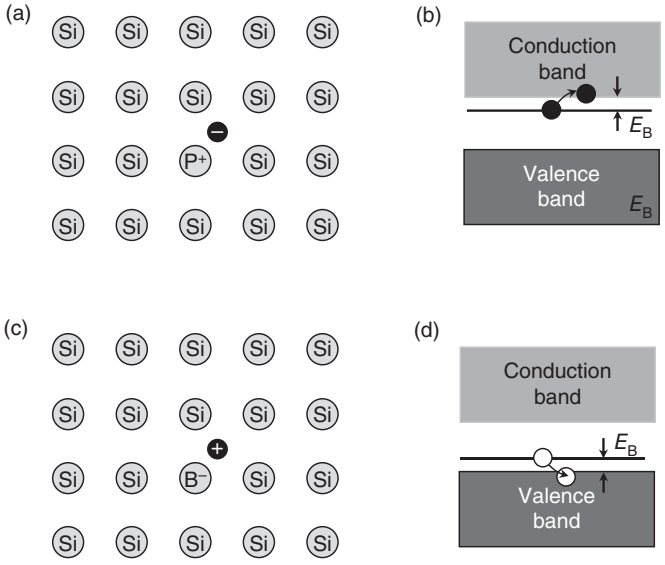

Figure 1.15 (a) Schematic representation of substitutional P doping of Si. (b) The donor level and free electron excitation in the conduction band of substitutional P defect. (c) Schematic representation of substitutional B doping of Si. (d) The acceptor level and free hole excitation in the valence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

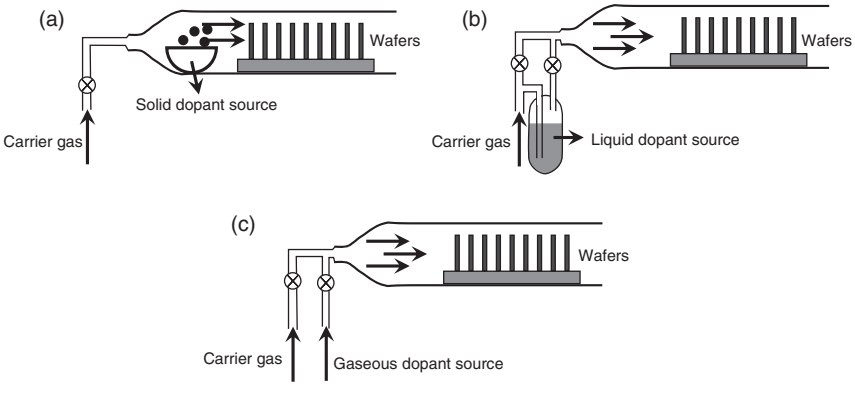

Figure 1.16 Schematic representations of diffusion systems with (a) solid, (b) liquid, and (c) gaseous dopants.

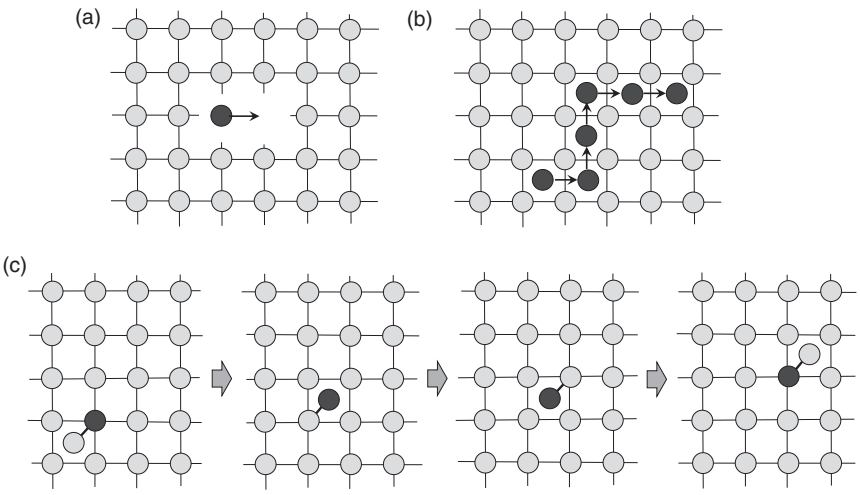

Figure 1.17 Diffusions of impurity atoms by (a) impurity jumping into a vacancy site, (b) the interstitial sites, and (c) impurity-defect pair (interstitialcy).

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

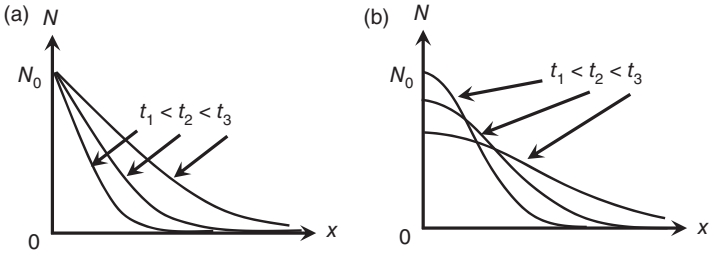

Figure 1.18 Impurity profiles with increasing time for (a) constant source diffusion and (b) constant dose diffusion.

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^+,

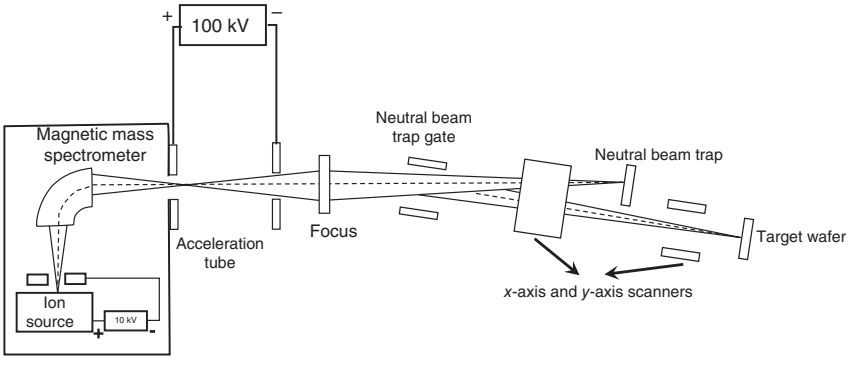

Figure 1.19 Schematic structure of an implantation system.

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

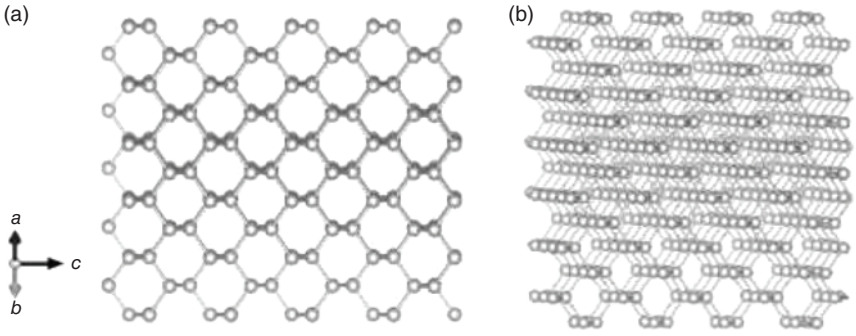

Figure 1.20 (a) Silicon crystal structure viewed from <110> direction with strong channeling effect during implantation. (b) Silicon structure tilted 10° from <110> direction.

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

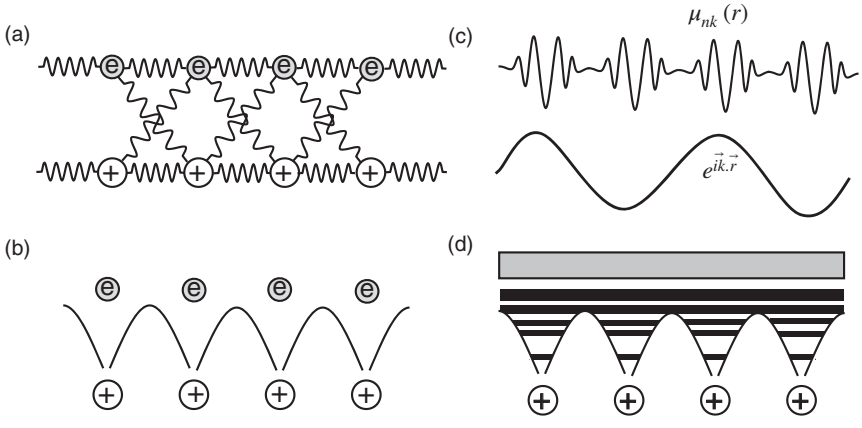

Figure 1.21 (a) Electron–ion, electron–electron, and ion–ion interactions in a crystal. (b) Mean-field potential experienced by electrons as independent particles with fixed ions in the Born–Oppenheimer approximation. (c) The periodic part and plane wave part of the Bloch wavefunction. (d) Filling of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

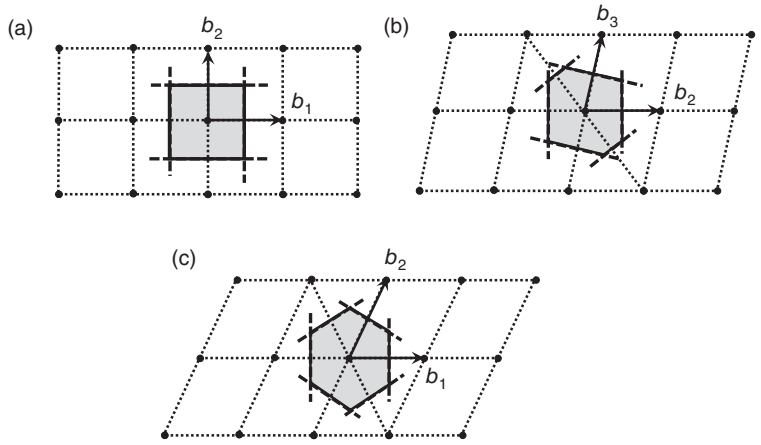

Figure 1.22 Construction of first Brillouin zones of (a) simple cubic lattice, (b) monoclinic lattice, and (c) hexagonal lattice by bisecting the reciprocal lattice vectors.

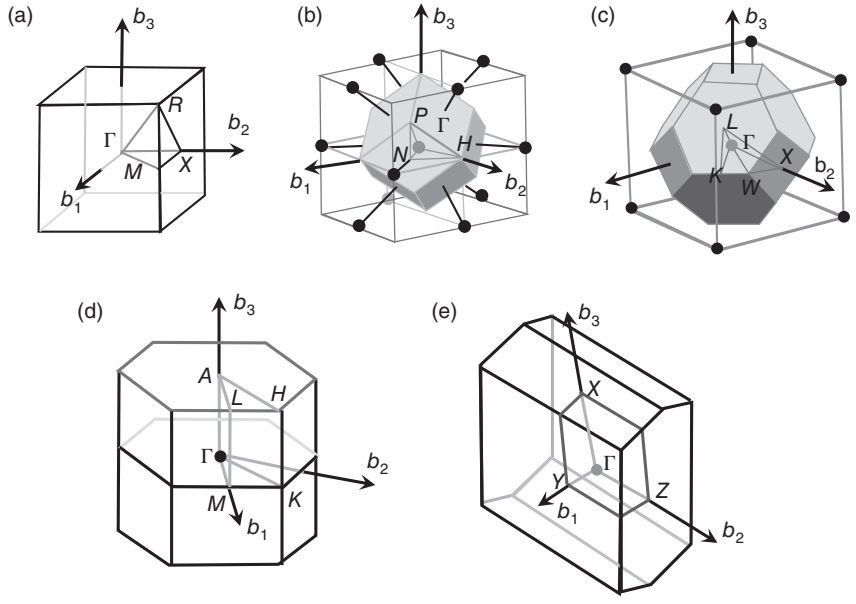

Figure 1.23 First Brillouin zones and high-symmetry point notations of (a) simple cubic lattice, (b) body-centered cubic lattice, (c) face-centered cubic lattice, (d) hexagonal lattice, and (e) monoclinic lattice.

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.

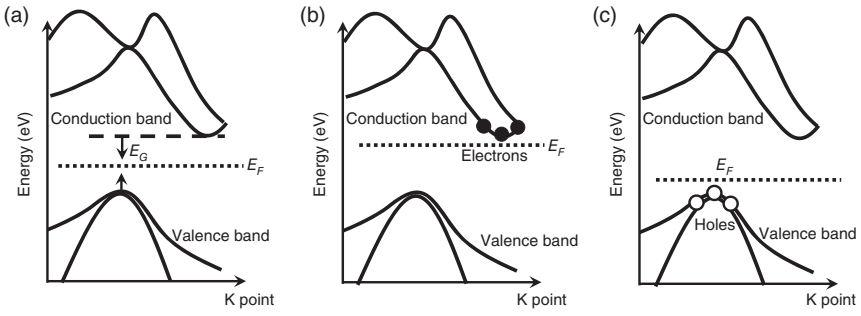

Figure 1.24 Schematic band diagrams and carrier types in (a) the intrinsic, (b) *n*-type, and (c) *n*-type semiconductors.

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

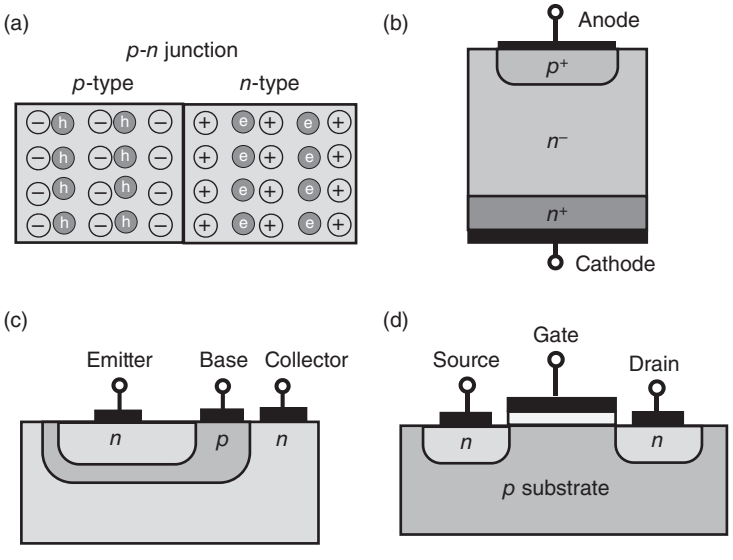

Figure 1.25 (a) Schematic of a PN junction formed by bringing n-type and n-type semiconductors into contact. Cross-sectional structures of (b) a $p^+ - n^- - n^+$ junction diode, (c) NPN bipolar junction transistor, and (d) an n-type metal–oxide–semiconductor field-effect transistor. The PN junctions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

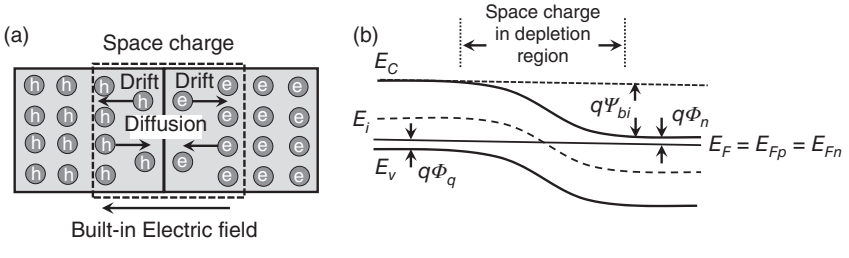

Figure 1.26 (a) The static equilibrium formed due to the balance of diffusion and drift currents in a PN junction. (b) Energy band diagram of a PN junction at equilibrium with aligned quasi-Fermi levels.

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:

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

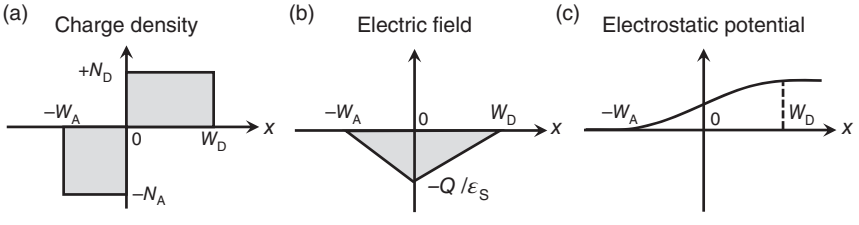

Figure 1.27 (a) Space-charge distribution in the depletion region of PN junction according to the box-profile approximation. (b) Electric field distribution generated by the space charge. (c) Electrostatic potential obtained by integration of the electric field.

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

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:

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

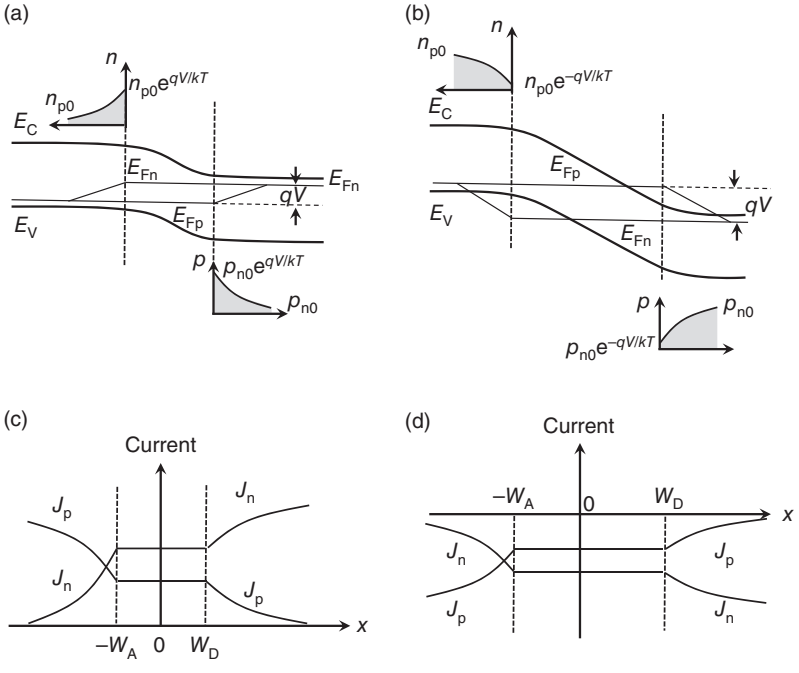

Figure 1.28 (a) Energy band diagram, quasi-Fermi level positions, and minority-carrier density at forward and reverse bias. (b) Energy band diagram, quasi-Fermi level positions, and minority-carrier density at reverse bias. Current contributions at (c) forward and (d) reverse biases.

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,

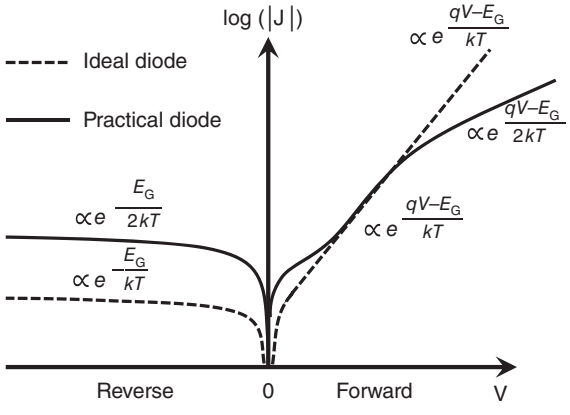

Figure 1.29 Current–voltage characteristics of the ideal diode and the practical diode.

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.

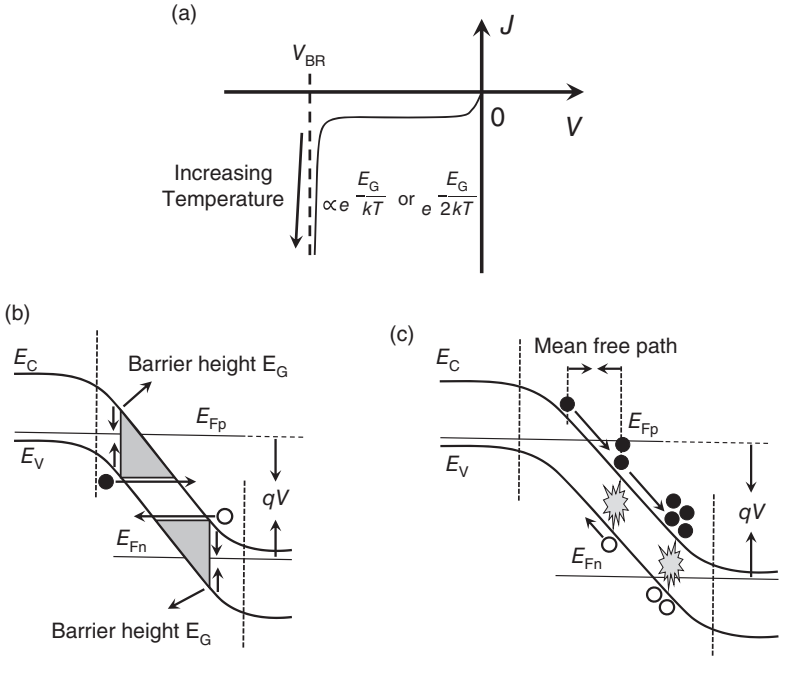

Figure 1.30 (a) Reverse current–voltage characteristics of thermal breakdown. (b) Energy band diagram of tunneling breakdown. (c) Energy band diagram of avalanche breakdown.

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.