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

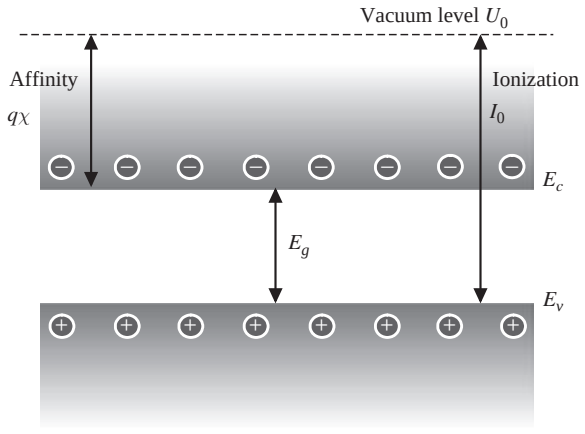

Figure 1.1 Main features of semiconductor bandstructure. E_g is the energy gap; E_c is the conduction band edge; E_v is the valence band edge.

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	$a_1 = a_2 \neq a_3, \alpha_1 = \alpha_2 = 90^\circ, \alpha_3 = 120^\circ$

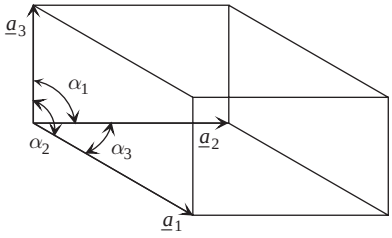

Figure 1.2 Semiconductor crystal structure: definition of the primitive cell.

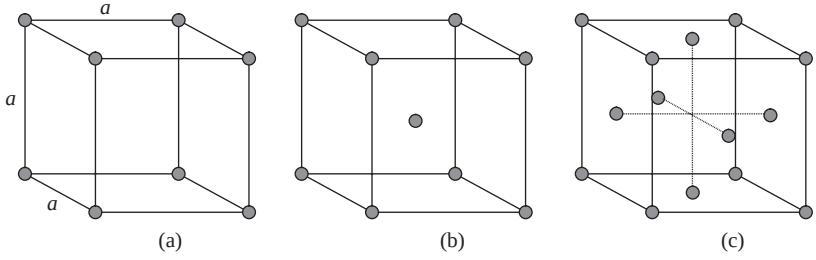

Figure 1.3 Cubic Bravais lattices: (a) simple, (b) body-centered, (c) face-centered.

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).

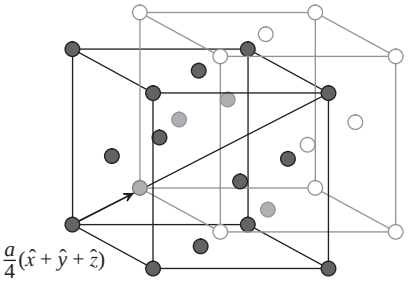

Figure 1.4 The diamond lattice as two cubic face-centered interpenetrating lattices. The pale and dark gray points represent the atoms falling in the basic cell.

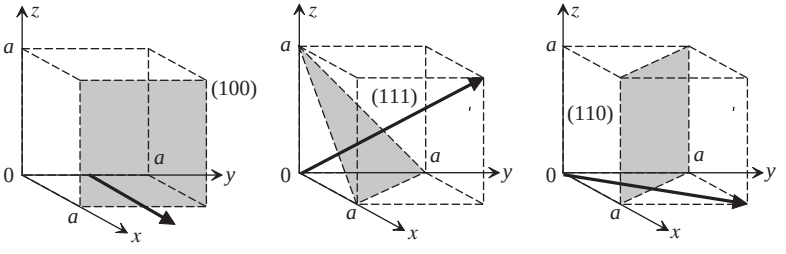

Figure 1.5 Examples of planes and directions according to the Miller notation.

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.

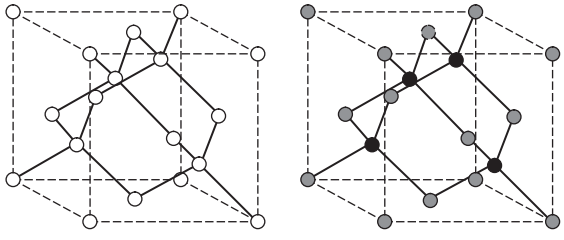

Figure 1.6 The diamond (left) and zinc-blende (right) lattices.

[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

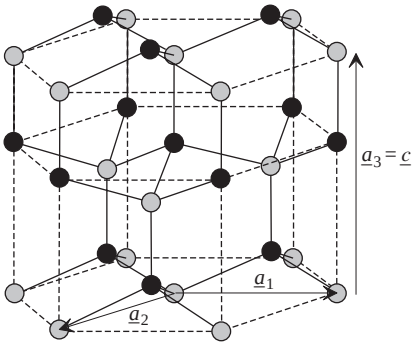

Figure 1.7 The hexagonal wurtzite cell. The c-axis corresponds to the direction of the $\underline{a}_3 = \underline{c}$ vector.

11.5 EPITAXIAL-GROWTH TECHNIQUES

In an epitaxial process, the substrate wafer acts as the seed crystal. Epitaxial processes are differentiated from the melt-growth processes described in previous sections in that the epitaxial layer can be grown at a temperature substantially below the melting point, typically 30–50% lower. The common techniques for epitaxial growth are chemical-vapor deposition (CVD) and molecular-beam epitaxy (MBE).

11.5.1 Chemical-Vapor Deposition

CVD, also known as vapor-phase epitaxy (VPE), is a process whereby an epitaxial layer is formed by a chemical reaction between gaseous compounds. CVD can be performed at atmospheric pressure (APCVD) or at low pressure (LPCVD).

Figure 19 shows three common susceptors for epitaxial growth. Note that the geometric shape of the susceptor provides the name for the reactor: horizontal, pancake, and barrel susceptors—all made from graphite blocks. Susceptors in the epitaxial reactors are analogous to the crucible in the crystal-growing furnaces. Not only do they mechanically support the wafer, but in induction-heated reactors they also serve as the source of thermal energy for the reaction. The mechanism of CVD involves a number of steps: (a) the reactants such as the gases and dopants are transported to the substrate region, (b) they are transferred to the substrate surface where they are adsorbed, (c) a chemical reaction occurs, catalyzed at the surface, followed by growth of the epitaxial layer, (d) the gaseous products are desorbed into the main gas stream, and (e) the reaction products are transported out of the reaction chamber.

CVD for Silicon

Four silicon sources have been used for VPE growth: silicon tetrachloride ($SiCl_4$), dichlorosilane (SiH_2Cl_2), trichlorosilane ($SiHCl_3$), and silane (SiH_4). Silicon tetrachloride has been the most studied and has the widest industrial use. The typical reaction temperature is 1200°C. Other silicon sources are used because of lower reaction temperatures. The substitution of a hydrogen atom for each chlorine atom from silicon tetrachloride permits about a 50°C reduction in the reaction temperature. The overall reaction of silicon tetrachloride that results in the growth of silicon layers is

$$SiCl_4\ (gas) + 2H_2\ (gas) \leftrightarrows Si\ (solid) + 4HCl\ (gas). \tag{22}$$

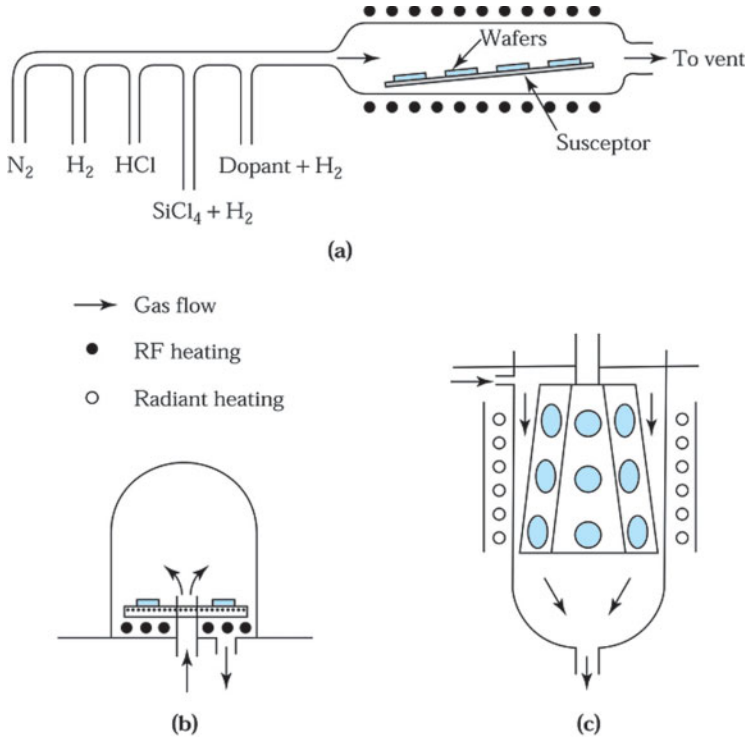

Fig. 19 Three common susceptors for chemical vapor deposition: (*a*) horizontal, (*b*) pancake, and (*c*) barrel susceptor.

An additional competing reaction is taking place along with that given in Eq. 22:

$$SiCl_4\ (gas) + Si\ (solid) \leftrightarrows 2SiCl_2\ (gas). \tag{23}$$

As a result, if the silicon tetrachloride concentration is too high, etching rather than growth of silicon will take place. Figure 20 shows the effect of the concentration of silicon tetrachloride in the gas on the reaction, where the *mole fraction* is defined as the ratio of the number of molecules of a given species to the total number of molecules.[14] Note that initially the growth rate increases linearly with increasing concentration of silicon tetrachloride. As the concentration of silicon tetrachloride is increased, a maximum growth rate is reached. Beyond that, the growth rate starts to decrease and eventually etching of the silicon will occur. Silicon is usually grown in the low-concentration region, as indicated in Fig. 20.

The reaction of Eq. 22 is reversible, that is, it can take place in either direction. If the carrier gas entering the reactor contains hydrochloric acid, removal or etching will take place. Actually, this etching operation is used for in-situ cleaning of the silicon wafer and coating on the reactor chamber wall prior to epitaxial growth.

The dopant is introduced at the same time as the silicon tetrachloride during epitaxial growth (Fig. 19*a*). Gaseous diborane (B_2H_6) is used as the *p*-type dopant, whereas phosphine (PH_3) and arsine (AsH_3) are used as *n*-type dopants. Gas mixtures are ordinarily used with hydrogen as the diluent to allow reasonable control of flow rates for the desired doping concentration. The dopant chemistry for arsine is illustrated in Fig. 21, which shows arsine being adsorbed on the surface, decomposing, and being incorporated into the growing layer. Figure 21 also shows the growth mechanisms at the surface, which are based on the surface adsorption of host atoms (silicon) as well as the dopant atom (e.g., arsenic) and the movement of these atoms toward the ledge sites.[15] To give these adsorbed atoms sufficient mobility for finding their proper positions within the crystal lattice, epitaxial growth needs relatively high temperatures.

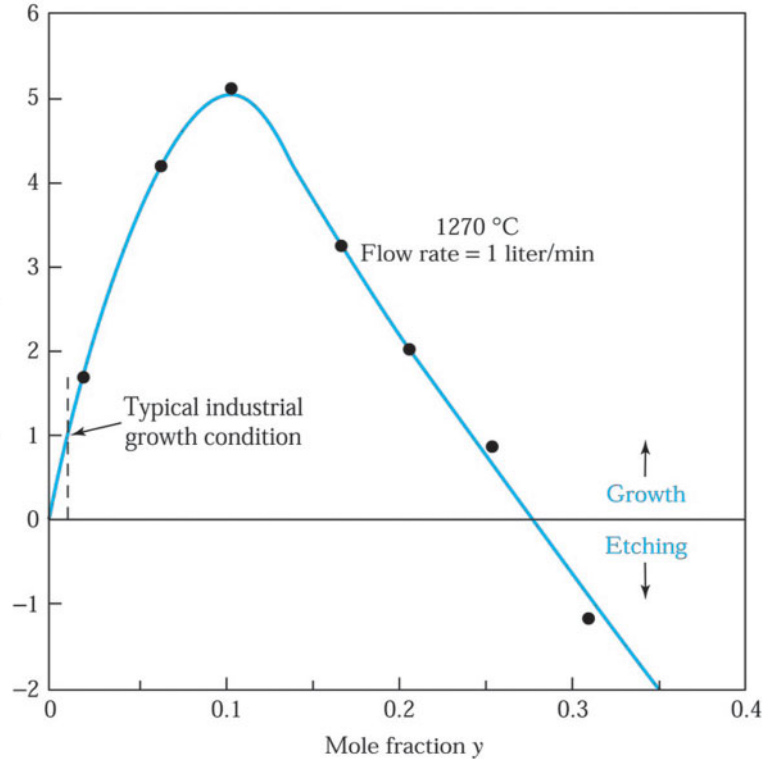

Fig. 20 **Effect of $SiCl_4$ concentration on silicon epitaxial growth.**[14]

CVD for GaAs

For gallium arsenide, the basic setup is similar to that shown in Fig. 19*a*. Since gallium arsenide decomposes into gallium and arsenic upon evaporation, its direct transport in the vapor phase is not possible. One approach is the use of As_4 for the arsenic component and gallium chloride ($GaCl_3$) for the gallium component. The overall reaction leading to epitaxial growth of gallium arsenide is

$$As_4 + 4GaCl_3 + 6H_2 \rightarrow 4GaAs + 12HCl. \tag{24}$$

The As_4 is generated by thermal decomposition of arsine (AsH_3):

$$4AsH_3 \rightarrow As_4 + 6H_2, \tag{24a}$$

and the gallium chloride is generated by the reaction

$$6HCl + 2Ga \rightarrow 2GaCl_3 + 3H_2. \tag{24b}$$

The reactants are introduced into a reactor with a carrier gas (e.g., H_2). Usually, the temperature for Eq. 24*b* is 800°C. The growth temperature of GaAs epilayer for Eq. 24 is below 750 °C. A two-zone reactor is needed for this exptaxial growth. Moreover, both reactions are exothermic: the epitaxy requires a reactor with hot walls. The reactions are near equilibrium condition, and process control is difficult. During the epitaxy, there must be sufficient arsenic overpressure to prevent thermal decomposition of the substrate and the growing layer.

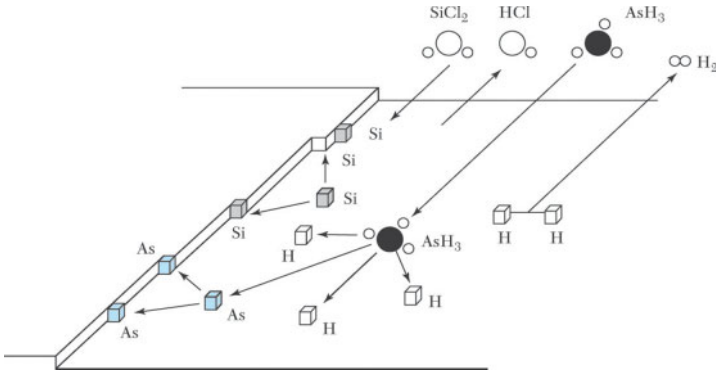

Fig. 21 **Schematic representation of arsenic doping and the growing processes.**[15]

Metalorganic CVD

Metalorganic CVD (MOCVD) is also a VPE process based on pyrolytic reactions. Unlike conventional CVD, MOCVD is distinguished by the chemical nature of the precursor. It is important for those elements that do not form stable hydrides or halides but that form stable metalorganic compounds with reasonable vapor pressure. MOCVD has been extensively applied in the heteroepitaxial growth of III-V and II-VI compounds.

To grow GaAs, we can use metalorganic compounds such as trimethylgallium $Ga(CH_3)_3$ for the gallium component and arsine AsH_3 for the arsenic component. Both chemicals can be transported in vapor form into the reactor. The overall reaction is

$$AsH_3 + Ga(CH_3)_3 \rightarrow GaAs + 3CH_4. \tag{25}$$

For Al-containing compounds, such as AlAs, we can use trimethylaluminum $Al(CH_3)_3$. During epitaxy, the GaAs is doped by introducing dopants in vapor form. Diethylzinc $Zn(C_2H_5)_2$ and diethylcadmium $Cd(C_2H_5)_2$ are typical *p*-type dopants and silane SiH_4 is an *n*-type dopant for III-V compounds. The hydrides of sulfur and selenium or tetramethyltin are also used for *n*-type dopants and chromyl chloride is used to dope chromium into GaAs to form semiinsulating layers. Since these compounds are highly poisonous and often spontaneously inflammable in air, rigorous safety precautions are necessary in the MOCVD process.

A schematic of an MOCVD reactor is shown[16] in Fig. 22. Due to the endothermic reaction, a reactor with a cold wall is used. Typically, the metalorganic compound is transported to the quartz reaction vessel by hydrogen carrier gas, where it is mixed with AsH_3 in the case of GaAs growth. The chemical reaction is induced by heating the gases to 600°–800°C above a substrate placed on a graphite susceptor using radio-frequency heating. A pyrolytic reaction forms the GaAs layer. The advantages of using metalorganics are that they are volatile at moderately low temperatures and there are no troublesome liquid Ga or In sources in the reactor. A single hot zone and nonequilibrium (one-way) reaction make the control of MOCVD easier.

11.5.2 Molecular-Beam Epitaxy

MBE[17] is an epitaxial process involving the reaction of one or more thermal beams of atoms or molecules with a crystalline surface under ultrahigh-vacuum conditions (~10^{-8} Pa).[§] MBE can achieve precise control in both chemical compositions and doping profiles. Single-crystal multilayer structures with dimensions on the order of atomic layers can be made using MBE. Thus, the MBE method enables the precise fabrication of semiconductor heterostructures having thin layers from a fraction of a micron down to a monolayer. In general, MBE growth rates are quite low, and for GaAs, a value of 1 μm/hr is typical.

[§] The international unit for pressure is the Pascal (Pa); 1 Pa =1 N/m^2. However, various other units have been used. The conversion of these units is: 1 atm = 760 mm Hg = 760 Torr = 1.013×10^5 Pa.

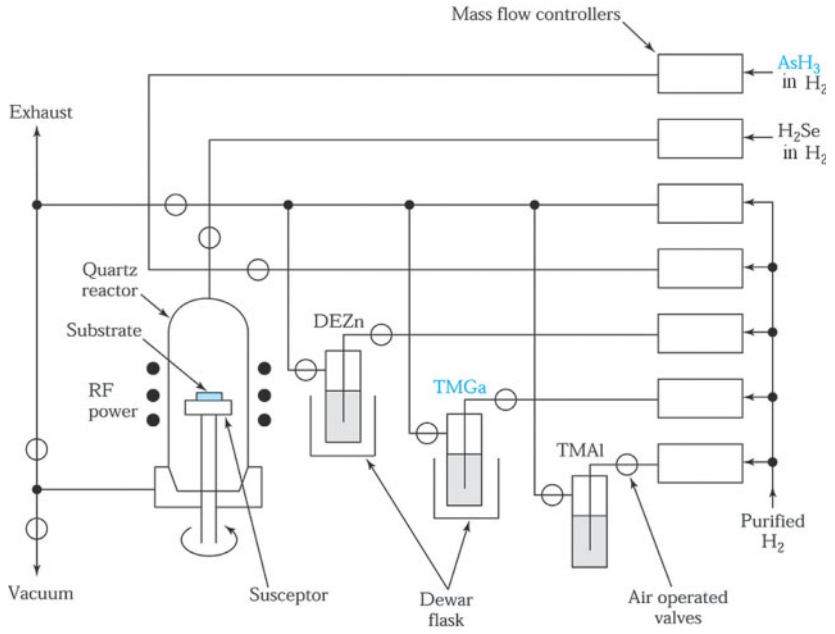

Fig. 22 Schematic diagram of a vertical atmospheric-pressure metalorganic chemical-vapor deposition (MOCVD) reactor.[16] DEZn is diethylozinc $Zn(C_2H_5)_2$, TMGa is trimethylgallium $Ga(CH_3)_3$, and TMAl is trimethylaluminum $Al(CH_3)_3$.

Figure 23 shows a schematic of an MBE system for gallium arsenide and related III-V compounds such as $Al_xGa_{1-x}As$. The system represents the ultimate in film deposition control, cleanliness, and in-situ chemical characterization capability. Separate effusion ovens made of pyrolytic boron nitride are used for Ga, As, and the dopants. All the effusion ovens are housed in an ultrahigh-vacuum chamber (~10^{-8} Pa). The temperature of each oven is adjusted to give the desired evaporation rate. The substrate holder rotates continuously to achieve uniform epitaxial layers (e.g., ±1% in doping variations and ±0.5% in thickness variations).

To grow GaAs, an overpressure of As is maintained, since the sticking coefficient of Ga to GaAs is unity, whereas that for As is zero, unless there is a previously deposited Ga layer. For a silicon MBE system, an electron gun is used to evaporate silicon. One or more effusion ovens are used for the dopants. Effusion ovens behave like small-area sources and exhibit a $\cos\theta$ emission, where θ is the angle between the direction of the source and the normal to the substrate surface.

MBE uses an evaporation method in a vacuum system. An important parameter for vacuum technology is the molecular impingement rate, that is, how many molecules impinge on a unit area of the substrate per unit time. The impingement rate ϕ is a function of the molecular weight, temperature, and pressure. The rate is derived in Appendix K and can be expressed as[18]

$$\phi = P(2\pi mkT)^{-1/2} \tag{26}$$

or

$$\boxed{\phi = 2.64\times10^{20}\left(\frac{P}{\sqrt{MT}}\right)\ \text{molecules / cm}^2\text{-s,}} \tag{26a}$$

where P is the pressure in Pa, m is the mass of a molecule in kg , k is Boltzmann's constant in J/K, T is the temperature in Kelvin, and M is the molecular weight. Therefore, at 300 K and 10^{-4} Pa pressure, the impingement rate is 2.7×10^{14} molecules/cm^2-s for oxygen (M = 32).

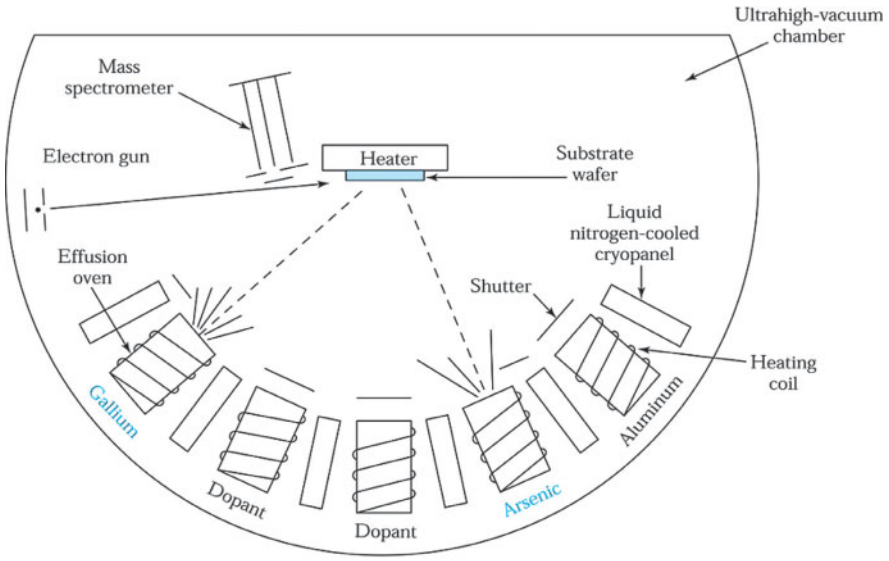

Fig. 23 Arrangement of the sources and substrate in a conventional molecular-beam epitaxy (MBE) system. (Courtesy of M. B. Panish, Bell Laboratories, Alcatel-Lucent Co.)

▶ EXAMPLE 3

At 300 K, the molecular diameter of oxygen is 3.64 Å, and the number of molecules per unit area N_s is 7.54×10^{14} cm^{-2}. Find the time required to form a monolayer of oxygen at pressures of 1, 10^{-4}, and 10^{-8} Pa.

SOLUTION The time required to form a monolayer (assuming 100% sticking) is obtained from the impingement rate:

$$t = \frac{N_s}{\phi} = \frac{N_s\sqrt{MT}}{2.64\times10^{20}P}.$$

Therefore,

$t = 2.8 \times 10^{-4} \approx 0.28$ ms at 1 Pa,
$= 2.8$ s at 10^{-4} pa,
$= 7.7$ hr at 10^{-8} pa.

To avoid contamination of the epitaxial layer, it is of paramount importance to maintain ultrahigh-vacuum conditions (~10^{-8} Pa) for the MBE process. ◀

During molecular motion, molecules will collide with other molecules. The average distance traversed by all the molecules between successive collisions with each other is defined as the mean free path. It can be derived from a simple collision theory. A molecule having a diameter d and a velocity υ will move a distance $\upsilon\delta t$ in the time δt. The molecule suffers a collision with another molecule if its center is anywhere within a distance d of the center of another molecule. Therefore, it sweeps out (without collision) a cylinder of diameter $2d$. The volume of the cylinder is

$$\delta V = \frac{\pi}{4}(2d)^2\upsilon\delta t = \pi d^2\upsilon\delta t. \tag{27}$$

Since there are n molecules/cm^3, the volume associated with one molecule is on the average $1/n$ cm^3. When the volume δV is equal to $1/n$, it must contain on the average one other molecule; thus, a collision would have occurred. Setting $\tau = \delta t$ as the average time between collision, we have

$$\frac{1}{n} = \pi d^2 \upsilon \tau, \tag{28}$$

and the mean free path λ is then

$$\lambda = \upsilon\tau = \frac{1}{\pi n d^2} = \frac{kT}{\pi P d^2}. \tag{29}$$

A more rigorous derivation gives

$$\lambda = \frac{kT}{\sqrt{2}\pi P d^2} \tag{30}$$

and

$$\boxed{\lambda = \frac{0.66}{P(\text{in Pa})} \text{ cm}} \tag{31}$$

for air molecules (equivalent molecular diameter of 3.7Å) at room temperature. Therefore, at a system pressure of 10^{-8} Pa, λ would be 660 km.

▶ EXAMPLE 4

Assume an effusion oven geometry of area A = 5 cm^2 and a distance L between the top of the oven and the gallium arsenide substrate of 10 cm. Calculate the MBE growth rate for the effusion oven filled with gallium arsenide at 900°C. The surface density of gallium atom is 6×10^{14} cm^{-2}, and the average thickness of a monolayer is 2.8 Å.

SOLUTION

On heating gallium arsenide, the volatile arsenic vaporizes first, leaving a gallium-rich solution. Therefore, only the pressures marked Ga-rich in Fig. 10 are of interest. The pressure at 900°C is 5.5×10^{-2} Pa for gallium and 1.1 Pa for arsenic (As_2). The arrival rate can be obtained from the impingement rate (Eq. 26*a*) by multiplying it by $A/\pi L^2$:

$$\text{Arrival rate} = 2.64 \times 10^{20} \left(\frac{P}{\sqrt{MT}}\right)\left(\frac{A}{\pi L^2}\right) \text{ molecules} / \text{cm}^2\text{-s}.$$

The molecular weight M is 69.72 for Ga and 74.92 × 2 for As_2. Substituting values of P, M, and T (1173 K) into the above equation gives

$$\begin{aligned} \text{Arrival rate} &= 8.2 \times 10^{14}/\text{cm}^2\text{-s} \qquad \text{for Ga,} \\ &= 1.1 \times 10^{16}/\text{cm}^2\text{-s} \qquad \text{for } As_2. \end{aligned}$$

The growth rate of gallium arsenide is found to be governed by the arrival rate of gallium. The growth rate is

$$\frac{8.2 \times 10^{14} \times 2.8}{6 \times 10^{14}} \approx 0.38 \text{ nm/s} = 23 \text{ nm/min}.$$

Note that the growth rate is relatively low compared with that of VPE. ◀

There are two ways to clean a surface in situ for MBE. High-temperature baking can decompose native oxide and remove other adsorbed species by evaporation or diffusion into the wafer. Another approach is to use a low-energy ion beam of an inert gas to sputter-clean the surface, followed by a low-temperature annealing to reorder the surface lattice structure.

MBE can use a wider variety of dopants than CVD and MOCVD, and the doping profile can be exactly controlled. However, the doping process is similar to the vapor-phase growth process: a flux of evaporated dopant atoms arrives at a favorable lattice site and is incorporated along the growing interface.

Fine control of the doping profile is achieved by adjusting the dopant flux relative to the flux of silicon atoms (for silicon epitaxial films) or gallium atoms (for gallium arsenide epitaxial films). It is also possible to dope the epitaxial film using a low-current, low-energy ion beam to implant the dopant (see Chapter 14).

The substrate temperatures for MBE range from 400°–900°C; and the growth rates range from 0.001 to 0.3 μm/min. Because of the low-temperature process and low growth rate, many unique doping profiles and alloy compositions not obtainable from conventional CVD can be produced in MBE. Many novel structures have been made using MBE, among them the *superlattice* and the heterojunction field-effect transistors discussed in Chapter 7.

A further development in MBE has replaced the group III elemental sources by metalorganic compounds such as trimethylgallium (TMG) or triethylgallium (TEG). This approach is called metalorganic molecular-beam epitaxy (MOMBE) and is also referred to as chemical-beam epitaxy (CBE). Although closely related to MOCVD, it is considered a special form of MBE. The metalorganics are sufficiently volatile that they can be admitted directly into the MBE growth chamber as a beam and are not decomposed before forming the beam. The dopants are generally elemental sources, typically Be for *p*-type and Si or Sn for *n*-type GaAs epitaxial layers.

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.

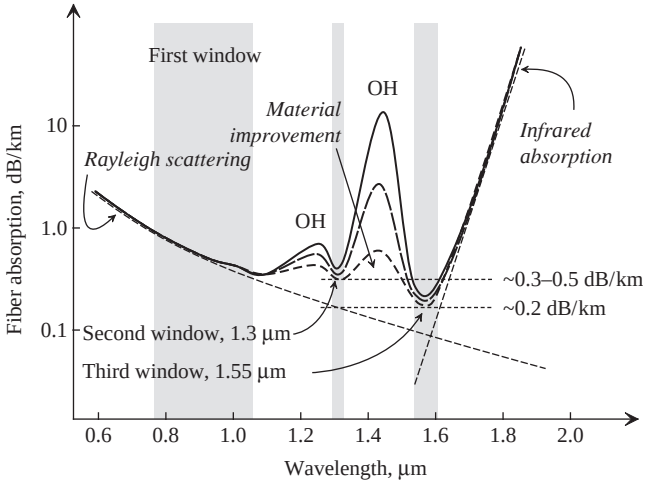

Figure 1.22 Absorption profile of a glass optical fiber.

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.

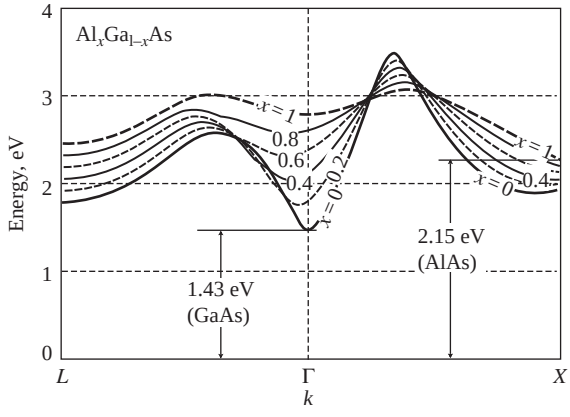

Figure 1.23 Evolution of the bandstructure of AlGaAs changing the Al content from 0 to 1.

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.

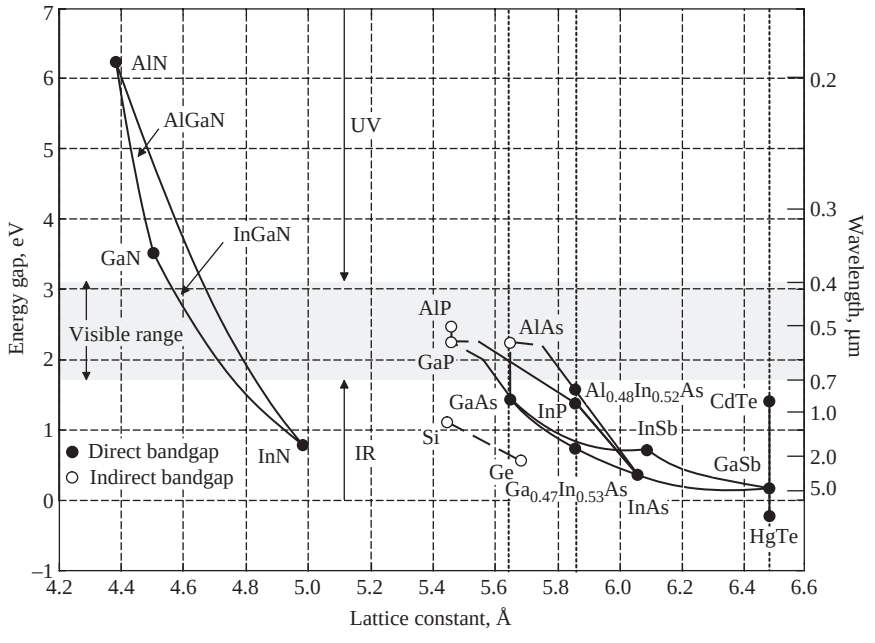

Figure 1.24 Some important alloys in the lattice constant–energy gap plane. In order of increasing gap and decreasing lattice constant: HgCdTe, InGaAsSb, InGaAsP, SiGe, AlGaAs, AlGaP, InGaN, AlGaN. For the widegap (wurtzite) nitrides (GaN, InN, AlN) the horizontal axis reports the equivalent cubic lattice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

1.19 Heterostructures

The ability to mix semiconductors of different chemical composition in a single crystal gives an important degree of freedom in device design. The combination of semiconductor materials of different stoichiometry in a single crystal is called a *heterostructure*.

Of foremost interest is the ability to change the energetic width of the forbidden gap – the bandgap energy, or band gap in short. This is shown in Figure 1.28 for some popular atomic and binary semiconductor materials. We note that changing the stoichiometry not only modifies the bandgap energy, but also the lattice constant, which can be understood as an average distance between the atoms in the crystal. This change in lattice constant is a major complication when designing devices, but we will exclude it for now by considering the material system (Al,Ga)As, where the lattice constant is almost independent of stoichiometry.

The ability to change bandgap through stoichiometry opens up several interesting design options.

- We may, for example, introduce a built-in electric field for one carrier species, but not for the other – schematically shown in Figure 1.29. This is a p-type semiconductor which has a smaller bandgap on the left-hand side ($E_{g,I}$) than on the right-hand side.

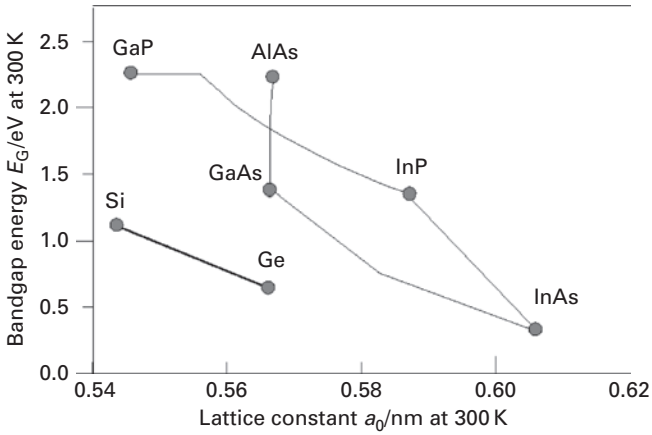

Fig. 1.28 Bandgap energy and lattice constant for several popular semiconductor materials.

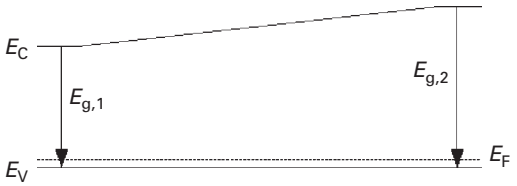

Fig. 1.29 Hypothetical band diagram of a p-doped graded heterostructure.

This could be done by starting with GaAs and gradually increasing the aluminium content while progressing to the right. Due to the p-type doping, the valence band stays approximately equidistant to the Fermi energy E_F (neglecting the change in the valence band density of states N_V), which is constant in thermodynamic equilibrium.[1] Due to the change in bandgap energy, the conduction band energy will change strongly and provide a built-in drift field for electrons, which in this schematic will be accelerated from right to left.

We will later use such a structure to accelerate the electrons in the base of a heterostructure bipolar transistor.

- Or we may abruptly change the bandgap by an abrupt modification of the stoichiometry (see Figure 1.30). In this case, we use an n-type semiconductor material, so the conduction bands remain approximately lined up horizontal (save for the stoichiometry-induced change in the conduction band density of states), but the change in bandgap results in a significant additional energy barrier for holes, which will keep them from moving from region 1 into region 2. This *carrier confinement* is at the heart of any semiconductor laser structure, and will also be used in heterostructure bipolar transistors. Note that a comparable energy barrier does not exist for electrons – an energy barrier has been created for one carrier species only, a feat possible only through the introduction of semiconductor heterostructures.

[1] The picture is a simplification because changing the stoichiometry also changes the density of states in the valence band, so there will be some variation in the E_F to E_V distance, which has been omitted.

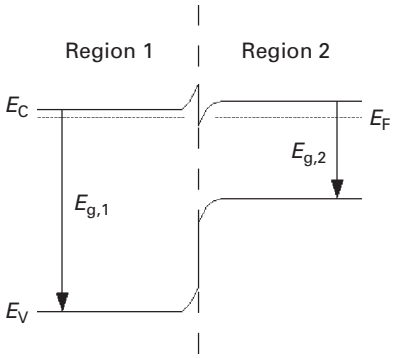

Fig. 1.30 Energy band diagram of an abrupt transition between two materials in a semiconductor heterostructure.

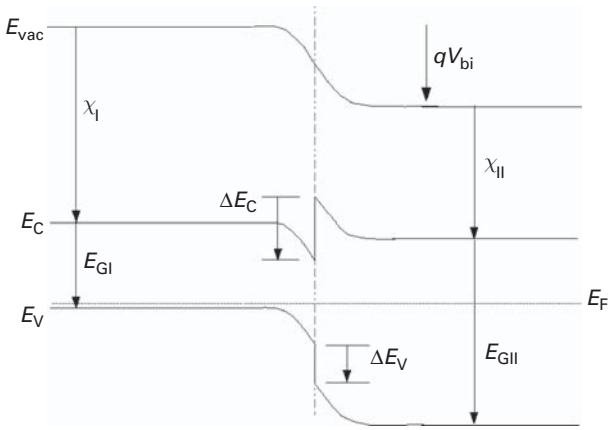

Fig. 1.31 Constructing heterostructure band diagrams using Anderson's rule.

1.19.1 Constructing heterostructure band diagrams

To efficiently use heterostructure band diagrams in the understanding of high-speed electronic and optoelectronic devices, we have to be able to construct their band diagrams. A simple procedure shall be described here.

It uses a modification of *Anderson's rule*, which has been a proven technique to draw the band diagram of the important SiO_2–Si interface. Anderson's rule postulates that the vacuum energy level is continuous and uses the electron affinities, i.e. the energetic spacing between the vacuum level and the conduction band, to calculate the band alignment in an abrupt heterostructure. The electron affinity χ is material-dependent.

When constructing the band diagrams, see Figure 1.31; we start out with the Fermi energy E_F, which in thermodynamic equilibrium is constant throughout the structure. We next draw conduction and valence bands *far away from the interface* between the two materials, which requires knowledge of the doping type and concentration, the bandgap energy and the densities of state as described in the introductory review section.

Next, we use knowledge of the electron affinities χ in the two materials to draw the vacuum level, again far away from the interface. Poisson's equation can be used to calculate the exact potential in the interfacial region, which is used to draw the continuous vacuum energy level. The material properties remain constant right up to the interface, so we can draw the conduction and valence bands to be perfectly parallel to the vacuum level.

The resulting conduction and valence bands show *discontinuities* at the interface between the two materials, in direct consequence of the the different electron affinities and the continuity of the vacuum energy level.

The built-in potential V_{bi} can be calculated as follows (refer again to Figure 1.31). We assume here for simplicity's sake that the Boltzmann equation can be used in lieu of the Fermi–Dirac function:

$$q \cdot V_{\mathrm{bi}} = \chi_{\mathrm{I}} - \chi_{\mathrm{II}} + E_{\mathrm{G,I}} + kT \cdot \ln\left(\frac{N_{\mathrm{V,I}} N_{\mathrm{C,II}}}{N_{\mathrm{A,I}} N_{\mathrm{D,II}}}\right). \tag{1.118}$$

Assume that you externally apply a voltage $-V_{\mathrm{bi}}$ to the structure, which compensates the built-in potential – the energy bands would then be constant within the two regions of the heterostructure (flatband condition); now you easily recognise that

$$\Delta E_{\mathrm{C}} = \chi_{\mathrm{I}} - \chi_{\mathrm{II}} \tag{1.119}$$

and

$$\Delta E_{\mathrm{V}} = E_{\mathrm{G,I}} - E_{\mathrm{G,II}} - \Delta E_{\mathrm{C}} = \Delta E_{\mathrm{G}} - \Delta E_{\mathrm{C}}. \tag{1.120}$$

However, in reality, the experimentally determined band discontinuities of heterostructures do not agree well with the values predicted using the electron affinities. This has to do with the different interface conditions between a free surface (used to determine the electron affinities) and a semiconductor heterostructure.

Anderson's rule can still be used, however. Note that only the difference of the electron affinities matters, not their absolute values. Hence, you can place the vacuum level at an arbitrary distance from the conduction band when drawing the band diagram, provided that the difference in the hypothetical electron affinities agrees with the *experimentally determined* ΔE_{C}.

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.

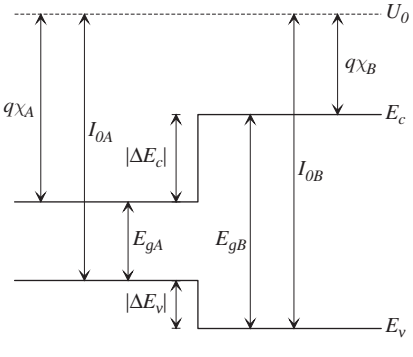

Figure 1.25 Heterostructure band alignment through application of the affinity rule to two materials having different bandstructures.

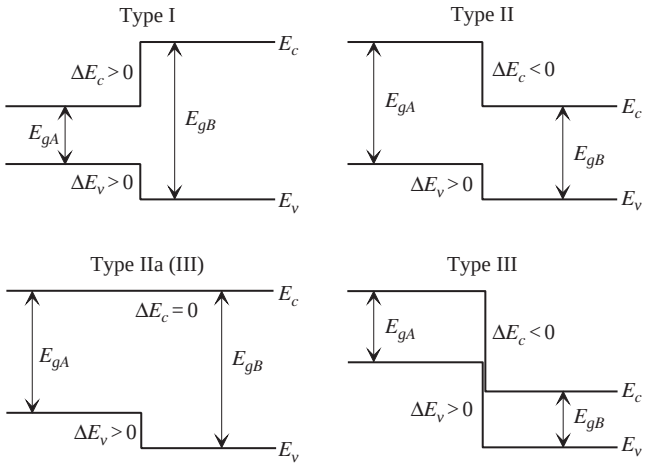

Figure 1.26 Classification of heterostructures according to band alignment; $\Delta E_j = E_{jB} - E_{jA}$.

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,

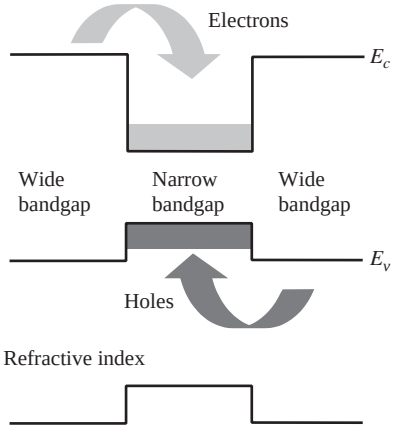

Figure 1.27 Example of carrier and light confinement in a $N\,i\,P$ double heterostructure in direct bias.

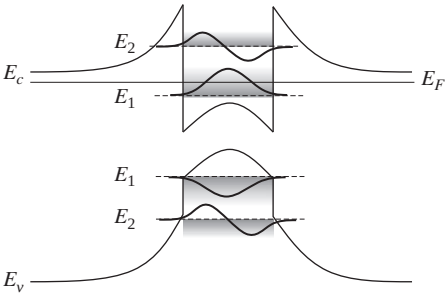

Figure 1.28 Qualitative example of quantization in a quantum well: in the conduction and valence bands, subbands arise with minimum energies corresponding to the levels $E_1, E_2, \ldots$; the total electron and hole wavefunction is given by the product of the 3D (bulk) wavefunction and of the envelope wavefunction shown. For simplicity, only the heavy hole subbands are shown.

no confinement) and obtain, by progressively restricting the degrees of freedom of the particle, the so-called *reduced dimensionality structures* corresponding to:

1. Confinement in one direction (x): particles are confined along x by a potential well but are free to move along y and z (**quantum well**).
2. Confinement in two directions (x and y): particles are confined along x and y, but they are free to move along z (**quantum wire**).
3. Confinement in three directions (x, y, z): particles are entirely confined and cannot move (**quantum dot**).

Summarizing:

Abrupt heterostructures can have three fundamental band alignments – refer to Figure 1.32 for a schematic representation:

(i) In a *type 1* heterojunction, the conduction band of the material with the lower bandgap is below the conduction band of the material with the higher bandgap, but the valence band of the lower-bandgap material is above the valence band of the higher-bandgap material. The smaller bandgap is hence fully within the larger bandgap.

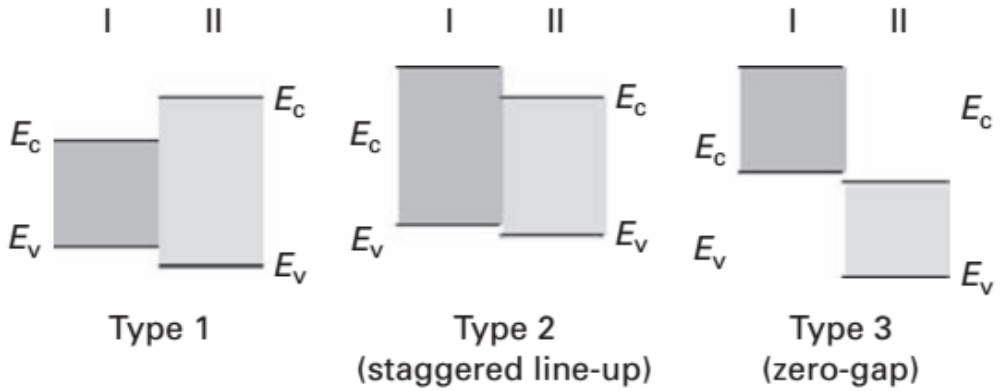

Fig. 1.32 Schematic representation of heterostructure line-ups.

(ii) In a *type 2* heterojunction, the conduction band and the valence band of one material are below their counterparts in the other material. This is sometimes also referred to as a *staggered line-up*.

(iii) Finally, in a *type 3* heterojunction, the valence band in one material is above the conduction band in the other. This is called a *zero-gap* configuration.

In current practical devices, the type 1 line-up is by far the most common.

Semiconductor Defects

Semiconductor devices have both unintended and intentional defects. Some unintentional defects are introduced due to either thermodynamic considerations or the presence of impurities during the crystal growth process. In general, defects in crystalline semiconductors can be characterized as i) point defects; ii) line defects; iii) planar defects and iv) volume defects. These defects are detrimental to the performance of electronic and optoelectronic devices and are to be avoided as much as possible.

Localized Defects

A localized defect affects the periodicity of the crystal only in one or a few unit cells. There are a variety of point defects, as shown in figure 1.13. Defects are present in any crystal.

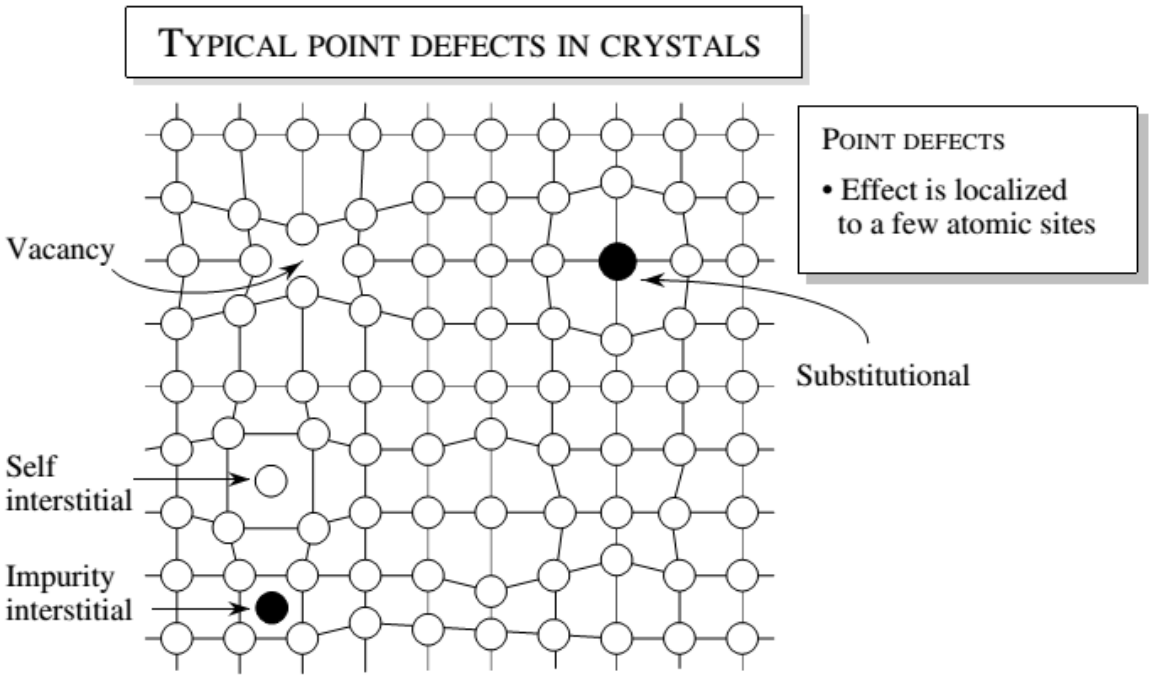

Figure 1.13: A schematic showing some important point defects in a crystal.

Dislocations

In contrast to point defects, line defects (called dislocations) involve a large number of atomic sites that can be connected by a line. Dislocations are produced if, for example, an extra half plane of atoms are inserted (or taken out) of the crystal as shown in figure 1.14. Such dislocations are called edge dislocations. Dislocations can also be created if there is a slip in the crystal so that part of the crystal bonds are broken and reconnected with atoms after the slip. In the nitride technology where alternate substrates are used, dislocation densities can be quite large.

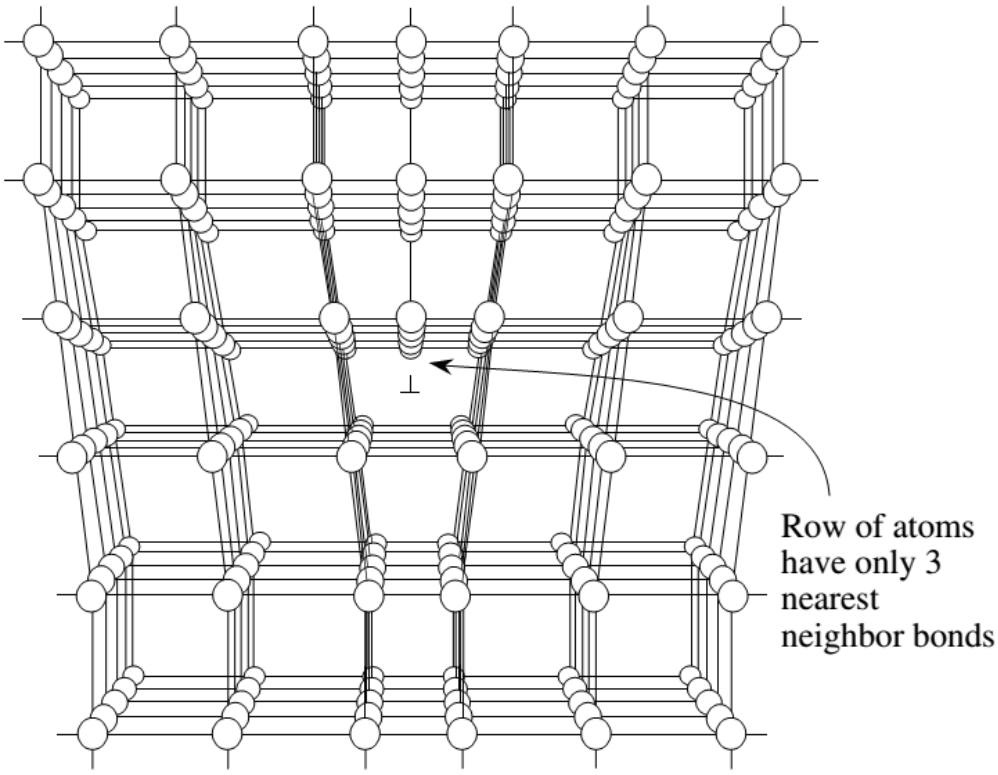

Figure 1.14: A schematic showing the presence of a dislocation. This line defect is produced by adding an extra half plane of atoms.

LATTICE MISMATCHED STRUCTURES

It is relatively easy to grow heterostructures where the overlayer lattice constant is the same or similar to that of the substrate. In such lattice matched epitaxy the interface quality can be very high with essentially negligible interface defects and atomically abrupt interface. However one often needs structures where there is lattice mismatch between the overlayer and the substrate. The motivation for lattice mismatched epitaxy is two fold:

i) Incorporation of built-in strain: When a lattice mismatched semiconductor is grown on a substrate and the thickness of the overlayer is very thin, the overlayer has a built-in strain. This built-in strain has important effects on the electronic and optoelectronic properties of the material and can be exploited for high performance devices. It can be exploited in nitride heterostructures to effectively dope structures. It can also be exploited in Si/SiGe systems.

ii) New effective substrate: High quality substrates are only available for Si, GaAs and InP (sapphire, SiC and quartz substrates are also available and used for some applications). Since most semiconductors are not lattice-matched to these substrates a solution that has emerged is to grow a thick overlayer on a mismatched substrate. If the conditions are right, dislocations are generated and eventually the overlayer forms its own substrate. This process allows a tremendous flexibility in semiconductor technology. Not only can it, in principle, resolve the substrate availability problem, it also allows the possibility of growing GaAs on Si, CdTe on GaAs, GaN on SiC etc. Thus different semiconductor technologies can be integrated on the same wafer.

In figure 1.15 we show a TEM image of an InP/InAs double-barrier resonant tunneling device (DBRT). The InP barriers are 5 nm wide, enclosing a 15 nm InAs quantum dot. The InP is coherently strained, with no dislocations created at the interfaces. The sharpness of the interfaces was determined to be 1-3 lattice spacings.

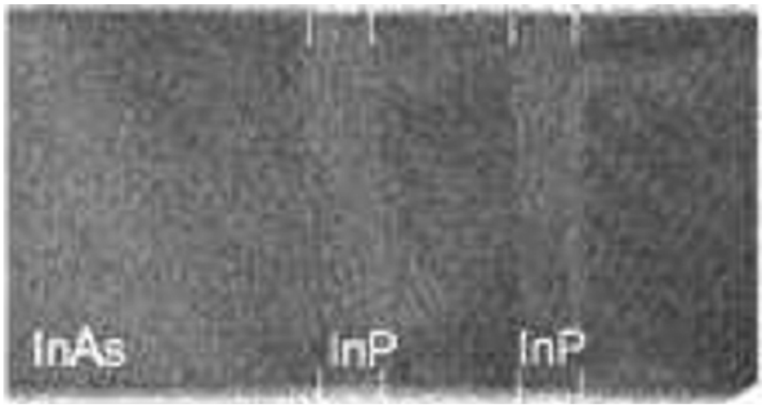

Figure 1.15: TEM image of an InP/InAs double-barrier resonant tunneling device (DBRT) consisting of 5 nm InP barriers surrounding a 15 nm InAs quantum dot. The InP is coherently strained, with no dislocations created at the interfaces. Image courtesy of M. Bjork, Lund University.

Coherent and Incoherent Structures

Consider situation shown schematically in figure 1.16 where an overlayer with lattice constant a_L is grown on a substrate with lattice constant a_S. The strain between the two materials is defined as

$$\epsilon = \frac{a_S - a_L}{a_L} \tag{1.3.1}$$

If the lattice constant of the overlayer is maintained to be a_L, it is easy to see that after every $1/\epsilon$ bonds between the overlayer and the substrate, either a bond is missing or an extra bond appears as shown in figure 1.16b. In fact, there would be a row of missing or extra bonds since we have a 2-dimensional plane. These defects are the dislocations discussed earlier.

An alternative to the incoherent case is shown in figure 1.16c. Here all the atoms at the interface of the substrate and the overlayer are properly bonded by adjusting the in-plane lattice constant of the overlayer to that of the substrate. This causes the overlayer to be under strain and the system has a certain amount of strain energy. This strain energy grows as the overlayer thickness increases. In the strained epitaxy, the choice between the state of the structure shown in figure 1.16b and the state shown in figure 1.16c is decided by free energy minimization considerations. The general observations can be summarized as follows:

For small lattice mismatch ($\epsilon < 0.03$), the overlayer initially grows in perfect registry with the substrate, as shown in figure 1.16c. However, as noted before, the strain energy will grow as the overlayer thickness increases. As a result, it will eventually be favorable for the overlayer

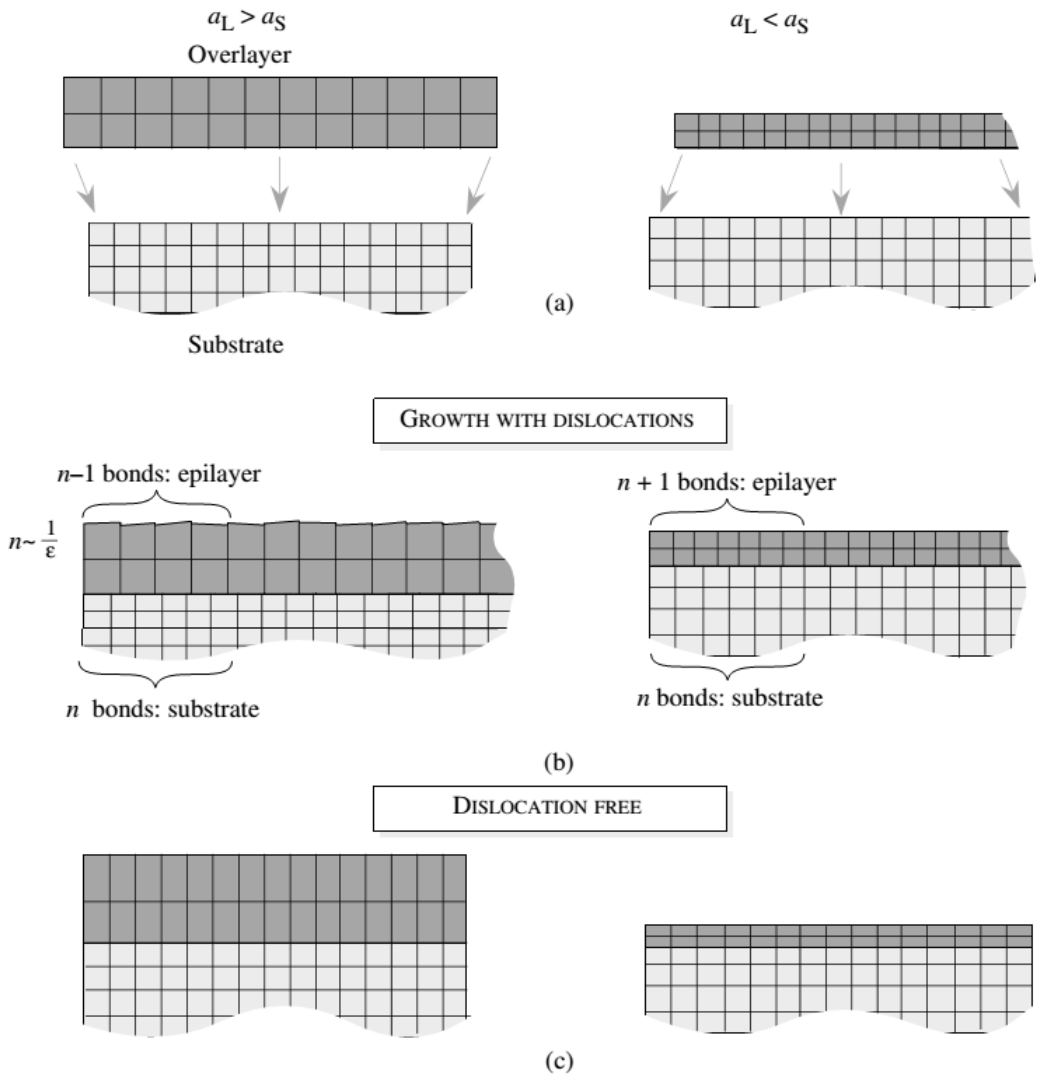

Figure 1.16: (a) An overlayer with one lattice constant is placed without distortion on a substrate with a different lattice constant. (b) Dislocations are generated at positions where the interface bonding is lost. (c) The case is shown where the overlayer is distorted so that no dislocation is free and coherent with the substrate.

to generate dislocations. In simplistic theories this occurs at an overlayer thickness called the critical thickness , d_c, which is approximately given by

$$d_c \cong \frac{a_S}{2|\epsilon|} \tag{1.3.2}$$

where a_S is the lattice constant of the substrate and ϵ the lattice mismatch. In reality, the point in growth where dislocations are generated is not so clear cut and depends upon growth conditions, surface conditions, dislocation kinetics, etc. However, one may use the criteria given by equation 1.3.2 for approximately characterizing two regions of overlayer thickness for a given lattice mismatch. Below critical thickness, the overlayer grows without dislocations and the film is under strain. Under ideal conditions above critical thickness, the film has a dislocation array, and after the dislocation arrays are generated, the overlayer grows without strain with its free lattice constant.

If the strain value is greater than 0.03 one can still have strained epitaxy but the growth occurs in the island mode where islands of the over-layer are formed. Such self-assembled islands are being used for quantum dot structures.

Epitaxy beyond the critical thickness is important to provide new effective substrates for new material growth. For these applications the key issues center around ensuring that the dislocations generated stay near the overlayer-substrate interface and do not propagate into the overlayer as shown in figure 1.17. A great deal of work has been done to study this problem. Often thin superlattices in which the individual layers have alternate signs of strain are grown to "trap" or "bend" the dislocations. It is also useful to build the strain up gradually.

In recent years, the GaN material system has seen much progress in electronic and optoelectronic applications. Since GaN substrates are still not readily available, it is typically grown on Al_2O_3 (sapphire) or SiC , neither of which are closely lattice matched to GaN. The resulting material is therefore highly dislocated. Many of the dislocations propagate upwards and are terminated at the surface. In figure 1.18a, we show a cross-sectional transmission electron microscope image of GaN grown on sapphire. The vertical lines propagating upwards from the highly defective interface are dislocations. Figure 1.18b is an atomic force microscope (AFM) image of the GaN surface. The black pits are dislocations that have propagated upwards. Also evident are the atomic steps that are typical of GaN surfaces. Such surface reconstructions were described in section 1.2.5. Note that these atomic steps are always terminated by a dislocation.

In figure 1.18c, we show an AFM image of the surface of dislocation-free GaN. In contrast to the dislocated material in figure 1.18b, there are no pits visible on the surface, and the surface step structure is smooth and continuous.

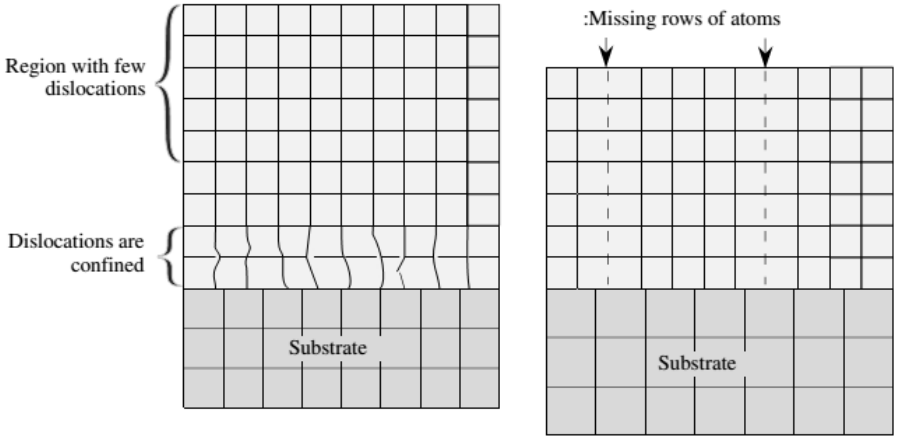

Figure 1.17: Strained epitaxy above critical thickness . The left hand side figure shows a desirable structure in which the dislocations are confined near the overlayer-substrate interface. On the right hand side, the dislocations are penetrating the overlayer.

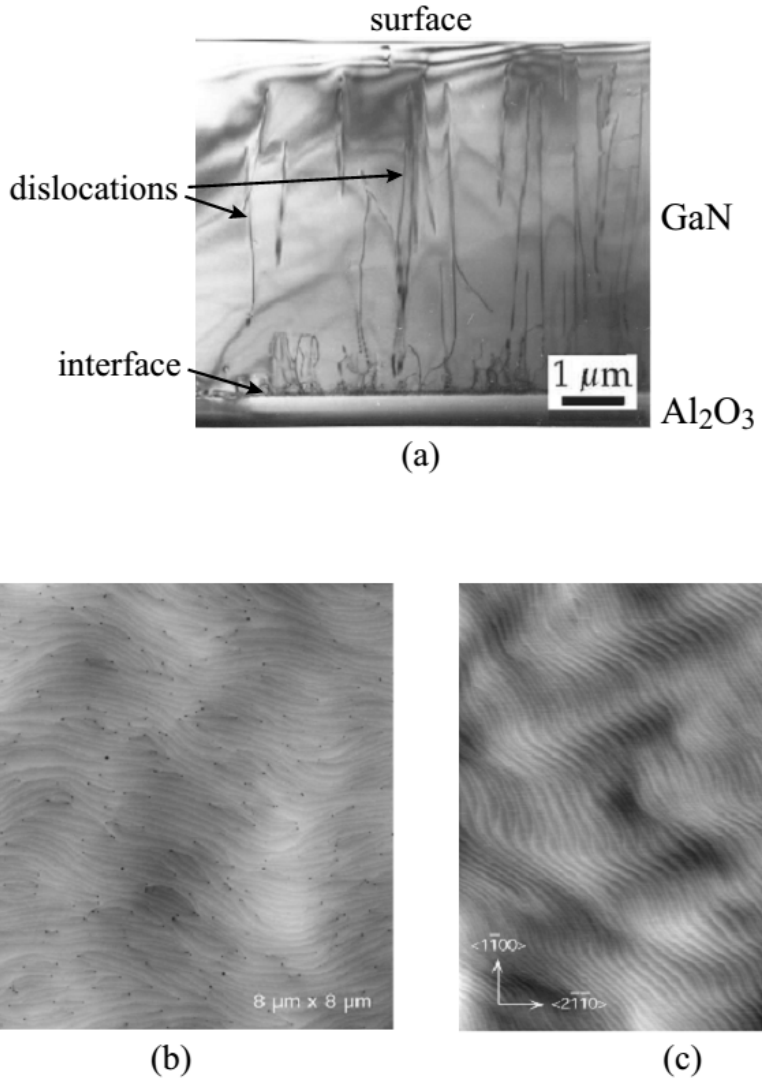

Figure 1.18: (a) Cross-sectional TEM image of GaN grown heteroepitaxially on sapphire, indicating the highly defective interface and the dislocations that propagate upwards. (b) AFM surface image of the dislocated GaN , showing the atomic step structure which is typical of GaN surfaces. The black dots are dislocations that have propagated upwards to the surface. (c) AFM surface image of non-dislocated GaN, exhibiting a smooth and continuous step structure. Images courtesy of P. Fini and H. Marchand of UCSB.

Summarizing:

As already observed in Figure 1.28, changing the stoichiometry will generally also modify the lattice constant. When combining materials with different lattice constants, the mismatch will create strong mechanical strain at the interface (experienced as tensile strain by the material with the smaller lattice constant and as compressive strain by other materials).

Please refer to Figure 1.33. We shall visualise in a schematic fashion the problem of lattice mismatch, taking the important Si/SiGe heterostructure as an example.

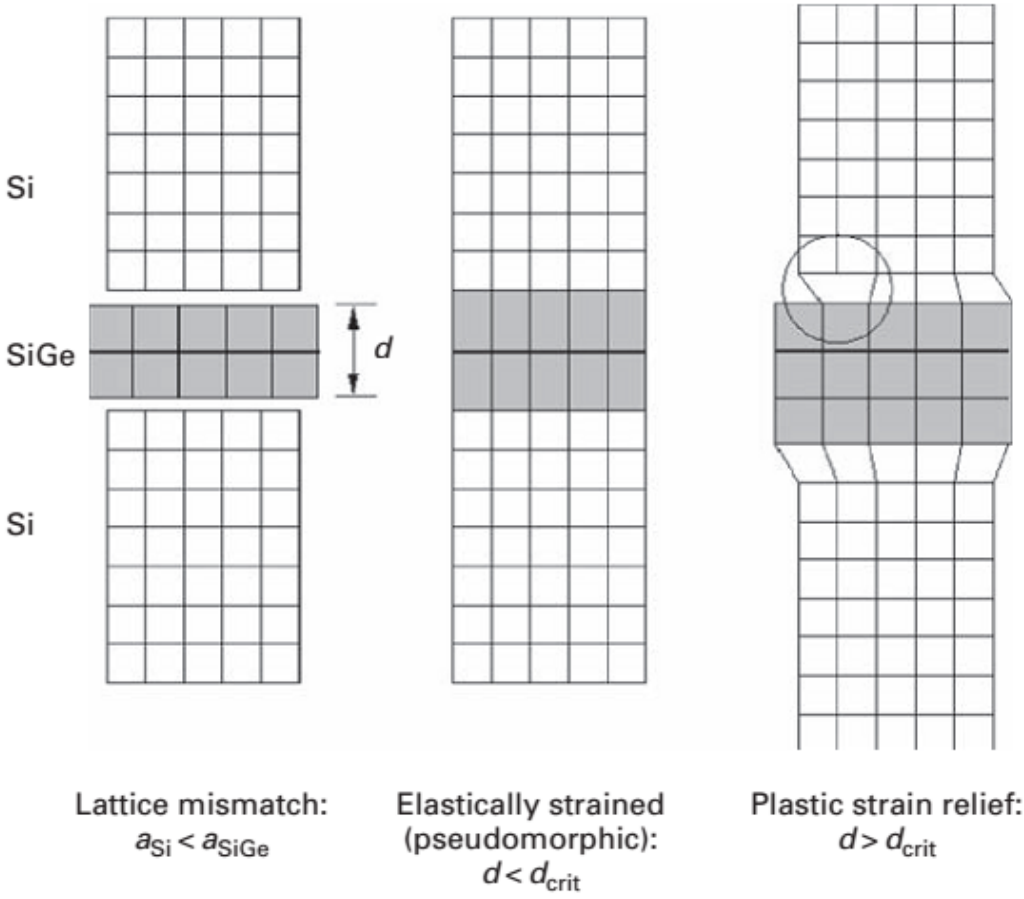

Fig. 1.33 Schematic representation of lattice mismatch in heterostructures.

In this example, a thin SiGe layer shall be sandwiched between two thick Si layers.[2] Provided that the thickness of the SiGe layer remains below a certain *critical thickness*, all of the lattice mismatch can be compensated for by *elastically* straining the thin layer (and a small fraction of the neighbouring layers). This case is also called the *pseudomorphic* case, a term which we will encounter later on, e.g. in the discussion of modern field effect transistor (FET) structures.

If, however, the strained layer thickness is significantly increased beyond the critical thickness, the mechanical forces at the interface become so large that they are able to break crystalline bonds – a crystal defect is created which will have detrimental effects in both optoelectronic and electronic structures and is hence to be avoided (but for a very few special cases, where this *plastic* strain relaxation is used deliberately in areas of the device devoid of mobile charge carriers).

Hence, the critical thickness is a parameter which must be carefully obeyed in heterostructure device design. It depends on both the lattice mismatch and the elements forming the heterostructure.

[2] 'Thin' here means that the whole layer can be deformed by the mismatch-generated strain, while 'thick' means that the largest part of the layer remains unstrained.

This introductory section on heterostructures concludes with what 2000 Nobel laureate Herbert Kroemer has called the *Central Design Principle* of semiconductor heterostructures:

> Heterostructures use energy gap variations in addition to electric fields as forces acting on holes and electrons to control their distribution and flow [7].

Our future treatment of high-speed electronic and optoelectronic devices will only exemplify this fundamental observation.

Buffer Layer Approaches

Buffer layer approaches involve the insertion of an epitaxial layer or layers in between the substrate and the device layer, solely for the purpose of reducing the dislocation density in the device layer. The buffer may be a single, uniform layer, a graded composition layer, or a superlattice or other multilayered structure, and all three types have been used with varying degrees of success.

7.2.1 Uniform Buffer Layers and Virtual Substrates

A uniform buffer layer has a constant composition throughout its thickness, and therefore the lattice mismatch with respect to the substrate is fixed at a constant value. It is usually reported that the threading dislocation density of such a buffer layer decreases with the reciprocal of its thickness. If the buffer layer is designed to be lattice-matched to the device layer, then this device layer may be grown on top of it without the introduction of new dislocations. In principle, then, the use of a sufficiently thick buffer layer will allow the growth of a device layer with a low dislocation density on a convenient lattice-mismatched substrate.

A thick, uniform buffer layer on a mismatched substrate is sometimes called a virtual substrate (VS). For example, a thick epitaxial layer of GaN on a sapphire substrate can serve as a virtual GaN substrate, even though conventional GaN substrates are not available in high quality at this time. If the GaN buffer layer is very thick, it will behave as a conventional GaN substrate in some, but not all, respects. It is expected that the virtual substrate will be relaxed at the growth temperature, and that it will have a low dislocation density. But, unless the thick buffer is exfoliated from its substrate, it will be constrained to mimic the thermal expansion of the substrate. All the same, virtual substrates open up possibilities for new materials, such as ternary or quaternary semiconductors, which cannot be readily manufactured in bulk.

The threading dislocation density at the surface of a uniform buffer layer decreases monotonically with its thickness. Usually, a reciprocal relationship is reported, and this is shown in the data of Figure 7.1. Moreover, the

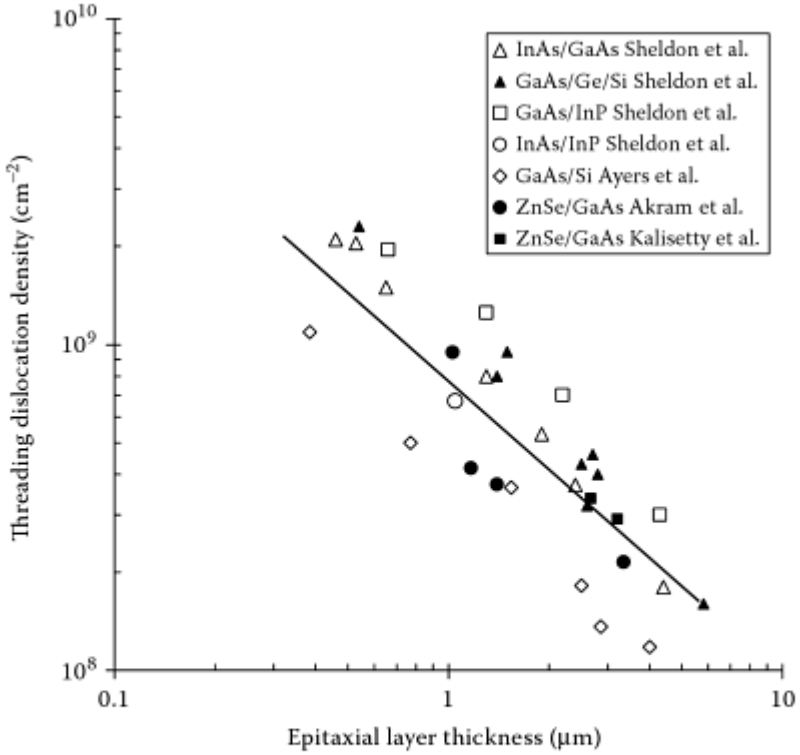

FIGURE 7.1
Threading dislocation densities in uniform buffer layers vs. the buffer layer thickness. The data are from Sheldon et al.,[70] Ayers et al.,[71] Akram et al.,[72] and Kalisetty et al.,[73] as indicated in the legend.

dislocation densities depend only weakly on the lattice mismatch. Therefore, layers of ZnSe/GaAs (001) have dislocation densities similar to those of InAs/GaAs (001) even though these systems differ in lattice mismatch by a factor of 1:30.