x} = \frac{qD_nn_{p0}}{L_n}\left(e^{\frac{qV}{kT}} - 1\right) \tag{1.94}$$

Similarly, for minority-carrier electrons in the n-type semiconductor region, the electron diffusion current density is

$$J_p = qD_p\frac{\mathrm{d}p_n}{\mathrm{d}x} = \frac{qD_pp_{n0}}{L_p}\left(e^{\frac{qV}{kT}} - 1\right) \tag{1.95}$$

From the aforementioned equation, we know that the diffusion current at forward bias is mainly caused by the minority-carrier injection (Figure 1.28c). The intensity of this diffusion current diminishes as carriers diffuse deeper into the semiconductor bulk. Assuming that carrier recombination and generation within the depletion region are negligible, the total current density simplifies to

$$J = J_n + J_p = J_0\left(e^{\frac{qV}{kT}} - 1\right) \tag{1.96}$$

where

$$J_0 = \frac{qD_nn_{p0}}{L_n} + \frac{qD_pp_{n0}}{L_p} \tag{1.97}$$

This expression represents the Shockley ideal diode equation. Although derived for forward bias, the same governing equation applies under reverse-bias conditions. In reverse bias, carrier generation within the neutral region follows an equation analogous to that describing recombination in forward bias, shown in Figure 1.28d. Given the acceptor N_A and donor N_D concentrations, J_0 can be reformulated as

$$J_0 = \frac{qD_n{n_i}^2}{L_nN_A} + \frac{qD_p{n_i}^2}{L_pN_D} \tag{1.98}$$

Temperature significantly influences PN junction diode current through several parameters. First, the terms $D_n/L_n = \sqrt{D_n/\tau_n}$ and $D_p/L_p = \sqrt{D_p/\tau_p}$ exhibit a temperature dependence proportional to $T^{\gamma/2}$, where γ is a material-specific exponent. Furthermore, the intrinsic carrier density n_i relates to temperature via $n_i \propto T^{3/2}\mathrm{e}^{-E_G/2kT}$. This exponential term overwhelmingly dominates the weaker power-law dependence. Consequently, for a constant forward bias voltage V, the current depends exponentially on temperature through $e^{\frac{qV-E_G}{kT}}$. Under reverse bias,

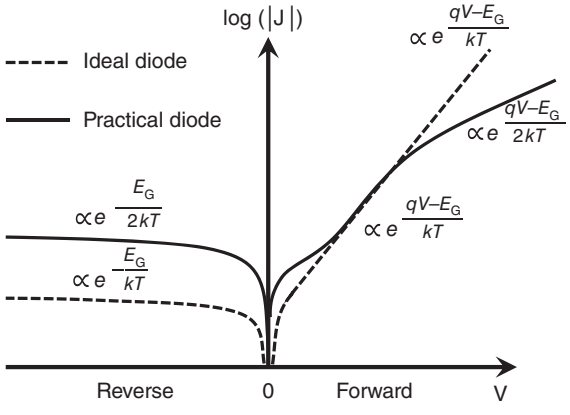

Figure 1.29 Current–voltage characteristics of the ideal diode and the practical diode.

however, the saturation current density scales with $e^{\frac{-E_G}{kT}}$. This analysis indicates that semiconductor PN junctions exhibiting smaller bandgaps support both larger forward and larger reverse saturation currents. For instance, Ge PN junctions with a bandgap of 0.66 eV exhibit higher currents than Si PN junctions, which possess a bandgap of 1.12 eV.

The Shockley equation can quite well predict the current–voltage of real PN junction qualitatively. However, because of the approximations that have been used during the derivation of Shockley equation, deviations from the ideal Shockley equation can be observed (Figure 1.29). The generation or recombination processes in the depletion region can lead to departures from the ideal Shockley equation even at small-injection conditions. At reverse bias, from the quasi-Fermi level positions, both the electron and hole densities in the depletion region is far smaller than the intrinsic carrier density, i.e., $p \ll n_i$ and $n \ll n_i$. The generation will dominate because $pn \ll n_i^2$. If the generation is mainly dominated by band-to-band transition, the generation rate will be proportional to $Rn_i{}^2 \propto e^{-\frac{E_G}{kT}}$. However, usually the trap-assisted generation is faster and more important in semiconductors, and the generation rate will be roughly proportional to $\sigma v_{th} N_t n_i/2 \propto e^{-\frac{E_G}{2kT}}$. Compared with the diffusion saturation current at reverse bias that is proportional to $e^{-\frac{E_G}{kT}}$, the trap-assisted generation can be more important in small-bandgap semiconductors, even at room temperature. For this reason, at room temperature, the reverse saturation current of Ge diode deviates more from the Shockley equation than Si diode. At forward bias, if carrier injection is comparable to the doping density or even larger, the trap-assisted recombination rate will be proportional to $\sigma v_{th} N_t n_i e^{\frac{qV}{2kT}}/2$, which depends differently on V than the $e^{\frac{qV}{kT}}$ term in the ideal Shockley equation. Considering this, an empirical form is usually used to describe the diode current at forward bias.

$$J_F \propto e^{\frac{qV}{\eta kT}} \tag{1.99}$$

where the ideality factor η equals 1 when diffusion currents dominate and 2 when recombination currents dominate. Note that under high-injection conditions, the

current also scales approximately with $e^{\frac{qV}{2kT}}$. However, such high currents typically induce significant resistive voltage drops due to series resistance, which must be considered.

1.7.3 Breakdown Mechanisms in PN Junctions

Under reverse bias, the PN junction current exhibits saturation. However, at sufficiently high bias voltages, a strong electric field develops within the depletion region. This field ultimately leads to PN junction breakdown, accompanied by a sudden current increase (Figure 1.30). Several mechanisms can cause breakdown. Thermally induced breakdown occurs when power dissipation within the depletion region elevates its temperature (Figure 1.30a). A substantial reverse bias generates significant power. Limited heat dissipation causes this power dissipation to increase the junction temperature. Higher temperature, in turn, increases the reverse saturation current, subsequently generating further heat. When the reverse-bias voltage is sufficiently large, this positive feedback loop becomes uncontrollable, resulting in breakdown. This thermally driven process is generally irreversible, typically destroying the device due to the resultant high current. Protective measures, such as incorporating current-limiting resistors, are therefore necessary. Thermally induced breakdown is more probable in small-bandgap semiconductors operating at relatively high ambient temperatures. At lower temperatures, alternative breakdown mechanisms may dominate.

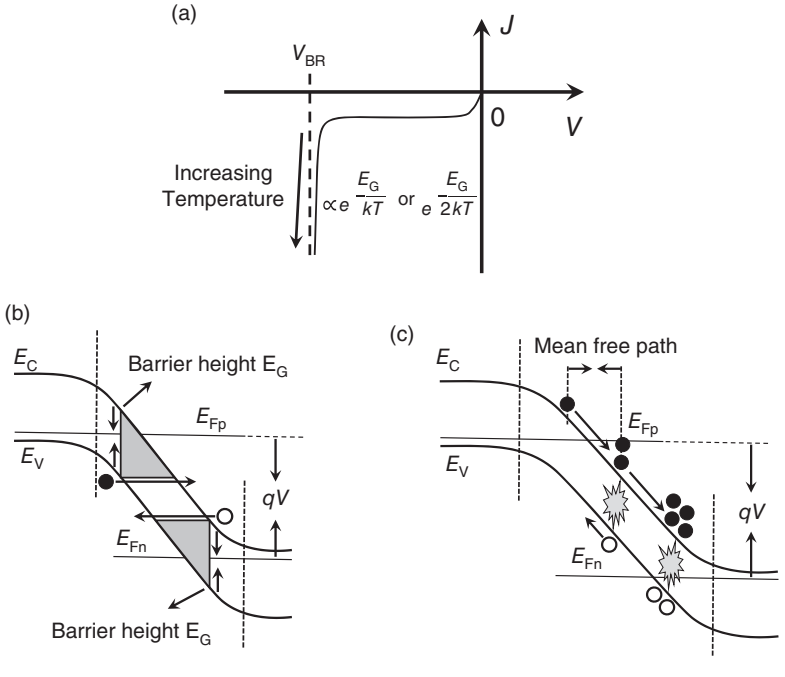

Figure 1.30 (a) Reverse current–voltage characteristics of thermal breakdown. (b) Energy band diagram of tunneling breakdown. (c) Energy band diagram of avalanche breakdown.

Besides thermal excitation, PN junction breakdown can also be induced by tunneling or avalanche effects (Figure 1.30b,c). This section presents both mechanisms and provides a systematic comparison. Tunneling results from a sufficiently strong electric field narrowing the tunneling barrier. As illustrated in Figure 1.30, electrons and holes are confined by the bandgap barrier within the conduction band and valence band, respectively. Under sufficiently large reverse-bias voltages, however, the barrier thickness decreases, as indicated by the triangular markers. Achieving tunneling necessitates an intense electric field, e.g., 10^6 V/cm for silicon. Such high field strengths require substantial doping concentrations to minimize the space-charge width. In contrast to band-to-band tunneling, breakdown can also result from impact ionization by high-energy carriers, the most prevalent mechanism observed experimentally in devices such as PN diodes, BJTs, and MOSFETs. When a free electron or hole gains sufficient kinetic energy ($>E_G$) during its free mean path before scattering within the depletion region, it can ionize the lattice and generate an electron–hole pair. These newly generated carriers may themselves be accelerated by the field to create further ionizations. Avalanche breakdown typically requires a critical field strength. Achieving this field alone is insufficient, as carrier acceleration and the resulting multiplicative impact ionization require a finite distance within the depletion region.

The tunneling breakdown and avalanche breakdown mechanisms differ in several aspects. Although both require a high electric field, tunneling breakdown typically occurs in heavily doped PN junctions and exhibits a relatively low breakdown voltage ($<4E_G/q$). In contrast, avalanche breakdown necessitates the maintenance of a high electric field over a distance and possesses a relatively high breakdown voltage ($>6E_G/q$). Another distinction lies in their temperature coefficients. Tunneling breakdown demonstrates a negative temperature coefficient for semiconductors such as Si and GaAs, meaning the breakdown voltage decreases with rising temperature. This phenomenon occurs because elevated temperatures typically reduce the bandgap, thereby lowering the barrier height and facilitating tunneling. Conversely, avalanche breakdown exhibits a positive temperature coefficient, resulting in an increased breakdown voltage at higher temperatures. This characteristic is attributed to increased scattering of energetic carriers, such as by phonons, at elevated temperatures. Such scattering diminishes the average kinetic energy of electrons and holes, consequently requiring a stronger accelerating electric field to reach the impact ionization energy threshold.

References

1 Wuttig, M., Schön, C.F., Schumacher, M. et al. (2022). *Adv. Funct. Mater.* 32: 2110166.

2 Cornil, J., Calbert, J.P., and Brédas, J.L. (2001). *J. Am. Chem. Soc.* 123: 1250–1251.

3 Radisavljevic, B., Radenovic, A., Brivio, J. et al. (2011). *Nat. Nanotechnol.* 6: 147–150.

4 Park, C.H., Cheong, B.H., Lee, K.H., and Chang, K.J. (1994). *Phys. Rev. B* 49: 4485–4493.
5 Ito, K., Hiramatsu, K., Amano, H., and Akasaki, I. (1990). *J. Cryst. Growth* 104: 533–538.
6 Stoumpos, C.C., Malliakas, C.D., Peters, J.A. et al. (2013). *Cryst. Growth Des.* 13: 2722–2727.