Heterostructure and Quantum Well Physics

William R. Frensley

May 15, 1998

[Ch. 1 of *Heterostructures and Quantum Devices*, W. R. Frensley and N. G. Einspruch editors, A volume of *VLSI Electronics: Microstructure Science*. (Academic Press, San Diego) Publication date: March 25, 1994]

Contents

I Introduction

Heterostructures are the building blocks of many of the most advanced semiconductor devices presently being developed and produced. They are essential elements of the highest-performance optical sources and detectors [1, 2], and are being employed increasingly in high-speed and high-frequency digital and analog devices [3, 4, 5]. The usefulness of heterostructures is that they offer precise control over the states and motions of charge carriers in semiconductors.

For the purposes of the present work, a *heterostructure* is defined as a semiconductor structure in which the chemical composition changes with position. The simplest heterostructure consists of a single *heterojunction*, which is an interface within a semiconductor crystal across which the chemical composition changes. Examples include junctions between GaSb and InAs semiconductors, junctions between GaAs and $Al_xGa_{1-x}As$ solid solutions, and junctions between Si and Ge_xSi_{1-x} alloys. Most devices and experimental samples contain more than one heterojunction, and are thus more properly described by the more general term *heterostructure*.

1 Atomic Structure of Heterojunctions

An ideal heterojunction consists of a semiconductor crystal (in the sense of a regular network of chemically bonded atoms) in which there exists a plane across which the identity of the atoms participating in the crystal changes abruptly. In practice, the ideal structure is approached quite closely in some systems. In high-quality $Al_xGa_{1-x}As$-GaAs heterojunctions it has been found that the interface is essentially atomically abrupt [6]. There is an entire spectrum of departures from the ideal structure, in the form of crystalline defects. The most obvious cause of such defects is mismatch between the lattices of the participating semiconductors. The lattice constants of GaAs and AlAs are nearly equal, so these materials fit together quite well. In contrast, the lattice constants of Si and Ge differ significnatly, so that over a large area of the heterojunction plane, not every Si atom will find a Ge atom to which to bond. This situation produces defects in the form of dislocations in one or the other of the participating semiconductors, and such dislocations usually affect the electrical characteristics of the system by creating localized states which trap charge carriers. If the density of such interfacial traps is sufficiently large, they will dominate the electrical properties of the interface. This is what usually happens at poorly controlled interfaces such as the grain boundaries in polycrystalline materials. The term *heterojunction* is usually

reserved for those interfaces in which traps play a negligible role.

From the above considerations one would logically conclude that closely matching the lattice constants of the participating semiconductors (good "lattice matching") is a necessary condition for the fabrication of high-quality heterojunctions. Indeed this was the generally held view for many years, but more recently high-quality heterojunctions have been demonstrated in "strained-layer" or pseudomorphic systems [7, 8]. The essential idea is that if one of the semiconductors forming a heterojunction is made into a sufficiently thin layer, the lattice mismatch is accommodated by a deformation (strain) in the thin layer. With this approach it has proved possible to make high-quality heterojunctions between Si and Ge_xSi_{1-x} alloys [4].

Heterostructures are generally fabricated by an epitaxial growth process. Most of the established epitaxial techniques have been applied to the growth of heterostructures. These include Molecular Beam Epitaxy (MBE) [6] and Metalorganic Chemical Vapor Deposition (MOCVD) [9]. Liquid Phase Epitaxy (LPE) is an older heterostructure technology, which has largely been supplanted by by MBE and MOCVD because it does not permit as precise control of the fabricated structure.

The examples of heterojunctions cited so far involve chemically similar materials, in the sense that both constituents contain elements from the same columns of the periodic table. It is possible to grow heterojunctions between chemically dissimilar semiconductors (those whose constituents come from different columns of the periodic table), such as Ge-GaAs and GaAs-ZnSe, and such junctions were widely studied early in the development of heterostructure technology [10]. There are, however, a number of problems with such junctions. Based upon simple models of the electronic structure of such junctions, one would expect a high density of localized interface states due to the under- or over-satisfied chemical bonds across such a junction [11, 12] . More significantly, perhaps, the constituents of each semiconductor act as dopants when incorporated into the other material. Thus any interdiffusion across the junction produces electrical effects which are difficult to control. For these reasons, most recent work has focused upon chemically matched systems.

If a heterojunction is made between two materials for which there exists a continuum of solid solutions, such as between GaAs and AlAs (as $Al_xGa_{1-x}As$ exists for all x such that $0 \leq x \leq 1$), the chemical transition need not occur abruptly. Instead, the heterojunction may be "graded" over some specified distance. That is, the composition parameter x might be some continuous function of the position. Such heterojunctions have desirable properties

for some applications.

II Electronic Structure of Semiconductors

1 Energy Bands

Heterostructures are able to improve the performance of semiconductor devices because they permit the device designer to locally modify the energy-band structure of the semiconductor and so control the motion of the charge carriers. In order to understand how such local modification of band structure can affect this motion, one needs to understand the energy bands of bulk semiconductors [13].

If a number of atoms of silicon, for example, are brought together to form a crystal, the discrete energy levels of the free atoms broaden into energy bands in the crystal. The reason for this is that the electrons are free to move from one atom to another, and thus they can have different amounts of kinetic energy, depending upon their motion. Each of the quantum states of the free atom gives rise to one energy band. The bonding combinations of states that were occupied by the valence electrons in the atom become the valence bands of the crystal. The anti-bonding combinations of these states become the conduction bands. The form of the wavefunctions of band electrons is specified by the Bloch theorem to be of the form $\psi_{n,\mathbf{k}}(\mathbf{x}) = u_{\mathbf{k}}(\mathbf{x})e^{-\mathbf{k}\cdot\mathbf{x}}$, where n labels the energy band, $\mathbf{k}$ is the wavevector of the state, and $u_{\mathbf{k}}(\mathbf{x})$ is a periodic function on the crystal lattice. Each such state has a unique energy $E_n(\mathbf{k})$, and a plot of this energy as a function of $\mathbf{k}$ represents the energy band structure. For most purposes we can confine the values of $\mathbf{k}$ to lie within a solid figure called the Brillouin zone. Perspective plots of the energy band structures derived from an empirical pseudopotential model [14] for Si and GaAs are plotted in Figures 1 and 2, respectively.

The dynamics of electrons in energy bands are described by two theorems [13]. The velocity of an electron with wavevector $\mathbf{k}$ is given by the group velocity:

$$\mathbf{v} = \nabla_{\mathbf{k}} E(\mathbf{k})/\hbar. \tag{1}$$

If a constant force $\mathbf{F}$ is applied to an electron, its wavevector will change according to

$$\frac{d\mathbf{k}}{dt} = \frac{\mathbf{F}}{\hbar}. \tag{2}$$

If the band structure is perfectly parabolic, $E \propto k^2$, these reduce to the ordinary Newtonian expressions. However, as shown in Figs. 1 and 2, there are large regions in the band structures of ordinary semiconductors were they are not parabolic.

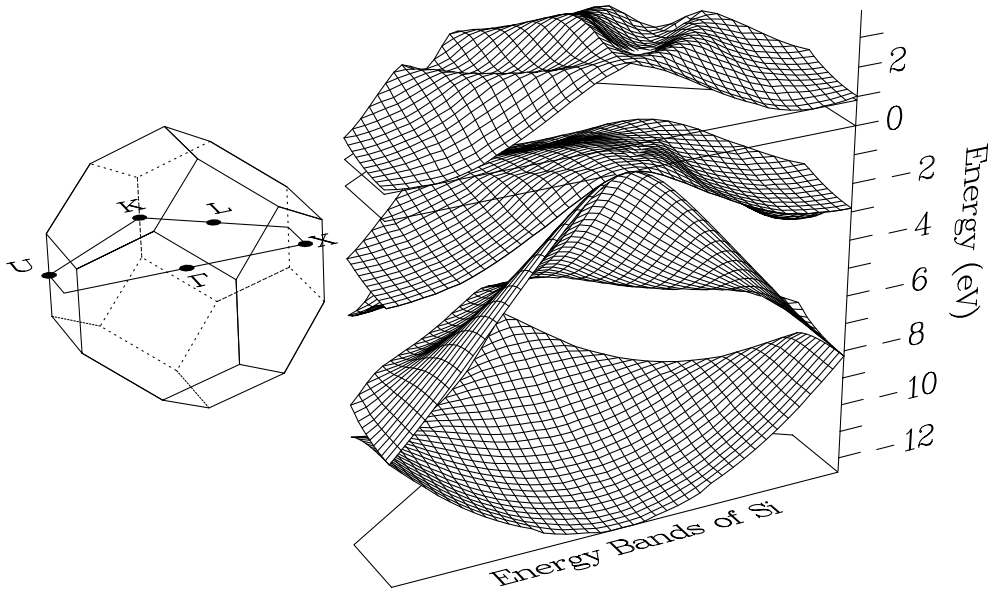

Figure 1: Perspective plot of the energy band structure of silicon. The figure to the left shows the Brillouin zone, and the two-dimensional section over which the energy bands are displayed. The energy bands are plotted to the right. The four surfaces lying below 0 eV are the valence bands, and the upper surface is the lowest conduction band. The maximum valence band energy occurs at $\mathbf{k} = 0$, which on this figure is the center of the front boundary of the Brillouin-zone section. The minimum conduction-band energy occurs along the front boundary of the section, near the left and right ends. Thus, Si has an indirect-gap band structure.

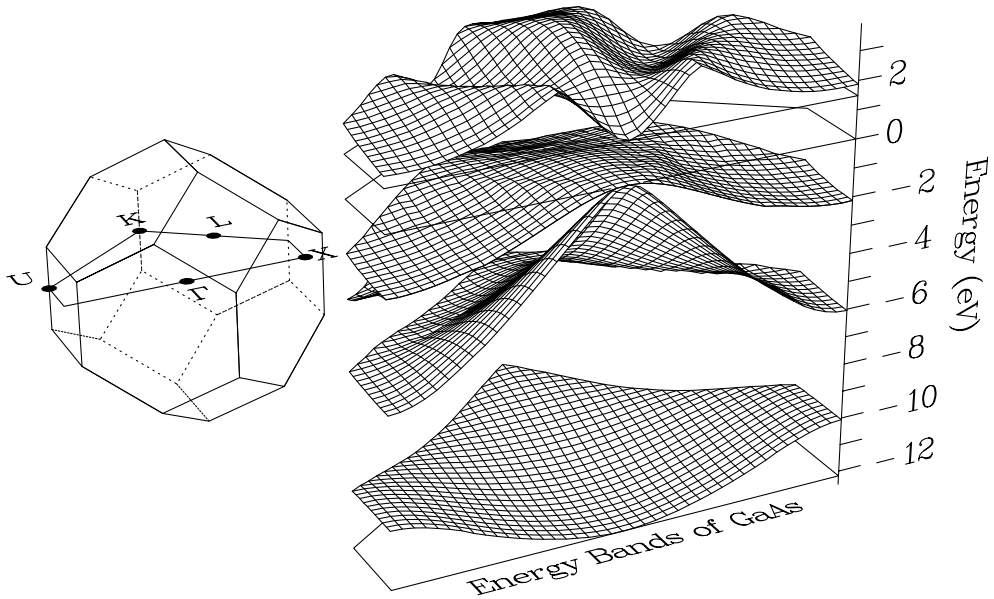

Figure 2: Perspective plot of the energy band structure of gallium arsenide. The conventions of the figure are the same as those of Fig. 1. The conduction-band minimum of GaAs occurs at $\mathbf{k} = 0$, and thus GaAs has a direct-gap band structure.

2 Effective Mass Theory

Energy-band theory is strictly applicable only to perfectly periodic crystals. This means, in particular, that it does not apply when macroscopic electric fields are present. Devices are not generally useful unless they contain such fields, so we need a formulation which can include them along with the crystal potential which produces the band structure. Such a formulation is provided by the effective-mass theorem [15, 16, 17]. This theorem provides a decomposition of the wavefunction into an atomic-scale part and a more slowly varying envelope function, and supplies a Schrödinger equation for the envelope function:

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2}\frac{\partial}{\partial x}\frac{1}{m^*}\frac{\partial}{\partial x}\Psi + [E_n - qV(x)]\Psi, \tag{3}$$

where Ψ is the envelope function, m^* is the effective mass, E_n is the energy at the edge of the nth band, and V is the electrostatic potential. The effects of the band structure are incorporated in the material-dependent parameters E_n and m^*. The standard picture of freely moving electrons and holes with material-dependent masses follows from the effective-mass theorem via the quantum-mechanical correspondence principle.

III Heterojunction Band Alignment

The central feature of a heterojunction is that the bandgaps of the participating semiconductors are usually different. Thus, the energy of the carriers at at least one of the band edges must change as those carriers pass through the heterojunction. Most often, there will be discontinuities in both the conduction and valence band. These discontinuities are the origin of most of the useful properties of heterojunctions.

As with all semiconductor devices, the key to understanding the behavior of heterojunctions is the energy-band profile which graphs the energy of the conduction and valence band edges versus position. The position-dependent band-edge energies are just the total potential appearing in (3), and we will use the symbols $U_C(x)$ and $U_V(x)$ to denote these quantities for the conduction and valence bands, respectively. Thus,

$$U_{V,C}(x) = E_{V,C}(x) - qV(x). \tag{4}$$

In a heterojunction, the dependence of U_C and U_V upon x are due to the combined effects of the electrostatic potential $V(x)$ and the energy-band discontinuities or shifts due to the heterostructure. In the earlier literature on heterojunctions, this latter effect is usually

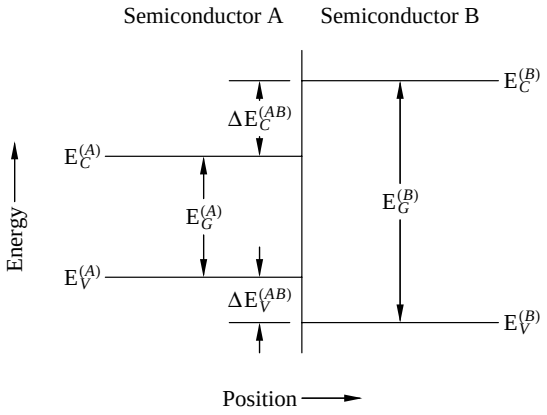

Figure 3: Definition of the quantities required to describe the band alignment of a heterojunction.

described in terms of the electron affinity χ [18, 10]. However, the electron affinity model is not a very accurate description of heterojunctions [19], so we will simply view the band-edge energies $E_{V,C}$ as fundamental properties of the semiconductors participating in the heterostructure. Thus, in a heterostructure, $E_{V,C}$ appears in the effective-mass Schrödinger equation (3) as a function of position. [The effective mass m^* is also a function of position, but the Hermitian form of (3) accounts for its variation.] The question of what is the appropriate reference energy for $E_{V,C}$ to permit a comparison of different semiconductors is the key question in the theory of the heterojunction band alignment. To begin our investigation of the band alignment, let us assume that the structure has been so designed that each semiconductor is precisely charge-neutral, and thus V will be constant and may be neglected. In such circumstances, we may focus upon the behavior of E_C and E_V in the vicinity of the heterojunction.

It has been found experimentally that there is no *a priori* relation between the band-edge energies of the two semiconductors forming a heterojunction, despite theoretical proposals of universal band alignments by Adams and Nussbaum [20] and by von Roos [21]. (These proposal were critiqued by Kroemer [22].) We therefore need a general scheme within which heterojunction band alignments may be described. The quantities used to describe the band alignment are defined in Fig. 3. The one quantity which is known with great certainty is the total bandgap discontinuity,

$$\Delta E_G = E_G^{(B)} - E_G^{(A)}, \tag{5}$$

where $E_G^{(A)}$ and $E_G^{(B)}$ are the energy gaps of materials A and B, respectively. The total

discontinuity is divided between the valence and conduction band discontinuities, defined by

$$\Delta E_V^{(AB)} = E_V^{(A)} - E_V^{(B)}, \tag{6a}$$

$$\Delta E_C^{(AB)} = E_C^{(B)} - E_C^{(A)}. \tag{6b}$$

Clearly, the individual discontinuities must add up to the total discontinuity,

$$\Delta E_G = \Delta E_V + \Delta E_C. \tag{7}$$

How the discontinuities are distributed between the valence and conduction bands is the major question to be answered by theory and experiment.

To illustrate the diversity of band alignments available, Figure 4 illustrates the best estimate of the band alignment for seven lattice-matched heterojunctions between group III-V semiconductors. Shown are the band alignments of (a) GaAs-$Al_xGa_{1-x}As$ in the direct-gap range [23], (b) $In_{0.53}Ga_{0.47}As$-InP [24], (c) $In_{0.53}Ga_{0.47}As$-$In_{0.52}Al_{0.48}As$ [24], (d) InP-$In_{0.52}Al_{0.48}As$ [25], (e) InAs-GaSb [26], (f) GaSb-AlSb [27], and (g) InAs-AlSb [28]. The topology of the band alignments are classified according to the relative ordering of the band-edge energies [29]. The most common (and generally considered to be the "normal') alignment is the *straddling* configuration illustrated in Figure 4 (a). The bandgaps need not entirely overlap, however. The conduction band of the smaller-gap material might lie above that of the larger-gap material, or its valence band might lie below that of the larger-gap material. Such a band alignment is called *staggered*, and is known to occur in the $In_xGa_{1-x}As$-$GaAs_{1-y}Sb_y$ system [26], as well as those of Figure 4 (d) and (g). The staggering might become so extreme that the bandgaps cease to overlap. This situation is known as a *broken gap*, and such a band alignment is observed in the GaSb-InAs system, Fig. 4 (e). Another nomenclature is occasionally employed, usually in describing superlattices, which are periodic heterostructures. If the extrema of both the conduction and valence bands lie in the same layers, the superlattice is referred to as "Type I," whereas if the band extrema are found in different layers the superlattice is "Type II." Aside from being rather uninformative, this notation makes no distinction between the staggered and broken-gap cases, and the more complete nomenclature described above should be preferred.

1 Theories of the Band Alignment

The problem of theoretically predicting heterojunction band alignments has attracted a good deal of attention in recent years. The electron-affinity model proposed by Anderson

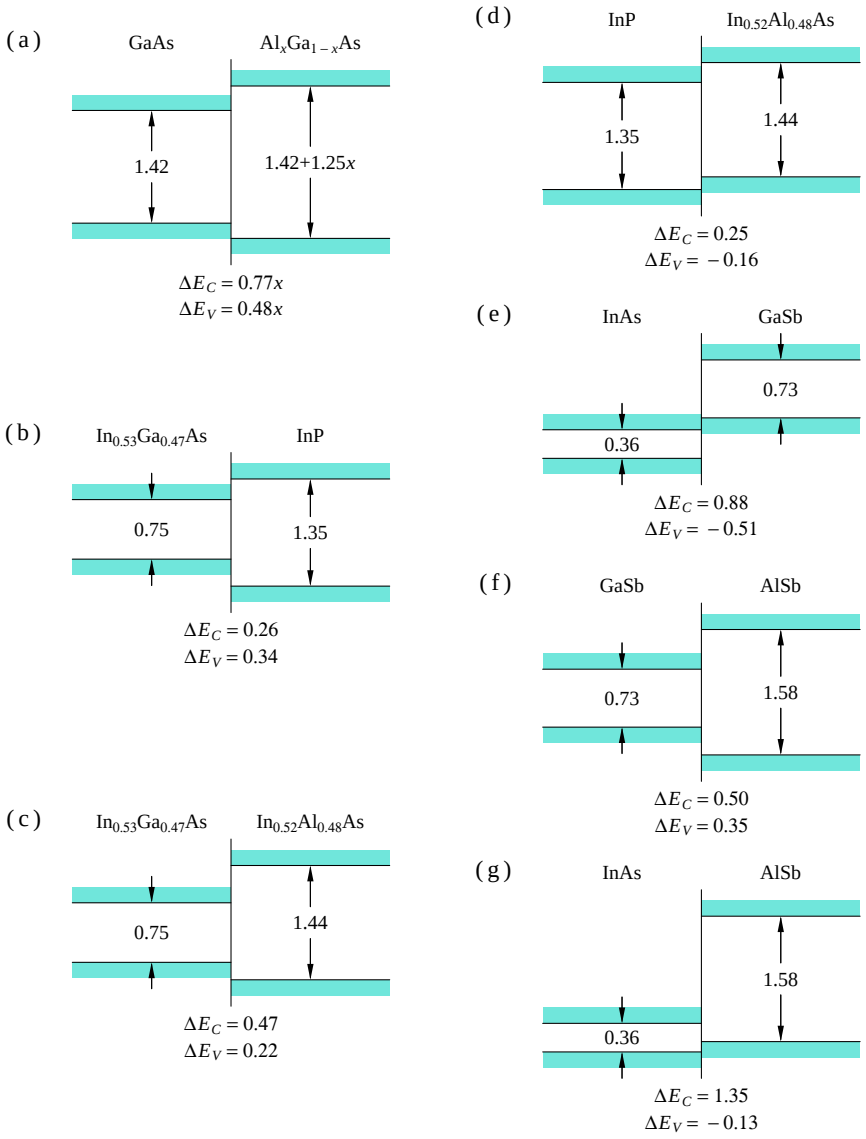

Figure 4: Experimentally determined band alignments for seven III-V heterojunctions, from a tabulation by Yu and co-workers. Energies are indicated in electron Volts. Cases (a), (b), (c), and (f) illustrate straddling alignments. Cases (d) and (g) illustrate staggered alignments, and case (e) illustrates a broken-gap alignment.

[18] was generally accepted until about 1976. The first attempts to predict band lineups for a variety of heterojunctions based upon microscopic models were those of Frensley and Kroemer [30, 31] and Harrison [32]. Since then, a large number of different approaches have been proposed and investigated. The interested reader should consult the reviews by Kroemer [29, 19], Tersoff [33], and Yu, McCaldin, and McGill [23].

Most theories of the band alignment conceptually divide the problem into two parts: the determination of the band-edge energies in the bulk with respect to some reference energy, and the determination of the difference (if any) between the reference energies across the heterojunction. An important question from both the theoretical and experimental point of view, is whether it is possible to define a universal scale for band energies which would always give the correct heterojunction band alignment. If this were the case, E_C and E_V would only depend upon the local chemical composition, and not upon the other material participating in the heterojunction under consideration. A useful concept by which this idea may be experimentally tested is "transitivity." Transitivity applies if one may predict the band alignment of a junction AC knowing the band alignments of junctions AB and BC, by

$$\Delta E_{V,C}^{(AC)} = \Delta E_{V,C}^{(AB)} + \Delta E_{V,C}^{(BC)}. \tag{8}$$

Most of the simpler theories of heterojunction band alignment possess transitivity, and it appears to be verified to within experimental uncertainties in lattice-matched heterostructure systems [28, 24].

If transitivity holds within a given set of materials, then there must exist a universal energy scale for semiconductor energy bands, at least for that set of materials. It makes absolutely no difference where the origin of this scale is chosen. It is often convenient, since the band discontinuities are the experimentally measured quantities, to choose a given band edge of a major material, such as the valence-band edge of GaAs, as the reference energy.

2 Measurement of the Band Alignment

There are a number of ways to measure the band alignment of a heterostructure, all of which are indirect (see chapters by several authors in Capasso and Margaritondo [34]). The reason for this situation will be discussed below. The result is that there remain significant uncertainties in the band alignments of many heterojunction systems. A comprehensive review of these issues has been prepared by Yu, McCaldin, and McGill [23], but the reader is cautioned to continue to consult the scientific literature on this subject, as further modifications to "known" band alignments are likely in the future.

One issue which arises in cases such as GaAs-$Al_xGa_{1-x}As$, where it is technologically convenient to make junctions involving any of a range of compositions x, is the question of how the band alignment changes with the solid-solution composition. It is often convenient to assume that the band energies vary linearly with composition, and then the band discontinuities may be expressed as fractions of the total band discontinuity ΔE_G. However, it is well known that the band gaps of such solid solutions often display significant nonlinearities as a function of composition [35], so a simple linear interpolation is rather suspect. The GaAs-$Al_xGa_{1-x}As$ band lineup has been studied over a wide range of compositions by Batey and Wright [36], who found that the valence-band discontinuity ΔE_V varied linearly with composition.

The difficulties involved in determining the band alignment at heterojunctions is vividly illustrated by the history of measurements of the GaAs-$Al_xGa_{1-x}As$ junction, which is certainly the most intensively studied system over the past two decades. The development of high-quality heterostructures grown by molecular beam epitaxy permitted the fabrication of "quantum wells" in which the electron and hole energies were size-quantized by the heterojunction energy barriers, and these quantum states were measured spectroscopically by Dingle, Wiegmann, and Henry in 1974 [37]. Fitting the observed spectra to a simple square-well model suggested that most of the discontinuity occurred in the conduction band, with $\Delta E_C = 0.85\Delta E_G$ [38], and this value was widely accepted until 1984. At that time, similar measurements were made by Miller, Kleinman, and Gossard [39] on quantum wells which were fabricated so that the potential profile was parabolic. In this case, the quantized energy levels are more sensitive to the value of the band discontinuity than in the square-well case. The parabolic-well experiments produced a value of $\Delta E_C \approx 0.57\Delta E_G$. The average value of more recent results is approximately $\Delta E_C = 0.60\Delta E_G$ [23].

Heterojunctions between Si and Ge_xSi_{1-x} alloys have attracted a great deal of attention recently. Because the lattice mismatch between Si and Ge is large (greater than 4%), one or the other of the materials participating in the heterojunction is generally highly strained. The band alignment depends rather sensitively on the strain, and is also complicated by the fact that the strain causes a splitting of the degenerate states at both the valence and conduction band edges. Further information may be found in the chapter by King [4]. Kasper and Schäffler [40] have also reviewed the work on this system.

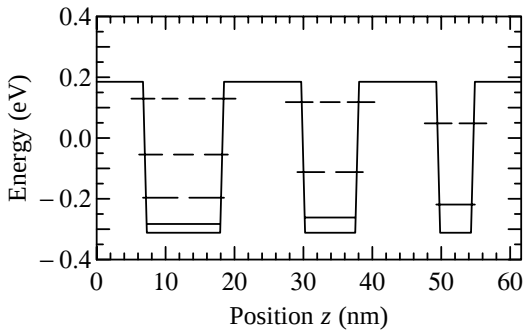

Figure 5: Energy-band profile of a structure containing three quantum wells, showing the confined states in each well. The structure consists of GaAs wells of thickness 11, 8, and 5 nm in $Al_{0.4}Ga_{0.6}As$ barrier layers. The gaps in the lines indicating the confined state energies show the locations of nodes of the corresponding wavefunctions.

3 Physical Interpretation of the Band Alignment

The significance of the effective-mass theorem to heterostructures is that it provides a precise definition of the idea of a "position-dependent band edge." On the surface, it would seem that an attempt to describe a band-edge energy as a function of position would violate the uncertainty principle, because the states which lie at the band edge are momentum eigenstates. There is, however, no conflict in the idea of a position-dependent potential, so the local band-edge energy should really be interpreted as that potential which appears in the appropriate effective-mass Schrödinger equation for a given heterostructure. The reason for the indirectness of experimental measurements of the band alignment is now apparent: The band discontinuities are not directly observable quantities, but rather parameters (albeit essential ones) of a particular level of theoretical abstraction.

IV Quantum Wells

If one makes a heterostructure with sufficiently thin layers, quantum interference effects begin to appear prominently in the motion of the electrons. The simplest structure in which these may be observed is a quantum well, which simply consists of a thin layer of a narrower-gap semiconductor between thicker layers of a wider-gap material [37]. The band profile then shows a "rectangular well," as illustrated in Fig. 5. The electron wavefunctions in such a well consist of a series of standing waves, such as might be found in a resonant

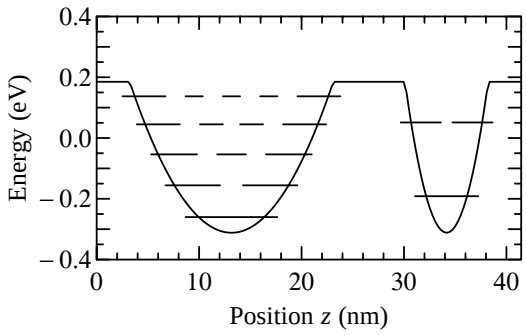

Figure 6: Energy band profile of a structure containing two parabolic quantum wells. The composition is similar to that of Fig. 5, and the overall width of the wells are 20 and 8 nm.

cavity in acoustic, optical or microwave technologies. The energy separation between these stationary states is enhanced by the small effective mass of electrons in the conduction bands of direct-gap semiconductors. With advanced epitaxial techniques, the potential profile of the quantum well need not be rectangular. Because the band-edge energy is usually linear in the composition, $E_{V,C}$ will follow the functional form of the composition. The quantum states in two parabolic wells [39] are illustrated in Fig. 6. Quantum well heterostructures are key components of many optoelectronic devices, because they can increase the strength of electro-optical interactions by confining the carriers to small regions [1, 2].

V Quasi-Equilibrium Properties of Heterostructures

1 Carrier Distribution and Screening

To understand how any heterostructure device operates, one must be able to visualize the energy-band profile of the device, which is simply a plot of the band-edge energies U_V and U_C as functions of the position x. These energies include the effects of the heterostructure energies $E_{V,C}(x)$ and the electrostatic potential $V(x)$. The electrostatic potential of course depends upon the distribution of charge within the device $\rho(x)$. In general, $\rho(x)$ depends upon the current flow within the device, and the evaluation of a self-consistent solution for the potential, carrier densities and current densities is the fundamental problem of device theory. However, in many cases, one may obtain an adequate estimate of the band profile by neglecting the current, and assuming that the device can be divided into different regions, each of which is locally in thermal equilibrium with a Fermi level set by the voltage of the

electrode to which that region is connected. We will refer to this as a quasi-equilibrium approximation. Such calculations are readily performed on computers of very modest capability. The formulation of the quasi-equilibrium problem of course holds exactly in the case of thermal equilibrium (no bias voltages applied to the device), and the equilibrium band profile of a heterojunction has been studied by Chatterjee and Marshak [41] and by Lundstrom and Schuelke [42].

It is fairly common for heterostructures to create regions in which the carrier densities become quantum-mechanically degenerate. One therefore needs to take degeneracy into account in evaluating the carrier densities. We will assume that the energy bands are parabolic, so that the quasi-equilibrium carrier densities are

$$p(x) = N_V(x)\mathcal{F}_{1/2}\{[E_V(x) - qV(x) - E_F(x)]/kT\}, \tag{9a}$$

$$n(x) = N_C(x)\mathcal{F}_{1/2}\{[E_F(x) - E_C(x) + qV(x)]/kT\}, \tag{9b}$$

where $\mathcal{F}_{1/2}$ is the Fermi-Dirac integral of order $\frac{1}{2}$,

$$\mathcal{F}_{1/2}(\eta) = \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{\xi^{1/2}\, d\xi}{1 + e^{\xi - \eta}}, \tag{10}$$

and the effective densities of states are

$$N_C(x) = 2\left[\frac{2\pi m_C^*(x)kT}{h^2}\right]^{3/2}, \tag{11a}$$

$$N_V(x) = 2\left[\frac{2\pi m_V^*(x)kT}{h^2}\right]^{3/2}. \tag{11b}$$

It is not particularly useful to express p and n in terms of the intrinsic density n_i and the intrinsic Fermi level E_i, because these quantities are not constant throughout a heterostructure. (Formulations which emphasize these quantities require the definition of an excessive number of auxiliary quantities to express the content of the heterostructure equations [42, 43].) Also, the usefulness of n_i in the elementary pn junction theory follows primarily from the mass-action law, $pn = n_i^2$, which is not valid in a degenerate semiconductor.

The net charge density includes contributions from the mobile carrier densities $n(x)$ and $p(x)$, and from the ionized impurity densities N_D^+ and N_A^-. If one takes into account the impurity statistics, the ionized impurity densities will depend upon the potential:

$$N_D^+(x) = \frac{N_D}{1 + g_D \exp\{[E_F(x) - E_D(x) + qV(x)]/kT\}}, \tag{12a}$$

$$N_A^-(x) = \frac{N_A}{1 + g_A \exp\{[E_A(x) - qV(x) - E_F(x)]/kT\}}. \tag{12b}$$

Here g_D and g_A are the degeneracy factors of the donors and acceptors, respectively, and the impurity state energies E_D and E_A are defined with respect to the same energy scale as $E_{V,C}$. The total charge density is then

$$\rho(x) = q[p(x) - n(x) + N_D^+(x) - N_A^-(x)], \tag{13}$$

Note that $\rho(x)$ depends upon V, E_F, and the band parameters E_V and E_C through equations (9) and (12).

With the above expressions for the charge density, the electrostatic potential is described by Poisson's equation, plus the appropriate boundary conditions. In a heterostructure, the dielectric constant will typically vary with semiconductor composition, so Poisson's equation must be written as

$$\frac{d}{dx}\epsilon(x)\frac{dV}{dx} = \rho(x). \tag{14}$$

This form guarantees the continuity of the displacement. The screening equation for a heterostructure is obtained by combining all of the equations in this section into (14). It is a nonlinear differential equation for $V(x)$, as the materials parameters are fixed by the design of the heterostructure, and the Fermi levels are fixed by the external circuit. The solutions to this nonlinear equation are well behaved and stable, however, because the charge density varies monotonically with V and has the screening property: making the potential more positive makes the charge density more negative and *vice versa*.

The boundary conditions to be applied to this screening equation follow from the condition that each semiconductor material must be charge-neutral far from the heterojunction. Let the boundary points be x_l and x_r. These can be taken to be $\pm\infty$ if one is solving for the potential analytically, but if numerical techniques are used x_l and x_r should be finite but deep enough into the bulk semiconductor that charge neutrality may be assumed. One then determines $V(x_l)$ and $V(x_r)$ simply by solving

$$\rho(x_l) = 0, \tag{15a}$$

$$\rho(x_r) = 0. \tag{15b}$$

The physical picture that is assumed in this formulation is that the Fermi energy (possibly different in different regions of the device) is set by the voltages on the terminals of the device. The terminals, together with the circuit node to which they are connected, are charge reservoirs whose chemical potential is just the Fermi level. The device and the circuit exchange charge, and the entire energy band structure, floats up or down until charge neutrality in the bulk is achieved. Thus the origin of the scale of V is set by the combined

choice of the energy scale for the band-structure energies E_V and E_C, and the choice of ground potential for the circuit voltages (and thus the Fermi levels). The Fermi energies on each side of the junction $E_F^{\pm}$ are determined by the externally applied voltages at the respective contacts. In fact, it is most convenient to define the Fermi energy with respect to the circuit ground potential so that

$$E_F^i = -qV_i, \tag{16}$$

where V_i is the voltage of the circuit node connected to the i'th device terminal.

If the carrier densities are neither degenerate nor closely compensated, the Fermi functions in (15) may be approximated by exponentials and one may directly solve for $V(x_{l,r})$ to obtain the more familiar expressions:

$$V(x_{l,r}) = \begin{cases} \{E_C(x_{l,r}) - E_{Fl,r} + kT\ln[N_D(x_{l,r})/N_C(x_{l,r})]\}/q, & \text{N-type;} \\ \{E_V(x_{l,r}) - E_{Fl,r} - kT\ln[N_A(x_{l,r})/N_V(x_{l,r})]\}/q, & \text{P-type.} \end{cases} \tag{17}$$

The diffusion voltage, which appears in the standard pn junction analysis, is just the magnitude of the potential difference across the heterojunction $V_d = |V(x_r) - V(x_l)|$.

The screening equation consisting of Poisson's equation (14) combined with the charge density expression (13) and subject to the boundary values obtained by solving (15) is a nonlinear differential equation for the electrostatic potential $V(x)$. It is best solved numerically for each specific case, due to the large number of band alignment topologies. An effective approach is to make a finite-difference approximation to the equation, reducing it to a set of simultaneous nonlinear algebraic equations, and solve these using Newton's method (see Selberherr [44]). The examples presented below were calculated using this approach.

If a given heterojunction is doped so as to achieve the same conductivity type on both sides of the junction (n-n or p-p), the junction is said to be *isotype*. If opposite conductivity types are achieved (p-n or n-p), it is an *anisotype* junction. Figures 7–9 illustrate a few of the many possible band profiles that can be obtained with heterojunctions. Figure 7 shows the band profile of an anisotype straddling junction in equilibrium. Apart from the band-edge discontinuities the profile resembles that of a pn homojunction. An isotype junction is shown in Fig. 8. Its band profile resembles that of a Schottky barrier. Figure 9 shows the profile of a broken-gap system. The bands are fairly flat, despite the fact that this is an anisotype junction.

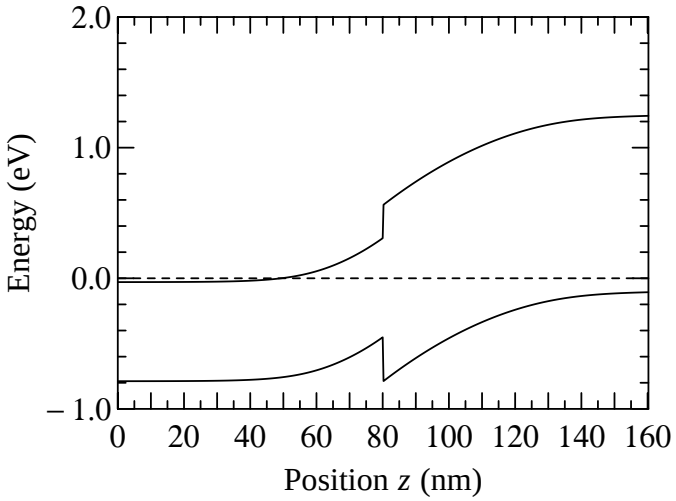

Figure 7: Self-consistent band profile of an anisotype straddling heterojunction in equilibrium. The $In_{0.53}Ga_{0.47}As$-InP heterojunction was chosen to emphasize the band discontinuities.

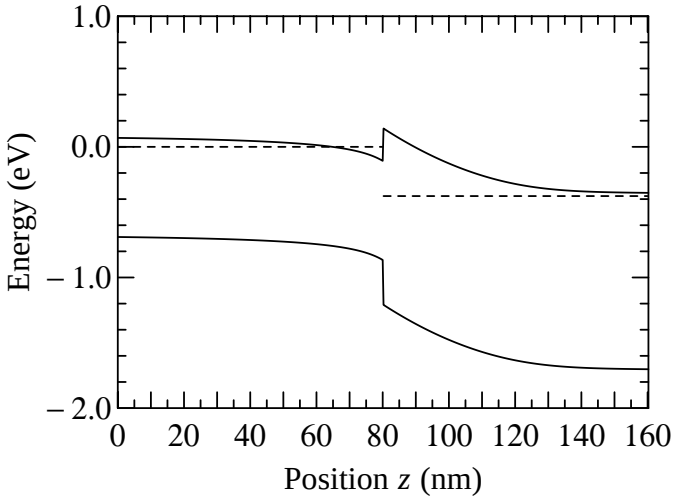

Figure 8: Self-consistent band profile of an isotype heterojunction under a small reverse bias. Again the $In_{0.53}Ga_{0.47}As$-InP is shown.

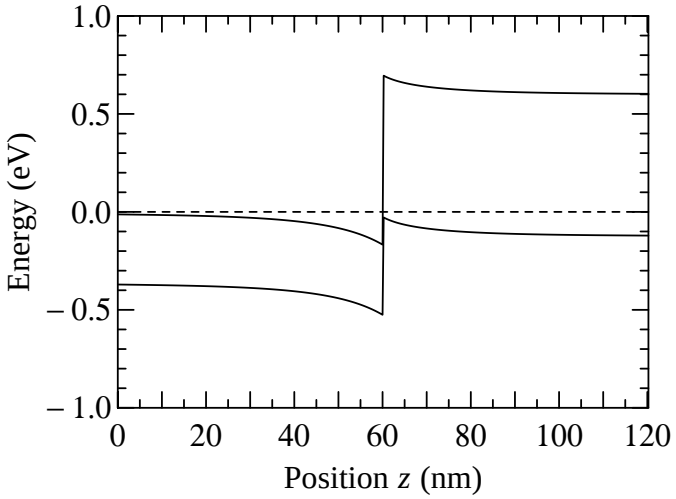

Figure 9: Self-consistent band profile of a broken-gap (N)InAs-(P)GaSb heterojunction in equilibrium. This doping configuration is the most easily fabricated.

VI Transport Properties

1 Drift-Diffusion Equation

In a heterostructure, the band structure necessarily varies with position. This variation requires that the drift-diffusion equation for the current density be modified. This is most easily demonstrated by considering the case of thermal equilibrium, where the total current density must be zero. If the electron density is non-degenerate it may be approximated by the Boltzmann distribution:

$$n(x) = N_C(x) \exp\{[E_F - E_C(x) + qV(x)]/kT\} \tag{18}$$

If we insert this into the ordinary expression for the diffusion current, we obtain an expression which must equal the negative of the drift current:

$$\begin{aligned} j_{\text{diff}} &= qD_n \nabla n \\ &= qD_n n \left(\frac{q}{kT} \nabla V - \frac{1}{kT} \nabla E_C + \frac{\nabla N_C}{N_C} \right) \\ &= -j_{\text{drift}}. \end{aligned} \tag{19}$$

The effective density of states N_C depends upon position through the effective mass m^*, which is a function of the semiconductor composition. Thus, from eq. (11a) for parabolic

bands,

$$\frac{\nabla N_C}{N_C} = \frac{3}{2}\frac{\nabla m_C^*}{m_C^*}. \tag{20}$$

Adding the drift and diffusion currents together, and making use of the Einstein relationship, we find that the electron current must be given by an expression of the form

$$J_n = -q\mu_n n\nabla V + \mu_n n\nabla E_C + qD_n\nabla n - \tfrac{3}{2}qD_n n\nabla \ln(m_C^*). \tag{21a}$$

By a similar argument one obtains an expression for the hole current:

$$J_p = -q\mu_p p\nabla V + \mu_p p\nabla E_V - qD_p\nabla p + \tfrac{3}{2}qD_p p\nabla \ln(m_V^*). \tag{21b}$$

The first and third terms of eqs. (21) are the usual drift and diffusion, respectively. The second and fourth terms are due to the spatial variability of the band structure. The second term resembles the drift term, but describes the carriers' response to changes in the band-edge energy, rather than to changes in the electrostatic potential. This effect is called a "quasi-electric field" [45], and is the origin of much of the usefulness of heterostructures. This term is readily understood on the basis that the carriers respond to gradients in the total band-edge energies U_C and U_V. The fourth term is more closely related to the diffusion term, and it describes the dynamical effects of a variable m^*. To visualize this effect, consider two materials, having different effective masses but equal potentials and equal temperatures, in intimate contact. The thermal energies in each material are equal, but the average thermal velocity will be larger in the material with the smaller m^*. Those carriers will diffuse across the interface between the materials faster than the heavier carriers, leading to a net flux of particles out of the region of smaller m^*. The heterostructure drift-diffusion equations (21) may also be derived microscopically, starting from the Bolzmann equation [46]. Equations (21) may also be written more compactly as

$$J_n = \mu_n n\nabla U_C + qD_n N_C\nabla(n/N_C), \tag{22a}$$

$$J_p = \mu_p p\nabla U_V - qD_p N_V\nabla(p/N_V), \tag{22b}$$

which is a more convenient form for subsequent manipulations.

Equations (22) may be solved analytically for the case of steady-state transport in one dimension, provided that recombination and generation may be neglected. The current density $J_{n,p}$ will then be independent of x. The carrier densities may be rewritten in terms of the quasi-Fermi levels, or, equivalently, one multiplies the drift-diffusion equation by an appropriate integrating factor. Let us consider the electron current first. Recognizing that both N_C and μ_n (and thus D_n) will be functions of the position x, the integrating factor

is $(\mu N_C)^{-1} e^{U_C/kT}$. Multiplying both sides of (22a) by this factor and integrating between points $x = a$ and $x = c$, where the electron density is presumed to be fixed, we find

$$J_n = \frac{kT}{F_n}\left[\frac{n(c)}{N_C(c)}e^{U_C(c)/kT} - \frac{n(a)}{N_C(a)}e^{U_C(a)/kT}\right] = \frac{kT}{F_n}\left[e^{E_F(c)/kT} - e^{E_F(a)/kT}\right], \tag{23a}$$

where

$$F_n = \int_a^c \frac{e^{U_C/kT}\,dx}{N_C\mu_n}. \tag{23b}$$

The drift-diffusion equation for holes may be similarly solved to yield

$$J_p = -\frac{kT}{F_p}\left[\frac{p(c)}{N_V(c)}e^{-U_V(c)/kT} - \frac{p(a)}{N_V(a)}e^{-U_V(a)/kT}\right] = \frac{kT}{F_p}\left[e^{-E_F(c)/kT} - e^{-E_F(a)/kT}\right], \tag{23c}$$

with

$$F_p = \int_a^c \frac{e^{-U_V/kT}\,dx}{N_V\mu_p}. \tag{23d}$$

This solution is mathematically valid even when there are discontinuities in the parameters such as U_C. It thus provides a convenient way to deal with abrupt heterojunctions. If one takes a and c to bound a differential element centered upon an abrupt heterojunction, one finds (not surprisingly) that the quasi-Fermi level should be continuous through the heterojunction. Equations (23) may also be used in numerical simulations, to evaluate the current density between discrete mesh points.

The heterostructure drift-diffusion equations (22) and their solutions (23) can be incorporated into the conventional pn junction theory to obtain expressions for the $I(V)$ characteristics of a heterojunction. The variety of band alignment topologies makes it difficult to write generally valid expressions. However, the general behavior of heterojunctions is easy to understand intuitively and to describe (neglecting the broken-gap or extremely staggered cases). The barrier for carriers in the wider-gap semiconductor to pass into the narrower-gap one is lowered as compared to the barrier for carriers to pass from the narrower-gap material to the wider-gap one. Thus the great majority of the forward current in a heterojunction consists of one type of carrier, or in the language of bipolar transistors, the injection efficiency is quite large. This effect is exploited in the heterojunction bipolar transistor (HBT) [4, 3].

Equations (23) also provide a model for the rather common case of current transport over an energy barrier. Suppose that $U_C(x)$ has a maximum in the interval (a, c) at $x = b$. Then, because of the exponential dependence upon U_C, most of the contribution to the integral F_n will come from the vicinity of the barrier at b. One may define an effective width

w_b for the barrier as that value such that $F_n = e^{U_C(b)/kT} w_b / N_C(b)\mu_n(b)$. The current density then becomes

$$J_n = \frac{kT\mu_n(b)N_C(b)}{w_b}\left\{e^{[E_F(c)-U_C(b)]/kT} - e^{[E_F(a)-U_C(b)]/kT}\right\}. \tag{24}$$

This demonstrates the exponential dependence upon applied voltage (through E_F) expected for barrier-limited current flow. If one considers very narrow barriers, the factor of w_b in the denominator leads to a very large pre-exponential factor. In such a case the energy band profile resembles that of a Schottky barrier, and the drift-diffusion equation is not the most appropriate model for current flow.

2 Abrupt Structures and Thermionic Emission

In structures with narrow barriers, the electrons will not travel far enough to suffer collisions as they cross the barrier. Under these circumstances, the thermionic emission theory is a more accurate representation of the current transport [47]. The current density is given by

$$J_n = A^* T^2 \left\{e^{[E_F(c)-U_C(b)]/kT} - e^{[E_F(a)-U_C(b)]/kT}\right\}, \tag{25}$$

where A^* is the effective Richardson constant given by

$$A^* = \frac{qm^* k_{\mathrm{B}}^2}{2\pi^2\hbar^3}.$$

If one compares the current density predicted by the diffusion theory (24) to that predicted by the thermionic-emission theory (25), one finds that the dependence upon the barrier height and the applied voltage is identical, and that the theories differ only in the pre-exponential factor. Moreover, if one evaluates the ratio of these factors one finds

$$\frac{J_n^{\text{diffusion}}}{J_n^{\text{thermionic}}} = \frac{\mu_n\sqrt{kTm^*}}{qw_b} = \frac{\lambda}{w_b},$$

where λ is the mean-free-path in one dimension. The processes modeled by diffusion and by thermionic emission are effectively in series, so that the current density is determined by that process which predicts the lower current density. On this basis, the diffusion theory is appropriate for barriers in which $w_b > \lambda$, while the thermionic emission theory is appropriate for barriers for which $w_b < \lambda$.

However, if the barrier becomes very narrow, current transport by quantum-mechanical tunneling becomes more prominent. In many semiconductor heterostructures significant

tunneling can occur through barriers of several nanometers thickness due to the low effective mass of the carriers. This may be observed in those heterojunctions which naturally form thin barriers, such as heavily-doped isotype junctions, or in thin heterostructure barriers designed to permit tunneling. The evaluation of the tunneling currents in heterostructures is described in detail in Chapter 9 of the present volume.

The ability to make abrupt steps in the band-edge energy using heterostructures is exploited in hot-electron transistors [5]. Electrons passing over such a barrier into a lower-potential region are suddenly accelerated to high kinetic energies, which can be sufficient to carry them across a sufficiently narrow base region.

3 Quantum-Mechanical Reflection

At an abrupt heterojunction, the sudden change in the wavevector of the quantum state will lead to a significant probability of reflection R for the electrons. For a simple abrupt junction R depends upon the velocities on the two sides of the junction:

$$R = \left| \frac{v_r - v_l}{v_r + v_l} \right|^2 . \tag{26}$$

This expression can be used to estimate the factor by which the thermionic emission current density will be reduced by reflection. For more complicated structures, the complete tunneling theory should be employed.

VII Summary

Heterostructures provide a wealth of physical phenomena and design options which may be exploited in advanced semiconductor devices, as the rest of the present volume attests. These advantages are traceable to the control which heterostructures provide over the motion of charge carriers. (In optoelectronic devices, the ability to confine the optical radiation is also extremely important.) This control can be exerted in the form of selective energy barriers (barriers for one carrier type different from that for the other) or quantum-scale potential variations.

An understanding of the physical properties of heterostructures is essential to their successful use in devices. The energy-band alignment is the most fundamental property of a heterojunction, and it determines the usefulness of various material combinations for different device applications. The band profile of a heterostructure is determined by the

combined effects of heterojunction discontinuities and carrier screening, and it determines many of the electrical properties of the structure. Transport through a heterostructure can be described at a number of different levels, depending upon the size and abruptness of the structure.

References

References

[1] G. M. Smith and J. J. Coleman, Chapter 7 of the present volume.

[2] J. C. Campbell, Chapter 8 of the present volume.

[3] P. M. Asbeck, M. F. Chang, K. C. Wang, and G. J. Sullivan, Chapter 4 of the present volume.

[4] C. A. King, Chapter 5 of the present volume.

[5] A. F. J. Levi, Chapter 6 of the present volume.

[6] R. J. Matyi, Chapter 2 of the present volume.

[7] G. C. Osbourn, J. Appl. Phys. **53**, 1586 (1982).

[8] T. P. Pearsall, editor, "Strained-Leyer Superlattices: Materials Science and Technology," Vol. 33 of "Semiconductors and Semimetals," (R. K. Willardson and A. C. Beer, series editors), Academic Press, San Diego, 1991.

[9] , P. D. Dapkus, Chapter 3 of the present volume.

[10] A. G. Milnes and D. L. Feucht, "Heterojunctions and Metal-Semiconductor Junctions," Academic Press, New York, 1972.

[11] G. A. Baraff, J. A. Appelbaum, and D. R. Hamann Phys. Rev. Lett. **38**, 237 (1976).

[12] W. E. Pickett, S. G. Louie, and M. L. Cohen, Phys. Rev. Lett. **39**, 109 (1977).

[13] , G. Burns, "Solid State Physics," ch. 10. Academic Press, Orlando, 1985.

[14] M. L. Cohen and T. K. Bergstresser, Phys. Rev. 141, 789 (1966).

[15] G. H. Wannier, Phys. Rev. **52**, 191 (1937).

[16] J. C. Slater, Phys. Rev. **76**, 1592 (1949).

[17] J. M. Luttinger and W. Kohn, Phys. Rev. **97**, 869 (1955).

[18] R. L. Anderson, Solid-State Electron. 5, 341 (1962).

[19] H. Kroemer, H., in "Molecular Beam Epitaxy and Heterostructures", (L. L. Chang and K. Ploog, eds.), p. 331, Martinus Nijhoff, Dordrecht, 1985.

[20] M. J. Adams, and A. Nussbaum, Solid-State Electron. 22, 783 (1979).

[21] O. von Roos, Solid-State Electron. 23, 1069 (1980).

[22] H. Kroemer, IEEE Electron Device Lett. EDL-4, 25 (1983).

[23] E. T. Yu, J. O. McCaldin, and T. C. McGill, in Solid State Physics, Advances in Research and Applications, (H. Ehrenreich and D. Turnbull, eds.), vol. 46, pp. 1–146, Academic Press, Boston, 1992.

[24] J. R. Waldrop, E. A. Kraut, C. W. Farley, and R. W. Grant, J. Appl. Phys. 69, 372 (1991).

[25] J. R. Waldrop, E. A. Kraut, C. W. Farley, and R. W. Grant, J. Vac. Sci. Technol. B 8, 768 (1990).

[26] H. Sakaki, L. L. Chang, R. Ludeke, C.-A. Chang, G. A. Sai-Halasz, and L. Esaki, Appl. Phys. Lett. 31, 211 (1977).

[27] U. Cebulla, G. Tränkle, U. Ziem, A. Forchel, G. Griffiths, H. Kroemer, and S. Subbanna, Phys. Rev. B 37, 6278 (1988).

[28] A. Nakagawa, H. Kroemer, and J. H. English, Appl. Phys. Lett. 54, 1893 (1989).

[29] H. Kroemer, Surface Sci. 132, 543 (1983).

[30] W. R. Frensley and H. Kroemer, J. Vac. Sci. Technol. 13, 810 (1976).

[31] W. R. Frensley and H. Kroemer, Phys. Rev. B 16, 2642 (1977).

[32] W. A. Harrison, J. Vac. Sci. Technol. 14, 1016 (1977).

[33] J. Tersoff, in "Heterojunction Band Discontinuities, Physics and Device Applications," (F. Capasso and G. Margaritondo, eds.), p. 3, North-Holland, Amsterdam, 1987.

[34] F. Capasso and G. Margaritondo, eds., Heterojunction Band Discontinuities, Physics and Device Applications," North-Holland, Amsterdam, 1987.

[35] H. C. Casey, Jr., and M.B. Panish, "Heterostructure Lasers, Part B: Materials and Operating Characteristics," Academic Press, New York, 1978.

[36] J. Batey, and S. L. Wright, J. Appl. Phys. 59, 200 (1986).

[37] R. Dingle, W. Wiegmann, and C. H. Henry, Phys. Rev. Lett. 33, 827 (1974).

[38] R. Dingle, in "Festkörperprobleme/Advances in Solid State Physics," (H.J. Queisser, ed.), Vol. 15, p. 21, Vieweg, Braunschweig, 1975.

[39] R. C. Miller, D. A. Kleinman and A. C. Gossard, Phys. Rev. B29, 7085 (1984).

[40] E. Kasper and F. Schäffler, in "Strained-Layer Superlattices: Materials Science and Technology," (T. P. Pearsall, ed.), Vol. 33 of "Semiconductors and Semimetals," p. 223, Academic Press, San Diego, 1991.

[41] A. Chatterjee and A. H. Marshak, Solid-State Electronics 24, 1111 (1981).

[42] M. S. Lundstrom and R. J. Schuelke, Solid-State Electronics 25, 683 (1982).

[43] M. S. Lundstrom and R.J. Schuelke, IEEE Trans. Electron Devices EDL-30, 1151 (1983).

[44] S. Selberherr, "Analysis and Simulation of Semiconductor Devices," Springer-Verlag, Wien, 1984.

[45] H. Kroemer, RCA Review 18, 332 (1957).

[46] A. H. Marshak and K. M. van Vliet, Solid-State Electronics 21, 417 (1978).

[47] E. H. Rhoderick and R. H. Williams, " Metal-Semiconductor Contacts," ch. 3. Clarendon Press, Oxford, 1988.