CAMBRIDGE www.cambridge.org/9780521763448

This page intentionally left blank

Semiconductor Devices for High-Speed Optoelectronics

Providing an all-inclusive treatment of electronic and optoelectronic devices used in high-speed optical communication systems, this book emphasizes circuit applications, advanced device design solutions, and noise in sources and receivers. Core topics covered include semiconductors and semiconductor optical properties, high-speed circuits and transistors, detectors, sources, and modulators. It discusses in detail both active devices (heterostructure field-effect and bipolar transistors) and passive components (lumped and distributed) for high-speed electronic integrated circuits. It also describes recent advances in high-speed devices for 40 Gbps systems. Introductory elements are provided, making the book open to readers without a specific background in optoelectronics, whilst end-of-chapter review questions and numerical problems enable readers to test their understanding and experiment with realistic data.

Giovanni Ghione is Full Professor of Electronics at Politecnico di Torino, Torino, Italy. His current research activity involves the physics-based and circuit-oriented modeling of high-speed electronic and optoelectronic components, with particular attention to III-N power devices, thermal and noise simulation, electrooptic and electroabsorption modulators, coplanar passive components, and integrated circuits. He is a Fellow of the IEEE and has authored or co-authored over 200 technical papers and four books.

Semiconductor Devices for High-Speed Optoelectronics

GIOVANNI GHIONE
Politecnico di Torino, Italy

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521763448

First published in print format 2009

ISBN-13 978-0-511-63022-4 eBook (Adobe Reader)

ISBN-13 978-0-521-76344-8 Hardback

To the memory of my parents

Contents

Preface

The development of high-speed fiber-based optical communication systems that has taken place since the early 1970s can be really considered as a technological wonder. In a few years, key components were devised (such as the semiconductor laser) with the help of novel technological processes (such as epitaxial growth) and found immediate application thanks to the development of low-loss optical fibers. New compound semiconductor alloys (namely, InGaAsP) were ready to provide their potential to emit the right wavelengths needed for long-haul fiber propagation. When electronic repeaters seemed unable to provide a solution to long-haul propagation, fiber amplifiers were developed that allowed for all-optical signal regeneration. And the list could be continued. A miracle of ingenuity from a host of researchers made it possible to assemble this complex puzzle in a few years, thus bringing optoelectronic technology to a consumer electronics level.

Increasing the system capacity by increasing the transmission speed was, of course, a main concern from the early stages of optical system development. While optoelectronic devices behave, on the electronic side, in a rather conventional way up to speeds of the order of 1 Gbps, for larger speeds (up to 40 Gbps and beyond) RF wave propagation has to be accounted for in designing and modeling optoelectronic devices. When speed increases, the distributed interaction between RF and optical waves becomes a useful, sometimes indispensable, ingredient in many optoelectronic devices, like modulators and (to a lesser extent) detectors. Similarly, the electronic circuits that interface light sources, modulators, and detectors should provide broadband operation up to microwave or millimeter-wave frequencies, thus making it mandatory to exploit compound semiconductor electronics (GaAs- or InP-based) or advanced Si-based solutions (like SiGe HBT integrated circuits or nanometer MOS processes).

Increasing speed beyond the 10 Gbps limit by improving device performance, however interesting it is from the research and development side, may in practice be less appealing from the market standpoint. The ultimate destiny of optoelectronic devices (such as sources, modulators, and detectors) optimized for 40 Gbps (or even faster) systems after the post-2000 market downturn still is uncertain, and research in the field has followed alternative paths to the increase of system capacity. At the same time, new application fields have been developed, for instance in the area of integrated all-Si optical signal processing systems, and also for integrated circuit level high-capacity communications. However, the development of high-speed optoelectronic devices has raised a number of stimulating (and probably lasting) design issues. An example is the

principle of the distributed interaction between optical and RF waves, which is common to a variety of high-speed components. Another relevant theme is the co-design and the (possibly monolithic) integration of the electronic and optoelectronic components of a system, not to mention the critical aspects concerning device packaging and interconnection in systems operating at 40 Gbps and beyond.

Taking the above into account, it is not surprising that the main purpose of the present book is to provide a kind of unified (or, perhaps, not too widely separated) treatment of high-speed electronics and optoelectronics, starting from compound semiconductor basics, down to high-speed transistors, ICs, detectors, sources and modulators. Part of the material was originally developed for a number of postgraduate and Master courses, and therefore has the ambition (but also the limitation) of providing a treatment starting from the very basics. It is hoped that this justifies both the presence of introductory material on semiconductors and semiconductor optical properties, and a treatment of high-speed electronics starting from a review of transmission lines and scattering parameters. From this standpoint, the text attempts to be as self-contained as possible. Of course, the choice of subjects is somewhat influenced by the author's personal tastes and previous research experience (not to mention the need to keep the page count below 500): some emphasis has been put on noise, again with an attempt to present a self-contained treatment of this rather difficult topic, and many important optoelectronic components have not been included (to mention one, semiconductor optical amplifiers). Yet another innovative subject that is missing is microwave photonics, where of course the RF and microwave and optoelectronic worlds meet. Nevertheless, the text is (in the author's opinion, at least) different enough from the many excellent textbooks on optoelectronics available on the market to justify the attempt to write it.

I wish to thank a number of colleagues (from Politecnico di Torino, unless otherwise stated) for their direct or indirect contribution to this book. Ivo Montrosset provided many useful suggestions on the treatment of optical sources. Incidentally, it was under the guidance of Ivo Montrosset and Carlo Naldi that (then an undergraduate student) I was introduced to the basics of passive and active optoelectronic devices, respectively; this happened, alas, almost 30 years ago. Helpful discussions with Gian Paolo Bava and Pierluigi Debernardi (Consiglio Nazionale delle Ricerche) on laser noise, with Simona Donati Guerrieri on the semiconductor optical properties and with Fabrizio Bonani and Marco Pirola on active and passive high-speed semiconductor electronic devices and circuits are gratefully acknowledged. Michele Goano kindly revised the sections on compound semiconductors and the numerical problems, and provided useful suggestions on III-N semiconductors. Federica Cappelluti prepared many figures (in particular in the section on photodetectors), initially exploited in lecture slides. Finally, Claudio Coriasso (Avago Turin Technology Center, Torino) kindly provided material on integrated electroabsorption modulators (EAL), including some figures. Additionally, I am indebted to a number of ME students who cooperated in research, mainly on lithium niobate modulators; among those, special mention goes to F. Carbonera, D. Frassati, G. Giarola, A. Mela, G. Omegna, L. Terlevich, P. Zandano. A number of PhD students also worked on subjects relevant to the present book: Francesco Bertazzi (now with Politecnico di Torino) on EM modeling of distributed electrooptic structures; Pietro Bianco,

on high-speed modulator drivers; Federica Cappelluti, on electroabsorption modulator modeling; Gloria Carvalho, on EAL modeling; Antonello Nespola (now with Istituto Superiore Mario Boella), on the modeling of distributed high-speed photodetectors. Part of the thesis work of Antonello Nespola and Federica Cappelluti was carried out within the framework of a cooperation with UCLA (Professor Ming Wu, now at University of California, Berkeley). Finally, I gratefully recall many helpful discussions with colleagues from the industry: among those, Marina Meliga, Roberto Paoletti, Marco Romagnoli, and Luciano Socci.

Giovanni Ghione
January 2009

1 Semiconductors, alloys, heterostructures

1.1 Introducing semiconductors

Single-crystal semiconductors have a particularly important place in optoelectronics, since they are the starting material for high-quality sources, receivers and amplifiers. Other materials, however, can be relevant to some device classes: polycrystalline or amorphous semiconductors can be exploited in light-emitting diodes (LEDs) and solar cells; dielectrics (also amorphous) are the basis for passive devices (e.g., waveguides and optical fibers); and piezoelectric (ferroelectric) crystals such as lithium niobate are the enabling material for a class of electrooptic (EO) modulators. Moreover, polymers have been recently exploited in the development of active and passive optoelectronic devices, such as emitters, detectors, and waveguides (e.g., fibers). Nevertheless, the peculiar role of single-crystal semiconductors justifies the greater attention paid here to this material class with respect to other optoelectronic materials.

From the standpoint of electron properties, semiconductors are an intermediate step between insulators and conductors. The electronic structure of crystals generally includes a set of allowed energy bands, that electrons populate according to the rules of quantum mechanics. The two topmost energy bands are the *valence* and *conduction* band, respectively, see Fig. 1.1. At some energy above the conduction band, we find the *vacuum level*, i.e., the energy of an electron free to leave the crystal. In *insulators*, the valence band (which hosts the electrons participating to the chemical bonds) is separated from the conduction band by a large energy gap E_g, of the order of a few electronvolts (eV). Due to the large gap, an extremely small number of electrons have enough energy to be promoted to the conduction band, where they could take part into electrical conduction. In insulators, therefore, the conductivity is extremely small. In *metals*, on the other hand, the valence and conduction bands overlap (or the energy gap is *negative*), so that all carriers already belong to the conduction band, independent of their energy. Metals therefore have a large conductivity. In *semiconductors*, the energy gap is of the order of 1–2 eV, so that some electrons have enough energy to reach the conduction band, leaving *holes* in the valence band. Holes are pseudo-particles with positive charge, reacting to an external applied electric field and contributing, together with the electrons in the conduction band, to current conduction. In pure (*intrinsic*) semiconductors, therefore, charge transport is *bipolar* (through electrons and holes), and the conductivity is low, exponentially dependent on the gap (the larger the gap, the lower the conductivity). However, impurities can be added (*dopants*) to provide large numbers of electrons to

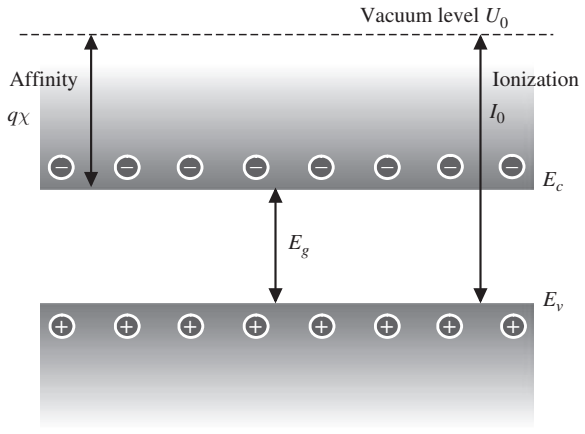

Figure 1.1 Main features of semiconductor bandstructure. E_g is the energy gap; E_c is the conduction band edge; E_v is the valence band edge.

the conduction band (*donors*) or of holes to the valence band (*acceptors*). The resulting doped semiconductors are denoted as n-type and p-type, respectively; their conductivity can be artificially modulated by changing the amount of dopants; moreover, the dual doping option allows for the development of pn junctions, one of the basic building blocks of electronic and optoelectronic devices.

1.2 Semiconductor crystal structure

Crystals are regular, periodic arrangements of atoms in three dimensions. The point set $\underline{r}$ defining the crystal nodes, corresponding to the atomic positions (Bravais lattice) satisfies the condition $\underline{r} = k\underline{a}_1 + l\underline{a}_2 + m\underline{a}_3$, where k, l, m are integer numbers and $\underline{a}_1$, $\underline{a}_2$, $\underline{a}_3$ are the *primitive vectors* denoting the *primitive cell*, see Fig. 1.2. Bravais lattices can be formed so as to fill the entire space only if the angles α_1, α_2, α_3 assume values from a discrete set (60°, 90°, 120°, or the complementary value to 360°). According to the relative magnitudes of a_1, a_2, a_3 and to the angles $\alpha_1, \alpha_2, \alpha_3$, 14 basic lattices can be shown to exist, as in Table 1.1. In semiconductors, only two lattices are technologically important at present, i.e. the *cubic* and the *hexagonal*. Most semiconductors are cubic (examples are Si, Ge, GaAs, InP…), but some are hexagonal (SiC, GaN). Both the cubic and the hexagonal structure can be found in carbon (C), where they are the diamond and graphite crystal structures, respectively.

Three kinds of Bravais cubic lattices exist, the simple cubic (sc), the face-centered cubic (fcc) and the body-centered cubic (bcc), see Fig. 1.3. The cubic semiconductor crystal structure can be interpreted as two *shifted* and *compenetrated* fcc Bravais lattices.

Let us consider first an elementary semiconductor (e.g., Si) where all atoms are equal. The relevant cubic lattice is the *diamond lattice*, consisting of two interpenetrating

Table 1.1 The 14 Bravais lattices.

Name	Bravais lattices	Conditions on primitive vectors
Triclinic	1	$a_1 \neq a_2 \neq a_3, \alpha_1 \neq \alpha_2 \neq \alpha_3$
Monoclinic	2	$a_1 \neq a_2 \neq a_3, \alpha_1 = \alpha_2 = 90^\circ \neq \alpha_3$
Orthorhombic	4	$a_1 \neq a_2 \neq a_3, \alpha_1 = \alpha_2 = \alpha_3 = 90^\circ$
Tetragonal	2	$a_1 = a_2 \neq a_3, \alpha_1 = \alpha_2 = \alpha_3 = 90^\circ$
Cubic	3	$a_1 = a_2 = a_3, \alpha_1 = \alpha_2 = \alpha_3 = 90^\circ$
Trigonal	1	$a_1 = a_2 = a_3, \alpha_1 = \alpha_2 = \alpha_3 < 120^\circ \neq 90^\circ$
Hexagonal	1	$a_1 = a_2 \neq a_3, \alpha_1 = \alpha_2 = 90^\circ, \alpha_3 = 120^\circ$

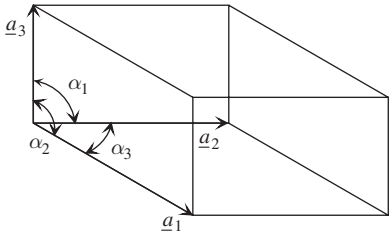

Figure 1.2 Semiconductor crystal structure: definition of the primitive cell.

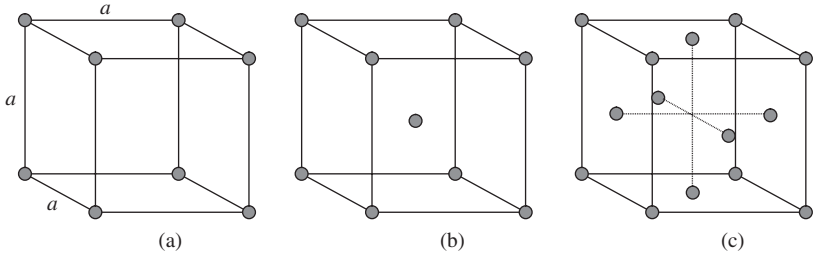

Figure 1.3 Cubic Bravais lattices: (a) simple, (b) body-centered, (c) face-centered.

fcc Bravais lattices, displaced along the body diagonal of the cubic cell by one-quarter the length of the diagonal, see Fig. 1.4. Since the length of the diagonal is $d = a\left|\hat{x} + \hat{y} + \hat{z}\right| = a\sqrt{3}$, the displacement of the second lattice is described by the vector

$$\underline{s} = \frac{a\sqrt{3}}{4}\frac{\hat{x} + \hat{y} + \hat{z}}{\sqrt{3}} = \frac{a}{4}\left(\hat{x} + \hat{y} + \hat{z}\right).$$

1.2.1 The Miller index notation

The Miller indices are a useful notation to denote planes and reference directions within a lattice. The notation (h, k, l), where h, k, l are integers, denotes the set of parallel planes that intercepts the three points $\underline{a}_1/h$, $\underline{a}_2/k$ and $\underline{a}_3/l$, or some multiple thereof, while $[h, k, l]$ in square brackets is the direction orthogonal to plane (h, k, l).

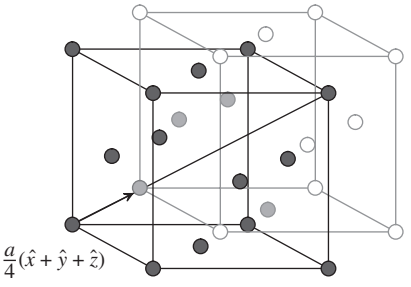

Figure 1.4 The diamond lattice as two cubic face-centered interpenetrating lattices. The pale and dark gray points represent the atoms falling in the basic cell.

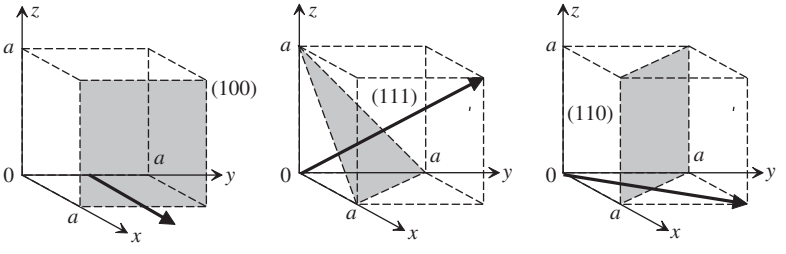

Figure 1.5 Examples of planes and directions according to the Miller notation.

Additionally, $\{h, k, l\}$ is a family of planes with symmetries and $\langle h, k, l\rangle$ is the related direction set. In cubic lattices, the primitive vectors coincide with the Cartesian axes and $a_1 = a_2 = a_3 = a$, where a is the lattice constant; in this case, we simply have $[h, k, l] \equiv h\hat{x} + k\hat{y} + l\hat{z}$ where $\hat{x}$, $\hat{y}$ and $\hat{z}$ are the Cartesian unit vectors.

To derive the Miller indices from the plane intercepts in a cubic lattice, we normalize with respect to the lattice constant (thus obtaining a set of integers (H, K, L)), take the reciprocal (H^{-1}, K^{-1}, L^{-1}) and finally multiply by a minimum common multiplier so as to obtain a set (h, k, l) such as $h : k : l = H^{-1} : K^{-1} : L^{-1}$. Notice that a zero index corresponds to an intercept point at infinity. Examples of important planes and directions are shown in Fig. 1.5.

Example 1.1: Identify the Miller indices of the following planes, intersecting the coordinate axes in points (normalized to the lattice constant): (a) $x = 4$, $y = 2$, $z = 1$; (b) $x = 10$, $y = 5$, $z = \infty$; (c) $x = 3.5$, $y = \infty$, $z = \infty$; (d) $x = -4$, $y = -2$, $z = 1$.

We take the reciprocal of the intercept, and then we multiply by the minimum common multiplier, so as to obtain an integer set with minimum module. In case (a), the reciprocal set is $(1/4, 1/2, 1)$, with minimum common multiplier 4, leading to the Miller indices $(1, 2, 4)$. In case (b), the reciprocals are $(1/10, 1/5, 0)$ with Miller indices $(1, 2, 0)$. In case (c), the plane is orthogonal to the z axis, and the Miller indices simply are $(1, 0, 0)$. Finally, case (d) is similar to case (a) but with negative intercepts; according to the Miller notation we overline the indices rather than using a minus sign; we thus have $(\overline{1}, \overline{2}, 4)$.

1.2.2 The diamond, zinc-blende, and wurtzite semiconductor cells

The cubic diamond cell includes 8 atoms; in fact, if we consider Fig. 1.6, the corner atoms each contribute to eight adjacent cells, so that only $8/8 = 1$ atom belongs to the main cell. The atoms lying on the faces belong half to the main cell, half to the nearby ones, so that only $6/2 = 3$ atoms belong to the main cell. Finally, the other (internal) 4 atoms belong entirely to the cell. Therefore, the total number of atoms in a cell is $1 + 3 + 4 = 8$. In the diamond cell, each atom is connected to the neighbours through a tetrahedral bond. All atoms are the same (C, Si, Ge...) in the diamond lattice, while in the so-called *zinc-blende lattice* the atoms in the two fcc constituent lattices are different (GaAs, InP, SiC...). In particular, the corner and face atoms are metals (e.g., Ga) and the internal atoms are nonmetals (e.g., As), or vice versa.

In the diamond or zinc-blende lattices the Miller indices are conventionally defined with respect to the cubic cell of side a. Due to the symmetry of the tetrahedral atom bonds, planes (100) and (110), etc. have two bonds per side, while planes (111) have three bonds on the one side, two on the other. Moreover, the surface atom density is different, leading, for example, to different etch velocities.

Some semiconductors, such as SiC and GaN, have the hexagonal *wurtzite* crystal structure. Hexagonal lattices admit many *polytypes* according to the stacking of successive atom layers; a large number of polytypes exists, but only a few have interesting semiconductor properties (e.g. 4H and 6H for SiC). The wurtzite cell is shown in Fig. 1.7, including 12 equivalent atoms. In the ideal lattice, one has

$$|\underline{a}_3| = c, \quad |\underline{a}_1| = |\underline{a}_2| = a, \quad \frac{c}{a} = \sqrt{\frac{8}{3}} \approx 1.633.$$

Some properties of semiconductor lattices are shown in Table 1.2.[1] It can be noted that wurtzite-based semiconductors are often anisotropic (uniaxial) and have two dielectric constants, one parallel to the c-axis, the other orthogonal to it.

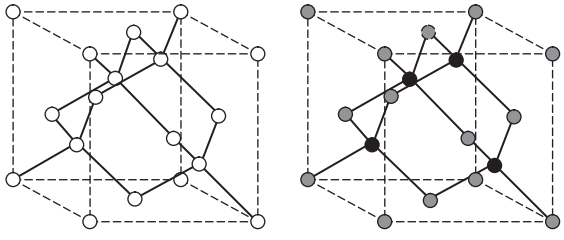

Figure 1.6 The diamond (left) and zinc-blende (right) lattices.

[1] Semiconductor properties are well documented in many textbooks; an excellent online resource is provided by the Ioffe Institute of the Russian Academy of Sciences at the web site [1].

Table 1.2 Properties of some semiconductor lattices: the crystal is D (diamond), ZB (zinc-blende) or W (wurtzite); the gap is D (direct) or I (indirect); $\epsilon_{\|}$ is along the c axis, $\epsilon_{\perp}$ is orthogonal to the c axis for wurtzite materials. Permittivities are static to RF. Properties are at 300 K.

Material	Crystal	E_g (eV)	D/I gap	ϵ_r or $\epsilon_{\|}$	$\epsilon_{\perp}$	a (Å)	c (Å)	Density, ρ (g/cm^3)
C	D	5.50	I	5.57		3.57		3.51
Si	D	1.12	I	11.9		5.43		2.33
SiC	ZB	2.42	I	9.72		4.36		3.17
Ge	D	0.66	I	16.2		5.66		5.32
GaAs	ZB	1.42	D	13.2		5.68		5.32
GaP	ZB	2.27	I	11.11		5.45		4.14
GaSb	ZB	0.75	D	15.7		6.09		5.61
InP	ZB	1.34	D	12.56		5.87		4.81
InAs	ZB	0.36	D	15.15		6.06		5.67
InSb	ZB	0.23	D	16.8		6.48		5.77
AlP	ZB	2.45	I	9.8		5.46		2.40
AlAs	ZB	2.17	I	10.06		5.66		3.76
AlSb	ZB	1.62	I	12.04		6.13		4.26
CdTe	ZB	1.47	D	10.2		6.48		5.87
GaN	W	3.44	D	10.4	9.5	3.17	5.16	6.09
AlN	W	6.20	D	9.14		3.11	4.98	3.25
InN	W	1.89	D	14.4	13.1	3.54	5.70	6.81
ZnO	W	3.44	D	8.75	7.8	3.25	5.21	5.67

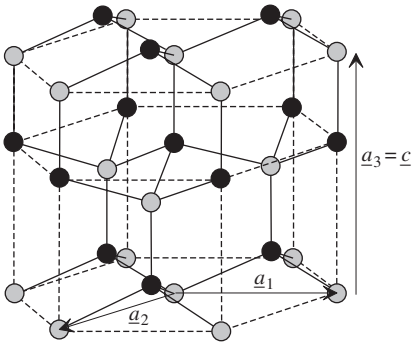

Figure 1.7 The hexagonal wurtzite cell. The c-axis corresponds to the direction of the $\underline{a}_3 = \underline{c}$ vector.

1.2.3 Ferroelectric crystals

Ferroelectric materials have a residual spontaneous dielectric polarization after the applied electric field has been switched off. The behavior of such materials is somewhat similar to that of ferromagnetic materials. Below a transition temperature, called the Curie temperature T_c, ferroelectric materials possess a spontaneous polarization or electric dipole moment. The magnitude of the spontaneous polarization is greatest at temperatures well below the Curie temperature, and approaches zero as the Curie

Table 1.3 Properties of some ferroelectric crystals. KDP stands for potassium dihydrogen phosphate. Data from [2], Ch. 13, Table 2.

Material class	Material	Curie temperature T_C (K)	Spontaneous polarization P_s ($\mu C/cm^2$)
KDP	KH_2PO_4	123	4.75
Perovskites	$BaTiO_3$	408	26
Perovskites	$LiNbO_3$	1480	71
Perovskites	$KNbO_3$	708	30

temperature is approached. Ferroelectric materials are inherently piezoelectric; that is, in response to an applied mechanical load, the material will produce an electric charge proportional to the load. Similarly, the material will produce a mechanical deformation in response to an applied voltage. In optoelectronics, ferroelectric materials are particularly important because of the excellent *electrooptic properties*, i.e., the strong variation of the material refractive index with an applied electric field. The crystal structure is often cubic face-centered, and the material is anisotropic and uniaxial. The most important ferroelectric crystal for optical applications is probably lithium niobate, $LiNbO_3$ (LN for short); some other materials (such as barium titanate) belonging to the so-called *perovskite* class are also sometimes used. The crystal structure of perovskites is face-centered cubic. Above the Curie temperature, the crystal is strictly cubic, and positive and negative ions are located in the cell so as to lead to zero dipole moment. Below the Curie temperature, however, a transition takes place whereby positive and negative ions undergo a shift in opposite directions; the crystal structure becomes tetragonal (i.e., the elementary cell height a_3 is different from the basis $a_1 = a_2$) and, due to the charge displacement, a net dipole moment arises. Table 1.3 shows a few properties of ferroelectric crystals, namely the spontaneous polarization P_s and the Curie temperature [2].

1.2.4 Crystal defects

In practice, the crystal lattice is affected by defects, either native (i.e., not involving external atoms) or related to nonnative impurities. Moreover, defects can be point defects (0D), line defects (1D), surface defects (2D), such as dislocations, and volume defects (3D), such as precipitates. Native point defects are *vacancies*, see Fig. 1.8, and *self-interstitials*, while *interstitials* are nonnative atoms placed in the empty space between the already existing lattice atoms. *Substitutional* defects involve an external atom, e.g., a dopant, which replaces one native atom. Typically, dopants act as donors or acceptors only if they are in a substitutional site; if they are in an interstitial site, they are inactive (chemically inactivated).[2]

[2] Dopants can also be electrically inactivated when they are not ionized.

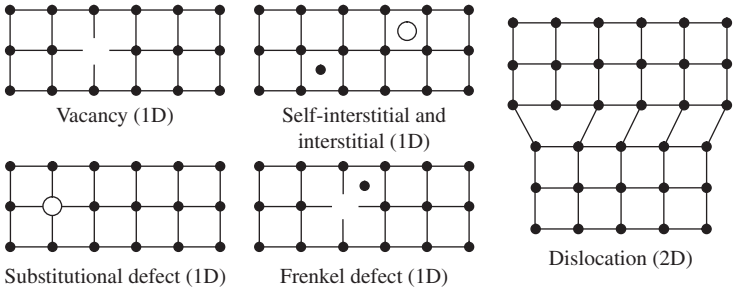

Figure 1.8 Point defects in a crystal (1D) and dislocations (2D).

1.3 Semiconductor electronic properties

1.3.1 The energy–momentum dispersion relation

A crystal is a periodic arrangement of atoms; since each positively charged nucleus induces a spherically symmetric Coulomb potential, superposition yields in total a periodic potential $U(\underline{r})$ such as

$$U(\underline{r}) = U(\underline{r} + \underline{L}),$$

where $\underline{L} = k\underline{a}_1 + l\underline{a}_2 + m\underline{a}_3$. In such a periodic potential, electrons follow the rules of quantum mechanics, i.e., they are described by a set of *wavefunctions* associated with allowed electron states. Allowed states correspond to allowed energy bands, which collapse into energy levels for isolated atoms; allowed bands are separated by forbidden bands. Low-energy electrons are bound to atoms, and only the two topmost allowed bands (the last, being almost full, is the *valence band*; the uppermost, almost empty, is the *conduction band*) take part in carrier transport. As already recalled, the vacuum level U_0 is the minimum energy of an electron free to move in and out of the crystal.

Electrons in a crystal are characterized by an energy–momentum relation $E(\underline{k})$, where the wavevector $\underline{k}$ is related to the electron momentum $\underline{p}$ as $\underline{p} = \hbar\underline{k}$. The *dispersion relation* $E(\underline{k})$ is defined in the $\underline{k}$ space, also called the reciprocal space; it is generally a multivalued function, periodic in the reciprocal space, whose fundamental period is called the *first Brillouin zone* (FBZ). A number of branches of the dispersion relation refer to the valence band, a number to the conduction band; the total number of branches depends on the crystal structure and is quite large (e.g., 12 for the conduction band and 8 for the valence band) in wurtzite semiconductors.

In cubic semiconductors, the FBZ is a solid with six square faces and eight hexagonal faces, as shown in Fig. 1.9. Owing to symmetries, only a portion of the FBZ, called the *irreducible wedge*, actually includes independent information; all the rest can be recovered by symmetry. Important points in the FBZ are the center (Γ point), the X point (center of the square face), and the L point (center of the hexagonal face).

The full details of the dispersion relation are not essential for understanding low-energy phenomena in semiconductors; attention can be restricted to the branches

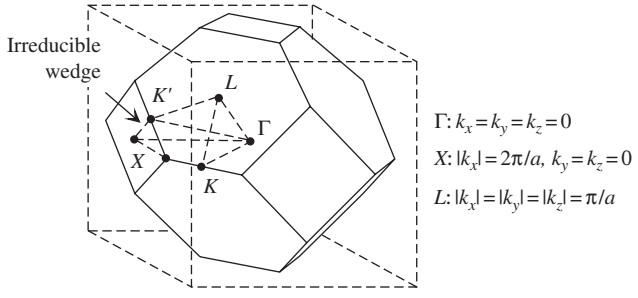

Figure 1.9 The first Brillouin zone (FBZ) in a cubic lattice (lattice constant a).

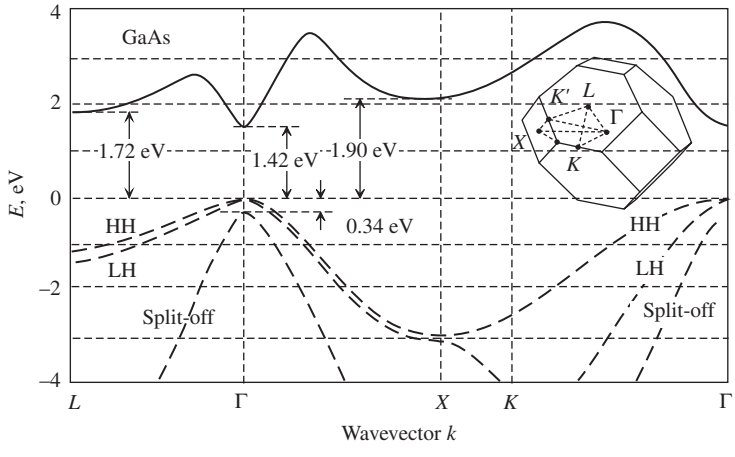

Figure 1.10 Simplified dispersion relation for GaAs.

describing low-energy electrons in the conduction band (around the *conduction band edge* E_c) and high-energy electrons (low-energy holes) in the valence band (around the *valence band edge* E_v). Valence band electrons are more efficiently described in terms of pseudoparticles (the *holes*) related to electrons missing from the valence band. Holes behave as particles with positive charge and potential energy opposite to the electron energy, so that the topmost branches of the dispersion relation (i.e., the branches describing low-energy holes) define the valence band edge.

As a relevant example, let us discuss the dispersion relation for a direct-bandgap semiconductor, GaAs. The term *direct bandgap* refers to the fact that the minimum of the conduction band and the maximum of the valence band (both located in the Γ point) correspond to the same momentum $\hbar\underline{k}$, in this case $\hbar\underline{k} = 0$. The dispersion relation shown in Fig. 1.10 is simplified, in the sense that only the lowest branch of the conduction band is shown, while three branches of the valence band appear, the *heavy hole* (HH), the *light hole* (LH), and the *split-off* band. Light and heavy hole bands are degenerate, i.e., they share the same minimum in the Γ point, and they differ because of the $E(\underline{k})$ curvature near the minimum, which corresponds to a larger or smaller hole effective mass. The split-off band enters some transport and optical processes but can be neglected in a first-order treatment. The conduction band has the lowest minimum at

the Γ point, and two secondary minima at the L and X points. The main gap is 1.42 eV, while the secondary gaps are 1.72 eV (L point) and 1.90 eV (X point). Only a section of the dispersion relation is presented, running from the L point to the Γ point (the center of the FBZ), and then from the Γ point to the X point and back to the origin through the K point.

Since electrons and holes have, at least in the absence of an applied field, a Boltzmann energy distribution (i.e., their probability to have energy E is proportional to $\exp(-E/k_BT)$, where $k_BT = 26$ meV at ambient temperature), most electrons and holes can be found close to the conduction band and valence band edges, respectively.

Consider now the lowest minimum of the conduction band or highest maximum in the valence band; the dispersion relation can be approximated (around the Γ point) by a parabola as

$$E_n - E_c \approx \frac{\hbar^2 k^2}{2m_n^*}, \quad E_v - E_h \approx \frac{\hbar^2 k^2}{2m_h^*},$$

where m_n^* and m_h^* are the electron and hole *effective masses*.[3] Therefore, the electron kinetic energy $E_n - E_c$ or hole kinetic energy $E_v - E_h$ (assuming the valence band edge energy E_v and the conduction band edge energy E_c to be the energy of a hole or of an electron, respectively, at rest) have, approximately, the same expression as the free-space particle kinetic energy, but with an effective mass m_n^* or m_h^* instead of the *in vacuo* inertial mass m_0. If the minimum is not located in the center of the first BZ (as for the conduction band of indirect bandgap semiconductors) the momentum (in a dynamic sense) can be defined "with respect to the minimum," so that the following approximation applies:

$$E_n - E_c \approx \frac{\hbar^2 \left|\underline{k} - \underline{k}_{\min}\right|^2}{2m_n^*}.$$

The effective mass can be evaluated from the inverse of the curvature of the dispersion relation around a minimum or a maximum. In general, the approximating surface can be expressed as

$$E_n - E_c = \frac{\hbar^2 k_a^2}{2m_{na}^*} + \frac{\hbar^2 k_b^2}{2m_{nb}^*} + \frac{\hbar^2 k_c^2}{2m_{nc}^*},$$

which is an ellipsoid; the coordinate system coincides with the principal axes. If the three effective masses are equal, the ellipsoid degenerates into a spherical surface, and we say that the minimum is *spherical*, with *isotropic* effective mass. This typically happens at Γ point minima. In indirect-bandgap semiconductors, the constant-energy

[3] Corrections to the parabolic approximation accounting for nonparaboliticity effects can be introduced (e.g., in the conduction band) through the expression:

$$E_k\,(1 + \alpha E_k) = \frac{\hbar^2 k^2}{2m_n^*},$$

where E_k is the electron kinetic energy $E_n - E_c$ and α is a nonparabolicity correction factor.

surfaces are rotation ellipsoids, and we can define two effective masses, one transversal m^*_{nt} (common to two principal directions) and one longitudinal m^*_{nl} (along the third principal direction). The electron effective mass increases with E_g, according to the fitting law (see (2.9)):

$$\frac{m^*_n}{m_0} \approx \frac{E_g\big|_{\mathrm{eV}}}{13}.$$

Due to degeneracy, the valence bands have a more complex behavior near the valence band edge, but can anyway be approximated with isotropic masses; however, since the heavy and light hole populations mix, a properly averaged effective mass has to be introduced; the same remark applies for electrons with anisotropic effective mass. The averaging law is related to the application, and is not unique; we can therefore have an effective mass for transport and also (as discussed later) an effective mass for the density of states that follow different averaging criteria. Concerning the density of states mass (denoted with the subscript D), we have for the electrons

$$m^*_{n,D} \triangleq \left(m^*_{na} m^*_{nb} m^*_{nc}\right)^{1/3} M_c^{2/3}.$$

The above expression refers to the general case of ellipsoidal minima with multiplicity M_c (more than one minimum in the FBZ); for a Γ point spherical minimum in the conduction band we have simply

$$m^*_{n,D} = m^*_n,$$

while for the rotation ellipsoid case in Si (where 6 equivalent minima are present in the FBZ) we obtain

$$m^*_{n,D} \triangleq 6^{2/3} (m^*_{nl})^{1/3} (m^*_{nt})^{2/3}.$$

For holes, in the case of degeneracy:

$$m^*_{h,D} \triangleq \left[(m^*_{hh})^{3/2} + (m^*_{lh})^{3/2}\right]^{2/3},$$

while of course $m^*_{h,D}$ reduces to m^*_{hh} or m^*_{lh} if degeneracy is removed (as in a strained quantum well, see Section 1.7). Concerning the effective masses for transport, since on average the electron moves along all three principal directions with the same probability, we have that the transport or conductivity average electron mass is given by

$$\frac{1}{m^*_{n,tr}} = \frac{1}{3m^*_{na}} + \frac{1}{3m_{nb}} + \frac{1}{3m_{nc}},$$

which reduces, for Si, to

$$\frac{1}{m^*_{n,tr}} = \frac{2}{3m^*_{nt}} + \frac{1}{3m_{nl}}.$$

In a spherical minimum (isotropic effective mass) we finally have

$$m^*_{n,tr} = m^*_n.$$

For holes, the situation is more complex, since heavy and light holes exist. It can be shown that the transport hole effective mass is given by a weighted average over the heavy and light holes as (see e.g., [3], Section 8.1.2)

$$\frac{1}{m^*_{h,tr}} = \frac{p_{hh}}{p m^*_{hh}} + \frac{p_{lh}}{p m^*_{lh}},$$

where p_{lh} and p_{hh} are the light and heavy hole densities and $p = p_{lh} + p_{hh}$ is the total hole density. At or near equilibrium, the HH and LH populations are related through the effective densities of states, so that

$$\frac{p_{hh}}{p} = \frac{{m^*_{hh}}^{3/2}}{{m^*_{lh}}^{3/2} + {m^*_{hh}}^{3/2}}, \quad \frac{p_{lh}}{p} = \frac{{m^*_{lh}}^{3/2}}{{m^*_{lh}}^{3/2} + {m^*_{hh}}^{3/2}};$$

it follows that

$$\frac{1}{m^*_{h,tr}} = \frac{{m^*_{hh}}^{1/2} + {m^*_{lh}}^{1/2}}{{m^*_{hh}}^{3/2} + {m^*_{lh}}^{3/2}}.$$

For instance, in Si we have $m^*_{hh} = 0.49\, m_0$, $m^*_{lh} = 0.16\, m_0$; thus:

$$\frac{m_0}{m^*_{h,tr}} = \frac{0.49^{1/2} + 0.16^{1/2}}{0.16^{3/2} + 0.49^{3/2}} \rightarrow m^*_{h,tr} = 0.37 m_0.$$

1.3.2 The conduction and valence band wavefunctions

Electrons and holes belonging to the conduction and valence bands are characterized, from the standpoint of quantum mechanics, by a wavefunction. According to the Bloch theorem, wavefunctions in a periodic potential (e.g., a crystal) can be generally expressed as

$$\psi_{\underline{k}}(\underline{r}) = \exp(-\mathrm{j}\underline{k} \cdot \underline{r}) u_{\underline{k}}(\underline{r}), \tag{1.1}$$

where $u_{\underline{k}}(\underline{r})$ is a periodic function in the crystal space, such as $u_{\underline{k}}(\underline{r}) = u_{\underline{k}}(\underline{r} + \underline{L})$, $\underline{L}$ being a linear combination (with integer indices) of the primitive lattice vectors. The functional form of the wavefunction in (1.1), called the *Bloch wave*, ensures that the probability associated with the wavefunction is indeed a periodic function in the crystal space. For $\underline{k} \approx 0$ (e.g., near the Γ point) one has $\psi_{\underline{k}}(\underline{r}) \approx u_0(\underline{r})$, where $u_0(\underline{r})$ follows single-atom-like wavefunctions (s-type or p-type, see Fig. 1.11).

Since the detailed spatial behavior of wavefunctions is relevant to optical properties, we recall that *conduction band wavefunctions are, near the Γ point, s-type*, i.e., they have a probability distribution with spherical constant-probability surfaces. On the other hand, *the valence band wavefunctions are p-type*, i.e., they are even with respect to two orthogonal directions and odd with respect to the third, see Fig. 1.11. For instance, p_x is even with respect to the y and z axes and odd with respect to the x axis. The detailed shape of the wavefunctions is much less important than their property of being even in all directions (the s-type wavefunction) or odd with respect to one direction.

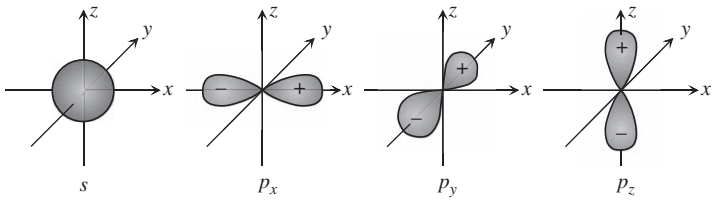

Figure 1.11 Conduction band (s-type) and valence band (p-type) wavefunctions: probability distribution and wavefunction sign (for p-type).

More specifically, it can be shown that heavy and light hole wavefunctions result from a superposition of p-type wavefunctions:

$$\phi_{HH}(x, y, z) = -\frac{1}{\sqrt{2}}\left(p_x \pm \mathrm{j}p_y\right) \tag{1.2}$$

$$\phi_{LH}(x, y, z) = -\frac{1}{\sqrt{6}}\left(p_x \pm \mathrm{j}p_y \mp 2p_z\right), \tag{1.3}$$

where the prefactors are introduced for normalization, see e.g., [4], Section 2.4.

1.3.3 Direct- and indirect-bandgap semiconductors

A simplified version of the dispersion relation, including the main conduction band minima and valence band maxima, is often enough to explain the electronic and optical behavior of a semiconductor. Such an example is shown in Fig. 1.12(a), for GaAs: the coincident maxima and minima in the Γ point make this semiconductor a typical example of *direct-bandgap* material. Direct-bandgap semiconductors are particularly important in optics, because they are able to interact directly with photons; in fact, those can provide an energy of the order of the energy gap, but negligible momentum. To promote an electron from the valence to the conduction band, an energy larger than the gap has to be provided, but, in GaAs, negligible momentum, since the valence band maximum and conduction band minimum are both at $\underline{k} = 0$. Since the interaction involves only one electron and one photon, the interaction probability is high.

Silicon, the most important semiconductor in electronics, is an example of an indirect-bandgap semiconductor, i.e., a material in which the valence band maximum and conduction band minimum occur at different values of $\underline{k}$, see Fig. 1.12(b). In particular, the main conduction band minimum is close to point X but within the FBZ, and six minima exist in the FBZ. The electron energy around such minima can be expressed as a function of the transverse (e.g., orthogonal to (100)) and of the longitudinal (e.g., parallel to (100)) wavenumbers:

$$E_n \approx E_c + \frac{\hbar^2 k_t^2}{2m_{nt}^*} + \frac{\hbar^2 k_l^2}{2m_{nl}^*}.$$

In Si, the electron–photon interaction leading to band-to-band processes requires a substantial amount of momentum, which has to be supplied by a further particle, typically a *lattice vibration* (phonon). The multibody nature of the interaction makes it less

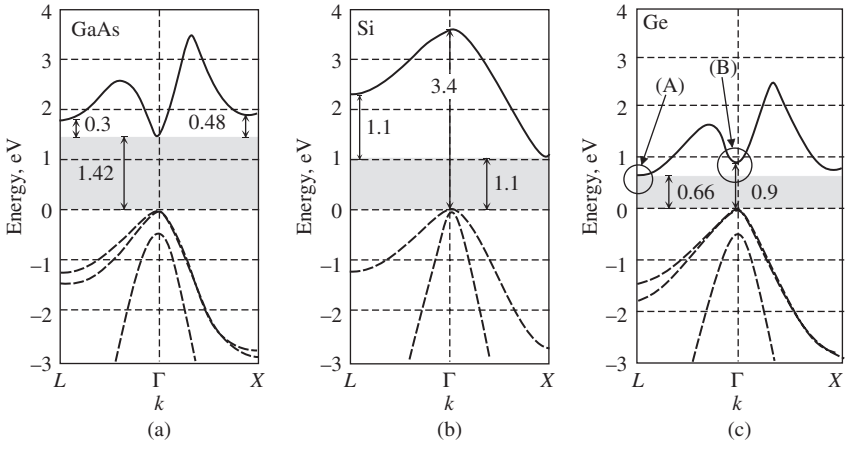

Figure 1.12 Simplified dispersion relation for (a) GaAs, (b) Si, (c) Ge. In Ge, the main conduction band minimum (A) is indirect, and has an impact on transport and low-energy optical properties; the secondary direct minimum (B) influences the optical properties at high photon energy.

probable, and therefore the interaction strength is lower. Typically, direct-bandgap semiconductors are able to absorb and emit light; indirect-bandgap semiconductors absorb light (albeit less efficiently) but are unable to operate as high-efficiency light emitters, particularly in lasers. Germanium (Ge), see Fig. 1.12(c), is an indirect-bandgap semiconductor; the lowest conduction band minimum is at point L, but a direct bandgap exists with a higher energy (0.9 eV) at point Γ. As a result, the main transport properties of Ge are typical of an indirect-bandgap material, but optical properties are influenced by the fact that high-energy photons can excite electrons directly from the valence band to the direct minimum. Some of germanium's optical properties (e.g., the absorption) exhibit both indirect- and direct-bandgap semiconductor features, depending on the photon energy.

In the above materials, the central minima can be characterized by isotropic or quasi-isotropic (as for the valence band) effective masses, while indirect bandgap minima are typically anisotropic and have to be described in terms of a longitudinal and transverse effective mass. A summary of the effective masses and other band properties in Si and GaAs is shown in Table 1.4.

Many III-V semiconductors have a bandstructure similar to GaAs. InP, see Fig. 1.13(a), has a slightly lower bandgap, but a larger difference between the central and the lateral minima. This has important consequences on transport properties, since it increases the electric field at which the electrons are scattered from the central minimum (characterized by high mobility, i.e., high electron velocity with the same applied electric field) to the lateral minima (with low mobility). This ultimately leads to a decrease of the average electron velocity with increasing field, see Fig. 1.14. The maximum velocity is larger in InP than in GaAs, allowing for the development of electron devices (such as transistors) with superior properties in terms of maximum speed. The peak electron velocity (corresponding to the onset of the negative differential mobility region) occurs at a field $\mathcal{E}_m$ related to the energy difference ΔE between the Γ and

Table 1.4 Main band properties of Si and GaAs. The electron mass m_0 is 9.11×10^{-34} kg.

Property	Si	GaAs
Electron effective masses	$m^*_{nl} = 0.98m_0$ $m^*_{nt} = 0.19m_0$	$m^*_n = 0.067m_0$
	$m^*_{n,D} = 1.08m_0$	$m^*_{n,D} = 0.067m_0$
	$m^*_{n,tr} = 0.26m_0$	$m^*_{n,tr} = 0.067m_0$
Hole effective masses	$m^*_{hh} = 0.49m_0$	$m^*_{hh} = 0.45m_0$
	$m^*_{lh} = 0.16m_0$	$m^*_{lh} = 0.08m_0$
	$m^*_{h,D} = 0.55m_0$	$m^*_{h,D} = 0.47m_0$
	$m^*_{h,tr} = 0.37m_0$	$m^*_{h,tr} = 0.34m_0$
Energy gap $E_g(T)$, T (K)	$1.17 - \dfrac{4.37 \times 10^{-4}T^2}{636 + T}$	$1.52 - \dfrac{5.4 \times 10^{-4}T^2}{204 + T}$
Electron affinity $q\chi$ (eV)	4.01	4.07

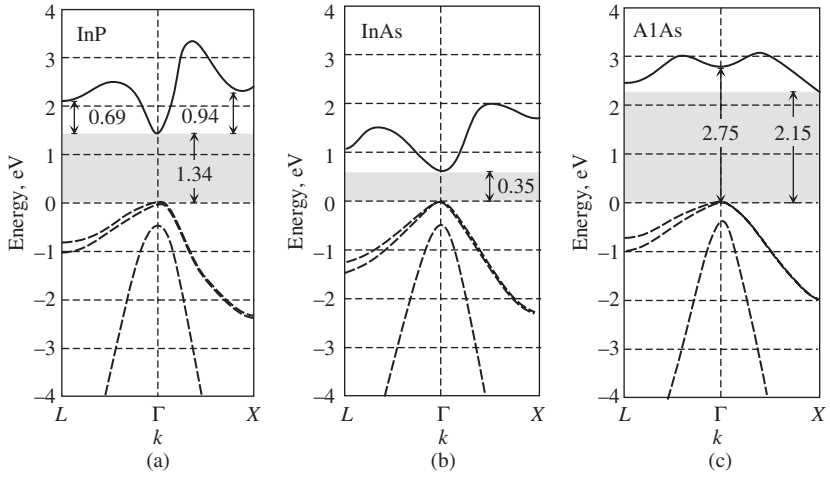

Figure 1.13 Simplified bandstructure of (a) InP, (b) InAs, (c) AlAs.

lateral minima; in GaAs $\Delta E \approx 300$ meV with $\mathcal{E}_m \approx 3.2$ kV/cm while in InP $\Delta E \approx 700$ meV with $\mathcal{E}_m \approx 10$ kV/cm.[4] InAs, see Fig. 1.13(b), has a very similar bandstructure, but with lower energy gap. For certain compound semiconductors, such as AlAs, see Fig. 1.13(c), the central minimum is higher than the lateral minima, thus making the material of indirect-bandgap type. InAs and AlAs are not particularly important per se, but rather as the components of semiconductor alloys. Some additional compound semiconductor properties are listed in Table 1.5, where v_s is the electron high-field saturation velocity (also denoted as $v_{n,\text{sat}}$), $v_{\max}$ is the maximum steady-state electron velocity. Notice that, while the saturation velocity is almost independent of doping, the maximum in the nonmonotonic velocity–field curve of most compound semiconductors

[4] In GaN, on the other hand, $\Delta E \approx 3$ eV, leading to a peak field in excess of 200 kV/cm, see, e.g., [5].

Table 1.5 Band properties of some important compound semiconductors. Mobility data are upper bounds referring to undoped material.

Property	$In_{0.53}Ga_{0.47}As$	GaAs	InP	AlAs	InAs
a (Å)	5.869	5.683	5.869	5.661	6.0584
E_g @300 K (eV)	0.717	1.424	1.34	2.168	0.36
$q\chi$ (eV)		4.07	4.37	3.50	4.90
m_n^*/m_0	0.041	0.067	0.077	0.150	0.027
m_{lh}^*/m_0	0.044	0.08	0.12	0.150	0.023
m_{hh}^*/m_0	0.452	0.45	0.6	0.76	0.60
$\epsilon(0)/\epsilon_0$	13.77	13.18	12.35	10.16	14.6
$\epsilon(\infty)/\epsilon_0$	11.38	10.9	9.52	8.16	12.25
$\mathcal{E}_{br}$ (kV/cm)	3.0	3.2	11		
μ_n (cm^2/Vs)	12000	8500	5500		40000
v_{max} (10^7 cm/s)	≈ 2.5	≈ 1.7	≈ 2.7		
v_s (10^7 cm/s)	0.7	1			

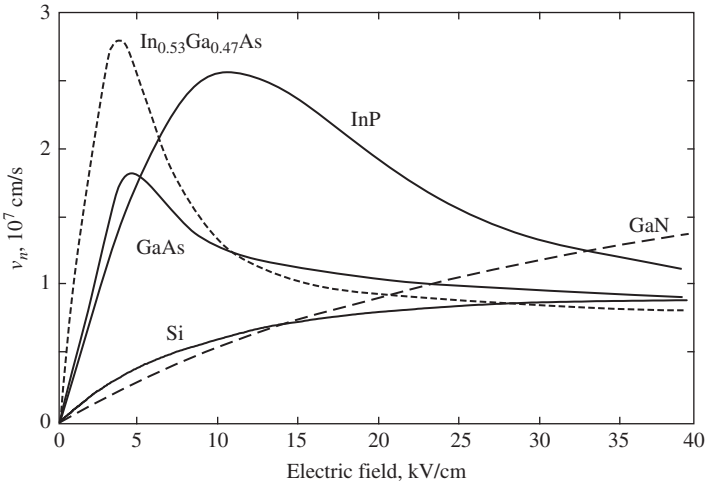

Figure 1.14 Electron drift velocity–field curves of Si, GaAs, InP, GaN, and InGaAs lattice matched to InP. The GaN velocity has a peak toward 200 kV/cm and then saturates with GaAs-like behavior. Adapted from [6], p. 13.

(see Fig. 1.14) depends on the low-field mobility and therefore on doping; the values provided (referring to intrinsic material) are therefore indicative.

Compound semiconductor families are classified according to the chemical nature of the metal and nonmetal components. If the metal component belongs to group III and the nonmetal to group V, we obtain a III-V compound. Examples of III-V compounds are GaAs, InP, GaSb, InAs (direct bandgap) and AlAs, GaP (indirect bandgap). III-V compounds with nitrogen such as GaN, InN, AlN are often referred to as III-N compounds. III-V compounds are probably the most important semiconductors for high-frequency electronics and optoelectronics. II-VI compounds include CdTe, HgTe,

ZnS, CdSe, ZnO (direct bandgap; note that HgTe has a negative bandgap and therefore has a metal rather than semiconductor behavior). IV-IV compounds are SiC and SiGe, both of them of indirect bandgap type. Finally, I-VII semiconductor compounds also exist, such as AgI and CuBr.

1.4 Carrier densities in a semiconductor

1.4.1 Equilibrium electron and hole densities

According to the picture drawn so far, a simplified representation of the semiconductor bandstructure includes two energy bands, the valence and conduction bands, separated by the energy gap E_g. Some electrons have large enough energy to be promoted from the valence to the conduction band, leaving behind positive charges called holes. Both electrons and holes can interact with an external electric field, and with photons or other particles. Further details of the bandstructure are introduced in Fig. 1.1, such as the electron affinity $q\chi$, i.e., the distance between the conduction band edge and the vacuum level U_0, and the ionization I_0, i.e., the distance between the valence band edge and the vacuum level. The electron and hole populations n and p depend on the number of electron and hole states per unit volume in the two bands (density of states N_c and N_v, respectively, both functions of the energy), and on how those states are populated as a function of the energy. According to statistical mechanics, electrons and holes follow at equilibrium the *Fermi–Dirac* distribution,[5] while the out-of-equilibrium distribution can be often approximated, in optoelectronic devices, by the so-called *quasi-Fermi* distribution.

In the effective mass approximation, the density of states (DOS) in a 3D (bulk) semiconductor can be shown to be

$$N_c(E) \equiv g_c(E) = \frac{4\pi}{h^3}(2m^*_{n,D})^{3/2}\sqrt{E - E_c}$$
$$N_v(E) \equiv g_v(E) = \frac{4\pi}{h^3}(2m^*_{h,D})^{3/2}\sqrt{E_v - E},$$

whose behavior is shown in Fig. 1.15. Owing to the effect of heavy holes, the valence band DOS typically is larger than the conduction band DOS.

The Fermi–Dirac distributions describing the electron and hole equilibrium occupation statistics are expressed as

$$f_n(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{k_B T}\right)} \tag{1.4}$$

$$f_h(E) = \frac{1}{1 + \exp\left(\dfrac{E_F - E}{k_B T}\right)}, \tag{1.5}$$

[5] Or by the *Boltzmann* distribution, an approximation of the Fermi–Dirac distribution holding for energies larger than the Fermi energy.

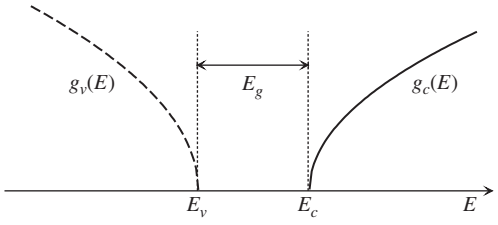

Figure 1.15 Valence (g_v) and conduction band (g_c) density of states in a bulk semiconductor.

where the Fermi level E_F is constant in the whole system. If the Fermi level is within the energy gap (this case corresponds to *nondegenerate semiconductors*) the Boltzmann approximation of the statistics holds:

$$f_n(E) \underset{E \gg E_F}{\approx} \exp\left(\frac{E_F - E}{k_B T}\right), \quad f_h(E) \underset{E \ll E_F}{\approx} \exp\left(\frac{E - E_F}{k_B T}\right).$$

The Boltzmann approximation applies, in fact, if the distance between E and E_F is larger than a few $k_B T$ units. In the *degenerate* case the Fermi level can fall into the conduction or valence bands, and this condition is violated; in such cases, the full Fermi–Dirac statistics has to be used.

The behavior of the two Fermi–Dirac distributions for electrons and holes is shown in Fig. 1.16. Integrating the product between the density of states and the statistical distributions (with the Boltzmann approximation) over all energies (i.e., from E_c to $\approx \infty$ for the conduction band and from $\approx -\infty$ to E_v for the valence band), we have

$$n = \int_{E_c}^{\infty} N_c(E) f_n(E)\, \mathrm{d}E = N_c \exp\left(\frac{E_F - E_c}{k_B T}\right)$$
$$p = \int_{-\infty}^{E_v} N_v(E) f_h(E)\, \mathrm{d}E = N_v \exp\left(\frac{E_v - E_F}{k_B T}\right),$$

where the *effective densities of states* are

$$N_c = 2\frac{(2\pi m^*_{n,D} k_B T)^{3/2}}{h^3}, \quad N_v = 2\frac{(2\pi m^*_{h,D} k_B T)^{3/2}}{h^3}. \tag{1.6}$$

In an *intrinsic* (undoped) semiconductor $p = n = n_i$, where

$$n_i = N_c \exp\left(\frac{E_{Fi} - E_c}{k_B T}\right) = p_i = N_v \exp\left(\frac{E_v - E_{Fi}}{k_B T}\right),$$

from which the intrinsic Fermi level can be derived; the intrinsic Fermi level is located at midgap, with a small (typically negative) correction related to the ratio $N_c/N_v = (m^*_{n,D}/m^*_{h,D})^{3/2}$:

$$E_{Fi} = k_B T \log\sqrt{\frac{N_c}{N_v}} + \frac{E_c + E_v}{2}.$$

Moreover, the intrinsic concentration can be directly related to the energy gap:

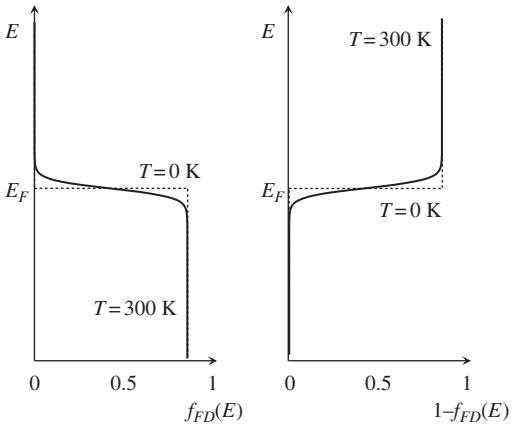

Figure 1.16 Fermi–Dirac distributions for electrons (left) and holes (right).

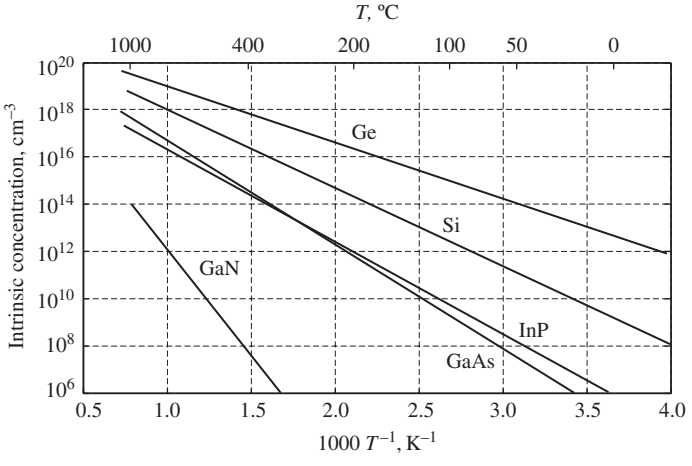

Figure 1.17 Intrinsic concentration for Ge, Si, GaAs, InP and GaN (wurtzite) as a function of the lattice temperature. Data from [1].

$$n_i p_i = n_i^2 = N_c N_v \exp\left(-\frac{E_g}{k_B T}\right). \tag{1.7}$$

The intrinsic concentration as a function of the temperature for Si, Ge, GaAs, InP and GaN is shown in Fig. 1.17. With increasing T, the intrinsic concentration increases exponentially; this is one of the main limitations in high-temperature semiconductor operation, since when the intrinsic concentration is of the order of the doping, doping becomes ineffective.

In equilibrium conditions, the product of the concentrations n and p does not depend on the position of the Fermi level, and is equal to the square of the intrinsic concentration (*mass action law*):

$$np = n_i^2. \tag{1.8}$$

1.4.2 Electron and hole densities in doped semiconductors

The mass action law also holds for doped semiconductors. A semiconductor can be doped with a donor (density N_D), an element able to provide an additional electron when substituting an atom of the native semiconductor lattice. Examples of donors in Si are As and P (both belonging to group V, and therefore with an extra electron in the outer shell vs. Si). The additional electron is weakly bound to the donor (ionization energy into the conduction band of the order of 10 meV for shallow donors) and therefore can easily be ionized and enter the conduction band, thus participating in conduction.[6] In this case, the semiconductor is called n-type. Semiconductors can also be doped with acceptors (concentration N_A). For instance, Si atoms have 4 electrons in the outermost shell; acceptors (e.g., B, a group III element) have 3 electrons in the outermost shell (i.e., one electron less than the substituted native atom) and can therefore attract an electron from the valence band, leaving behind a hole (again with a ionization energy of the order of 10 meV). The semiconductor in this case is called p-type.

If donors and acceptors are fully ionized one has, also taking into account the mass action law (1.8):

$$n \approx N_D^+ \approx N_D, \qquad p \approx n_i^2/N_D \quad n\text{-type semiconductor}$$
$$p \approx N_A^- \approx N_A, \qquad n \approx n_i^2/N_A \quad p\text{-type semiconductor.}$$

In a doped semiconductor, the carrier concentration evolves with temperature according to a three-region behavior; the relevant intervals are the freeze-out, the saturation, and the intrinsic range.

At extremely low temperature, most carriers do not have enough energy to ionize into the conduction band, and the carrier population decreases with T well below the value $n \approx N_D$ (*freeze-out range*). The intermediate range (called the *saturation range*), corresponding to normal device operation, begins at a temperature such as $(3/2)k_BT \approx 20$ meV, i.e., $T \approx 150$ K (this is just an indicative value, since the donor or acceptor ionization energy depends on the doping and semiconductor materials), and ends at a temperature such as $n_i(T) \approx N_D$ (in n-type Si with $N_D = 10^{15}$ cm^{-3} this corresponds to $T \approx 200^{\circ}$C). In the saturation range, $n \approx N_D$ or $p \approx N_A$; the maximum operating temperature increases with increasing gap. Finally, above the saturation range we find the *intrinsic range*: at high temperature the intrinsic concentration becomes large enough to flood the semiconductor with electrons not originating from the donors (or holes not originating from the acceptors).

[6] A donor or acceptor introduces an isolated energy level in the forbidden band. Shallow donors have an energy level E_D close to the conduction band edge (typically a few meV), while for shallow acceptors the energy level E_A is close to the valence band. Deep donors and acceptors have energy levels close to the center of the gap and act more as electron or hole *traps* (or *recombination centers*) than as dopants, since their ionization (or electrical activation) is low. Ionized dopants follow electron- or hole-like Fermi statistics: donors are almost 100% activated if the Fermi level is *below* the donor level, while acceptors are almost 100% activated if the Fermi level is *above* the acceptor level. This implies, for example, that a deep donor is not ionized in an n-type semiconductor, and even the activation of shallow donors ultimately drops for extremely large n-type doping, since for increasing donor concentration the Fermi level finally becomes larger than the donor level.

From the expressions for the electron and hole densities, the Fermi level can easily be evaluated. In n-type semiconductors, the Fermi level increases vs. E_{Fi}, becoming closer to the conduction band edge, while for p-type semiconductors the Fermi level decreases and becomes closer to the valence band edge. For very high doping (e.g., in excess of 10^{19} cm^{-3}), donors and acceptors cannot be assumed to be 100% ionized (or electrically activated) any longer, but their ionization is related to the very position of the Fermi level and typically decreases, as already remarked, when the Fermi level becomes larger than the donor or smaller than the acceptor energy level.

In a degenerate semiconductor, the Fermi level (or the quasi-Fermi level out of equilibrium) is very close to the conduction or valence band edges or even falls within one of the two bands. Typically, a semiconductor cannot be made degenerate by doping, but degeneracy is a condition that can be achieved out of equilibrium (e.g., in a direct-bias pn junction under high carrier injection).

1.4.3 Nonequilibrium electron and hole densities

To address the out-of-equilibrium statistics in a simplified way, we note that deviations from thermodynamic equilibrium may imply two quite different consequences: disequilibrium between the electron and the hole populations, and disequilibrium in carrier populations due to an applied (electric) field.

In equilibrium, the electron and hole populations follow the mass action law, any deviation from this being compensated for by generation–recombination (GR) processes whereby electron–hole (e-h) pairs are generated or disappear by recombination. The excess charge n' or p' (with respect to equilibrium) is removed according to the time behavior

$$n'(t) = n'(0)\exp(-t/\tau_n),$$

with a characteristic time (called the excess lifetime, τ_n or τ_h for electrons and holes, respectively) whose order of magnitude can range from a few milliseconds to nanoseconds according to the restoring mechanism. Recombination processes basically involve an exchange of energy and momentum with other particles, e.g., phonons (lattice vibrations, corresponding to the so-called thermal GR process), photons (radiative GR), other electrons and holes (Auger recombination and impact generation). If the carrier population deviation with respect to equilibrium is *maintained by an external cause* (e.g., a photon flux leading to radiative generation of e-h pairs) the resulting out-of-equilibrium condition can be characterized by a slightly modified form of the equilibrium probability distribution (called the *quasi-Fermi distribution*).

A second nonequilibrium situation derives from the effect of an applied electric field. While the average carrier velocity is zero at equilibrium, and therefore the *carrier distribution* in the velocity space is symmetrical with respect to the origin, application of an electric field leads to an increase of the average velocity and to a nonsymmetrical velocity distribution. For very large fields, the change in shape of

the distribution with respect to the equilibrium may become dramatic and a simple quasi-Fermi approach will not be sufficient. However, this form of extreme field–carrier disequilibrium is not essential in the analysis of most optoelectronic devices, and therefore a simplified discussion based on the static carrier velocity–field properties will suffice.

To describe electron–hole imbalance with respect to the equilibrium, we therefore introduce the so called quasi-Fermi statistics, where the single Fermi level is replaced by two separate *quasi-Fermi levels* E_{Fn} and E_{Fh} according to the following formulae:

$$f_n(E, E_{Fn}) = \frac{1}{1 + \exp\left(\dfrac{E - E_{Fn}}{k_B T}\right)} \underset{E \gg E_{Fn}}{\approx} \exp\left(\frac{E_{Fn} - E}{k_B T}\right) \tag{1.9}$$

$$f_h(E, E_{Fh}) = \frac{1}{1 + \exp\left(\dfrac{E_{Fh} - E}{k_B T}\right)} \underset{E \ll E_{Fh}}{\approx} \exp\left(\frac{E - E_{Fh}}{k_B T}\right), \tag{1.10}$$

where the relevant Boltzmann approximations have also been introduced. Within the Boltzmann approximation the carrier densities become

$$n = N_c \exp\left(\frac{E_{Fn} - E_c}{k_B T}\right), \quad p = N_v \exp\left(\frac{E_v - E_{Fh}}{k_B T}\right),$$

while the mass action law can be modified to allow for a difference in the two quasi-Fermi levels (in equilibrium $E_{Fn} = E_{Fh} = E_F$):

$$np = n_i^2 \exp\left(\frac{E_{Fn} - E_{Fh}}{k_B T}\right). \tag{1.11}$$

In particular,

$$np > n_i^2 \quad \text{for } E_{Fn} > E_{Fh} \text{ (carrier injection)}$$
$$np < n_i^2 \quad \text{for } E_{Fn} < E_{Fh} \text{ (carrier depletion).}$$

In the degenerate case, the Boltzmann approximation is invalid and we have to express the charge density with the help of special functions (the Fermi–Dirac integrals):

$$n = \frac{2}{\sqrt{\pi}} N_c \mathcal{F}_{1/2}\left(\frac{E_{Fn} - E_c}{k_B T}\right), \quad p = \frac{2}{\sqrt{\pi}} N_v \mathcal{F}_{1/2}\left(\frac{E_v - E_{Fh}}{k_B T}\right).$$

The computation of the Fermi–Dirac integral can be performed through suitable analytical approximations; an example is given by the Joyce–Dixon (inverse) formulae:

$$E_{Fn} \approx E_c + k_B T \left[\log \frac{n}{N_c} + \frac{1}{\sqrt{8}} \frac{n}{N_c}\right] \tag{1.12}$$

$$E_{Fh} \approx E_v - k_B T \left[\log \frac{p}{N_v} + \frac{1}{\sqrt{8}} \frac{p}{N_v}\right]. \tag{1.13}$$

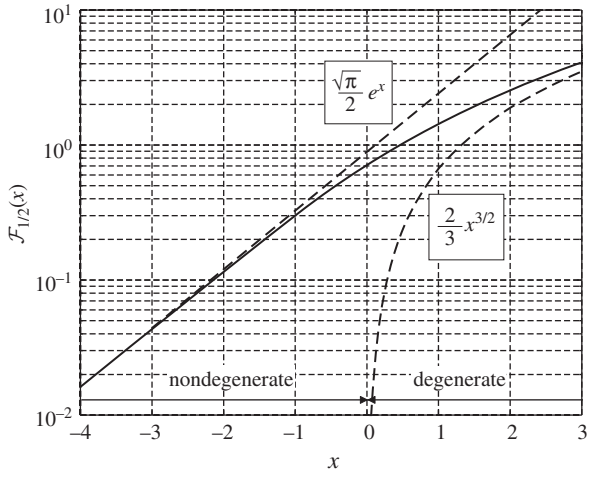

Figure 1.18 Behavior of the Fermi–Dirac integral ($\mathcal{F}_{1/2}$) in the degenerate and nondegenerate ranges.

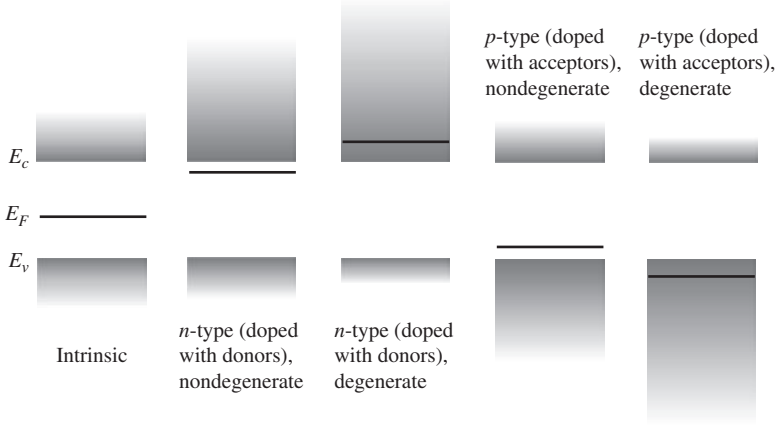

Figure 1.19 Examples of the position of the Fermi level in several semiconductors at equilibrium. In practice, semiconductors cannot be made degenerate by doping.

The overall behavior of the Fermi integral in the two ranges (nondegenerate and degenerate) is shown in Fig. 1.18. For extreme degeneration, the following polynomial approximation holds:

$$n \approx \frac{\sqrt{2}m_{n,D}^{3/2}}{\pi^2\hbar^3}\frac{2}{3}(E_{Fn}-E_c)^{3/2}\,, \qquad p \approx \frac{\sqrt{2}m_{h,D}^{3/2}}{\pi^2\hbar^3}\frac{2}{3}(E_v-E_{Fh})^{3/2}\,.$$

A summary of some possible equilibrium bandstructures is shown in Fig. 1.19; notice that the *n*-type and *p*-type degenerate cases are purely theoretical, since increasing the doping level beyond a certain level makes $E_F > E_D$ or $E_F < E_A$, thus decreasing the donor or acceptor activation. This implies that the degenerate condition cannot practically be obtained at equilibrium. Finally, Fig. 1.20 concerns examples out of equilibrium; degeneracy arises in these cases from the high-injection condition.

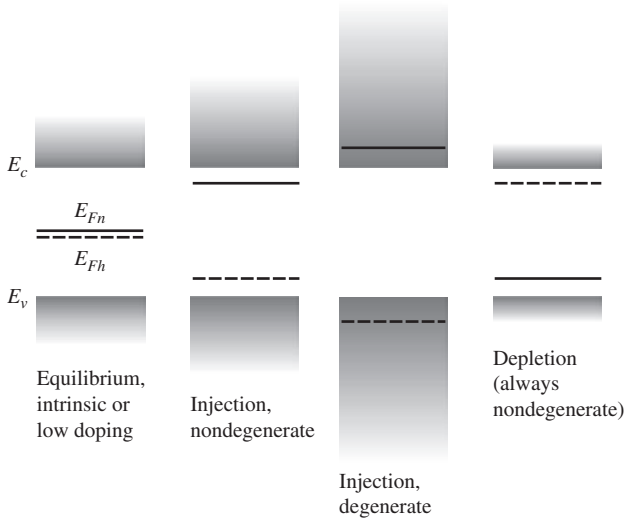

Figure 1.20 Examples of transition from the equilibrium (left) to the out-of-equilibrium bandstructure for degenerate and nondegenerate semiconductors.

1.5 Heterostructures

Crystals with different lattice constants grown on top of each other by epitaxial techniques are affected by interface defects called *misfit dislocations*. Such defects operate as electron or hole traps, and therefore the resulting structure is unsuited to the development of electron devices. However, if the lattice mismatch between the substrate and the heteroepitaxial overlayer is low or zero, an ideal or almost ideal crystal can be grown, made of two different materials. The resulting structure is called a *heterostructure*, and, since the electronic properties of the two layers are different, we also refer to it as a *heterojunction*. The material discontinuity arising in the heterojunction leads to important electronic and optical properties, such as confinement of carriers (related to the discontinuity of the conduction or valence bands) and confinement of radiation (due to the bandgap discontinuity and to the related refractive index step).

Heterostructures can be lattice-matched (if the two sides have the same lattice constant) or affected by a slight mismatch (indicatively, the maximum mismatch is of the order of 1%), which induces tensile or compressive strain. In this case, we talk about *pseudomorphic* or *strained* heterostructures, see Fig. 1.21. A small amount of strain in the heterostructure can be beneficial to the development of electronic or optoelectronic devices, since it leads to additional degrees of freedom in the band structure engineering, and in many cases allows for an improvement of the material transport or optical properties.

A double heterojunction made with a thin semiconductor layer (the thickness should be typically of the order of 100 nm) sandwiched between two layers (e.g., AlGaAs/GaAs/AlGaAs) creates a potential well in the conduction and/or valence band and is often referred to as a quantum well (QW). A succession of weakly interacting

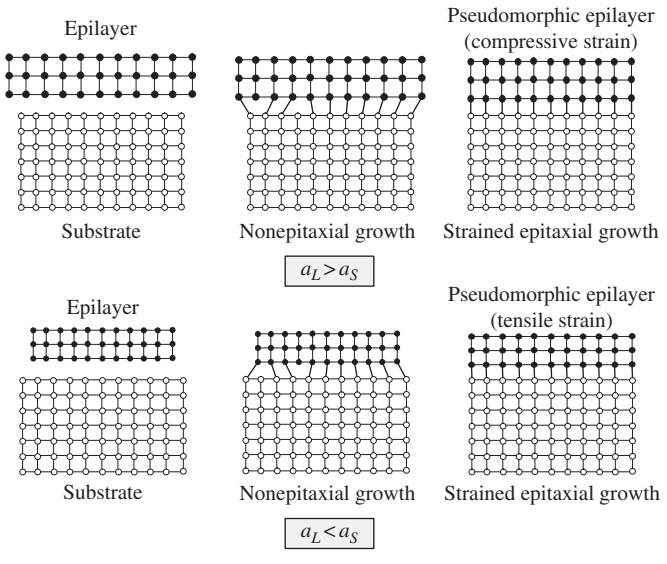

Figure 1.21 Pseudomorphic or strained growth. Above, the epilayer lattice constant is larger than that of the substrate: nonepitaxial growth with interface misfit dislocations and strained epitaxy. Below, the epilayer lattice constant is smaller than that of the substrate.

quantum wells is called a multi quantum well (MQW); if the MQW has many layers, with significant overlapping between the wavefunctions of adjacent wells, we finally obtain a superlattice (SL). The artificial periodicity imposed by the superlattice over the natural periodicity of the crystal introduces important modifications in the electronic properties.

1.6 Semiconductor alloys

Heterostructures are largely based on semiconductor alloys. The idea behind alloys is to create semiconductors having intermediate properties with respect to already existing "natural" semiconductors. Among such properties are the lattice constant a and the energy gap E_g. In several material systems, both a and E_g approximately follow a linear law with respect to the individual component parameters. The motivation to tailor the lattice constant is of course to achieve lattice matching to the substrate; tailoring the energy gap gives the possibility to change the emitted photon energy, thus generating practically important wavelengths, such as the 1.3 or 1.55 μm wavelengths needed for long-haul fiber communications (since they correspond to minimum fiber dispersion and absorption, respectively, see Fig. 1.22). Examples are alloys made of two components and three elements (called *ternary alloys*: e.g., AlGaAs, alloy of GaAs and AlAs) and alloys made of four components and elements (called *quaternary alloys*, e.g., InGaAsP, alloy of InAs, InP, GaAs, GaP). By proper selection of the alloy composition, semiconductor alloys emitting the right wavelength and matched to the right substrate can be generated.

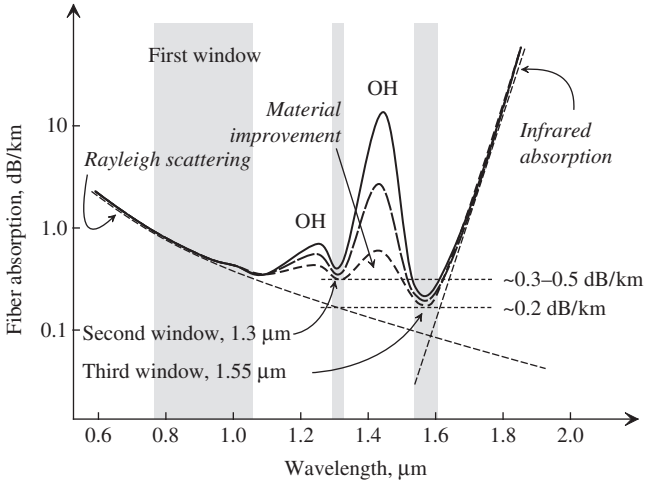

Figure 1.22 Absorption profile of a glass optical fiber.

In order to quantitatively define an alloy, we have to consider that compound semiconductors (CS) are polar compounds with a metal M combined with a nonmetal N in the form MN. Two different CSs sharing the same metal or nonmetal give rise to a ternary alloy or compound:

$$(M_1N)_x\,(M_2N)_{1-x} = M_{1x}M_{2(1-x)}N, \qquad \text{e.g., } \mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$$
$$(MN_1)_y\,(MN_2)_{1-y} = MN_{1y}N_{2(1-y)}, \qquad \text{e.g., } \mathrm{GaAs}_y\mathrm{P}_{1-y},$$

where x and $1-x$ denote the mole fraction of the two metal components, and y and $1-y$ denote the mole fraction of nonmetal components. Four different CSs sharing two metal and two nonmetal components yield a quaternary alloy or compound. In the following formulae, M and m are the metal components, N and n are the nonmetal components, and $\alpha+\beta+\gamma=1$:

$$(MN)_\alpha\,(Mn)_\beta\,(mN)_\gamma\,(mn)_{1-\alpha-\beta-\gamma} = M_{\alpha+\beta}m_{1-\alpha-\beta}N_{\alpha+\gamma}n_{1-\alpha-\gamma}$$
$$= M_x m_{1-x} N_y n_{1-y} \quad (\text{e.g., } \mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}_y\mathrm{P}_{1-y}).$$

Most alloy properties can be derived from the component properties through (global or piecewise) linear interpolation (Vegard law), often with second-order corrections (Abeles law); examples are the lattice constant, the energy gap, the inverse of the effective masses, and, in general, the bandstructure and related quantities. Varying the composition of a ternary alloy (one degree of freedom) changes the gap and related wavelength, but, at the same time, the lattice constant; in some cases (AlGaAs) the two components (AlAs and GaAs) are already matched, so that alloys with arbitrary Al content are lattice matched to the substrate (GaAs).

On the other hand, varying the composition of a quaternary alloy (two degrees of freedom) independently changes both the gap and the lattice constant, so as to allow for lattice matching to a specific substrate, e.g., InGaAsP on InP.

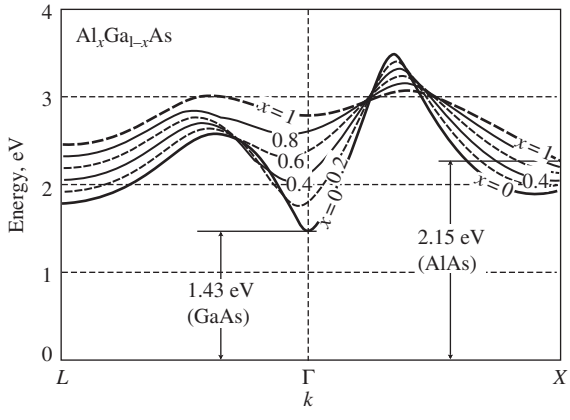

Figure 1.23 Evolution of the bandstructure of AlGaAs changing the Al content from 0 to 1.

The Vegard or Abeles laws must be applied with care in some cases. As an example, consider the $Al_xGa_{1-x}As$ alloy and call P an alloy parameter, such as the energy gap. The Vegard law can be written as:

$$P(x) = (1-x)P_{GaAs} + xP_{AlAs};$$

by inspection, this yields a linear interpolation between the two constituent parameters. However, this law fails to accurately reproduce the behavior of the AlGaAs energy gap because GaAs is direct bandgap, and AlAs is indirect. To clarify this point, let us consider the simplified bandstructure of the alloy as shown in Fig. 1.23. We clearly see that the main and secondary (X point) minima have the same level for $x = 0.45$; for larger Al mole fraction, the material becomes indirect bandgap. Since the composition dependence is different for the energy levels of the Γ and X minima, a unique Vegard law fails to approximate the gap for any alloy composition, and a piecewise approximation is required:

$$E_g \approx 1.414 + 1.247x, \quad x < 0.45$$
$$E_g \approx 1.985 + 1.147(x - 0.45)^2, \quad x > 0.45.$$

The same problems arise in the InGaAsP alloy, since GaP is indirect bandgap; thus, a global Vegard approximation of the kind

$$P_{InGaAsP} = (1-x)(1-y)P_{GaAs} + (1-x)yP_{GaP} + xyP_{InP} + x(1-y)P_{InAs}$$

(by inspection, the approximation is bilinear and yields the correct values for the four semiconductor components) may be slightly inaccurate.

1.6.1 The substrate issue

Electronic and optoelectronic devices require to be grown on a suitable (typically, semiconductor) substrate. In practice, the only semiconductor substrates readily available are those that can be grown into monocrystal ingots through Czochralsky or Bridgman

techniques – i.e., in order of decreasing quality and increasing cost, Si, GaAs, InP, SiC, and a few others (GaP, GaSb, CdTe). Devices are to be grown so as to be either lattice matched to the substrate, or slightly (e.g., 1%) mismatched (pseudomorphic approach). The use of graded buffer layers allows us to exploit mismatched substrates, since it distributes the lattice mismatch over a larger thickness. This approach is often referred to as the *metamorphic* approach; it is sometimes exploited both in electronic and in optoelectronic devices. Metamorphic devices often used to have reliability problems related to the migration of defects in graded buffer layers; however, high-quality metamorphic field-effect transistors with an InP active region on a GaAs substrate have recently been developed with success.

1.6.2 Important compound semiconductor alloys

Alloys are often represented as a straight or curved segment (for ternary alloys) or quadrilateral area (for quaternary alloys) in a plane where the x coordinate is the lattice constant and the y coordinate is the energy gap; see Fig. 1.24. The segment extremes and the vertices of the quadrilateral are the semiconductor components. In Fig. 1.24 some important alloys are reported:

- AlGaAs, lattice-matched for any composition to GaAs, direct bandgap up to an Al mole content of 0.45.
- InGaAsP, which can be matched either to GaAs or to InP substrates; InP substrate matching includes the possibility of emitting 1.55 or 1.3 μm wavelengths;[7] the alloy is direct bandgap, apart from around the GaP corner, whose gap is indirect.
- InAlAs, which can be lattice matched to InP with composition $Al_{0.48}In_{0.52}As$.
- InGaAs, a ternary alloy matched to InP with composition $Ga_{0.47}In_{0.53}As$; it is a subset of the quaternary alloy InGaAsP.
- InGaAsSb, the antimonide family, a possible material for long-wavelength devices, but with a rather underdeveloped technology vs. InGaAsP.
- HgCdTe, a ternary alloy particularly relevant to far infrared (FIR) detection owing to the very small bandgap achievable.
- SiGe, an indirect bandgap alloy important for electronic applications (heterojunction bipolar transistors) but also (to a certain extent) for detectors and electroabsorption modulators;
- III-N alloys, such as AlGaN and InGaN, with applications in short-wavelength sources (blue lasers) but also in RF and microwave power transistors. AlGaN can be grown by pseudomorphic epitaxy on a GaN virtual substrate; GaN has in turn no native substrate so far, but can be grown on SiC, sapphire (Al_2O_3) or Si. The InGaN alloy is exploited in optoelectronic devices such as blue lasers and LEDs, besides being able to cover much of the visible spectrum.[8]

[7] InGaAsP lattice-matched to InP can emit approximately between 0.92 and 1.65 μm.

[8] The InN gap is controversial, and probably is much smaller than the previously accepted value around 2 eV. The nitride data in Fig. 1.24 are from [8], Fig. 3.

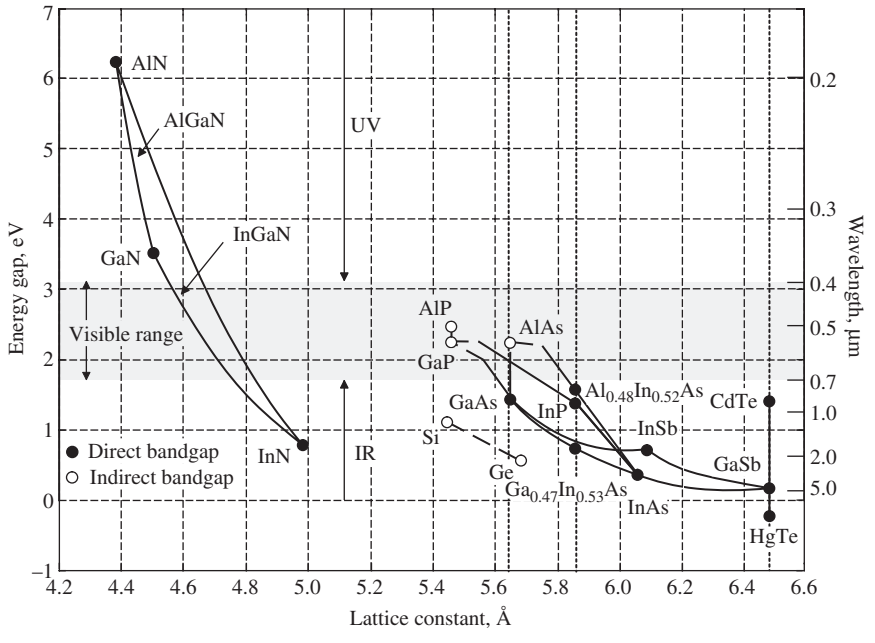

Figure 1.24 Some important alloys in the lattice constant–energy gap plane. In order of increasing gap and decreasing lattice constant: HgCdTe, InGaAsSb, InGaAsP, SiGe, AlGaAs, AlGaP, InGaN, AlGaN. For the widegap (wurtzite) nitrides (GaN, InN, AlN) the horizontal axis reports the equivalent cubic lattice constant. The InGaAsP, AlGaAs and InGaAsSb data are from [7]; the GaN, AlN and InN data are from [8], Fig. 3.

Since GaN, AlN, and InN have the wurtzite (hexagonal) crystal structure, an equivalent lattice constant $a_{C,eq}$ has to be defined for comparison with cubic crystals, so as to make the volume of the wurtzite cell V_H (per atom) equal to the volume of a cubic cell (per atom); taking into account that the wurtzite cell has 12 equivalent atoms, while the cubic cell has 8 equivalent atoms, we must impose:

$$\frac{1}{12}V_H = \frac{1}{12}\frac{3\sqrt{3}}{2}ca_H^2 = \frac{1}{8}a_{C.eq}^3 \rightarrow a_{C,eq} = \left(\sqrt{3}ca_H^2\right)^{1/3},$$

where V_H is the volume of the wurtzite cell prism of sides a and c. For GaN $a_H = 0.317$ nm, $c = 0.516$ nm; it follows that

$$a_{C,eq} = \left(\sqrt{3}ca_H^2\right)^{1/3} = \left(\sqrt{3}\cdot 0.516\cdot 0.317^2\right)^{1/3} = 0.448 \text{ nm}.$$

1.7 Bandstructure engineering: heterojunctions and quantum wells

Although the bandstructure of a semiconductor depends on the lattice constant a, which is affected by the operating temperature and pressure, significant variations in the bandstructure parameters cannot be obtained in practice. Nevertheless, semiconductor alloys

enable us to generate new, "artificial" semiconductors with band properties intermediate with respect to the components. A more radical change in the bandstructure occurs when heterojunctions are introduced so as to form quantized structures. A deep variation in the density of states follows, with important consequences in terms of optical properties (as we shall discuss later, the absorption profile as a function of the photon energy mimics the density of states). Moreover, strain in heterostructures allows for further degrees of freedom, like controlling the degeneracy between heavy and light hole subbands.

Heterojunctions are ideal, single-crystal junctions between semiconductors having different bandstructures. As already recalled, lattice-matched or strained (pseudomorphic) junctions between different semiconductors or semiconductor alloys allow for photon confinement (through the difference in refractive indices), carrier confinement (through potential wells in conduction or valence bands), and quantized structures such as superlattices, quantum wells, quantum dots, and quantum wires. An example of a heterostructure band diagram is shown in Fig. 1.25, where the band disalignment derives from application of the *affinity rule* (i.e., the conduction band discontinuity is the affinity difference, the valence band discontinuity is the difference in ionizations). In many practical cases, however, band disalignments are dominated by interfacial effects and do not follow the affinity rule exactly; for instance, in the AlGaAs-GaAs heterostructure one has

$$|\Delta E_c| \approx 0.65 \Delta E_g, \quad |\Delta E_v| \approx 0.35 \Delta E_g. \tag{1.14}$$

More specifically, the valence and conduction band discontinuities as a function of the Al fraction are (in eV) [1]:

$$|\Delta E_v| = 0.46x$$
$$|\Delta E_c| = \begin{cases} 0.79x, & x < 0.41 \\ 0.475 - 0.335x + 0.143x^2, & x > 0.41. \end{cases}$$

According to the material parameters, several band alignments are possible, as shown in Fig. 1.26; however, the most important situation in practice is the Type I band alignment in which the energy gap of the narrowgap material is included in the gap of the widegap material.

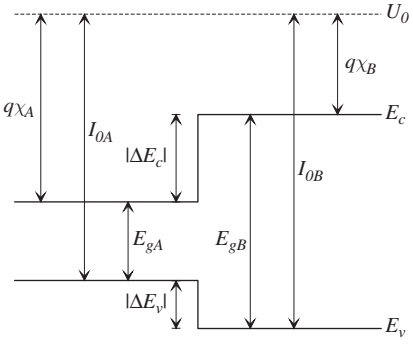

Figure 1.25 Heterostructure band alignment through application of the affinity rule to two materials having different bandstructures.

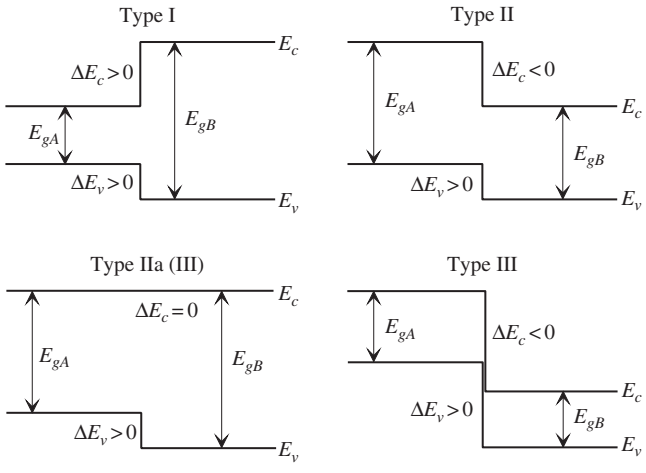

Figure 1.26 Classification of heterostructures according to band alignment; $\Delta E_j = E_{jB} - E_{jA}$.

Heterojunctions can be made with two n-type or p-type materials (*homotype heterojunctions*) so as to form a pn junction (*heterotype heterojunctions*). Often, the widegap material is conventionally denoted as N or P according to the type, the narrowgap material as n or p. According to this convention, a heterotype heterojunction is, for example, Np or nP and a narrowgap intrinsic layer sandwiched between two widegap doped semiconductors is NiP.

Single or double heterostructures can create potential wells in the conduction and/or valence bands, which can confine carriers so as to create conducting channels (with application to electron devices, such as field-effect transistors), and regions where confined carriers achieve high density and are able to recombine radiatively. In the second case, the emitted radiation is confined by the refractive index step associated with the heterostructure (the refractive index is larger in narrowgap materials). An example of this concept is reported in Fig. 1.27, a NiP structure in direct bias that may operate like the active region of a light-emitting diode or a semiconductor laser.

Carriers trapped by the potential well introduced by a double heterostructure are confined in the direction orthogonal to the well, but are free to move in the two other directions (i.e., parallel to the heterojunction). However, if the potential well is very narrow the allowed energy levels of the confined electrons and holes will be quantized. The resulting structure, called a quantum well (QW), has a different bandstructure vs. bulk, where sets of energy subbands appear (see Fig. 1.28). Also the density of states is strongly affected.

The quantum behavior of carriers in narrow (conduction or valence band) potential wells originated by heterojunctions between widegap and narrowgap semiconductors can be analyzed by applying the Schrödinger equation to the relevant particles (electron or holes) described in turn by a 3D effective mass approximation. Solution of the Schrödinger equation enables us to evaluate the energy levels and subbands, given the well potential profile. In a rectangular geometry, we start from bulk (3D motion possible,