CHAPTER 5. QUANTUM-EFFECT AND HOT-ELECTRON DEVICES

S. Luryi

Dept. of Electrical Engineering
SUNY at Stony Brook
Stony Brook, NY, U.S.A. 11794

A. Zaslavsky

Div. of Engineering
Brown University
Providence, RI, U.S.A. 02912

TABLE OF CONTENTS PAGE

5.1 INTRODUCTION

Quantum mechanics underpins all of semiconductor physics at both the atomic level of electrons interacting with the periodic potential of the semiconductor material and at the envelope function level appropriate, for example, to metal-semiconductor contacts or metal-oxide-semiconductor interfaces. Still, the vast majority of semiconductor devices can be treated as classical systems of carriers near equilibrium. Thus, in bipolar and field effect transistors, quantum and hot-electron effects manifest themselves either as minor corrections to the fundamentally classical operation principles or as undesirable phenomena that limit device performance and reliability. The past two decades have witnessed considerable research interest and effort in semiconductor structures that could exploit quantum and hot-electron phenomena to perform circuit functions. Even though, to date, none of these structures has evolved beyond laboratory demonstration, continuing interest in the device research community has been maintained by several mutually-reinforcing factors.

First, there exists near-universal recognition that transistor-based microelectronics, the basis of all modern computing and much of modern communications, will at some point cease to improve at the device level. The current evolution of silicon technology towards ever-denser design of ever-faster devices is following a virtually one-dimensional path: the scaling of device dimensions by reducing the minimal size of lithographic features. Lithography-driven performance gains can be expected to continue for the next one or two decades. Smaller device dimensions will yield faster carrier transit times at lower operating voltages and currents, leading to higher maximum frequencies at lower power per device.[1] This evolution faces numerous technological hurdles, described elsewhere in this book — from currently unavailable deep-submicron lithographic techniques with sufficient throughput to the wiring delays and power dissipation problems anticipated in future microelectronic circuits. Another constraint is the rapidly escalating fabrication costs that may divert investment in device-level technology to the more profitable software and circuit-design arena. However, even if both the technological and economic constraints are overcome, it appears evident that device-level performance enhancement — higher speeds, lower power consumption, or increased device functionality — will require new operational concepts. As described elsewhere in this book, the minimum critical dimension (channel length L) of a scaled Si MOSFET operated at room temperature bottoms out not too far below $L \approx 0.1$ μm.

The second, related reason for the continued interest in quantum mechanical and hot-electron phenomena has been their deleterious effect on the operation of semiconductor devices even before the scaling limit is reached. Thus, as the oxide thickness d in complementary metal-oxide-semiconductor (CMOS) technology is scaled down to accommodate shorter L, electron tunneling into the oxide conduction band leads to leakage current that eventually results in gate breakdown. Further, at the smallest oxide thickness, $d \approx 30$ Å, required near the $L \approx 0.1$ μm scaling limit, direct tunneling from the MOSFET inversion layer into the gate is expected to become a limiting factor. Similarly, carrier heating near the drain junction due to the high lateral electric fields is a principal reliability issue, since hot carriers interact with the oxide and degrade the device lifetime. The anticipated importance of these constraints on the reliability of highly dense microelectronic circuitry of the next several generations has prompted much research on the phenomenology of tunneling and carrier heating in semiconductor devices, as described in Chapter 3. While much of the effort has concentrated on sidestepping these effects on the road to the scaling limit, the device know-how aimed at characterizing and controlling tunneling and carrier heating has also been applied to devices based on precisely such effects.

The final and perhaps most important factor driving active research in quantum-effect and hot-electron devices is the rapidly expanding semiconductor bandgap-engineering capability provided by modern epitaxy. Molecular beam epitaxy (MBE) and metalorganic chemical vapor deposition (MOCVD) of III-V semiconductors, silicon, and silicon-based alloys (SiGe, SiGeC) provide exceptional control over semiconductor layer thickness, doping, and composition. This control gives the device designer unprecedented freedom in specifying regions of carrier localization and transport, tailored electric fields and potential barriers, and precise amounts of built-in strain. Thus, near-monolayer layer control has introduced size quantization, reduced dimensionality of carriers and the many attendant effects — from changes in the density of states to high carrier mobilities produced by modulation doping — into the parameter space of proposed devices. Further progress on bandgap engineering in directions other than the epitaxy axis, either via conventional processing in the deep submicron regime or via additional epitaxial regrowth on nonplanar substrates, is a subject of much current research. This research promises to extend device physics to full two- or three-dimensional quantum confinement (quantum wires and dots). Multi-dimensional confinement in these low-dimensional structures has long been predicted to alter significantly the transport and optical properties compared to the bulk or planar heterostructure results.[2] More recently, the effects of charge quantization on transport in small semiconductor quantum dots[3] have stimulated much research in single-electron devices, in which the transfer of a single electron is sufficient to control the device.

Having briefly enumerated the scientific and technological factors that have been driving quantum-effect and hot-electron device research, let us turn to the performance advantages that such devices offer, at least in theory. Speed is often cited as a primary benefit. Quantum mechanical tunneling, on which most quantum-effect devices rely, is an intrinsically fast process. Analogously, many hot-electron devices employ ballistic transport of carriers moving at velocities considerably in excess of the Fermi velocity v_F. However, the very high speeds achieved in the active regions of the device often do not translate directly into device performance because of various delays elsewhere — for example, the *RC* time delays that accompany electrode biasing. Frequently, a more significant advantage is the higher functionality of quantum and hot-electron devices, that is their capability to perform an operation with a greatly reduced device count. Higher functionality is made possible either by strong, tunable nonlinearities in their current-voltage (*I-V*) characteristics or by unusual electrode symmetries. As a result, these devices can perform relatively complex circuit functions, replacing large numbers of transistors or passive circuit components. Examples covered in this chapter include multistate memory and logic implementations using small numbers of tunneling devices, as well as single-device logic gates fashioned from three-terminal real-space transfer devices.

Finally, although quantum-effect and hot-electron devices face a long struggle with room-temperature operation and large-scale integration before they become technologically viable for general-purpose semiconductor circuitry, even today they appear poised to take over in certain niche applications. Thus, the extreme constraints on device uniformity and operating temperature inherent in single-electron devices may render them ill-suited for large-scale logic, but the robustness of charge quantization effects in a single device at cryogenic temperatures appears ideal for extremely precise current sources in metrological applications. Similarly, the recently demonstrated quantum-cascade laser requires stringent epitaxial precision at the limit of MBE capabilities, but the absence of competing semiconductor lasers in the near-infrared makes it technologically attractive all the same.

In this chapter, the basic device structures and operating principles are discussed in Sections 5.2 and 5.3 for quantum-effect and hot-electron devices respectively. These sections also include a simple introduction to the underlying physics of quantization effects on the carrier density of states and quantum mechanical tunneling, as well as hot-carrier production, ballistic transport, and real-space transfer. The various proposed device implementations — ranging from memories and logic circuits, to specialized applications — are presented in Section 5.4. The chapter concludes with a brief overview of the prospects of quantum-effect and hot-electron devices, incorporating both the positive impact of probable technological advances and the anticipated capabilities of the rapidly evolving silicon technology.

5.2 RESONANT TUNNELING (RT) STRUCTURES

5.2.1 Quantum Mechanical Tunneling

The resonant tunneling mechanism arises from two quantum mechanical consequences of the Schrödinger equation that have no classical analog. First, if a particle is confined by some potential V(**r**) on a scale comparable to its de Broglie wavelength, the particle's momentum $\hbar\mathbf{k}$ is quantized. The continuous energy spectrum $E(\mathbf{k}) = \hbar^2\mathbf{k}^2/2m$ corresponding to free motion (m is the particle mass) is broken up into energy subbands $E_n(\mathbf{k})$. Second, as long as the confining potential V(**r**) is not infinite, the particle has a finite probability of being in the classically forbidden region, where its energy E is lower than the local value of the potential. Both of these effects are most easily illustrated in the case of one-dimensional (1D) motion in a finite potential well of width L_W and height $V(z) = V_0$ ($|z| \geq L_W/2$) shown in Fig. 1a. The 1D Schrödinger equation for the wavefunction $\chi(z)$ can be written as follows,

$$H\,\chi(z) = \left[\frac{-\hbar^2}{2m}\frac{d^2}{dz^2} + V(z)\right]\chi(z) = E\,\chi(z)\,, \qquad (1)$$

where H is the Hamiltonian defined by the terms in the square brackets and $\hbar$ is the reduced Planck's constant. Equation 1 can be solved in each of the three regions and by imposing continuity conditions on $\chi(z)$ and $d\chi/dz$ one obtains discrete energy levels E_n, as well as the explicit form of the corresponding $\chi_n(z)$. The normalized $\chi_n(z)$ are related to the probability of finding the particle at some coordinate $z = z_0$ by $P(z_0) = |\chi(z_0)|^2$. In the infinite potential well limit (V_0 ∅), the eigenfunctions $\chi_n(z)$ must go to zero at $|z| = L_W/2$ and the energy levels are given by

$$E_{\mathrm{n}} = \frac{\hbar^2 \check{s}^2 n^2}{2m\, L_{\mathrm{W}}^2}\,, \qquad (2)$$

where n is an integer. In the more relevant finite potential well case of Fig. 1a, one finds (cf. Problem 1) that the well contains a finite number of energy levels E_{n}, which do not have the rapidly increasing n^2 dependence on level number (every 1D potential well contains at least one level). Furthermore, the corresponding wavefunctions $\chi_{\mathrm{n}}(z)$ penetrate into the potential barriers according to

$$\chi_{\mathrm{n}}(z) \sim e^{-\kappa_{\mathrm{n}}|z|} \qquad (3)$$

where $\kappa_{\mathrm{n}} = [2m(V_0 - E_{\mathrm{n}})/\hbar^2]^{1/2}$ and the other mathematically possible solution in the barrier, $\chi_{\mathrm{n}}(z) \sim e^{\kappa_{\mathrm{n}}|z|}$ can be excluded on the physical grounds that it diverges as $|z|$ ∅ . Although the barrier penetration is described by an exponentially decreasing function, Eq. 3 implies that a carrier in the state characterized by $\chi_{\mathrm{n}}(z)$ has a finite probability of being found in the classically forbidden barrier region $|z| > L_{\mathrm{W}}/2$.

A similar treatment of a particle characterized by kinetic energy E incident from one side on a 1D potential barrier of finite height V_0 and width L_{B}, shown in Fig. 1b, suffices to illustrate the basic mechanism of quantum mechanical tunneling. Classically, if $E < V_0$ the particle would be reflected regardless of barrier width, but barrier penetration analogous to Eq. 3 ensures a finite transmission probability $T(E)$ that depends on V_0 and L_{B}. Indeed, the solutions of the Schrödinger equation to the left of the barrier, $z < 0$,

$$\chi(z) \sim \mathrm{A}e^{\mathrm{i}kz} + \mathrm{B}e^{-\mathrm{i}kz} \qquad (4)$$

where A and B are constants, can be naturally associated with the incident ($\chi(z) \sim \mathrm{A}e^{\mathrm{i}kz}$) and reflected ($\chi(z) \sim \mathrm{B}e^{-\mathrm{i}kz}$) particles. The analogous solutions to the right of the barrier, $z > L_{\mathrm{B}}$, can be taken as $\chi(z) \sim \mathrm{C}e^{\mathrm{i}kz}$, where C is a constant and we assume that the particle was originally incident from the left. Solving the Schrödinger equation in the barrier region, $0 \leq z \leq L_{\mathrm{B}}$ and imposing the continuity conditions at the barrier boundaries, one associates the ratios $|\mathrm{B/A}|^2$ and $|\mathrm{C/A}|^2$ with reflection $R(E)$ and transmission $T(E)$ probabilities respectively,[4] with $R + T = 1$. The result is that if the incident energy E is such that $e^{-\kappa L_B} << 1$, where $\kappa = [2m(V_0 - E)/\hbar^2]^{1/2}$, the transmission probability is approximately (cf. Problem 2):

$$T(E) - e^{-2\kappa L_B} \tag{5}$$

It is apparent from Eq. 5 that in the single barrier case the transmission probability $T(E)$ for $E < V_0$ is a monotonically increasing function of incident energy E. This rather uninspiring result changes drastically when the same particle is incident on two potential barriers separated by the well of width L_W, as shown in Fig. 2a. The explicit double-barrier transmission $T(E)$ probability can be obtained[5] by repeated application of Eq. 1, but it is instructive to consider the physics of the situation. Unless the potential barriers are very narrow or the energy of a state approaches V_0, the energy levels in the quantum well will coincide approximately with those of the finite potential well in Fig. 1a. On the other hand, semiclassically[4] a particle occupying one the energy levels E_n oscillates between the barriers with velocity $v_z = \hbar k_z/m$ and, in effect, is incident on a barrier twice in each period of oscillation $2L_W/v_z$. Every incidence involves some probability $T(E_n)$ of tunneling out of the double-barrier confining potential, making the energy levels metastable with a finite lifetime τ_n with respect to tunneling out, and hence a finite energy width $\Delta E_n = \hbar/\tau_n$ (Problem 3).

If a particle is incident on the double-barrier potential with energy E that does not coincide with one of the levels E_n, the total transmission probability is given by the product of the individual transmission probabilities of the first (emitter) and second (collector) barriers, $T(E) = T_E T_C$ — an exponentially small quantity given reasonably opaque barriers with T_E, $T_C << 1$. On the other hand, if the incident energy matches one of the energy levels E_n, the amplitude of the wavefunction builds up in the well as the reflected waves cancel, just as in a Fabry-Perot resonator, and the resulting transmission probability[5]

$$T(E=E_n) = \frac{4T_E\, T_C}{(T_E + T_C)^2} \tag{6}$$

reaches unity in a symmetric structure with $T_E = T_C$. Hence, the transmission probability $T(E)$ is a sharply peaked function of incident energy, illustrated in Fig. 2b, with the energy width of the transmission peaks obtaining from finite lifetime of the discrete levels.[4]

In principle, an imaginary ideal device where monoenergetic 1D particles impinge on a double-barrier potential, whose energy levels E_n are tunable by some voltage V, would exhibit a sharply peaked current-voltage I-V characteristic, replicating the $T(E)$ shown in Fig. 2b. Several constituent parts of such a device can be implemented in semiconductors. As discussed elsewhere in this book, semiconductor heterostructures can provide the required double-barrier potential. To a good approximation, electrons and holes in direct-gap semiconductors like GaAs obey parabolic dispersions of the form $E = \hbar^2\mathbf{k}^2/2m^*$, differing from free electrons only by virtue of the effective mass m^* rather than the free electron mass m_0. On the other hand, monoenergetic carrier distributions and independent voltage control of $T(E)$ are more difficult to arrange. But an even more fundamental difference between the idealized 1D scenario described by Eqs. 1 to 6 and semiconductor heterostructures lies in the existence of other spatial degrees of freedom. True 1D wires, where carrier dispersion is given by Eq. 1, are difficult to fabricate and even more difficult to use, because they are limited in their current-carrying capacity. Instead, a typical semiconductor implementation of the structure in Fig. 2a has the double-barrier potential along the epitaxial direction V(z), with free transverse motion in the (x, y) plane. If the in-plane motion can be separated from motion along the epitaxial (tunneling) direction, the total wavefunction $\Psi(\mathbf{r})$ of an electron in one of the metastable quantum well levels $\chi_n(z)$ depends on in-plane momentum $\mathbf{k}_\perp$ and can be written as

$$\Psi_{n,\mathbf{k}_\perp}(\mathbf{r}) = \mathrm{N}\,\chi_n(z)\,e^{i\mathbf{k}_\perp\cdot\mathbf{r}_\perp} \tag{7}$$

where N is a normalization factor. Given an isotropic effective mass, the corresponding total energy is given by

$$E = E_n + \frac{\hbar^2\,\mathbf{k}_\perp^2}{2m^*}\ . \tag{8}$$

Equation 8 indicates that each of the quantized energy levels gives rise to a subband and, in contrast to the 1D situation, there are no gaps in the energy spectrum above the lowest-lying subband E_1. An electron at an energy $E > E_n$ can belong to any of the n subbands from E_1, in which case it would have a large in-plane kinetic energy) to E_n. Further, in the presence of scattering, the degenerate states belonging to different subbands become mixed and the factorization of the wavefunction in Eq. 7 generally breaks down. However, since the coupling between degenerate states belonging to different subbands involves a finite change in $\mathbf{k}_\perp$, the single-subband states $\chi_n(z)$ may be sufficiently long-lived to treat their coupling as a perturbation leading to intersubband scattering.

The in-plane motion drastically changes the effective densities of states both in the quantum well and in the regions outside the double-barrier potential, $|z| > (L_W + 2L_B)/2$. Instead of discrete levels in the well, each subband E_n contributes a constant 2D density of states, $g^{2D}(E) = m^*/\pi\hbar^2$. At the same time, in real devices the tunneling carriers arrive from carrier reservoirs outside the double-barrier potential. Typically, the states in the emitter and collector carrier reservoirs can be taken as 3D. As discussed in Appendix A, the appropriate 3D density of states is

$$g^{3D}(E) = \frac{(2m^*)^{3/2}}{2\text{š}^2\hbar^3} E^{1/2} \tag{9}$$

and one can determine the Fermi level E_F in the reservoir in terms of the 3D carrier density n_{3D}, temperature T, and the Fermi-Dirac occupational probability $f_{FD}(E)$:

$$n_{3D} = \square g^{3D}(E) f_{FD}(E\text{-}E_F)\, dE, \quad f_{FD}(E) = (e^{-E/kT} + 1)^{-1} . \tag{10}$$

Instead of the idealized monoenergetic 1D carriers, the incident particles will have a relatively broad energy distribution of at least E_F in the simplest case of degenerately doped electron tunneling reservoirs at low temperatures, such that E_F separates occupied from unoccupied states in the emitter. As long as the factorization of the wavefunction into in-plane and tunneling-direction components remains valid, in-plane degrees of freedom do not complicate the situation unduly. The in-plane momentum $\mathbf{k}_\perp$ remains a constant of motion that is conserved as the carrier tunnels from the emitter reservoir through the 2D subbands E_n with a transmission probability $T(E_z)$ that depends on the energy of motion in the tunneling direction E_z. The total tunneling current density J can be computed by integrating over the electron distribution in the emitter reservoir:

$$J = \frac{e}{2\text{š}\,\hbar} \square N(E_z)\, T\,(E_z)\, \mathrm{d}E_z \tag{11}$$

where $N(E_z)$ is the number of electrons with the same E_z per unit area. At a given temperature T, the quantity $N(E_z)$ is easily evaluated by integrating the product of the constant 2D density of states $g^{2D}(E)$ (see Appendix A) with the Fermi-Dirac occupational probability $f_{FD}(E)$, yielding

$$N(E_z) = \frac{kT\,m^*}{\text{š}\,\hbar^2} \ln\left[1 + e^{(E_F - E_z)/kT}\right] . \tag{12}$$

While these equations are a reasonable starting point for considering realistic semiconductor resonant tunneling (RT) structures, it is important to recognize that their validity depends to a great extent on the absence of scattering. Clearly, even in the 1D picture of Fig. 2, the build-up of near-unity transmission probabilities of Eq. 6 by the Fabry-Perot mechanism requires the electron to retain phase coherence over a very large number of bounces (proportional to $(T_E + T_C)^{-1}$, the inverse of the single-barrier transmission coefficients[6]). The phases of the many multiply reflected amplitudes combining to cancel the net reflected wave. Any interaction that changes the phase of the wavefunction, whether elastic — like impurity scattering. or inelastic — like scattering by phonons or other electrons, will destroy the overall cancellation of the reflected wave. In a realistic three-dimensional structure with in-plane degrees of freedom, elastic impurity scattering relaxes $\mathbf{k}_\perp$ conservation. More generally, scattering unavoidably mixes in-plane and tunneling direction motion, qualitatively changing the wavefunction penetration into the barrier.[7] In the absence of scattering, in-plane motion is separable from tunneling. The wave-function penetration into the barrier then depends on the quantized energy of motion E_n in the tunneling direction, regardless of the in-plane kinetic energy $\hbar^2\mathbf{k}_\perp^2/2m^*$. Extending Eq. 3 to the case of a more general barrier potential $V(z)$, one finds:

$$\Psi_n(\mathbf{r}) \sim \exp\Big[-\hbar^{-1} \Box \; \{2m^*(\mathrm{V}(z) - E_n)^{1/2}\} \, \mathrm{d}z \Big] \qquad (13)$$

Scattering mixes in-plane motion with tunneling and barrier penetration asymptotically approaches

$$\Psi_n(\mathbf{r}) \sim \exp\Big[-\hbar^{-1} \Box \; \{2m^*(\mathrm{V}(z) - E\,)^{1/2}\} \, \mathrm{d}z \Big] \qquad (14)$$

where the energy that enters into the exponential decay of the wavefunction is the total energy $E = E_n + \hbar^2\mathbf{k}_\perp^2/2m^*$. The transition from Eq. 13 to Eq. 14 is described by a pre-exponential factor that depends on the specific scattering mechanism.[7]

All of the effects associated with scattering and limited phase coherence significantly alter the idealized sharply peaked current-voltage *I-V* characteristic that we would obtain from the ideal transmission through the double-barrier potential illustrated in Fig. 2. The many orders of magnitude peak-to-valley ratios predicted by coherent $T(E)$ calculations have not been observed experimentally in double-barrier RT structures, even at low temperatures. In fact, in realistic semiconductor RT structures, scattering limitations, and the energy width of the incident electron distributions are such that an alternative sequential tunneling model[8] predicts the *I-V* characteristics equally well. In this model, current transport is described by carriers tunneling into the 2D density of states in the well followed by uncorrelated tunneling out into the collector. The *I-V* nonlinearities arise from E and $\mathbf{k}_\perp$ conservation without recourse to near-unity transmission coefficients of the double-barrier potential in the coherent limit. These issues, as well as other effects that become relevant in optimizing RT structures for device applications, such as maximizing peak current densities or reducing the temperature sensitivity of the *I-V* characteristics, are discussed in the next section.

5.2.2 Two-Terminal RT Structures

The first experimental realization of the double-barrier RT device using semiconductor heterostructures dates back to 1974, when current peaks corresponding to electrons tunneling through the lowest two subbands in a GaAs quantum well confined by $Al_xGa_{1-x}As$ barriers were observed at low temperatures.[9] Improvements in epitaxial material quality and device design since then have led to RT diodes with very sharp low-temperature *I-V* characteristics,[10] as illustrated in Fig. 3. The inset of Fig. 3 shows the epitaxial layer sequence of the device, with $\mathrm{Al_xGa_{1-x}As}$ barriers, a narrow L_W = 56 Å GaAs well and heavily doped ($N_D = 2\mathrm{x}10^{17}$ cm^{-3}) n^+-GaAs electrodes. The device exhibits a strong negative differential resistance (NDR) characteristic above the peak ($V > V_P$) of the *I-V* curve with the peak-to-valley current ratio (PVR) reaching ~30 at low temperatures. Still, the measured PVR and overall *I-V* lineshape are quite different from the theoretical prediction of Fig. 2b based on the calculated coherent transmission coefficient $T(E)$. The valley current is much larger than predicted by simple theory because of nonresonant processes, such as scattering or phonon-assisted tunneling. It turns out that coherence and the Fabry-Perot model are not required to explain the *I-V* characteristics of realistic RT structures. It suffices to impose energy E and transverse momentum $\mathbf{k}_\perp$ conservation on carriers tunneling into the 2D subband E_n, with the only constraint imposed on the coherence of the oscillating wavefunction in the well being that it should be sufficient to produce a well-resolved 2D subband. The tunneling out of the well into the collector may then occur in a second step that may be completely uncorrelated with the tunneling into the well, resulting in current transport by a sequential tunneling scheme.

The sequential tunneling model[8] is illustrated in Fig. 4, using the n-$\mathrm{Al_xGa_{1-x}As}$/GaAs double-barrier RT structure of Fig. 3 as an example. At flatband, when no bias is applied to the device, the lowest 2D subband E_1 in the well lies above the emitter E_F, making E and $\mathbf{k}_\perp$ conserving tunneling into E_1 states impossible. As the bias V increases, E_1 is lowered with respect to the emitter E_F, as shown in the self-consistent potential distribution of Fig. 4. Resonant tunneling becomes possible once the bias brings E_1 into alignment with the occupied states in the emitter, at which point a subset of the occupied emitter states — their measure is denoted by the supply function $N(V)$ — can tunnel into the well conserving both E and $\mathbf{k}_\perp$. At low temperature, a simple geometrical evaluation of the supply function can be constructed by noting that the occupied states in the emitter can be characterized in terms of E and $\mathbf{k}_\perp$ as follows,

$$E = E_z + \frac{\hbar^2 \mathbf{k}_\perp^2}{2m^*}\,, \quad 0 \text{ Š } E_z \text{ Š } E_F \tag{15}$$

whereas the available states in the well lie on a single dispersion $E = E_1(V) + \hbar^2\mathbf{k}_\perp^2/2m^*$. Taking the bottom of occupied states in the emitter as the energy reference, the emitter states form a parabolic solid of revolution in (E—$\mathbf{k}_\perp$) phase space, filled up to E_F with carriers. The supply function can be geometrically described by the intersection of the available 2D states in the well and the occupied emitter states, see Fig. 4. It can be easily shown (cf. Problem 4) that $N(V)$ ˜ $[E_F - E_1(V)]$ as long as $E_1(V)$ does not fall below the bottom of the occupied states in the emitter, at which point the supply function drops to zero. The current into the well due to E and $\mathbf{k}_\perp$ conserving tunneling is given by

$$J = \frac{q}{2\text{š}\,\hbar} N(V) T_E(V) \tag{16}$$

to which one must add all other current components. Examples of additional current components are direct tunneling into collector states through both barriers (J ˜ $T_E T_C$), phonon-assisted tunneling, impurity or interface roughness-assisted tunneling that conserves E but not $\mathbf{k}_\perp$, and so forth. In addition, at sufficiently high V tunneling through the second 2D subband E_2 also becomes possible. Those current components that are not cut off by $\mathbf{k}_\perp$ conservation once $E_1(V)$ is biased below the emitter states contribute to the valley current. For example, the strong electron-optical phonon coupling in GaAs leads to a phonon-assisted replica peak when $E_1(V)$ is biased below the bottom of the occupied states in the emitter by optical phonon energy $\hbar\omega_{opt} = 36$ meV.[11] The quantitative modeling of nonresonant current components is not well developed and typically relies on adjustable parameters.[12] This is unfortunate since the valley current plays an important role in the minimum power dissipation of RT-based devices.

According to the sequential tunneling model, the relevant transmission coefficient that determines the current density is T_E (typically, $T_E << 1$) and the NDR is a consequence of E and $\mathbf{k}_\perp$ conservation that governs carrier tunneling into the well. In contrast, the idealized coherent model of resonant tunneling involves the total transmission coefficient $T(E_z)$ of the double-barrier potential given by Eq. 6, which goes to unity if the emitter and collector barrier transmission coefficients are equal, $T_E = T_C$, at operating bias $V \approx V_P$. Surprising though it might appear, the two models predict essentially the same *I-V* characteristics for realistic RT structures, in which both the bias required to observe the current peak V_P and the width of the tunneling carrier energy distribution E_F are much larger than the 2D level widths ΔE_n.[6,13] The essential point is that while the total transmission coefficient $T(E_z)$ of the coherent model is exponentially large compared to the single barrier coefficients, it is also exponentially narrow as shown in Fig. 2b. Indeed, if the incident energy E_z is close to matching a 2D subband energy E_n, the transmission coefficient of Eq. 5 can be expanded in terms of the small parameter (E_z - E_n) as follows,[13]

$$T(E_z \sim E_n) - \frac{4T_E\, T_C}{(T_E + T_C)^2}\frac{{}^2\, E_n^2}{(E_z - E_n)^2 + {}^2\, E_n^2} \qquad (17)$$

where $\Delta E_n = \hbar/\tau_n$ is the lifetime of subband E_n with respect to tunneling out. The total current through the device is obtained by averaging over Eq. 17. Since $E_F >> \Delta E_n$, the Lorentzian factor in Eq. (17) reduces to a δ-function,* $\pi \Delta E_n\, \delta(E_z\text{-}E_n)$. The δ-function, in turn, cancels one of the ($T_E + T_C$) factors in the denominator of Eq. 17, reducing the average transmission coefficient for the carrier ensemble. To first order, the two pictures predict the same current density,[13] except in exotic limits.** Hence, the choice of coherent or sequential tunneling model might appear immaterial. However, the geometric interpretation of Fig. 4 implicit in the sequential model is useful in predicting the *I-V* characteristics of more complicated structures, for instance those with nonparabolic in-plane carrier dispersions $E(\mathbf{k}_\perp)$ or different dispersions in the emitter and well (cf. Problem 5). More importantly, the sequential tunneling approach provides a more natural framework for discussing three-terminal RT structures that rely on the NDR in the *I-V* characteristics provided by tunneling into a restricted density of states in a quantum well without an attendant second tunneling step. For this reason, the subsequent discussion will be based on the *sequential tunneling model*.

With the exception of direct tunneling to the collector through both barriers, all of the various current components and particularly the E and $\mathbf{k}_\perp$ conserving term of Eq. 16 depend sensitively on the alignment of the 2D subbands with the occupied states in the emitter. Yet in a standard, vertical two-terminal implementation of RT structures, this alignment can only be controlled by changing the applied bias V, only a fraction of which lowers the 2D subbands (see Fig. 4). If one ignores the penetration of the electric field into the emitter and collector regions, one immediately obtains that $V_P^{(n)} = 2E_n$ (see Problem 4). This frequently cited result typically fails to describe realistic double-barrier RT structures, where undoped spacer regions around the double-barrier structure and relatively low electrode doping lead to significant potential drops in the emitter and collector regions. This is especially pertinent to RT devices designed for high-frequency operation, where low emitter-collector capacitance is often achieved by a large collector spacer region.[15] A self-consistent calculation of the potential distribution over the device, including the voltage drops in the accumulation and depletion regions as shown in Fig. 4, is therefore necessary to predict $V_P^{(n)}$ as a function of device parameters. An additional complication is the dynamically stored charge density σ_W in the well under bias, given by

$$\sigma_W = J\tau_n \tag{18}$$

where τ_n is the lifetime for subband E_n. The effect of σ_W is to increase the electric field in the second barrier and hence reduce the bias-induced lowering of the subbands $E_n(V)$ for a given V. In double-barrier structures with symmetric barriers, σ_W is typically small because $T_C >> T_E$ at V_P and $\tau_n \sim T_C^{-1}$, since only tunneling out to the collector is possible under bias (see Fig. 4). If the collector barrier is made larger, σ_W can be increased and in the case of highly asymmetric barriers

* Here we make use of the well-known identity, $\lim_{{}^2 \to 0} \frac{{}^2}{x^2 + {}^2\,2} = š\ \delta(x)$.

** In a hypothetical device where either E_F or $qV_P < \Delta E_n$ the situation would be different, with the coherent model predicting much greater current densities.[14] Such RT structures have not been realized to date.

the result can be a bistable *I-V* characteristic shown in Fig. 5. The arrows in Fig. 5, which is the other bias polarity *I-V* of the same RT device as in Fig. 3, indicate the direction of the bias sweep. If V is increased from flatband, V_P occurs at a higher bias because of significant dynamic charge storage σ_W, whereas if the resonant alignment is approached by decreasing V from above V_P, the valley current and hence σ_W is small and the *I-V* characteristic switches to the high-current branch at a lower bias. In effect, the resulting intrinsically bistable *I-V* arises from the feedback of the dynamically stored charge in the well σ_W on the alignment of $E_1(V)$ with the occupied emitter states.[16] At least in principle, the bistable *I-V* offers the possibility of constructing a single-device two-state semiconductor memory.

Double-barrier structures implemented in the *n*-AlGaAs/GaAs material system have been very useful in clarifying the relevant physics of resonant tunneling at low temperatures, but their *I-V* characteristics are less suitable for real electronic devices. First, the sharp NDR characteristic offered by the RT structures needs to survive at room temperature. At T = 300 K, the peak current density J_P remains essentially unchanged, but the valley current is supplemented by thermionic emission over the barriers and thermally-assisted tunneling through higher-lying subbands. At room temperature, both of these valley current components can significantly degrade the available PVR. Clearly, thermionic emission over the barriers can be exponentially reduced by increasing the barrier height, which in the context of AlGaAs/GaAs heterostructures implies the use of pure AlAs barriers, maximizing V_0. Yet for high-speed operation there exists the conflicting requirement of maximizing J_P, since high current densities are necessary for rapid charging of the various device and circuit capacitances — ideally $J_P \geq 10^5$ A/cm^2. The use of very narrow AlAs barriers is therefore indicated but, even so, thermally-assisted tunneling through higher-lying subbands remains a problem. For this reason, the fastest reported GaAs/AlAs double-barrier RT oscillators[17] with high $J_P \approx 10^5$ A/cm^2 exhibit a room-temperature PVR of only 3.

A considerable improvement in the PVR and J_P figures of merit in two-terminal RT devices has been obtained by moving to the *n*-$In_{0.53}Ga_{0.47}As$/AlAs material system, which is lattice matched to InP substrates. The physics of device operation is still described by Fig. 4, but the maximum barrier heights are larger and, more importantly, the lower m^* of electrons in the InGaAs well leads to higher subband separation (E_2 - E_1). The room temperature *I-V* characteristic of a double-barrier n-$In_{0.53}Ga_{0.47}As$/AlAs device is shown in Fig. 6, with $J_P > 10^5$ A/cm^2 and PVR $\approx$ 8.[18] This RT structure included a large undoped collector spacer region to reduce the emitter-collector capacitance, hence the high V_P. Because of the sharp NDR when the device is biased beyond V_P, the biasing circuit becomes unstable over a range of voltages $V_P < V \leq 2.5$ V. The circuit oscillations are rectified by the RT device, leading to the characteristic discontinuous jumps in the *I-V* curve.[19] Note that the AlAs barrier layers are not lattice matched to the substrate, but since their thickness can be kept very narrow, about three monolayers for the structure of Fig. 6, they can be deposited pseudomorphically without generating large numbers of dislocations. Further improvements in the PVR of the first resonant peak can be obtained in this material system by growing a narrow InAs layer in the center of the InGaAs well, which has the effect of further separating the lowest two subbands (cf. Problem 6).

Another variant of two-terminal RT devices involves GaSb/AlSb/InAs heterostructures. These heterostructure are known as polytype because of their staggered bandgap alignment, wherein AlSb barriers separate the InAs conduction band from the GaSb valence band edges.[20] The schematic diagram of a double-barrier polytype RT structure with GaSb electrodes, AlSb barriers, and an InAs well is shown in Fig. 7. Under bias, the current can be described in terms of holes tunneling from the GaSb emitter into the InAs well conserving E and $\mathbf{k}_\perp$ in the usual fashion — a geometrical evaluation of the supply function simply requires inverting the emitter dispersion in Fig. 4.[21] The polytype structure represents an RT version of the Esaki tunnel diode. Its advantage lies in the band-gap blocking beyond V_P, since for $V > V_P$ the emitter states line up with the bandgap of the InAs well, as shown in Fig. 7b. As a result, impurity-assisted $\mathbf{k}_\perp$-nonconserving tunneling into the well is completely suppressed, removing one of the major valley current contributions (the band lineup is similar to the standard tunnel diode). Very good PVR has been achieved in polytype structures, albeit at relatively modest peak current densities.

Further, since the bandgap blocking mechanism is independent of subband quantization in the well and the electron effective mass in InAs is very small, $m^* = 0.023m_0$, RT designs with very wide quantum wells $L_W \approx 1000$ Å are realizable without compromising PVR.[22]

Alongside other combinations of III-V semiconductors, RT structures have been fabricated in $Si_{1-x}Ge_x/Si$ heterostructures.[23] In addition to the availability of high-quality substrates and oxides, Si-based quantum-effect devices are interesting because of their potential integration with the dominant silicon technology. Unfortunately, the lattice mismatch hinders the epitaxy of $Si_{1-x}Ge_x$ layers with large Ge contents and the available bandgap difference is rather small: in RT structures strained to Si substrates, a barrier $V_0 \approx 200$ meV is available in the valence band and no appreciable barrier appears in the conduction band.* Because of low V_0 and relatively small 2D subband separation in the $Si_{1-x}Ge_x$ well, where heavy and light hole branches of the dispersion give rise to separate subbands, no room temperature NDR has been observed in p-$Si_{1-x}Ge_x/Si$ RT structures to date, although PVR ≈ 4 has been observed at cryogenic temperatures.[25] Consequently, while the strained p-$Si_{1-x}Ge_x/Si$ RT structures have been employed for spectroscopic probing of anisotropic hole dispersions[26] and strain relaxation in microstructures,[27] the prospects of their integration into mainstream technology appear remote.

All of the RT structures discussed thus far had been produced by epitaxial growth of a sequence of layers, with the necessary double-barrier potential arising from the bandgap offsets of the heterostructure constituents. While this approach has been the dominant one, there has been some research into fabricating lateral tunneling structures by depositing electrostatic gates on the surface of a modulation-doped 2D electron gas (2DEG) heterostructure. By applying a gate potential V_G with respect to the 2DEG, electrons can be electrostatically depleted underneath the gates (analogously to gate control of an FET). Figure 8 shows a schematic diagram of a

* A conduction band barrier can be obtained in structures strained to $Si_{1-x}Ge_x$ substrates (actually thick relaxed $Si_{1-x}Ge_x$ buffer layers grown on Si substrates). Low-temperature NDR has been observed in such n-$Si_{1-x}Ge_x/Si$ RT structures.[24]

lateral RT structure. The potential advantages include: excellent electronic properties of the 2DEG; tunability of the double-barrier potential by V_G, including control over barrier asymmetry by separate gate control of the two barriers; and planar device layout, compatible with FET technology. The main drawback is the relative weakness of the electrostatically created confining potentials in the plane of the 2DEG. The minimum geometric separation of the metal gates is set by lithographic limitations and far exceeds the monolayer control available by epitaxial techniques. Furthermore, regardless of the surface gate geometry, the double-barrier potential parameters L_B, L_W cannot be reduced below the spacer layer thickness (see Fig. 8). Finally, since the confining potential arises from the self-consistent electrostatics, the barrier height V_0 produced in the plane of the 2DEG is proportional to barrier thickness L_W — high barriers are necessarily broad. Hence, the *I-V* characteristics of lateral RT structures produced by electrostatic gating exhibit weak NDR and only at cryogenic temperatures.[28] If sharp confining barriers in a lateral RT structure could be produce by some means, interesting device possibilities would result. An approach using epitaxial regrowth will be discussed in the next Section.

5.2.3 Three-Terminal RT Structures

All of the previously discussed double-barrier RT structures are two-terminal devices, potentially useful in oscillator and frequency multiplier circuits, but ill-suited for more general circuitry. The addition of a third terminal to control the *I-V* characteristics of an RT structure, either with a small current as in a bipolar transistor or a gate voltage V_G as in an FET, has been attempted in a number of schemes.

Current-controlled variants of three-terminal RT structures involve a separate contact to the quantum well that can source or sink a "base" current large enough to alter the alignment of the 2D subbands E_n and the emitter E_F. In principle, the base current can be of the same or opposite polarity as the tunneling carriers. The band diagram of a bipolar structure is illustrated in Fig. 9 for an *n*-type double-barrier RT device with a separate contact to a *p*-type quantum well. This implementation is preferable to unipolar versions both because of improved isolation between the controlling base current and the tunneling electron current and because of fabrication constraints.

It is easier to contact the narrow quantum well without shorting to the nearby emitter and collector layers if the well doping is of the opposite polarity[29] (see Fig. 9). If the RT structure is biased close to V_P, a small hole current in the well can turn off the collector current I_C by biasing the 2D subband below the occupied emitter states, giving rise to negative transconductance.

A significant constraint on current-controlled three-terminal RT structures like that in Fig. 9 is the effective base resistance. To have significant 2D subband separation and hence strong NDR in the tunneling *I-V* characteristic, the quantum-well width L_W must be small. At the same time, the lateral base resistance is inversely proportional to L_W. Setting the benchmark for a truly competitive high-speed device at 1 ps, we can estimate the $R_B C$ time delays associated with charging either the emitter-well or well-collector capacitance:

$$R_B C \approx \varepsilon_s L^2 R_S / L_B \tag{19}$$

where ε_s is the semiconductor dielectric constant; L is the characteristic lateral extent of the device limited by lithographic resolution to $L \geq 500$ Å for the foreseeable future; L_B is the emitter or collector barrier thickness; R_S is the sheet resistance of the base, $R_S = (q n_B \mu)^{-1}$ where μ is the majority carrier mobility and n_B is the charge density per unit base area. In the case of the well-collector capacitance, the barrier thickness L_B can be augmented by an undoped collector spacer, as shown in Fig. 4, but increasing the well-collector separation beyond 1000 Å introduces transit-time delays on the order of 1 ps. From Eq. 19, the resulting constraint on R_S is about 10^3 Ω per square. As in heterojunction bipolar transistors, heavy doping of the quantum well appears to resolve all difficulties, but since sharp 2D quantization requires narrow L_W ~ 100 Å quantum wells, doping sufficient to achieve $R_S \leq 10^3$ Ω is problematic. First, as can be seen in Fig. 9, holes can tunnel from the quantum well into the emitter, contributing a current that will increase with emitter-well bias regardless of the 2D subband alignment with emitter states — in other words, a nonresonant current component that will reduce the PVR. The heavier hole m^* makes the corresponding tunneling transmission smaller, but if the hole density in the well exceeds the electron supply function by many orders of magnitude, the nonlinear *I-V* can be washed out completely. Second, the very existence of a large impurity density in the quantum well introduces substantial scattering, inhomogeneously broadening the subband energy width ΔE_n to much larger values than the lifetime broadening due to tunneling out of the well. As a result, current-controlled three-terminal devices with separately contacted quantum wells appear most promising in polytype GaSb/AlSb/InAs RT structures, where bandgap blocking of the tunneling current and low m^* in the InAs well results in good PVR even for very wide quantum wells, $L_W \approx 1000$ Å.[22] We will encounter similar structures in the discussion of hot-electron devices, where no quantization in the well is required and operation depends on ballistic electron transport from the emitter to the collector.

Interestingly, a similar vertical structure with a separate contact to the quantum well can be employed to produce a unipolar, voltage-controlled tunneling transistor — essentially by designing the quantum well to perform the functions of a collector. Consider the schematic band diagram shown in Fig. 10, where the second barrier is designed to be so high and wide as to eliminate tunneling out of the well. The voltages are applied with respect to the quantum well contact. At some emitter bias V_E, the tunneling current into the well can be evaluated using Eq. 16: once again, it depends on the alignment of the 2D subband E_1 with the emitter E_F. This current is extracted from the well laterally. Three terminal operation is achieved by applying a gate bias to the remaining electrode, as in Fig. 10. The gate bias V_G shifts E_1 by two mechanisms. First, the electric field in the second barrier changes the effective confining potential of the quantum well, shifting E_1 down from its position at $V_G = 0$.[30] Second, the 2DEG in the well does not screen the V_G-induced electric field completely because of the quantum capacitance effect.[31] The latter is a consequence of the Pauli exclusion principle, by which no two electrons can occupy the same quantum state. Because of this, extra kinetic energy is required to fill a given density of states with electrons. The 2D density of states in the quantum well results in a quantum capacitance C_Q per unit area, given at low temperatures by:

$$C_Q = m^* q^2/\pi\hbar^2 \, . \tag{20}$$

As a result, it becomes energetically favorable (cf. Problem 7) for part of the V_G-induced field to penetrate into the emitter barrier, inducing additional charge in the emitter and altering the alignment of E_1 with occupied emitter states. The importance of quantum capacitance depends on the relative magnitudes of C_Q and the geometric capacitances $C^{(1,2)} = \varepsilon_s/L_B^{(1,2)}$, where $L_B^{(1,2)}$ are the barrier thicknesses in Fig. 10. In RT structures with low effective mass, gate control due to quantum capacitance can be significant. Further, negative transconductance $g_m + \partial(I\text{-}V_E,V_G)/\partial V_G < 0$ is expected when, at fixed V_E, the gate bias V_G lowers E_1 below the bottom of the occupied states in the emitter and $\mathbf{k}_\perp$ conservation cuts off the tunneling current as in a standard RT diode. The main obstacle to the fabrication of such transistors lies in implementing good lateral contacts to the quantum well and keeping the gate capacitance $C^{(2)} = \varepsilon_s/L_B^{(2)}$ large without causing significant gate leakage. To date, gate control of tunneling has been demonstrated at low temperatures[32] but with a transconductance too small to make such structures practical.

An alternative route to a three-terminal RT structure is voltage control by means of a sidewall gate electrode adjacent to the active region of a standard RT structure of sufficiently small lateral extent L that a gate bias V_G can effectively control the *I-V* curve. The vertical pillar geometry of epitaxially grown RT structures makes the fabrication rather difficult. One approach has been the self-aligned *p*-type implantation with the top metal contact of an *n*-type RT diode serving as a mask.[33] The result is a lateral *pn* junction in the plane of the active RT region, shown in Fig. 11a. Reverse bias can then be used to deplete the RT structure from the side, controlling its effective electrical size. In addition to gate leakage currents, the difficulty with this approach is the lateral straggle of the implantation that becomes an issue for submicron lateral device diameter L. An alternative scheme involves the self-aligned deposition of an in-plane metal Schottky gate* directly adjacent to the RT diode pillar.[34] Gate bias V_G on the Schottky electrode can then be employed to deplete the effective lateral size of the RT structure.[35] By employing an undercut RT pillar profile to avoid shorting the gate to the top contact, as illustrated in Fig. 11b, structures exhibiting room temperature control of the *I-V* have been fabricated:[36] the *I-V*,V_G curve of a GaAs/AlGaAs RT stripe geometry device is shown in Fig. 11(c). Gate control of the resonant *I-V* peak is achieved with reasonably small gate leakage. The reason for the observed peak position V_P shift towards higher bias as V_G is increased cannot be unambiguously identified. The V_G-induced lateral potential distribution will be different in the undoped active region and the doped emitter, changing the relative alignment of emitter E_F and the quantized subbands E_n, but contact series resistance could also play a role. Note that the side-gating geometry of Fig. 11a,b sacrifices the effective transconductance g_m unless the pillar diameter is extremely narrow, resulting in formidable fabrication difficulties regardless of the gate electrode fabrication technique.

A long-proposed alternative to the external gating of standard RT structure is illustrated in Fig. 12 for the GaAs/AlGaAs system.[37] The original epitaxial structure follows the double-barrier potential layer sequence, but with very large undoped spacers on both sides of the active regions. The function of these spacers is to prevent RT currents from flowing through the bulk at low source-drain voltages V. An angled interface through the double-barrier sequence is etched and an AlGaAs gate insulator is deposited, followed by a metal gate electrode. A positive gate bias V_G induces 2DEG in the undoped GaAs layers as in a standard FET, with the usual nearly triangular potential V(x). In Fig. 12 we assume that only the lowest subband E_1 is occupied, which holds for moderate 2DEG densities. In the well, subband quantization arises from the AlGaAs double-barrier potential V(z) in the direction of current flow combined with the FET confining potential V(x) under the gate, so the lowest 1D subband E_1' lies above the Fermi level in the 2DEG. A potential difference V between the 2DEG's above and below the double-barrier potential will produce a tunneling current subject to the usual E and $\mathbf{k}_\perp$ conservation. The supply function can be determined as before, with the only difference that the conserved $\mathbf{k}_\perp = k_y$, which describes free 1D motion along the quantum wire. As a result, sharply nonlinear I-V characteristics similar to

* Low temperature operation of a p-$Si_{1-x}Ge_x$/Si three-terminal RT transistor with an oxide-isolated gate electrode has also been reported.[38]

standard two-terminal RT diodes is expected, but with effective gate control.

Figure 12 shows that the gate bias V_G controls the emitter 2DEG density and hence the magnitude of the RT current. More interestingly, V_G can also be used to tune V_P, because the fringing electric field penetrates into the double-barrier region, shifting E_1' with respect to E_F for the same source-drain bias V. As a result, $g_m < 0$ can be achieved. If the 2DEG depletion in the collector region is ignored, the electrostatic problem reduces to a parallel plate capacitor with a slit of width $(2L_B + L_W)$ and the electric field distribution can be solved exactly by conformal mapping techniques.[38] The magnitude of the transconductance can then be explicitly calculated and, given sufficiently narrow gate insulator thickness, V_G can be nearly as effective as the source-drain voltage V in shifting the relative alignment of the emitter 2DEG and the 1D subband E_1'. In realistic devices, some depletion of the 2DEG adjacent to the collector barrier can be expected, leading to smaller fringing fields in the plane of the well for a given V_G and hence lower transconductance.[39]

The main difficulty in fabricating the three-terminal structure of Fig. 12 is the creation of a clean interface through the pregrown epitaxial structure that can support a 2DEG. The most obvious approach of etching through the double-barrier structure and subsequent epitaxial deposition of the gate layer would result in oxide formation (especially in the Al-containing barrier layers) on the interface. Proof of concept has been achieved by cleaved edge regrowth,[40] where the pregrown heterostructure is physically cleaved in the growth chamber immediately prior to deposition of the AlGaAs insulating layer on the cleaved edge. The I-V,V_G curves for various V_G of the resulting device[41] are shown in Fig. 13a (the structure differs slightly from Fig. 12 in that modulation-doping during regrowth is used to produce a 2DEG under the gate even at $V_G = 0$ resulting in a "depletion mode" transistor). Negative transconductance is indeed observed, as shown in Fig. 13(b), albeit at cryogenic temperatures. Demonstration of room temperature operation and, more importantly, the fabrication of such devices by technological means, such as *in-situ* etching followed by in-vacuum transfer to the epitaxy chamber, is yet to be reported.

In addition to the severe fabrication problems faced by all of the discussed three-terminal RT devices, it is not clear that the negative transconductance they promise can be usefully applied for computation. Although it has been suggested that such devices can, in principle, perform complementary functions,[38,42] no RT transistor circuit analogous to a CMOS inverter has been demonstrated to date. The difficulty lies in the fact that in complementary CMOS transistors current is due to carriers of opposite polarity. This makes it possible to connect the drains of two transistors, rather than the source of one to the drain of the other. Consider the CMOS inverter logic gate, shown in Fig. 14. The source of the *n*-channel transistor is connected to ground, the source of the *p*-channel transistor is connected to V_{DD}, and the same input voltage V_{IN} is applied to the gates of both transistors. As V_{IN} increases, the current in the *n*-channel transistor increases, while the current in the *p*-channel transistor decreases. When the input switches between a low and a high voltage, one of the pair turns on and the other turns off. The output voltage thus switches between V_{DD}, when the *n*-channel device is off, and ground, when the *p*-channel device is off. If the input is steady, there is no current path between V_{DD} and ground, so the circuit consumes very little power. During switching, on the other hand, one of the transistors — the one that is being turned on — provides the necessary transient current to charge or discharge the output node. Note that the output node is connected to the drains of both transistors.

It might appear that by virtue of the negative transconductance exhibited by three-terminal RT structures like the one in Fig. 13, both transistors in the CMOS pair can be directly replaced by RT devices. Unfortunately, it is not always sufficient to have transconductances of complementary polarity to implement CMOS functions, at least not in a straightforward way. In RT devices, the current depends on the alignment of the emitter and quantum well densities of states and hence on the emitter bias V_E. If the *p*-channel transistor in Fig. 14 were replaced by an *n*-type RT transistor with negative transconductance, its emitter bias would itself vary between a high and a low state, rather than remaining at a constant potential. In other words, the gate voltage V_{IN} is referenced to V_{OUT}, rather than V_{DD}. This makes it difficult to design a useful circuit.

5.2.4 Cascaded RT And Superlattice-Based Structures

The *I-V* characteristic of a single double-barrier RT structure exhibits one or more resonant current peaks depending on the number n of quantized 2D subbands E_n. Several proposed applications of RT devices require a multipeak *I-V*, but with the current peaks approximately equal in magnitude and regularly spaced in voltage by ΔV_P. Neither condition is fulfilled by a typical RT structure: the subband separation (E_n - E_{n-1}) changes with n, so the peak voltages $V_P^{(n)}$ are not evenly spaced, while the peak currents increase rapidly with n because the emitter transmission coefficient T_E increases exponentially as the barrier height drops. However, the desired *I-V* curve can be obtained from a cascaded RT structure, in which N double-barrier potentials are epitaxially grown on top of one another. If these RT potentials are separated by doped layers, the resulting band diagram is shown in Fig. 15. The distribution of the total applied bias V can be calculated self-consistently, including current continuity once the RT diodes are biased above threshold, $V \geq NV_{th}$. As V is increased further, one of the diodes will be biased beyond V_P. In a perfect structure this would happen at the anode because of dynamic charge accumulation in the RT quantum wells. Realistically, variation in quantum well thickness L_W or cladding layer doping can cause one of the RT diodes to have a lower V_P than the rest. Regardless of which RT diode goes off resonance first, it suddenly presents a high resistance to the biasing circuit and the total *I-V* exhibits an NDR region. The crucial point is that if V is increased still further, current continuity requires that almost all of the increase drop over the off-resonance diode, until it begins to conduct through the next 2D subband E_2 — this is the situation illustrated in Fig. 15. This process then is repeated with other diodes, with the result that a high-field domain consisting of diodes biased into the second resonant peak expands through the structure. As each diode is biased off resonance, another current peak appears in the *I-V*, for a total of N peaks that are approximately evenly spaced in V.

The maximum number of diodes that can be cascaded in this manner depends on the PVR required in the *I-V* characteristic. For a given RT diode that is biased beyond V_P, the other diodes act as a series resistance R_S. As discussed above in the context of bistability, the *I-V* acquires hysteretic loops as R_S increases. If the R_S is sufficient to shift V_P beyond the NDR region, it also reduces the PVR. By improving RT design to increase the peak current density J_P and doping the RT regions as well as the cladding regions (which reduces the PVR of individual diodes because of increased impurity scattering, but also reduces the R_S due to other diodes in series), many RT diodes can be cascaded.[43] The room temperature *I-V* curve of an $N = 8$ cascaded RT structure is shown in Fig. 16. The scatter in the I_P, voltage spacing, and PVR of the eight current peaks is not great and can be attributed to monolayer variations in the barrier or well layers during epitaxial growth.

If the quantized subbands in different RT quantum wells are allowed to interact, for example by removing the doped cladding regions in Fig. 15, the result is a superlattice (SL) of period $d = L_B + L_W$ shown in Fig. 17. Consider the wavefunctions $\Psi(z)$ along the SL direction z. If the barriers are infinitely high, V_0 ∅ , we simply have isolated quantum wells. These wells contain the usual quantized levels E_n described by wavefunctions $\chi_n^{(m)}(z)$, where m labels the quantum well. Since $\chi_n^{(m)}(z)$ do not penetrate into the barriers, each 2D subband has a degeneracy of $2N$ including spin. If the barrier height is finite, the $\chi_n^{(m)}(z)$ wavefunctions penetrate into the barriers according to Eq. 3, allowing the wavefunctions in neighboring wells to interact. The previously degenerate levels E_n will broaden into minibands of width Δ_n. In a bulk semiconductor, according to the Bloch theorem, an electronic state can be described by a product of a plane wave and a function periodic in the lattice potential. Analogously, in a superlattice, a state in the nth miniband can be described by linear combinations of wavefunctions periodic in the SL period d, $\varphi_n^{(m)}(z) \;\; + \varphi_n(z - md)$ multiplied by a plane wave[44]

$$\Psi^{k_z}(z) = \sum_{m=1}^{N} e^{ik_z md}\, \varphi_n^{(m)}(z) \qquad (21)$$

Equation 21 is a restatement of the Bloch theorem for superlattices. As long as $\Delta_n << (E_n - E_{n-1})$ $\varphi_n^{(m)}(z)$ are to a good approximation built up from combinations of $\chi_n^{(m)}(z)$. For some range of barrier parameters V_0 and L_B, only interactions with adjacent wells are significant and the problem simplifies drastically because the periodic components of Eq. 21 can be taken as the ordinary single-well wavefunctions $\chi_n^{(m)}(z)$.* The dispersion $E(k_z)$ for motion along the SL axis becomes

$$E^{SL}{}_n(k_z) = E_n + S_n + 2T_n\cos(k_z d) \tag{22}$$

where the shift integral S_n is defined as

$$S_n \equiv \int \chi_n^{(m)}(z)\, V_0'(z)\, \chi_n^{(m)}(z)\, dz \tag{23}$$

and the transfer integral T_n as

$$T_n \equiv \int \chi_n^{(m)}(z)\, V_0'(z)\, \chi_n^{(m+1)}(z)\, dz\ . \tag{24}$$

The potential $V_0'(z)$ employed in the calculation of the shift and transfer integrals, Eqs. 23 and 24, includes all potential wells other than the *m*th — see Fig. 17. From Eq. 22 it follows that the width of the *n*th miniband $\Delta_n = 4T_n$. The allowed values of k_z can be obtained by imposing periodic boundary conditions on Eq. 21: $k_z = 2\pi p/Nd$, where $p = 0, 1, 2, \ldots (N - 1)$, so each miniband contains exactly $2N$ states.

The dispersion of the lowest miniband for motion along the SL direction is plotted in Fig. 18a, while the SL density of states, including the transverse degrees of freedom described by $\mathbf{k}_\perp$ is shown in Fig. 18b. It is evident that the effective mass along the SL, $m^* = \hbar^2(k_z^{-1}\partial E/\partial k_z)^{-1}$, is a strongly varying function of k_z: starting with a "band-edge" value $m^*_{SL} + m^*(k_z = 0)$, the mass becomes heavier as k_z increases, diverges at $k_z = \pi/2d$ (the inflection point in $E(k_z)$, see Fig. 18a), and becomes negative thereafter.

If a constant electric field $\mathcal{E}$ is applied along the SL direction and no scattering is present, the semiclassical equation of motion $\hbar(\partial k_z/\partial t) = q\mathcal{E}$ implies that k_z changes linearly with time. Since

* This is known as the tight-binding approximation and is a reasonable description of semiconductor superlattices if the barriers are not too narrow. The single-well wavefunctions $\chi_n^{(m)}(z)$ from different wells are not quite orthogonal and Eq. 22 is valid only to the extent that the overlap between $\chi_n^{(m)}(z)$ and $\chi_n^{(m+1)}(z)$ can be neglected.

$v(k_z)$ is periodic, carriers execute oscillatory motion. After reaching the miniband edge at $k_z = \pi/d$ they are Bragg-reflected to $k_z = -\pi/d$ by the periodic SL potential (see Fig. 18a and Problem 9). These Bloch oscillations exist for any periodic potential, including that of the original semiconductor lattice a_0. They cannot be observed in bulk semiconductors because collisions typically return the carriers to the bottom of the band long before they can complete one period. That is, the scattering time τ is insufficiently long for realistic electric fields $\mathcal{E}$. A periodic potential with period $d >> a_0$ is required to relax the constraint on τ. Early on, high-frequency ultrasonic waves propagating through the semiconductor were suggested as a possible realization of such a potential[45] and it was the pioneering suggestion of Esaki and Tsu to employ superlattices for this purpose that opened the modern era of heterostructure bandgap engineering.[46] In that celebrated paper, the effects of a finite scattering time τ on the average drift velocity v_D of electrons propagating in a 1D superlattice with dispersion given by Eq. 22 was evaluated classically:

$$v_D = \int_{t=0}^{\infty} e^{-t/\tau} a(t) dt \quad , \tag{25}$$

where $a(t) + a[k_z(t)]$ is the acceleration of the miniband electron in the superlattice direction. Using the tight-binding approximation of Eq. 22, they obtained v_D in terms of $\mathcal{E}$, τ, SL period d, and m^*_{SL} (see Appendix B),

$$v_D = \frac{\hbar}{m^*_{SL} d} \frac{\xi}{1 + \xi^2} \tag{26}$$

where $\xi = q\mathcal{E}\tau d/\hbar$. The average drift velocity peaks at $\xi = 1$, that is when the electric field $\mathcal{E} = \hbar/q\tau d$. Beyond this point, increasing $\mathcal{E}$ results in a lower v_D because, on average, more and more carriers reach the negative-mass region of $E(k_z)$. As a result, the *I-V* characteristic should exhibit NDR.

Although it might appear that the NDR regime can be reached simply by increasing $\mathcal{E}$, this is not the case. The above analysis breaks down in high electric fields, where, due to Zener tunneling, the single-band approximation is no longer valid. It is, however, true that the constraint on τ for observing NDR is easier to achieve by a factor of 2π (Problem 9) than for Bloch oscillations. Nonetheless, these effects have proved elusive in *I-V* measurements, because of scattering, Zener tunneling between different minibands, and, particularly, electric field domain formation due to space-charge instabilities associated with the nonlinear current flow through the SL.[47] For this reason, while *I-V* nonlinearities in SL transport have been observed and attributed to the excursion of carriers into the negative-mass regions of the dispersion,[48] these nonlinearities have not been used in devices to date.

Thus far our discussion has considered electric fields $\mathcal{E}$ that are weak in the sense that the carriers are essentially delocalized along the SL and obey the dispersion of Eq. 22. In higher fields, the *n*th miniband breaks up into a set of discrete levels, separated by energy intervals $q\mathcal{E}d$, with wavefunctions centered in different wells and extending over $\Delta_n/q\mathcal{E}d$ periods.[49,50] This so-called Wannier-Stark ladder of states, illustrated in Fig. 19a, forms for all values of $\mathcal{E}$ but becomes physically meaningful only when adjacent ladder states can be resolved: $q\mathcal{E}d > \hbar/\tau$ which is the Bloch oscillation criterion again. As soon as the extent of the Wannier-Stark wavefunctions falls below N periods, they no longer reach from one end of the superlattice to another. For any dc current to flow some scattering process becomes necessary. The current will remain small until $\mathcal{E}$ brings into resonance Wannier-Stark states arising from different minibands, that are by then confined to individual wells:[51] $q\mathcal{E}_j d = E_j - E_1$, j = 2, 3 ... . At these sharply defined values of $\mathcal{E}_j$, the current can flow by sequential tunneling between different Wannier-Stark states in adjacent wells, followed by relaxation to a lower-lying state (see Fig. 19b). Ignoring possible series resistance outside the SL, the *I-V* curve should then exhibit peaks at $V = Nq\mathcal{E}_j d$, followed by NDR regions where current flow once again requires scattering or some inelastic mechanism.

A particularly interesting case of the latter is photon emission in the regime where $\mathcal{E} > \mathcal{E}_j$, which was proposed by Kazarinov and Suris decades ago as a system capable of voltage-tunable lasing.[51] The scheme is shown in Fig. 19c and the photon energy is $\hbar\omega = q(\mathcal{E} - \mathcal{E}_j)d$, tunable in the infrared by the applied voltage and appropriate choice of SL parameters. The problem with this exciting possibility is the same as with observing *I-V* peaks at $V = Ne\varepsilon_j d$, and even the Esaki-Tsu NDR at low ε: all of these schemes rely on a uniform electric field ε extending through the superlattice. At the same time, current flow through the SL leads to dynamically stored space-charge densities in the various quantum wells that produce nonuniform ε. Devices that operate in the NDR regions of their *I-V* characteristics are particularly susceptible to the electric field breaking up into high- and low-field domains.[48] For this reason, voltage-controlled lasing illustrated in Fig. 19(c) has not been observed and it is not clear whether it can be observed even in principle. Another problem with experimental measurements on current-carrying superlattices has been the impedance matching of the SL to the Ohmic contacts, discussed in Appendix C.

On the other hand, the alignment provided by ε_j between different Wannier-Stark states in adjacent wells can also turn the biased SL into a lasing medium, provided that at least some fraction of the $E_2 \varnothing E_1$ relaxation processes is radiative — cf. Fig. 19b. The voltage tunability of the emitted radiation is now lost, since $\hbar\omega = (E_2 - E_1)$, which is set by the SL parameters. Also, since the lower level in the radiative transition is called upon to supply the higher level in the downstream well, population inversion is difficult to achieve. On the other hand, the device need not operate in the NDR region of the *I-V* curve, so the problem of maintaining uniform alignment between adjacent periods of the SL becomes more tractable. Recently, infrared lasing in a conceptually similar device — the quantum cascade laser (QCL), based on intersubband transitions in a modified SL structure — has been achieved.[52] A more detailed discussion of the QCL, including its output characteristics as well as the structural design required to overcome domain formation and establish population inversion, will be discussed in Section 5.4.4.

5.2.5 RT Nanostructures And The Coulomb Blockade

If a double-barrier RT structure with well width L_W is etched into a sufficiently narrow pillar or biased to a narrow effective size by a lateral gate, new effects come into play. The first and more obvious is the possibility of lateral size quantization in the quantum well. Current fabrication techniques can only produce structures with lateral extent $L >> L_W$, while gate-induced electrostatic confining potentials are much weaker than the heterostructure potential V(z), so lateral quantization will be much weaker. To a good approximation each of the 2D subbands E_n in the well will give rise to a series of fully quantized, atomic-like states E_{nm}, where m labels the different states due to the lateral confining potential V(x,y). In principle, the same lateral quantization into 1D subbands could apply to the doped emitter and collector regions. However states in these regions are broadened by impurity scattering and their description by a 3D density of states is often a good approximation. Moreover, the confining potential in the emitter and collector regions is usually much weaker than that in the undoped well due to the screening by charged impurities.*

The RT transport resulting from alignment of the occupied emitter states with discrete quantum dot levels in the well can be treated within the usual sequential tunneling formalism[8,53] but with a new effect. The charging energy U required to transfer even a single electron from the emitter into the well becomes significant for small L. If the charging energy is ignored, the situation is shown in Fig. 20a. Since the lateral confining potential V(x, y) changes between the emitter and well, $\mathbf{k}_\perp$ is no longer a conserved quantity and only energy conservation holds as a tunneling selection rule. As the bias V lowers E_{11} below E_F in the emitter, tunneling through this single state becomes possible — this defines the threshold V_{th}. At higher V, additional tunneling channels open up. The resulting I-V will exhibit a rising staircase of step-like features, with bias spacing corresponding to the energy separation of the levels.[54] The strength of these features depends on the transmission coefficient of the emitter barrier $T_E(V)$ and also on the degeneracy of the E_{1m} states, which may be large if V(x, y) is approximately parabolic. Finally, NDR is not expected in the I-V characteristic, since $\mathbf{k}_\perp$ conservation no longer cuts off the tunneling through higher-lying E_{1m} states when E_{11} drops below the bottom of the occupied states in the emitter. Instead, the I-V

* The lateral confining potential V(x,y) is determined by a self-consistent electrostatic potential with boundary conditions set either by the pinning of the Fermi level at the semiconductor-air interface in etched pillars or by V_G in gated structures (in addition to the Schottky barrier in metal-gated structures or the built-in pn junction voltage in implanted structures).

should become nonlinear whenever the density of levels changes appreciably, for example when E_{2m} levels arising from the second subband become accessible.

This picture of tunneling into a quantum dot would be quite unpromising from the device standpoint. However, the charging energy U associated with electronic transport through an RT nanostructure is important. If L is small, the energy $U = q^2/2C_w$ associated with the tunneling of a single electron into the well, where C_w is the effective capacitance of the quantum dot, can appreciably alter the alignment of E_{nm} with emitter E_F. This effect is illustrated in Fig. 20b. A simple, geometric estimate of the capacitance is $C_w \approx \varepsilon_s L^2/d$, where d is the effective collector barrier thickness (if the depletion in the collector electrode is negligible, $d = L_B$). For current to flow, at least one electron must tunnel into the dot. So, V_{th} shifts to higher bias by the single-electron charging energy U. The shift in the bias of other step-like features depends on the average occupation of the well by electrons, which is determined by the transmission ratio T_E/T_C of the emitter and collector barriers.[55] If $T_E/T_C << 1$ (the case for a symmetric double-barrier structure once it is biased to $V > V_{th}$), the occupation of the well by more than one electron at a time is rare and all the step-like features in the *I-V* corresponding to additional channels coming into resonance will be shifted by U. On the other hand, if $T_E/T_C >> 1$, each available level is occupied most of the time, so the opening of every additional channel requires sufficient biasing to overcome the charging energy — this is the so-called Coulomb blockade of tunneling. For example, the second current step requires V to lower E_{12} below emitter E_F by at least $2U$ to surmount the energy barrier due to the simultaneous occupation of the well by two electrons. An additional complication is that the charging energy will vary with electron number because of electron-electron interactions in the dot and also because the effective dot size L changes. Since both the empty ($T_E/T_C << 1$) and occupied ($T_E/T_C >> 1$) regimes are accessible in the same asymmetric double-barrier RT nanostructure by changing the bias polarity, such devices have been studied to probe the energy spectrum of quantum dots with and without electron-electron interactions.[54,55]

It is the charging energy required to change the electron occupation of the dot together with the possibility of tuning the energy alignment of the dot levels with respect to the emitter electron reservoir via a third terminal that makes RT nanostructures promising for devices. Schematically, a three-terminal nanostructure involves nothing but the addition of a gate electrode that can change the potential between the quantum dot and the emitter (see Fig. 20), but is sufficiently isolated from the dot to prevent any possibility of electron transfer from the gate. Then, if the device is biased by V_E near a voltage step corresponding to the addition of another electron to the dot, a small change in V_G can tune the occupation of the dot. This controls the current through the dot, resulting in a single-electron transistor. Because of the fabrication difficulties associated with vertical RT nanostructures, gate control of single electron tunneling has proved easier in the planar geometry. The dot and the controlling electrode are defined by electrostatic metal gates deposited on top of a high-mobility 2D electron gas heterostructure. A top view of the gated structure is shown in the inset of Fig. 21. The outside gates are biased into deep depletion, forming a small island of 2DEG connected to the reservoirs by tunneling barriers. As we had seen in the context of planar double-barrier RT structures (see Fig. 8), these islands are necessarily large and the electrostatic barriers are wide and low, so the energy quantization in the island is weak. But, this is an advantage in the context of Coulomb blockade devices, because the energy spectrum of the dot is now entirely defined by the charging energy U. The gate electrode can alter the effective size and capacitance of the island. As long as the island capacitance C_w is small while the island size is relatively large, L ~ 0.1—1 μm, the I-V characteristic as a function of V_G should show regularly spaced steps corresponding to the adding of electrons to the island. At low temperatures very regular conductance ($G + \partial I/\partial V$) peaks have been observed in such structures,[3] an example is shown in Fig. 21.

In principle, precise single-electron control over electron occupation or the tunneling transport in small quantum dots or islands has led to many proposals of logic and memory circuits based on single-electron transistors (SETs) and other devices.[56] To some extent, single electron devices can be considered the logical endpoint of miniaturization-driven semiconductor technology. The main difficulty from the practical standpoint is posed by the extremely stringent fabrication requirements on large-scale SET circuitry, especially at non-cryogenic temperatures. Currently, SET characteristics, like the data in Fig. 21, are measured at $T < 1$ K, to ensure the condition $U = q^2/2C_\mathrm{w} >> kT$. Clearly, device sizes will need to be reduced by orders of magnitude before higher temperature operation can be contemplated. For $T = 4.2$K the charging energy must be certainly larger than 1 meV. This requires a capacitance $C_\mathrm{w} < 10^{-16}$ F, a very stringent condition. It is imperative that there be no parallel capacitance due to leads or other electrodes. Note that a simple thin wire has an intrinsic capacitance of about 10^{-16} F per micron. It is also not clear that semiconductor SET realizations have any advantages over metal tunnel junctions for most proposed devices: the first observation of Coulomb-blockade phenomena[57] and the first SET with voltage gain[58] both employed small Al tunnel junction capacitors. One specific application for which the single-electron transistor appears promising is the construction of precision current standards. In a gated 2DEG island, by sequentially lowering and raising the emitter and collector barriers at small V_E in the Coulomb-blockade regime, where only one excess electron can occupy the island, the transfer of one electron per cycle of barrier biasing can be achieved.[59] If the barriers are cycled at frequency f, the emitter-collector current is given by $I = qf$, making for a very precise current source. It is anticipated that such a device may provide a new metrological current standard, although single-electron transfer along a chain of small metallic islands may prove a more successful implementation.[60]

5.3 HOT-ELECTRON STRUCTURES

5.3.1 Hot Electrons In Semiconductors

In many of the resonant tunneling structures discussed in the preceding section electrons (or holes) are injected into the collector region with energies that are several kT above the collector Fermi energy E_F, where T is the lattice temperature. These electrons are clearly not in thermal equilibrium with the lattice and their occupation of available states is not described by the standard Fermi-Dirac function $f_{FD}(E)$. Further, their velocity distribution in the direction of current flow is strongly peaked, at least in the immediate vicinity of the collector barrier, making up a "ballistic" electron packet. As the electrons propagate into the collector, the velocity distribution broadens due to scattering, resulting in a distribution that can be taken as Maxwellian and parametrized by an effective temperature $T_e > T$. In either case, the electrons are "hot" with respect to the lattice.

Another possible reason for carrier heating is a strong electric field $\boldsymbol{\mathcal{E}}$ in some region of the device. Depending on the energy relaxation time, a large fraction of the carriers can be accelerated into states of high kinetic energy. As in the course of ongoing miniaturization device dimensions shrink at a faster rate than various electrode voltages, the internal fields rise and carrier heating becomes more significant. Thus, oxide damage by hot electrons accelerated by the large lateral $\boldsymbol{\mathcal{E}}$ at the channel drain has become a major reliability issue as silicon FET's are scaled down. Undesirable though carrier heating may be in standard silicon technology, a number of devices based on hot electrons have been proposed. In this chapter we will focus on injection devices, where the hot carriers are physically transferred between adjacent semiconductor layers.* As we shall see, although the first hot-electron injection devices[61] were proposed as far back as 1960, the abrupt heterojunction interfaces and doping profiles made possible by modern epitaxy have greatly widened hot-electron device possibilities.

Let us consider the two principal techniques for producing hot carriers by electric current — ballistic injection and electric field heating — and the resulting carrier distributions in more detail. Ballistic injection by thermionic emission from a wider-bandgap semiconductor and by tunneling into states of high kinetic energy is illustrated in Figs. 22a and 22b respectively. Immediately upon injection, the corresponding velocity distribution is also shown (cf. Problem 11). As the carriers propagate away from the injection point, their energy and velocity distribution will change and broaden. Given the initial velocity and momentum distribution $f(\mathbf{r},\boldsymbol{v},t = 0)$ and ignoring the possibility of interband transitions (like electron-hole recombination), the evolution of $f(\mathbf{r},\boldsymbol{v},t)$ with time as a function of spatial position can be determined by solving the Boltzmann transport equation. In the simplest case of parabolic dispersion, the Boltzmann equation has a physically transparent form:

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla_{\mathbf{r}} f + \boldsymbol{a} \cdot \nabla_{\boldsymbol{v}} f = \left(\frac{\partial f}{\partial t}\right)_{\mathrm{coll}} \tag{27}$$

with semiclassical equations of motion given by

$$\boldsymbol{v} = \hbar\mathbf{k}/m^* \qquad\qquad m^*\boldsymbol{a} = q(\boldsymbol{\mathcal{E}} + \boldsymbol{v} \text{ x } \boldsymbol{B}) \tag{28}$$

where $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{B}$ are the electric and magnetic fields, m^* is the effective mass in $E(\mathbf{k}) = \hbar^2\mathbf{k}^2/2m^*$, and $\boldsymbol{a}$ is the acceleration.* The collision term on the right of Eq. 27 represents all scattering processes, including phonon emission and absorption, impurity scattering, electron-electron interaction, and so on. It can be formally defined as the integral of the scattering probability W($\mathbf{k}'$,$\mathbf{k}$) between states characterized by wavevectors $\mathbf{k}$ and $\mathbf{k}'$ over the first Brillouin zone multiplied by the appropriate occupation probabilities. A great simplification results in the relaxation time approximation, which replaces the entire collision term by $-(f-f_0)/\tau$, where τ characterizes the time it takes for the distribution to relax to its equilibrium value f_0 — the Boltzmann equation is then no longer an integral equation. This is rarely possible, however, because different characteristics of the distribution function relax at different rates. Because of

* Transferred electron devices, like Gunn oscillators, employ hot-electron scattering into lower mobility satellite valleys of the semiconductor dispersion $E(\mathbf{k})$ — they are covered in Chapter 6.

this,one usually defines separately the momentum $\tau_{\mathbf{k}}$ and energy τ_e relaxation times in terms of W(**k**',**k**).[62] Still another complication is that the transition matrix element W(**k**',**k**) depends on the electron energy in different ways for different scattering mechanisms. For states of higher energy new scattering mechanisms set in: optical phonon emission, impact ionization, and so on. Finally, since the collision integral extends over the Brillouin zone, it depends on the density of states available for scattering (i.e. on the explicit form of $E(\mathbf{k})$, leading to ever greater complexity as energy increases and larger sections of the Brillouin zone become accessible), and even on **k** orientation. For this reason, Eqs. 27 and 28 are rarely tractable analytically and numerical Monte Carlo techniques are often employed.[63]

Of course, there exists one limiting case which avoids the difficulty altogether: ballistic motion, in which collisions are negligible. This limits the critical dimension of any device to $v_z\tau_{\mathbf{k}}$, where v_z is the (high) injected electron velocity and $\tau_{\mathbf{k}}$ is the momentum relaxation time. This is the preferred operating regime of ballistic hot-electron transistors (HETs), in which electrons are injected into a narrow base layer of length L_B. Control over injection energy in heterostructures (see Fig. 22) can generate a narrow hot-electron distribution, centered around a high velocity normal to the base layer, as shown in Fig. 23. As long as $L_B < v_z\tau_{\mathbf{k}}$, a large fraction of the hot electrons will traverse the base without scattering.

The energy and velocity distribution of hot carriers created by a strong electric field $\boldsymbol{\mathcal{E}}$ is necessarily rather different. Before the electric field is applied, the carriers are in equilibrium with the lattice. The field accelerates the carriers according to Eq. 28, shifting the distribution function $f(\mathbf{r},\boldsymbol{v},t)$ away from equilibrium. Since the scattering mechanisms depend on the carrier energy E, the same difficulty in solving the Boltzmann equation arises. However, momentum relaxation times $\tau_{\mathbf{k}}$ are generally much shorter than τ_e, so numerous electron-electron collisions can establish a quasi-equilibrium within the electronic system that is effectively decoupled from

* In the case of an arbitrary dispersion $E(\mathbf{k})$, the equations of motion become $\boldsymbol{v} = \hbar^{-1}\ {}_{\mathbf{k}}E(\mathbf{k})$ and $\hbar\partial\mathbf{k}/\partial t = q(\boldsymbol{\mathcal{E}} + \boldsymbol{v} \times \boldsymbol{B})$, with corresponding changes in Eq. 27.

the lattice. In this limit, one can define the effective electron temperature T_e from the average energy $<E>$ of the electron ensemble: $<E> = 3kT_e/2$. The resulting hot-electron distribution is shown in Fig. 23.

Evidently, the effective electron temperature T_e depends on the magnitude of the electric field in a complicated fashion given by the solution of Eqs. 27 and 28. Once again, Monte Carlo simulations are generally required, especially in the presence of heterostructure barriers that are necessary for real-space transfer (RST) devices based on field-induced electron heating.[64] In these devices, carriers are heated by an electric field applied parallel to a heterostructure barrier V_0. If T_e becomes sufficiently high, some fraction of the electron distribution will acquire enough energy to spill over the barrier and transfer to a different region of the structure, which may have a different mobility or a separate electrical contact. Even though only electrons in the high-energy tail of the hot-carrier distribution function can surmount V_0 for a given heating field $\boldsymbol{\mathcal{E}}$, they are quickly replenished (on a scale of τ_e), so the RST process can be fast and efficient.

It should be emphasized that both ballistic injection and real-space transfer devices involve the injection of nonequilibrium carriers over (or through) heterostructure barriers into adjacent layers of the structure. Conceptually, the real distinction lies in the different hot-carrier distribution functions illustrated in Fig. 23. Operationally, the three-terminal implementations of ballistic and RST devices employ rather different controlling electrode geometries. As will become clear, there is a strong parallel between ballistic hot-electron transistors and standard bipolar transistors: the hot-electron current across the base is controlled by a smaller base current of equilibrium carriers moving in different portions of **k** space. Base transparency for ballistic carriers (i.e., the base transport factor α_T of a bipolar transistor) relies on the short time required to traverse the narrow base compared to the momentum relaxation time $\tau_{\mathbf{k}}$. Also, as in bipolar transistors, speed limitations arise from the base traversal time $\approx L_B/v_z$ (here hot-electron transistors really shine by virtue of large injected v_z) and the R_BC delay associated with charging the base-emitter and base-collector capacitances through a finite lateral base resistance.

Real-space transfer devices, on the other hand, have no ready analog among standard transistors. Two-terminal versions, which rely on RST of hot carriers to a lower-mobility region of the structure to produce NDR in the two-terminal *I-V* characteristic, are essentially similar to Gunn oscillators. In three-terminal versions, RST of hot carriers to a region that can be contacted separately is employed to control the transferred current by a heating field. As we shall see, three-terminal RST devices possess unusual terminal symmetries, arising from the insensitivity of hot-carrier distributions to the heating field polarity, that can be exploited for increased functionality.

5.3.2 Ballistic Injection Structures

Figure 24 shows a schematic band diagram of a ballistic HET based on tunneling injection and implemented in GaAs/AlGaAs.[65] Hot electrons are injected at an energy $\Delta \approx qV_{\mathrm{BE}}$ with respect to the Fermi level in the heavily doped base (held at ground potential), traverse the base and are collected after surmounting the collector barrier, which is a function of collector bias V_{CB}. There is a clear analogy between this type of HET and the current-controlled three-terminal RT structure of Fig. 9. However, since quantization in the base region of Fig. 24a is not required, HET base width can considerably exceed the $L_{\mathrm{B}} \leq 100$ Å strong-quantization condition. This leads to lower base resistance R_{B} and, hence, reduces the time delays associated with charging the base-emitter and base-collector capacitance. Of course, Eq. 19 for the $R_{\mathrm{B}}C$ time delay still applies, so ballistic HET design involves a trade-off between low base resistance, which requires large L_{B} and base doping, and high base transport factor α_{T}, which requires short L_{B} and minimal base scattering. For amplification, the highest α_{T} is achievable for $V_{\mathrm{CB}} > 0$. In that configuration, electrons that have experienced scattering in the base can still be collected. If V_{BE} is much larger than the collector barrier height, the emission of one or more optical phonons still leaves the electron with enough kinetic energy to reach the collector, provided the direction of its velocity has not changed. Fortunately, optical phonon emission by high-energy electrons is a predominantly forward scattering process and the effective $\tau_{\mathbf{k}}$ is not as short as it is for electrons just above the optical phonon emission threshold. Still, the very high α_{T} required to produce competitive differential current gain $\beta + \alpha_T/(1-\alpha_T)$ is difficult to realize in ballistic HETs even with the shortest L_B allowed by Eq. 19.

First, there is the matter of quantum mechanical reflection at the collector barrier. In our discussion of tunneling we found that given a collector barrier height Φ_C, there is a non-zero reflection probability $R(E_z)$ for incident electrons with kinetic energy $E_z > \Phi_C$. In the case of rectangular heterostructure barriers, $R(E_z)$ remains significant unless $E_z >> \Phi_C$, except at some special values of E_z that depend on the barrier parameters (cf. Problem 2). This difficulty can be circumvented to some extent by grading the collector barrier (Problem 12, see also Fig. 24a for an example), but reducing $R(E_z)$ to nearly zero at moderate hot-electron injection energies is problematic.*

It might appear that α_T can be improved by increasing V_{BE} and hence the injection energy, since this should reduce the collector reflection coefficient and the base transit time simultaneously. Unfortunately, this approach runs into the second important limitation of HET performance. If the electron kinetic energy exceeds the energy of the satellite valleys in the dispersion (e.g., the L valley in GaAs, which lies 0.3 eV above the Γ conduction band minimum), the high density of final states leads to very efficient intervalley scattering by phonons and impurities. At electron energies below the intervalley scattering threshold, but above the optical phonon emission threshold (36 meV in GaAs) the dominant form of phonon interaction is emission of polar optical phonons, a process that has only a weak dependence on the carrier energy. At the same time, ionized impurity scattering in the heavily doped base is minimized with increased injection energy. Consequently, the optimum V_{BE} for high gain is just below the threshold for intervalley scattering. In GaAs/AlGaAs HETs analogous to Fig. 24(a), this has limited the highest observed gain to $\beta \approx 10$ at low temperatures,[65] corresponding to $\alpha_T \approx 0.9$. Higher gains of $\beta \approx 30$ ($T = 77$ K) have been observed in similar HET structures with a narrow $L_B = 200$ Å pseudomorphic InGaAs base, because of larger the Γ-L energy separation.[66] More recently, similar structures have yielded $\beta \approx 10$ at $T = 300$ K,[67] which may be approaching the limit for HET structures grown on GaAs or InP substrates.

Compared to heterojunction bipolar transistors (HBTs), ballistic injection HETs suffer from relatively low base transport factors. Furthermore, the upper limit Eq. 19 places on the lateral base resistance R_B is more stringent in the case of HETs with tunneling injection, since the emitter barrier must be fairly thin to keep the tunneling current high. This renders HETs noncompetitive for most device applications. They have proved valuable for research into non-equilibrium carrier transport, however. The use of an injection HET as a hot-electron spectrometer is illustrated in Fig. 24a, while a representative set of hot-electron energy distributions is shown in Fig. 24b. The idea is to measure the collector current I_C as a function of $V_{CB} < 0$ at a fixed V_{BE} (which sets the injection energy). In a certain range of V_{CB} the collector barrier height varies linearly, $\delta\Phi_C \sim \delta V_{CB}$, and $\partial I_C/\partial V_{CB}$ is proportional to the number of carriers arriving at the collector barrier with $E_z = \Phi_C$. To the extent that V_{CB} does not affect the hot-electron energy distribution in the base, the injected distribution, and the above-barrier quantum mechanical reflection, one can deduce the mean-free-path l as a function of injection energy and correlate the dynamics of energy loss with various scattering mechanisms.[65,69] For example, the main peak in Fig. 24b corresponds to hot-electrons arriving at the collector barrier without a single phonon-emission event, with $l \approx 1000$ Å. This is quite remarkable considering that these electrons had to traverse not only the doped GaAs base but also the AlGaAs collector barrier. Lateral hot-electron spectrometers have been constructed in 2DEG using electrostatic barriers,[70] as in Fig. 8. The minimal scattering in 2DEG at low temperatures leads to considerably longer $l \approx 0.5$ μm. As a result, very high $\beta > 100$ was measured at $T = 4.2$ K in devices with $L_B \approx 1700$ Å. More importantly, similar structures have provided the laboratory for studying the physics of ballistic transport in small systems.[71]

[*] It should be emphasized that this envelope function approach to heterostructure barrier transmission is only valid if the heterojunction semiconductors are lattice matched and have similar dispersions in **k** space. An example is the GaAs/$Al_xGa_{1-x}As$ heterostructure with $x \leq 0.45$, where electrons in both materials move in the same Γ valley of the conduction band. Although frequently used, the applicability of envelope function transmission calculations to other types of heterostructures is questionable.[68]

Finally, it should be noted that hot-carrier injection has been employed to good effect in HBTs by designing structures with a wider bandgap emitter, schematically illustrated in Fig. 25 for an AlInAs/InGaAs HBT lattice-matched to InP. The advantages conferred by hot-electron effects are several. Electrons are injected by thermionic emission over the emitter-base barrier at an energy $\Delta \approx 0.5$ eV above the conduction band edge in the *p*-InGaAs base. Here the purpose of ballistic injection is to shorten the base traversal time by replacing the relatively slow diffusive motion by faster ballistic propagation. Since there is no collector barrier to surmount, scattering does not degrade the transport factor α_T, as it does in unipolar hot-electron transistors. Further, the fact that the injected velocity distribution is sharply peaked in the direction perpendicular to the base aids device scaling by minimizing lateral excursion into the extrinsic base region. As discussed in more detail elsewhere in Chapter 1, hot-electron HBTs exhibit high gain and high-speed operation, with f_T exceeding 100 GHz at room temperature.[72] Note that in optimized transistors the base is so thin ($L_B << 1000$ Å) that the diffusive transport time across the base is also shorter than 1 ps and hence high-speed HBT operation is not in itself evidence of ballistic transport.*

If the transport across the HBT base is truly ballistic and the velocity distribution of injected hot electrons is sufficiently narrow, $\Delta v_z/v_z << 1$, the result is a coherent transistor that is predicted to have gain above f_T.[73] When the injection from the emitter is varied at some frequency *f*, an electron density wave of wavelength $\lambda = v_z/f$ is set up in the base. The minority-carrier density wave is screened by majority carriers and the base remains neutral everywhere. Such is the situation with all bipolar transistors. Neglecting recombination, the base current in the HBT flows only to neutralize variations in the overall number of minority carriers. At low frequencies, $\lambda >> L_B$, the minority charge in the base increases and decreases in phase with injection. This leads to a characteristic frequency roll-off of the current gain $\beta \sim 1/f$ and the characteristic value for the $\beta = 1$ frequency cutoff $2\pi f_T = v_z/L_B$. The functional form of this roll-off begins to change when the wavelength of minority-carrier density wave becomes comparable to the base width. If L_B is an integer multiple of λ, which corresponds to *f* being an integer multiple of v_z/L_B, there is no change in the total minority charge, as the electron density wave goes through the base. Hence, the collector current modulation is accomplished, in this case, with no high-frequency base current input, leading to $\beta\ \varnothing$. Obviously, in a real device β will be limited by the recombination current, scattering-induced damping of the density wave, and the finite velocity distribution width Δv_z.

Still, as long as a high degree of coherent transport is maintained, β(*f*) will peak at integer multiples of $2\pi f_T$, leading to current gain above the usual cut-off frequency (we present a more detailed discussion of this effect in Appendix D). Moreover, the transistor power gain has been predicted[73] to peak at multiples of πf_T, exhibiting two peaks for each peak in β. The difficulty with implementing such a transistor lies in maintaining coherence across the base. Since optical phonon emission is a very effective scattering mechanism, the injection energy $\Delta = m^* v_z^2/2$ should remain below the optical phonon energy but still much larger than kT, implying cryogenic operation. Further, device parasitics can wash out the gain peaks above f_T. To date, the proposed coherent transistor has not been experimentally characterized.

5.3.3 Real-Space Transfer Structures.

Proposals of generating NDR in the two-terminal *I-V* characteristics of a device by real-space transfer (RST) between semiconductor layers of high and low mobility date back several decades.[74] These ideas received further development in the context of modulation-doped multiquantum well GaAs/AlGaAs heterostructures[75] and first experiments using such structures were carried out in the early 1980's.[76] A schematic illustration of the two-terminal GaAs/AlGaAs RST structure is shown in Fig. 26. If longitudinal electric field $\mathcal{E}_x$ is small, electrons reside in the undoped GaAs quantum wells and the source-drain I-V_D depends linearly on $\mathcal{E}_x$ with the slope

[*] It is an exceedingly difficult matter to demonstrate the ballistic nature of transport in a given transistor and claims of ballistic HBT operation have always been controversial. An easier, but indirect approach involves measurements of gain in a series of transistors with variable base width L_B, which allows one to discriminate between ballistic and diffusive transport mechanisms by appealing to a theoretical model.

determined by high GaAs mobility. However, as $\mathcal{E}_x$ is increased, the power input into the electron distribution exceeds the rate of energy loss into the lattice by phonon emission, and the electrons heat up to some field-dependent temperature T_e. At sufficiently high T_e, there will be partial transfer to the doped AlGaAs layers over the heterostructure barrier Φ, where the mobility is much lower due to heavy doping and higher m^*. The two-terminal I-V_D then exhibits NDR with the peak-to-valley ratio determined by the magnitude of the transferred electron density and the ratio of the mobilities in the GaAs and AlGaAs layers (cf. Problem 13 for an analytically tractable model). The analogy to the Gunn effect, where high-mobility carriers are scattered to low-mobility valleys in momentum space, is obvious. In fact, Gunn effect and RST mechanisms are competing processes that depend on the relative magnitude of the barrier height Φ and the valley separation.

In addition to the interplay between transfer mechanisms, a realistic treatment of electron heating in an RST structure involves the formation of longitudinal electric field domains, redistribution of electrons both vertically and laterally, the self-consistent electric fields $\mathcal{E}_z$ in the transfer direction, and quantum mechanical reflections at heterostructure interfaces. As long as the electron ensemble in a given GaAs channel of Fig. 26 can be described by a local temperature $T_e(x)$, which is a function of position between source and drain, the density of the RST electron current $J(x)$ can be estimated by the thermionic emission formula

$$J(x) - \frac{qn(x)v(T_e)}{L_W}\, e^{-\Phi/kT_e} \tag{29}$$

where $n(x)$ is the sheet density in the quantum well, L_W is the well thickness and $v(T_e) = (kT_e/2\pi m^*)^{1/2}$. Obviously, this current depends exponentially on T_e. A semiclassical treatment of RST between two layers with a conduction band discontinuity Φ (e.g., one of the GaAs/AlGaAs heterointerfaces in Fig. 26a, involves solving the appropriate Boltzmann equation, on either side of the junction, with the appropriate boundary conditions that include quantum reflection by the barrier. The transverse electric field $\mathcal{E}_z$ must be calculated self-consistently from the Poisson equation that includes the electron density $n(z)$ and the fixed charges that are present (e.g., ionized impurities in the AlGaAs). If the band bending due to $\mathcal{E}_Z$ is much smaller than the heterostructure barrier Φ, the $\mathcal{E}_Z$-dependent terms in the Boltzmann equation may be dropped, leaving $\mathcal{E}_X$ as the only electric field in the problem. Still, the collision integral on the right of Eq. 27 compels the use of Monte Carlo techniques. The results of such calculations as applied to RST structures are collected in Ref. 64.

Like two-terminal resonant tunneling diodes, two-terminal RST structures illustrated in Fig. 26 are potentially useful as high-frequency oscillators. The figures of merit in this case are the speed with which electrons cycle between high (GaAs) and low (AlGaAs) mobility layers and the magnitude of the resulting NDR in the *I-V* characteristic. Unfortunately, while the hot-electron transfer time to the AlGaAs can be quite short, the return process of relatively cold electrons by thermionic emission over the space-charge potential barrier of the ionized donors is much longer.[77] An additional difficulty is the formation of macroscopic traps in the AlGaAs due to fixed charge inhomogeneities: these potential pockets effectively collect transferred hot electrons and present a higher barrier to their return. In fact, the RST transfer times are typically longer than the momentum-space transfer times. Further, the maximum high-frequency power that can be extracted from an RST oscillator is limited by the peak-to-valley ratio of the NDR, but a large PVR is obtained only if the mobilities of the two layers differ by orders of magnitude (Problem 13). Such a large mobility ratio is no easier to engineer in RST structures than in homogeneous multiple valley semiconductors. For these reasons, two-terminal RST oscillators do not appear to offer any significant advantages over Gunn oscillators and have not benefited from much experimental development. What makes RST structures considerably more interesting from the device standpoint is the possibility of extracting the transferred hot-carrier current via a third terminal, resulting in an RST transistor (RSTT).[78]

Figure 27 shows a schematic cross-section and the corresponding band diagram of an RSTT device implemented in GaAs/AlGaAs. The source and drain contacts are to a high-mobility GaAs channel, while the collector contact is to a doped GaAs conducting layer that is separated from the channel by a large heterostructure barrier. An electron density is induced in the source-drain channel by a sufficient positive collector bias V_C with respect to the grounded source, but no collector current I_C flows because of the AlGaAs barrier at $V_\mathrm{D} = 0$. As V_D is increased, however, a drain current I_D begins to flow and the channel electrons accordingly heat up to some effective temperature $T_\mathrm{e}(V_\mathrm{D})$. This electron temperature determines the RST current injected over the collector barrier and the injected electrons are swept into the collector by the V_C-induced electric field, giving rise to I_C. Thus, transistor action results from control of the electron temperature T_e in the source-drain channel that modulates I_C flowing into the collector electrode. In contrast to the two-terminal device of Fig. 26, the RST current is removed from the drain current loop, leading to very strong NDR in the I_D-V_D curve, with room-temperature PVR reaching 160 in GaAs/AlGaAs devices[79] similar to Fig. 27. Subsequent improvements in RSTT design included structures with InGaAs channels, either lattice-matched to InP substrates or pseudomorphically strained to GaAs, taking advantage of the lower electron effective mass and higher Γ-L valley separation in InGaAs. More importantly, these devices used epitaxial rather than alloyed contacts to the source-drain channel.[80] The drain I_D and collector I_C characteristics of such an RSTT are shown in Fig. 28: the PVR in the I_D-V_D characteristics reaches 7000 at $T =$ 300 K. Another version had a top-collector design with self-aligned collector regions[81] that avoid the vertical overlap between the source and drain contacts evident in Fig. 28. The result is reduced parasitic capacitance between the source, drain, and the collector: a current gain cut-off frequency $f_\mathrm{T} > 50$ GHz was reported. Finally, there have been recent reports of δ-doped pseudomorphic InGaAs/GaAs RSTT[82] with high channel mobility, a PVR > 10^5, and a transconductance of 23.5 S/mm at $T = 300$ K, as well as silicon-compatible hole-based RSTT with SiGe channel and collector layers.[83]

As might be expected from our discussion of the RST process between two layers, theoretical modeling of RSTT devices is extremely involved and Monte Carlo simulations are required for quantitative comparison with experiment.[64,84] Some qualitative insight can be gained by assuming that the RST current $J(x)$ exists only within a certain domain of the concentrated longitudinal electric field along the channel. In this high-field domain we will take the electron temperature T_e as uniform and assume that the channel carriers move at their saturation velocity v_{sat}, so $I_D = qn(x)Wv_{sat}$, where W is the device width and the diffusion component of I_D is ignored. Current continuity in the source-drain loop gives $dI_D/dx = -J(x)W$. Substituting Eq. 29 for $J(x)$ one obtains that both $n(x)$ and $I_D(x)$ decrease exponentially with a characteristic length

$$\lambda = \frac{v_{sat} L_W}{v(T_e)} e^{\Phi/k_B T_e} \tag{30}$$

For high T_e, λ becomes quite short, making the diffusion component non-negligible, but the exponential decay of $I_D(x)$ remains qualitatively unchanged (Problem 14). The remaining difficulty is the estimation of T_e for a given drain bias V_D. This has been done semi-analytically by assuming a uniform electric field $\mathcal{E}_x$ in the channel and taking into account only two energy-loss mechanisms: optical phonon emission and electron RST to the collector.[85] In this model, at sufficiently high V_D the region of high T_e becomes much larger than λ and, by virtue of Eq. 30, I_D becomes vanishingly small. However, since Monte Carlo studies show that **k** space transfer into lower mobility satellite valleys is a dominant scattering mechanism at high fields, simple energy-loss models are of limited applicability.

An interesting aspect of RSTT devices is the nature of their intrinsic speed limitations. The two contributing factors are the time-of-flight delays associated with space-charge-limited current and the finite time required to establish the hot-carrier ensemble of temperature T_e. Significantly, the relevant length entering the time-of-flight delay is not the gate length of a standard FET, because once the high-field domain is established the speed with which T_e can be modulated is not limited by the source-drain transit time. Instead, the relevant length is the extent of the high-field domain in the channel plus the thickness of the potential barrier separating the channel from the collector. Barrier thicknesses in RSTTs are ~10^3 Å, as are the high-field domains when T_e is large (Problem 14). As a result, the time-of-flight delay should be in the 1 ps range, competitive with state-of-the-art conventional transistors. As for the establishment of an effective T_e in the hot-carrier distribution, the relevant mechanisms are optical phonon emission, **k** space scattering, and electron-electron interaction, which might be the dominant mechanism for reaching quasi-equilibrium in the high-energy tail of the distribution. Once again, Monte Carlo simulations[88] of RSTT structures indicate that equilibration of hot-electron ensembles takes less than 1 ps, at least at high electron concentrations and operating voltages.

Let us briefly consider possible applications of RSTTs. Obviously they can be used as conventional high-speed transistors, in which case the figures of merit are the transconductance $g_m + \partial I_C/\partial V_D$ (at fixed V_C) and current-gain cut-off frequency f_T. Like resonant tunneling devices, RSTT combinations can be used for memory and logic elements by virtue of the strong NDR in the source-drain circuit. Further, since the source and drain contacts of an RSTT are fully symmetric, these devices have additional logic functionality. A single RSTT like that shown in Fig. 27 can perform an exclusive-OR (XOR) function, because the collector current I_C flows if source and drain are at different logic values, regardless of which is "high". This and related logic implementations will be discussed later in the chapter. Finally, by changing the doping and design of the collector region, light-emitting operation of RSTT structures has been demonstrated in the InGaAs/InAlAs material system.[86] The only change from Fig. 27 is the opposite doping in the *n*-InGaAs channel and the *p*-InGaAs active region grown on top of the p^+-InGaAs collector. As in a standard RSTT, hot electrons are injected by RST over the InAlAs barrier, but then they recombine radiatively with holes in the active region. As long as radiative recombination in the channel is negligible, the optical output is insensitive to the parasitic leakage of collector holes into the channel. This means that the optical on-off ratio is directly determined by I_C and hence the device works like a light-emitting diode with built-in logic functionality.

5.3.4 Resonant Hot-Electron And Bipolar Transistors.

As we have seen, three-terminal RT structures in which the control electrode directly modulates the alignment of the resonant subband and the emitter are difficult to fabricate. An alternative approach is the incorporation of a double-barrier RT potential into the emitter of a hot-electron transistor.[87] A schematic band diagram of the resonant hot-electron transistor (RHET) is shown in Fig. 29. Its operation essentially combines the resonant emitter I_E-V_{BE} characteristic with the current gain β available in the HET. Consider the collector current I_C as a function of base-emitter voltage V_{BE} at some fixed base-collector voltage V_{BC}. At small V_{BE} the emitter RT structure is below threshold, the emitter current is negligible and the collector current I_C consists of the small thermionic emission current over the collector barrier Φ_C. At larger V_{BE} a resonant current flows through the emitter, injecting hot electrons at $\Delta \approx qV_E$ above E_F in the base. Given proper design, with $\Phi_C < \Delta \leq \Gamma$-L energy separation in the base, a large fraction α_T of the injected electrons traverses the base and contributes to I_C. The large Γ-L separation makes InGaAs heterostructures on InP substrates advantageous for the implementation of RHETs, as discussed in Section 5.3.1. As before, the current gain $\beta + \alpha_T/(1 - \alpha_T)$ is limited by the hot-electron mean-free-path in the heavily doped base, but room-temperature $\beta \approx 10$ has been reported in InGaAs/AlAs/InGaP RHET structures on InP substrates.[67] Finally, as V_{BE} biases the emitter RT diode beyond V_P, the emitter current drops. The corresponding PVR in I_C will approximately reproduce the PVR of the emitter diode, although changes in α_T as a function of injection energy might alter this result somewhat. Peak-to-valley ratios of approximately 10 have been reported in the I_C-V_{BE} characteristics of RHETs at both $T = 300$ K and $T = 77$ K.[67,87]

Very similar characteristics can be obtained by inserting a double-barrier RT diode or several cascaded RT diodes on the emitter side of the emitter-base junction in an *npn* bipolar transistor.[88] Such structures were fabricated in InGaAs/AlInAs: the operation is analogous to RHET, except that the emitter bias V_{BE} divides between the RT diodes in the emitter and the emitter-base *np* junction to maintain current continuity. As long as $V_{BE} < V_{bi}$ of the *np* junction, the emitter current increases as in a conventional bipolar transistor, with somewhat higher emitter series resistance due to the RT diodes, and the current gain β is large. Beyond flatband, $V_{BE} \geq V_{bi}$, most of the additional V_{BE} drops in the RT diodes and I_E exhibits one or more NDR regions when the diodes are biased beyond V_P. Consequently, I_C also exhibits peaks as a function of V_{BE}. The gain in the NDR regions is typically lower than in low V_{BE} characteristic, because the hole current into the emitter keeps increasing with V_{BE}. Since the electron current into the base drops at $V_{\mathrm{BE}} > V_{\mathrm{P}}$, the result is lower emitter efficiency. The multipeaked I_{C} characteristic of a bipolar transistor with two RT diodes in the emitter has been used as a frequency multiplier. By driving the base with an ac signal of frequency f and sufficient amplitude to bias both RT diodes through their resonances, signals at $3f$ (for sawtooth input) and $5f$ (for sinusoidal input) were generated with reasonable conversion efficiency.[89]

Like RSTTs, resonant hot-electron and bipolar transistors exhibit higher logic functionality in a single device, illustrated schematically in Fig. 30. Given a common-emitter I_{C}-V_{BE} characteristic with reasonable PVR shown in Fig. 30a, the output I_{C} can be high when $V_{\mathrm{BE}} = V_{\mathrm{high}} < V_{\mathrm{P}}$, but low when $V_{\mathrm{B}} = 0$ or $2V_{\mathrm{high}}$ (where the RHET is in the negative transconductance regime). As a result, an exclusive-NOR (XNOR) function can be easily implemented in a single device, as shown in Fig. 30b. With the emitter grounded and two inputs to the RHET base, V_{OUT} will be high when one of the base inputs is high and low otherwise. Room temperature XNOR gate operation with a reasonable V_{OUT} voltage swing has been demonstrated, using a device layout similar to Fig. 30b.[72] In addition to the necessary resistor network, a drawback of these designs is that unless the PVR in the I_{C} characteristic is very large, there is still power dissipated in the collector resistor when both base inputs are high. In a single device this added power dissipation can be minimized simply by downscaling the area and reducing I_{C}, but this is not possible in large circuits where I_{C} must charge interconnect capacitances.

5.4 DEVICE APPLICATIONS

5.4.1 RT Oscillators.

Despite the high speed and functionality offered, in principle, by quantum-effect and hot-electron devices, their technological applications thus far have been few. In some cases, the main obstacle has been room-temperature operation, in others the difficulty of large-scale fabrication or the integration of nonconventional devices into standard technology. For these reasons, many of the device applications discussed below exist only as laboratory demonstrations. While possibly relevant in the relatively distant future, when their edge over conventional devices might become compelling — at very small device dimensions *L*, cryogenic temperatures *T*, or whatever other design criteria future technology may require — few quantum and hot-electron devices offer a sufficient advantage today. One happy exception is the use of RT diodes as solid-state high-frequency oscillators. The advantages of two-terminal RT oscillators include relative ease of fabrication, reasonable output power, and high maximum oscillation frequencies $f_{\max}$

compared to competing microwave tunnel and transit-time diodes.

Figure 31 shows the simplest equivalent circuit of a two-terminal diode oscillator with a static *I-V* characteristic that includes an NDR region described by peak (V_P, I_P) and valley (V_V, I_V) points in both voltage and current. This equivalent circuit has been successful in the analysis of tunnel diodes with *I-V* characteristics similar to the RT diode of Fig. 6. The real part of the equivalent circuit impedance R_{eq} is given by:

$$R_{eq} = R_S + \frac{-R_D}{1 + (\omega R_D C_D)^2} \qquad (31)$$

where $-R_D = (V_V - V_P)/(I_V - I_P)$ is the negative diode resistance; C_D is the diode capacitance; and R_S is the series lead resistance. For steady-state oscillation, R_{eq} must be negative, so from Eq. 31 the cut-off frequency f_{max} is found to be

$$f_{max} = \frac{1}{2\text{š}\, R_D C_D} \sqrt{\frac{R_D}{R_S} - 1} \ . \qquad (32)$$

To increase f_{max}, the quantities to minimize are then the parasitic series resistance R_S and diode capacitance C_D. A sharp current drop after V_P and a high PVR are also helpful in minimizing R_D and hence increasing f_{max}, but there is the competing requirement of maximizing high-frequency output power P_{max}. Although the exact value of P_{max} depends on the actual *I-V*) in the $V_P < V < V_V$ region, generally $P_{max} \sim (V_P - V_V)(I_P - I_V)$, making both PVR and a high current density essential for good oscillator performance.

By analogy with tunnel diodes, Eqs. 31 and 32 have been employed in the design of RT diode oscillators with empirical parameters (e.g., taking for C_D the measured two-terminal emitter-collector capacitance) and extended to include collector transit- and tunneling-time effects.[19] However, the equivalent circuit of Fig. 31 is physically unsatisfactory. The current flowing in an RT diode depends on the alignment of the emitter and the 2D subband in the well, with the tunneling current densities into and out of the well balancing in steady state, $J_{in} = J_{out}$. It is difficult to construct a useful equivalent circuit for RT diodes either in the coherent or in the sequential picture. The main difficulty lies in the unknown energy distribution of the dynamically stored charge density $\sigma_W = qn_W$ in the well, which makes it impossible to describe J_{out} as a unique function of the electrostatic potential difference V_C between the well and the collector.*

A reasonable and tractable model arises if one assumes that carriers equilibrate in the well. Then, the collector current can be described by a function of σ_W and V_C, and its small variations about a steady state can be written in the form:**

$$\delta J_{out} = \delta\sigma_W/\tau + \delta V_C/R_C \qquad (33)$$

where τ is the lifetime of the carriers in the well, while the collector resistance R_C reflects thedependence of tunneling rate on the well-collector potential difference due to changes in the collector barrier shape. Variation in the stored charge density σ_W and its time dependence obey Gauss' and Kirchhoff's laws:

$$\delta\sigma_W = C_E\,\delta V_E - C_C\,\delta V_C, \quad \partial(\delta\,\sigma_W)/\partial t = \delta J_{in} - \delta J_{out} \qquad (34)$$

where C_E and C_C are the emitter-well and well-collector capacitances, respectively. By definition, $\delta V_E + \delta V_C = \delta V$, the variation in total emitter-collector bias V.

Not far from the tunneling resonance, J_{in} is a unique function of V_E, which determines the emitter-well alignment, $\delta J_{in} = \delta V_E/R_E$. Combining this with Eqs. 33 and 34 for J_{out}, one obtains

$$\partial(\delta\sigma_W)/\partial t = -\delta\sigma_W/\tau_{eff} + C_G\delta V/\tau_G \qquad (35)$$

where the geometric quantities C_G and τ_G are defined as

$$\tau_G = \frac{R_E R_C(C_E + C_C)}{R_E + R_C} \qquad C_G = \frac{R_E C_E - R_C C_C}{R_E + R_C} \qquad (36)$$

and τ_{eff} is given by

$$\tau_{eff} = \frac{\tau_G\tau}{\tau_G + \tau} \quad . \qquad (37)$$

* The Fermi level difference — which is the true meaning of "voltage" — between the well and the collector cannot even be meaningfully defined without some assumption about the electron distribution in the well.

** This model has been suggested by P. Solomon in a private communication (1995) to one of us (SL). Rigorously, parameters τ and R_C can be defined as follows: $\tau^{-1} + (\partial J/\partial\sigma_W)$ at fixed V_C and $R_C^{-1} + (\partial J/\partial V_C)$ at fixed σ_W. Their interpretation as lifetime and resistance is only approximate.

It is the effective time constant of Eq. 37 that determines the diode dynamics. If the applied voltage V changes abruptly by δV, the charge density in the well will evolve exponentially towards the new steady-state value $(\sigma_W + \delta\sigma_W)$ with a time constant given by τ_{eff}. The magnitude of $\delta\sigma_W$ is described by an effective capacitance C_{eff}:

$$\delta\sigma_W = C_{\mathrm{eff}}\delta V = (C_G\tau_{\mathrm{eff}}/\tau_G)\delta V \tag{38}$$

Evidently, C_{eff} can be either positive or negative, depending on the sign of C_G in Eq. 36. The value of this capacitance is irrelevant to the dynamics of the variation.

If the $\delta\sigma_W/\tau$ component of J_{out} were absent from Eq. 33, the RT diode would be in a true linear response regime and could be rigorously described by an equivalent circuit consisting of R_E paralleled with C_E in series with R_C paralleled with C_C. Solving such an equivalent circuit would give Eq. 35, but with τ_{eff} replaced by τ_G. This, however, is not a good approximation for real RT diode oscillators.* In the operating regime V_C is large, the value of τ is shorter than τ_G, and hence $\tau_{\mathrm{eff}} \approx \tau$. This is particularly true in structures with large undoped spacer regions on the collector side of the double barrier. In such structures, the collector barrier transparency becomes only weakly dependent on V_C, effectively making R_C in Eq. 33 very large. The key parameter for high speed is τ, which should be minimized by making the collector barrier as transparent as possible while keeping the sharpness of the 2D quantization sufficient for NDR in the *I-V* characteristic.

Figure 32 summarizes experimentally measured, room-temperature oscillator performance of high-speed RT oscillators fabricated in different material systems: GaAs/AlAs, InGaAs/AlAs, and InAs/AlSb.[19] While the power density P_{max} available in GaAs/AlAs is limited by the relatively low PVR at $T = 300$ K, InGaAs/AlAs RT oscillators exhibit good output power, while

* Once again, the point is that these diodes are operated at high V_C and hence far from equilibrium. By contrast, near equilibrium J_{out} can be described by the Landauer formula, $\delta J_{\mathrm{out}} = \delta V_C/R_C$, where R_C is a function of the collector barrier transmission coefficient T_C. But at high V_C it is the lifetime τ and not R_C that describes the tunneling rate of the carriers in the quantum well.

InAs/AlSb devices show promise for submillimeter wave ($f > 300$ GHz) performance and hold the record for solid-state oscillator frequency at 712 GHz.[90] No other solid-state sources generate coherent power at submillimeter fundamental frequencies. One possible application of such devices is for low-noise local oscillators in high-sensitivity radiometers. A more detailed discussion of microwave diode performance and applications is available in Chapter 6 of this book.

5.4.2 Memories.

Several approaches have been pursued in constructing memory circuits from quantum-effect and hot-electron devices. Single device memories can be constructed from asymmetric two-terminal RT diodes with a bistable I-V characteristic shown in Fig. 5 by biasing the device below V_P in the bistable region and changing the memory state using voltage pulses. Alternatively, an ordinary RT diode with an NDR I-V characteristic in series with a load resistor R_L can be dc biased into a regime with two stable bias points, as shown in Fig. 33a. Once again, voltage pulses can be used to change the memory state. The drawback of such memories is that at least one of the memory states corresponds to a high current through the RT diode. The resulting power dissipation is prohibitively large compared to the larger and more complex conventional memory designs. One approach for overcoming the power dissipation problem is to increase the functionality of the RT memory by employing a multistate design. As discussed in Section 5.2.3, proper design of a cascaded RT structure with N diodes results in a multipeaked I-V characteristic with N current peaks of approximately equal magnitude evenly spaced by ΔV_P (see Fig. 16 for a cascaded RT structure with $N = 8$ and $\Delta V_P \approx 0.95$ V). By biasing such a structure with a constant operating current I_{OP} supplied by an FET, as shown in Fig. 33b, the output node V_{OUT} can be at any of the $(N + 1)$ stable voltage points. Switching between V_{OUT} states is performed by setting an input voltage via a momentarily enabled write line. As soon as the write line is disabled, the cascaded RT will adjust to the nearest stable V_{OUT} value and maintain it indefinitely, leading to an $(N + 1)$-state memory.[44] However, this type of multistate memory still dissipates $P_{OUT} =$

$I_{OP}V_{OUT}$ of power, with average $<P_{OUT}> \approx I_{OP}\Delta V_P N/2$. Minimum power dissipation requires high PVR, since $I_{OP} > I_V$, where I_V is the worst-case valley current among the N peaks, and small ΔV_P. Also, $<P_{OUT}>$ increases with N and the accumulated series resistance from the N RT diodes enters into the RC switching time delay. Furthermore, practical implementation of circuits based on multistate RT memories requires stringent reproducibility of characteristics between different devices. Ultimately, the quantifiable advantage of a multistate memory is the reduction of the number of elements necessary to store the same amount of information by a factor of $\log_2(N + 1)$ for an $(N + 1)$-state device replacing a binary flip flop.

A different approach is the series connection of two negative-resistance devices, which can be RT diodes, RSTTs, RHETs, or any other device with an NDR *I-V* characteristic. In fact, many of these devices were proposed in the 1960's with tunnel diodes in mind.[91] If the total applied bias V_{DD} exceeds roughly twice the critical voltage V_P for the onset of NDR in one device, the voltage division between the two devices becomes unstable because of current continuity. One of the two devices takes on most of the applied bias, thereby determining the voltage of the middle node V_{OUT}. This is illustrated by the load-line construction in Fig. 34: operating points A and C are stable, while B is unstable. As V_{DD} is ramped up beyond $2V_P$, the system will choose one of the two stable points depending on which of the devices goes into NDR first — either because of a fluctuation or, realistically, because of a slight difference in the *I-V* characteristics. Switching between the two states can be accomplished either by controlling the parameters of the two devices (in the case of three-terminal RT structures or RSTTs) or by changing the middle node bias via an additional electrode. Significantly, the current flowing through the two NDR devices connected in series when $V_{DD} > 2V_P$ depends on the valley current, (see Fig. 34). If the PVR of the devices is large, the current will be small regardless of whether the circuit is in state A or C.

A schematic memory constructed from two RT diodes in series with an additional control electrode separated from the middle node by a tunnel barrier is shown in Fig. 35a. Such devices have been fabricated in the InAs/AlSb/GaSb material system,[92] which provides good room-temperature PVR in the *I-V* characteristics, after their original demonstration at $T = 77$ K using InGaAs/AlAs/InP RHETs.[93] As in Fig. 34, biasing the RT diodes in series with $V_{DD} > 2V_P$ switches one of them to the valley region, so a small current $I << I_P$ flows and V_{OUT} is close to either V_{DD} or ground. To change V_{OUT}, a voltage V_{IN} is applied to the sub-collector control electrode, causing current I_C to flow between the middle mode and the sub-collector. Since this current flows by tunneling through the sub-collector barrier, it increases rapidly with the potential difference $|V_{OUT} - V_{IN}|$. When the sub-collector current reaches I_P, V_{OUT} switches, resulting in hysteresis in the V_{OUT} *vs.* V_{IN} characteristic, shown schematically in Fig. 35b. The magnitude of the V_{OUT} voltage swing depends in the RT diode *I-V* characteristic and can nearly reach V_{DD} if the PVR of the diodes is high enough. Conversely, the required switching bias V_{IN} depends on the single barrier I_C-($V_{OUT} - V_{IN}$) curve. A smaller subcollector barrier requires a smaller $|V_{OUT} - V_{IN}|$ difference to reach I_P and switch the middle node, in effect squeezing the hysteretic loop in Fig. 35b along the horizontal axis. Crucially, until a switching V_{IN} pulse is applied to the control electrode, the total current flowing through the memory is limited from below by the valley current in the RT *I-V* characteristic, since the additional subcollector leakage current can be made very small by designing an appropriate subcollector tunnel barrier.

Memory cells based on two RT diodes or RHETs in series, along the lines of Fig. 35, are smaller than standard CMOS designs. Thus, the RHET version[97] operating at $T = 77$ K claims an order of magnitude in area savings, while the room-temperature RT diode implementation[96] offers area savings of 2-4 depending on whether the diodes are laid out horizontally or stacked vertically. The remaining issue for large-scale memory arrays is power dissipation. Since a reasonable I_P is needed to charge up the interconnect capacitance and the standby power dissipation depends on the valley current, the relevant figure of merit is the available PVR. By using polytype InAs/GaSb/AlSb RT diodes, a room temperature PVR of nearly 20 has been achieved,[96] but much higher PVR appears necessary to achieve acceptable power dissipation. As a result, NDR-based memories with their exotic materials appear unlikely to challenge CMOS in high-density memory applications. On the other hand, they may be suitable for applications that require small amounts of memory and can afford higher static power consumption.

5.4.3 Logic Elements.

In addition to memory devices, the use of RHETs and RSTTs for logic elements has been proposed and, in some cases, demonstrated by a number of groups. In particular, the compact XNOR functionality of RHETs, illustrated in Fig. 30b, has been employed in the design of elementary logic components, such as latches and full-adders.[94] A typical building block in such designs is the three-input majority logic gate, shown in Fig. 36, which uses three RHETs. By using a four resistor summing network connected to the emitter-base diode of the first RHET, the operating point lies below V_P in the I_C-V_{BE} characteristic if fewer than two of the three inputs is high and above V_P if two or three inputs are high — cf. Fig. 30a. The second RHET senses whether the output of the first RHET is above or below V_P. The third RHET, which is larger and delivers higher I_C, increases the output current drive of the logic gate. By combining this majority logic gate with two XNOR gates made of two RHETs each, a full adder operating at T = 77 K was demonstrated.[99] Room temperature operation of a hybrid full-adder incorporating bipolar transistors with and without RT diodes in the emitter-base junction has also been reported.[95] Such designs accomplish the required logic function with a reduced number of transistors. Note, however, that the reduced transistor counts available in RHET and resonant bipolar logic designs come at the expense of fabricating additional resistors. But since thin-film resistor fabrication in microelectronic technology requires additional processing steps and real estate, it is not clear that such circuits provide great area-saving advantages. Further, the impact of all these resistors on the switching speed and propagation delay in such circuits has not been characterized to date. Finally, the integration of these circuits with conventional silicon technology is problematic, while the possibility of a stand-alone quantum device logic circuitry built in III-V semiconductors competing with the ever-advancing silicon CMOS logic is extremely remote.

Integration of high functionality devices with conventional logic circuitry is considerably easier when they are built in Si/SiGe heterostructures. As discussed in Section 5.2.2, silicon-based RT diodes and transistors[23-25,37] perform acceptably at low temperatures only, because of the low Si/SiGe barriers. On the other hand, there has been recent progress in Si/SiGe real-space transfer devices. The drain I_D-V_D and collector I_C-V_D characteristics of a p-Si/SiGe RSTT at room temperature[83] are shown in Fig. 37. The structure of the device is identical to the GaAs/AlGaAs RSTT of Fig. 27, but with $Si_{0.7}Ge_{0.3}$ layers comprising the channel and collector regions, separated by an undoped 3000 Å Si barrier. Negative collector bias induces a hole density in the channel, while V_D drives a source-drain current and heats the holes. As V_C increases, the drain characteristic exhibits RST-induced NDR, with PVR slightly exceeding two at V_C = -5.5 V. A further increase in V_C results in increasing leakage current due to cold hole tunneling. While the PVR is greatly inferior to that available in III-V RSTTs, it is sufficient to implement a single-device XOR gate: with V_C = -4.0 V and V_S, V_D = 0 or -4 V for low and high inputs, the gate has a 10 dB on/off ratio at T = 300 K and a 65 dB on/off ratio at T = 77 K. For a source-drain separation L = 0.5 μm, this device had a current-gain cut-off frequency f_T = 6 GHz. Finally, simulations indicate that there is considerable room for improvement of the drain circuit PVR by reducing the barrier thickness and fine-tuning some of the structural parameters.[83]

Since the input source and drain terminals in an RSTT are completely symmetric, even higher logic functionality can be obtained by increasing the number of input terminals. For example, three input terminals permit a single-device implementation of an ORNAND gate. Depending on whether the control input is high or low, the output current behaves as either a NAND or an OR function of the other two inputs.[96] The device structure is illustrated in Fig. 38a, where the control input V_3 is subject to periodic boundary conditions for ORNAND functionality. The logic operation of this device at T = 77 K, V_C = -5 V and V_{low}, V_{high} = 0, -3 V respectively, is shown in Fig. 38b.[87] The same device is expected to perform at room temperature if either the cold hole-leakage current or the channel length L is reduced.

In principle, Si/SiGe RSTTs are compatible with silicon microelectronics, although the epitaxial deposition of pseudomorphic SiGe layers for the active regions obviously requires additional fabrication steps and reduces the thermal budget available for subsequent processing. The trade-off between the added fabrication complexity and the area savings due to the higher functionality will decide the future of silicon-based RSTTs. If technological evolution eventually brings silicon technology to cryogenic (T = 77 K) operating temperatures, the chances of silicon-based quantum-effect and hot-carrier devices will improve dramatically.*

5.4.4 Quantum Cascade Laser.

As we have seen, prospects of quantum-effect or hot-electron devices replacing conventional semiconductor technologies — whether digital logic and memory chips or analog amplifiers and switches — are hampered by difficulties with room temperature operation, device reproducibility, and fabrication complexities. The advantages of novel devices, typically higher functionality and speed, have not and for the foreseeable future will not displace standard FET and bipolar technologies. On the other hand, such devices as submillimeter RT diode oscillators and Coulomb blockade current sources are poised for success in niche applications, precisely because conventional solid-state alternatives do not exist. Yet another, potentially more significant area where quantum-effect devices are about to make their mark is solid-state laser sources in the mid-infrared, operating in the λ = 4—12 μm wavelength range, where current technology relies on low-power and low-yield lead-salt devices. The recently developed quantum cascade laser (QCL),[52,97] combines resonant tunneling and hot-electron aspects in a device structure that makes full use of heterostructure bandgap engineering. The lasing occurs in an intersubband transition and λ is tunable in the infrared region via the quantum well design.

Figure 39 shows a partial band diagram of the QCL gain region together with its output characteristics. The entire QCL structure is grown by MBE, lattice-matched to an n-InP

* Any silicon-based heterostructure with higher barriers than the 0.2-0.3 eV available in Si/SiGe, would greatly brighten the prospects of such devices.

substrate. The total gain region comprises 25 stages of an InGaAs/AlInAs coupled-quantum-well active region followed by a superlattice Bragg reflector. The undoped, coupled-quantum-well active region is designed for the following 2D subband structure under the operating applied bias of ~10^5 V/cm: an upper-lying E_3 subband with a wavefunction $|\chi_3(z)|^2$ concentrated in the first well, and two lower-lying subbands E_2 and E_1 concentrated in the first and second well respectively (see Fig. 39). The radiative transition is $E_3 \varnothing E_2$, so the laser output energy is $\hbar\omega = E_3 - E_2$. As in all lasers, the radiative transition has to compete with other $E_3 \varnothing E_2$ relaxation mechanisms. In the QCL, the main nonradiative relaxation mechanism involves optical phonon emission. This process is relatively slow, however, because $\hbar\omega >> \hbar\omega_{\mathrm{opt}}$ and hence $E_3 \varnothing E_2$ relaxation requires large in-plane momentum transfer. On the other hand, since $E_2 - E_1 \approx 30$ meV $\approx \hbar\omega_{\mathrm{opt}}$, $E_2 \varnothing E_1$ relaxation by optical phonon emission is very fast. The superlattice (SL) downstream of the coupled quantum well completes the set of conditions necessary for population inversion between E_3 and E_2. In the InGaAs/AlInAs SL Bragg reflector region, the well and barrier widths are adjusted in pairs[98] to compensate for the applied bias and give rise to an approximately flat miniband structure shown in Fig. 39. The sequence that contributes to photon emission is as follows: an electron in the E_3 state of a given active region relaxes radiatively via the $E_3 \varnothing E_2$ transition, then relaxes via the $E_2 \varnothing E_1$ transition by phonon emission, tunnels from the E_1 state into the lowest SL miniband of the Bragg reflector, and finally tunnels into the E_3 state of the next active region downstream. There the process is repeated, until the electron cascades down all of the 25 stages and is collected in the doped optical cladding layers that sandwich the active region.

In order to achieve optical gain, population inversion between the E_3 and E_2 states in the QCL active regions is needed. This requires rapid removal of electrons from the lower E_2 state and long nonradiative lifetime in the upper E_3 state. As described above, optical phonon emission vacates the E_2 state very quickly, but is much slower to vacate the E_3 state. The other factor required for a long nonradiative lifetime in the E_3 state is the prevention of direct tunneling out of the active region. However, as shown in Fig. 39, direct tunneling out of the E_3 state into the SL region is blocked because E_3 lines up with the SL minigap. That is, the SL acts as a Bragg reflector (see discussion in Section 5.2.3).

Finally and crucially, the layers near the middle of the SL region are doped in the 10^{17} cm^{-3} range to provide carriers for injection into the coupled quantum well regions and ensure the overall charge neutrality under operating conditions, when a current density J flows through the structure. To avoid space charge accumulation associated with J, a reservoir of fixed positive charge is needed to compensate the current-carrying electrons in each QCL stage. The role of the SL regions is best appreciated by comparing the QCL structure of Fig. 39 with the conceptually similar SL structure of Fig. 19. Even if SL were biased into resonance between adjacent wells, rather than into the NDR regime originally proposed by Kazarinov and Suris, tunneling from the higher-lying E_2 states into the continuum would work against population inversion. Also, a constant electric field in an undoped SL would be impossible to maintain in the presence of significant current. Introduction of a doped SL region was the key design innovation that led to the successful implementation of the first QCL.[53] Subsequent QCL designs relied on the SL regions to suppress tunneling from the upper level of the radiative transition into the continuum. To this end, the SL regions were designed to serve as electronic Bragg reflectors with minigap in the range of energy near the E_3 state, cf. Fig. 39. Note that since effective Bragg reflection requires very accurate grading of layer widths in the SL regions, this elegant approach places stringent demands on band structure modeling and layer control by molecular beam epitaxy.

The lasing characteristics shown in Fig. 39 at lower left corresponds to a $\lambda \approx 4.5$ µm laser with cleaved facets operated in pulsed mode, but continuous mode operation at $T = 140$ K and pulsed operation at room temperature has recently been reported in an optimized QCL structure.[99] The power output is quite high, but the threshold current density J_{th} increases rapidly with temperature, reaching 3×10^3 A/cm^2 at $T = 100$ K. If the $E_3 \varnothing E_2$ radiative transition is treated as an atomic two-level system, the degradation of performance at higher T can be attributed to reduced population inversion due to temperature-induced backfilling of the E_2 level from the electrons in the doped SL regions.[99] The realistic situation is certainly more complicated. A recent theoretical analysis of gain in QCL pointed to the importance of hot-electron effects in the presence of in-plane subband nonparabolicity.[100] Indeed, not only do electrons tunnel into E_3 with a considerable spread in energy of in-plane motion, but those that relax nonradiatively to the E_2 subband are initially very hot — on average, $(\hbar\omega - \hbar\omega_{opt}) \approx 250$ meV = 3000 K above the bottom of the subband for the $\lambda = 4.5$ μm transition. If the in-plane subband dispersion is nonparabolic (certainly true that far above the subband minima), $\hbar\omega = (E_3 - E_2)$ changes as a function of in-plane energy and therefore the gain depends on the difference between hot-electron distributions in these subbands. The shapes of these distributions are radically different in the limits of low and high sheet-carrier concentrations n_D per QCL period, provided by the doped SL regions. For $n_D << 10^{11}$ cm^{-2}, the rate of electron-electron collisions is low and the distribution functions are not Maxwellian. The dominant scattering process is then due to optical-phonon emission within the same subband. It is reasonable to assume that the distribution of electrons tunneling into the upper E_3 subband from the SL miniband is in quasi-equilibrium with the lattice temperature T.* After a nonradiative $E_3 \varnothing E_2$ intersubband transition, the lower subband electrons are in states of high kinetic energy. Subsequently, they cascade down emitting optical phonons and partially escaping into the SL miniband reservoir. The resulting distribution is given by a quasi-discrete ladder with the occupation probabilities decreasing towards the bottom of the E_2 subband, as if the effective temperature were negative.

The calculated gain spectra for low n_D are shown in Fig. 40a for several lattice temperatures T. The peak gain is substantial even at $T = 300$ K. Note that no overall population inversion between the E_3 and E_2 subbands is assumed, $\xi + n_3/n_2 = 1$. In the absence of lasing, ξ is determined by non-radiative kinetics as the ratio of the $E_3 \varnothing E_2$ nonradiative transition rate and the rate of carrier removal from the E_2 subband. In the low-concentration regime, the peak wavelength in

* This assumption implies a sufficiently rapid energy relaxation in transport between the QCL stages.

the gain spectra does not depend on temperature. To our knowledge, this regime has not yet been realized experimentally.

On the other hand, the gain spectra calculated for the high n_D limit,[101] where it is safe to assume Maxwellian hot-electron distributions, also show a range of positive gain, but the peak gain is much lower. Moreover, the peak shifts to longer wavelengths at higher T. These effects have been observed experimentally in existing QCL structures. In the high-concentration regime, the range of positive gain for $n_2 > n_3$ arises entirely from the nonparabolicity. In fact, if quasi-Fermi levels E_{F3} and E_{F2} are introduced to characterize the hot-electron distributions in the two relevant subbands, positive gain occurs when $\hbar\omega < (E_{F3} - E_{F2})$, a condition that is familiar from the theory of conventional semiconductor lasers. By contrast, the existence of positive gain in the low-concentration limit does not rely on nonparabolicity and persists to concentrations far from overall population inversion.[100] Room temperature gain spectra calculated for several values of ξ at low n_D are shown in Fig. 40b. Implementation of the low-concentration regime appears to be a promising strategy for maximizing QCL performance.

5.5 SUMMARY AND FUTURE TRENDS

In this chapter, we have reviewed some of the recent research in the area of quantum-effect and hot-electron devices. It has not escaped the reader that while many of these devices are quite successful according to some (but not all) benchmarks, none has found large-scale commercial application to date. A decade or two ago this situation could perhaps be attributed to the immaturity of the field and the need for further development. But today, a quarter century after the first experimental demonstrations of both resonant tunneling diodes and hot-electron transistors, this excuse is no longer available. It is imperative to confront the basic issue: what are these devices good for?

To be sure, exotic device research can be proud of its scientific accomplishments. Fascinating new physics has been discovered, with the fractional[102] quantum Hall effect serving as a prime example, and many previously obscure issues have been elucidated. The basic effects relevant to electronic devices, such as tunneling in heterostructures, ballistic transport, carrier heating, and charge injection across potential barriers, are no longer manifested by hardly discernible blips in low-temperature characteristics. They are now available as robust and reproducible phenomena, with on/off ratios quite adequate for the implementation of useful devices. Despite these successes, or perhaps precisely because of them, the general attitude toward the exotic device research has hardened into a widespread skepticism. If these devices have not made it, despite the considerable world-wide effort, why should we throw good research funding after the bad?

In our opinion, there is, indeed, little chance that either resonant tunneling or hot-electron devices will form the basis of a successful stand-alone technology. On the other hand, they have significant potential in connection with other technologies, such as optoelectronic integrated circuits that are likely to benefit from the introduction of ultra-fast functional elements, based on resonant tunneling or hot-electron effects. The recently developed quantum cascade laser appears particularly hopeful in this regard, since it promises good performance in the mid-infrared wavelength range ($\lambda > 4$ μm). Moreover, superb frequency characteristics can be expected from this class of lasers, with modulation frequencies exceeding 100 GHz. It is generally believed that light will eventually replace electrical current as the carrier of information signals, both in computer and communications applications, currently the main drivers of innovation in semiconductor devices. Still, the "dark" age of electronics is far from over, and it is interesting to contemplate possible application of exotic devices within the context of the future evolution of circuits which operate without emitting, absorbing, or transforming light.

The evolution of semiconductor electronics has always been intimately connected with advances in materials science and technology. The first revolution in electronics, which replaced vacuum tubes with transistors, was based upon doped semiconductors and relied on newly discovered methods of growing pure crystals. Prior to the 1950s, semiconductors could not be properly termed "doped" — they were impure. Today, semiconductors routinely used in devices are cleaner, in terms of the concentration of undesired foreign particles, than the vacuum of vacuum tubes.

Subsequent evolution of transistor electronics has been associated with the progress in two areas: miniaturization of device design rules, brought about by advances in the lithographic resolution and doping by ion implantation; and development of techniques for layered-crystal growth and selective doping, culminating in such technologies as MBE and MOCVD, that are capable of monolayer precision in doping and chemical composition.

Of these two areas, the first has definitely had a greater impact in the commercial arena, whereas the second has been mainly supplying the device physics field with new systems to explore. These roles may well be reversed in the future. Development of new and exotic lithographic techniques with nanometer resolution will set the stage for the exploration of various physical effects in mesoscopic devices, while epitaxially grown devices, particularly heterojunction transistors integrated with optoelectronic elements, will be gaining commercial ground. When and whether this role reversal will take place, will be determined perhaps as much by economic as by technical factors. We believe that most significant applications of heterostructure electronics will be associated with its use in silicon electronics.

The logic of industrial evolution will motivate new paths for a qualitative improvement of system components, other than the traditional path of a steady reduction in fine-line feature size.

Miniaturization progress faces diminishing returns in the future, when the speeds of integrated circuits and the device packing densities will be limited primarily by the delays and power dissipation in the interconnects rather than individual transistors. Further progress may then require circuit operation at cryogenic temperatures and/or heavy reliance on high bandwidth optical and electronic interconnects. Implementation of optical interconnects within the context of silicon microelectronics requires hybrid-material systems with islands of foreign heterostructures grown or grafted on Si substrates. In terms of the old debate on Si *versus* GaAs, our view is that silicon is the ultimate customer for GaAs. In this scenario, the current noncompetitiveness of quantum and hot-electron devices for general purpose digital and analog electronics could give way to novel devices serving as small, highly functional application specific components that add significant value to main blocks of microelectronic circuitry.

Appendix A. Densities Of States And Fermi Integrals

Consider the phase space of a single particle in d dimensions (d = 1, 2, or 3). It contains 2d

axes, corresponding to the d coordinates and p_d momenta of the particle. It is a basic tenet of quantum statistics that a hypervolume V^d in the phase space contains $2V^d/2\pi\hbar)^d$ distinct states, where the factor of two in the numerator arises from the spin degeneracy. Thus, the density of states in the phase space is given by:

1D: $2\,\mathrm{d}L\ \mathrm{d}p/(2\pi\hbar)$ (A1a)

2D: $2\,\mathrm{d}A\ \mathrm{d}^2p/(2\pi\hbar)^2$ (A1b)

3D: $2\,\mathrm{d}V\ \mathrm{d}^3p/(2\pi\hbar)^3$ (A1c)

where dL, dA, and dV are elements of length, area, and volume, respectively.

Band theory of solids retains the same expressions A1a-c. They now describe the density of states in each band. Of course, $\boldsymbol{p}$ is no longer the electron momentum, but the crystal momentum (in terms of the wavevector $\boldsymbol{k}$ used in the chapter, $\boldsymbol{p} = \hbar\boldsymbol{k}$). Inasmuch as the occupation of different states in equilibrium depends only on their energy, it is convenient to express the density of states as a function of energy. If we define N(E) as the number of states in a given band with energy less than E, then the density of states for various dimensionalities d is given by:

1D: $g^{1\mathrm{D}}(E) + L^{-1}\ \mathrm{d}N/\mathrm{d}E$ (A2a)

2D: $g^{2\mathrm{D}}(E) + A^{-1}\ \mathrm{d}N/\mathrm{d}E$ (A2b)

3D: $g^{3\mathrm{D}}(E) + V^{-1}\ \mathrm{d}N/\mathrm{d}E$ (A2c)

Note that the density of states has different unites for different dimensionalities (cm^{-d} eV^{-1} in d dimensions). Closed-form expressions for $g(E)$ can be obtained only for simplest band structures, e. g., for isotropic bands, $E(\boldsymbol{p}) = E(p)$. For isotropic and parabolic bands, $E = p^2/2m^*$, where m^* is some effective mass, N(E) can be found explicitly by counting the states from the bottom of the band up to some crystal momentum p:

1D: $\mathrm{N}(E) = 2pL/(2\pi\hbar) = L(2m^*E)^{1/2}/\pi\hbar$ (A3a)

2D: $\mathrm{N}(E) = 2\pi p^2A/(2\pi\hbar)^2 = A(m^*E)/\pi\hbar^2$ (A3b)

3D: $\mathrm{N}(E) = 2(4\pi p^3/3)V/(2\pi\hbar)^3 = V(2m^*E)^{3/2}/3\pi^2\hbar^3$ (A3c)

Substituting Eqs. A3a-c into the appropriate expressions for $g(E)$ one obtains the following densities of states:

1D: $$g^{1D}(E) = (m^*/2E)^{1/2}/\pi\hbar \tag{A4a}$$

2D: $$g^{2D}(E) = m^*/\pi\hbar^2 \tag{A4b}$$

3D: $$g^{3D}(E) = m^{*3/2}(2E)^{1/2}/\pi^2\hbar^3 \tag{A4c}$$

The actual density n of electrons in the system (per unit length, area, or volume, as appropriate) is found by integrating the density of states multiplied by the Fermi-Dirac occupation probability $f_{FD}(E\text{-}E_F)$, as in Eq. 10. The resulting general equation,

$$n = \int_0^{\infty} g(E)\ f_{FD}(E\text{-}E_F)\ \mathrm{d}E, \quad f_{FD}(E) = (\mathrm{e}^{-E/kT} + 1)^{-1}, \tag{A5}$$

provides a relation between n and the Fermi level E_F. In general, this relation contains a Fermi integral of order s,

$$F_s(E_F/kT) \equiv \frac{1}{\Gamma(s+1)} \int_0^{\infty} \frac{E^s\,\mathrm{d}E}{1 + \mathrm{e}^{(E - E_F)/kT}}, \tag{A6}$$

where Γ is the gamma function: $\Gamma(1/2) = \pi^{1/2}$, $\Gamma(1) = 1$, $\Gamma(s + 1) = s\Gamma(s)$. It follows from Eqs. A4a-c that the Fermi integrals for 3D, 2D, and 1D systems are of order s = 1/2, 0, and -1/2 respectively. Analytic solutions of the Fermi integral are available only for integer s, so for 3D and 1D systems Eq. A5 must be evaluated numerically. In 2D, on the other hand, the density of states is constant and the appropriate Fermi integral is $F_0(\eta) = \ln[1 + \mathrm{e}^{\eta}]$, yielding

$$n = (m^*kT/\pi\hbar^2)F_0(E_F/kT) = (m^*kT/\pi\hbar^2)\ln[1 + \mathrm{e}^{(E - E_F)/kT}], \tag{A7}$$

where the prelogarithmic factor $(m^*kT/\pi\hbar^2)$ is the effective density of states in the 2D subband. A similar calculation gives Eq. 12. Note that the lower limit of integration in Eq. A5 refers to an appropriate zero energy point. In a bulk semiconductor this would be the bottom of the conduction band, in a quantum well this would be the bottom of a given 2D subband (cf. Eq. 12).

Appendix B. Drift Velocity In A Superlattice With Scattering

Take a superlattice (SL) in the tight-binding approximation, such that the dispersion along the SL direction z is given by Eq. 22. Consider the motion of an electron initially at rest at $k_z = 0$ in a constant electric field $\mathcal{E}$. As a function of k_z, the band velocity is

$$v(k_z) = - \frac{2Td}{\hbar} \sin(k_z d) \; , \qquad \text{(B1)}$$

where d is the SL period and T is the transfer integral defined by Eq. 24 (one can show by direct calculation that T is a negative quantity). Since $\hbar(\mathrm{d}k_z / \mathrm{d}t) = q\mathcal{E}$, the acceleration $a(t)$ is given by:

$$a(t) = \frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}k_z}\frac{\mathrm{d}k_z}{\mathrm{d}t} = - \frac{2Td^2}{\hbar^2} \cos(k_z d) q\mathcal{E} \; . \qquad \text{(B2)}$$

On the other hand, the effective mass $m^*(k_z)$, defined by $\hbar k_z = m^*(k_z)v(k_z)$, is

$$m^*(k_z) = - \frac{\hbar^2 k_z}{2Td\sin(k_z d)} \; . \qquad \text{(B3)}$$

Taking the k_z ∅ 0 limit of Eq. B3, we obtain the effective mass m^*_{SL} at the bottom of the miniband, $m^*_{\mathrm{SL}} = -(\hbar^2/2Td^2)$. Note that the miniband width Δ in terms of m^*_{SL} is given by

$$\Delta = 2\hbar^2/m^*_{\mathrm{SL}}d^2 \; . \qquad \text{(B4)}$$

Substituting Eq. B3 into Eq. B2, we have

$$a(k_z) = \frac{1}{m^*_{\mathrm{SL}}} \cos(k_z d)\, q\mathcal{E} \; . \qquad \text{(B5)}$$

Finally, inserting Eq. B5 into the Esaki-Tsu expression for the average drift velocity v_{D}, Eq. 25, one finds:[46]

$$v_{\mathrm{D}} = \underset{t=0}{\overset{\square}{\square}} e^{-t/\tau} a[k_z(t)]\, dt = \frac{q\mathcal{E}}{m^*_{\mathrm{SL}}} \underset{t=0}{\overset{\square}{\square}} e^{-t/\tau} \cos(\frac{q\mathcal{E}d}{\hbar}\, t)\, dt$$

$$= \frac{q\mathcal{E}\tau}{m^*_{\mathrm{SL}}} \; \frac{1}{1 + (q\mathcal{E}\tau d/\hbar)^2} \; . \qquad \text{(B6)}$$

This result is equivalent to Eq. 26. Note that, in the absence of scattering (τ ∅ 0), the average drift velocity goes to zero. The particle is localized and performs purely oscillatory motion. These are the famous Bloch oscillations discussed in Section 5.2.4.

Appendix C. Contacts And Superlattices

Consider a superlattice (SL) of N identical periods sandwiched between doped contact electrodes. Suppose a voltage is applied between the electrodes and a current flows through the

SL. Given a sufficiently small current, the space-charge effects associated with the current can be neglected. One might imagine that a uniform electric field exists across the SL, as shown in Fig. 41a. The real situation is shown in Fig. 41b: most of the applied voltage drops over the first and last barriers, while the superlattice in-between exerts little if any influence on the *I-V* characteristic of the device. This is a manifestation of the dramatic difference between coherent transmission and incoherent decay in quantum mechanics.

In order to appreciate this difference, consider two apparently related processes illustrated in Fig. 42. Figure 42a shows a symmetric coupled-quantum-well system, with the wells separated by a tunneling barrier of height V_0 and width L_B. In the absence of interwell tunneling (e.g., in the V_0 ∅ limit), each well would contain a quantized level E_0 as shown. Tunneling between the wells splits E_0 into symmetric and antisymmetric states. The doublet splitting $\hbar\omega$ between the symmetric (lower) and antisymmetric (upper) states is, approximately, $\hbar\omega \approx E_0 e^{-\kappa L_B}$, where $\kappa = [2m^*(V_0 - E_0)/\hbar^2]^{1/2}$ and m^* is the electron mass. If, at time $t = 0$, an electron is placed in the left well, it will oscillate between the wells with a characteristic frequency ω. After a "short" time $\tau_1 = \pi/\omega$, the electron will be in the right well with unity probability.

Next, consider the escape process of an electron initially placed in the metastable state E_0 of a single quantum well separated from the continuum by the same barrier of height V_0 and width L_B, as shown in Fig. 42b. Due to the possibility of escape, the state has a finite lifetime τ_2 and hence a finite energy width $\Gamma = \hbar/\tau_2 \approx E_0 e^{-2\kappa L_B}$, cf. Eq. 5. Since $e^{-\kappa L_B}$ is the small parameter in our tunneling problem, typically $\Gamma << \hbar\omega$. Therefore, the lifetime τ_2 can be "long", perhaps orders of magnitude longer than τ_1.

The energy splitting $\hbar\omega$ in the coupled-quantum-well system is similar to the miniband width Δ in the superlattice problem. It describes the resonant transmission rate between discrete states, which is much faster than the seemingly analogous incoherent decay process into the continuum. In order to achieve a constant electric field in the superlattice, shown in Fig. 41a, the first and last barriers of the superlattice must be impedance-matched by making the $2\kappa L_B$ of the first and last barriers equal to the κL_B of the internal superlattice barriers. To first order, this can be achieved by making the first and last barriers narrower by a factor of approximately two. Failure to do so has been a common problem in many experimental studies of superlattice transport.[103]

Appendix D. Coherent Transistor Base Transport

In general, every bipolar or hot-electron ballistic transistor is characterized by a base transport factor α which is a complex function of frequency ω:

$$\alpha \equiv \left(\frac{\check{Z}I_{\mathrm{C}}}{\check{Z}I_{\mathrm{E}}}\right)_{V_{\mathrm{BC}}} = \mathrm{e}^{-i\omega\tau}\ |\alpha|\ . \qquad \text{(D1)}$$

The time τ which enters into the phase of α is the base transit time. In practice, all transistors operate at frequencies sufficiently low that $\omega\tau << 1$. In a well-designed bipolar transistor the deviation of $|\alpha|$ from unity is negligible at low frequencies, since the base is much narrower than the diffusion length. Consider the case $|\alpha| = 1$ more closely. The complex current gain $\beta(\omega)$ becomes:

$$\beta = \frac{\alpha}{1 - \alpha} = \frac{\mathrm{e}^{-i\omega\tau}}{1 - \mathrm{e}^{-i\omega\tau}} = \frac{\mathrm{e}^{-i\omega\tau/2}}{2i}\ \frac{1}{\sin(\omega\tau/2)}\ . \qquad \text{(D2)}$$

At low ω, the frequency dependence of the current gain is, therefore,

$$|\beta| = \frac{1}{2\sin(\omega\tau/2)}\ -(\ \omega\tau)^{-1}\ . \qquad \text{(D3)}$$

The magnitude of the current gain rolls off as ω^{-1}. This type of roll-off (typically referred to as 10 dB per decade or 3 dB per octave) is normally observed in microwave characterization of transistors. Extrapolating Eq. D3 to unity gain, one obtains the cut-off frequency $f_{\mathrm{T}} = (2\pi\tau)^{-1}$.

Note, however, that Eqs. D2 and D3 predict regions of high gain above f_{T}. These are the "coherent" gain peaks, corresponding to integer numbers of minority-carrier density wave periods in the base. The necessary condition for observing these peaks is the persistence of near-unity $|\alpha|$ at high frequencies. In fact, all that is required for $|\alpha| > 1$ at $f = 2\pi f_{\mathrm{T}}p$, where p is an integer, is $|\alpha| > 0.5$ at that frequency.[73] This condition, however, is extremely difficult to realize. If the base transport is diffusive, $|\alpha| << 1$ above f_{T}. As discussed in Section 5.3.2, ballistic transport offers one possibility of circumventing this problem. Another possibility is to replace random diffusive transport in the base by directed drift in specially graded base structures.[104]

REFERENCES

[1] At the time of writing, the Semiconductor Research Council roadmap predicts continuous

improvement in device performance to the year 2015, at which point the minimum lithographic size would reach below 1000 Å and the DRAM memory size would reach 16 Gb. See J. F. Freedman, "Comments on the National Technology Roadmap for semiconductors," in: S. Luryi, J. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology*, Kluwer, Dordrecht, 1996.

[2] H. Sakaki, "Scattering suppression and high-mobility effect of size-quantized electrons in ultrafine semiconductor wire structures," *Japan. J. Appl. Phys.*, **19**, L735 (1980); Y. Arakawa and H. Sakaki, "Multidimensional quantum well laser and temperature dependence of its threshold current," *Appl. Phys. Lett.*, **40**, 439 (1982).

[3] U. Meirav, M. A. Kastner, and S. J. Wind, "Single-electron charging and periodic conductance resonances in GaAs nanostructures," *Phys. Rev. Lett.*, **65**, 771 (1990); L. P. Kouwenhoven, N. C. van der Vaart, A. T. Johnson, W. Kool, C. J. P. M. Harmans, J. G. Williamson, A. A. M. Staring, and C. T. Foxon, "Single electron charging effects in semiconductor quantum dots," *Z. Phys. B.*, **85**, 367 (1991).

[4] This problem is treated in all textbooks on quantum mechanics. For a particularly thorough discussion, see C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics*, vol. 1, Chapter I, Wiley-Interscience, New York, 1977.

[5] B. Ricco and M. Ya. Azbel, "Physics of resonant tunneling: the one-dimensional double-barrier case," *Phys. Rev. B*, **29**, 1970 (1984).

[6] M. Jonson and A. Grincwajg, "Effect of inelastic scattering on resonant and sequential tunneling in double-barrier heterostructures," *Appl. Phys. Lett.*, **51**, 1729 (1987).

[7] S. V. Meshkov, "Tunneling of electrons from a two-dimensional channel into the bulk," *Zh. Eksp. Teor. Fiz.*, **91**, 2252 (198) [*Sov. Phys. JETP*, **64**, 1337 (1986)].

[8] S. Luryi, "Frequency limit of double-barrier resonant-tunneling oscillators," *Appl. Phys. Lett.*, **47**, 490 (1985).

[9] L. L. Chang, L. Esaki, and R. Tsu, "Resonant tunneling in semiconductor double barriers," *Appl. Phys. Lett.*, **24**, 593 (1974).

[10] A. Zaslavsky, D. C. Tsui, M. Santos, and M. Shayegan, "Magnetotunneling in double-barrier heterostructures," *Phys. Rev. B*, **40**, 9829 (1989).

[11] V. J. Goldman, D. C. Tsui, and J. E. Cunningham, "Evidence for LO-phonon-emission-assisted tunneling in double-barrier heterostructures," *Phys. Rev. B*, **36**, 7635 (1987).

[12] N. S. Wingreen, K. W. Jacobsen, and J. W. Wilkins, "Resonant tunneling with electron-phonon interaction: an exactly solvable model," *Phys. Rev. Lett*., **61**, 1396 (1988);
F. Chevoir and B. Vinter, "Calculation of phonon-assisted tunneling and valley current in a double-barrier diode," *Appl. Phys. Lett.*, **55**, 1859 (1989).

[13] M. C. Payne, "Transfer Hamiltonian description of resonant tunneling," *J. Phys. C*, **19**, 1145 (1986);
T. Weil and B. Vinter, "Equivalence between resonant tunneling and sequential tunneling in double-barrier diodes," *Appl. Phys. Lett.*, **50**, 1281 (1987).

[14] An extensive discussion is available in M. Büttiker, "Coherent and sequential tunneling in series barriers," *IBM J. Res. Develop.*, **32**, 63 (1988).

[15] See, for example, S. K. Diamond, E. Ozbay, M. Rodwell, D. M. Bloom, Y. C. Pao, E. Wolak, and J. S. Harris, "Fabrication of 200-GHz f_{max} resonant-tunneling diodes for integrated circuit and microwave applications," *IEEE Electron Dev. Lett.*, **10**, 104 (1989).

[16] V. J. Goldman, D. C. Tsui, and J. E. Cunnigham, "Observation of intrinsic bistability in resonant-tunneling structures," *Phys. Rev. Lett.*, **58**, 1256 (1987).

[17] E. Ozbay and D. M. Bloom, "110-GHz monlithic resonant-tunneling-diode trigger circuit," *IEEE Electron Dev. Lett.*, **12**, 480 (1991).

[18] E. R. Brown, C. D. Parker, A. R. Kalawa, M. J. Manfra, and K. M. Molvar, "A quasioptical resonant-tunneling-diode oscillator operating above 200 GHz," *IEEE Trans. Microwave Theory Tech.*, **41**, 720 (1993).

[19] E. R. Brown, "High-speed resonant-tunneling diodes," in: N. G. Einspruch and W. R. Frensley, eds., *Heterostructures and Quantum Devices*, Academic Press, San Diego, 1994.

[20] M. Sweeny and J. Xu, "Resonant interband tunneling diodes," *Appl. Phys. Lett.*, **54**, 546 (1989);
J. R. Söderstrom, D. H. Chow, and T. C. McGill, "New negative differential resistance device based on resonant interband tunneling," *Appl. Phys. Lett.*, **55**, 1094 (1989).

[21] E. E. Mendez, J. Nocera, and W. I. Wang, "Conservation of momentum, and its consequences, in interband resonant tunneling," *Phys. Rev. B*, **45**, 3910 (1992).

[22] R. Beresford, L. Luo, K. Longenbach, and W. I. Wang, "Resonant interband tunneling

through a 110 nm InAs quantum well," *Appl. Phys. Lett.*, **56**, 551 (1990).

[23] H. C. Liu, D. Landheer, M. Buchanan, and D. C. Houghton, "Resonant tunneling in Si/Si$_{1-x}$Ge$_x$ double-barrier structures," *Appl. Phys. Lett.*, **52**, 1809 (1988).

[24] Z. Matutinovic-Krstelj, C. W. Liu, X. Xiao, and J. C. Sturm, "Evidence for phonon-absorption-assisted electron resonant tunneling in Si/Si$_{1-x}$Ge$_x$ diodes," *Appl. Phys. Lett.*, **62**, 603 (1993).

[25] G. Schuberth, G. Abstreiter, E. Gornik, F. Schäffler, and J. F. Luy, "Resonant tunneling of holes in Si/Si$_{1-x}$Ge$_x$ quantum-well structures," *Phys. Rev. B*, **43**, 2280 (1991).

[26] U. Gennser, V. P. Kesan, D. A. Syphers, T. P. Smith III, S. S. Iyer, and E. S. Yang, "Probing band structure anisotropy in quantum wells via magnetotunneling," *Phys. Rev. Lett.*, **67**, 3828 (1991).

[27] A. Zaslavsky, K. R. Milkove, Y. H. Lee, B. Ferland, and T. O. Sedgwick, "Strain relaxation in silicon-germanium microstructures observed by resonant tunneling spectroscopy", *Appl. Phys. Lett.*, **67**, 3921 (1995).

[28] S.-Y. Chou, D. R. Allee, R. F. W. Pease, and J. Harris, Jr., "Observation of electron resonant tunneling in a lateral dual-gate resonant tunneling field-effect transistor," *Appl. Phys. Lett.*, **55**, 176 (1989);
K. Ismail, D. A. Antoniadis, and H. I. Smith, "Lateral resonant tunneling in a double-barrier field effect transistor," *Appl. Phys. Lett.*, **55**, 589 (1989).

[29] M. A. Reed, W. R. Frensley, R. J. Matyi, J. N. Randall, and A. C. Seabaugh, "Realization of a three-terminal resonant tunneling device: the bipolar quantum resonant tunneling transistor," *Appl. Phys. Lett.*, **54**, 1034 (1989).

[30] A. R. Bonnefoi, D. H. Chow, and T. C. McGill, "Inverted base-collector tunnel transistors," *Appl. Phys. Lett.*, **47**, 888 (1985).

[31] S. Luryi, "Quantum capacitance devices," *Appl. Phys. Lett.*, **52**, 501 (1988).

[32] F. Beltram, F. Capasso, S. Luryi, S. N. G. Chu, and A. Y. Cho, "Negative transconductance via gating of the quantum well subbands in a resonant tunneling transistor," *Appl. Phys. Lett.*, **53**, 219 (1988).

[33] C. J. Goodings, H. Mizuta, J. Cleaver, and H. Ahmed, "Variable-area resonant tunneling diodes using implanted in-plane gates," *J. Appl. Phys.*, **76**, 1276 (1994).

[34] T. K. Woodward, T. C. McGill, and R. D. Burnham, "Experimental realization of a resonant

tunneling transistor," *Appl. Phys. Lett.*, **50**, 451 (1987).

[35] M. Dellow, P. H. Beton, M. Henini, P. C. Main, L. Eaves, S. P. Beaumont, and C. D. W. Wilkinson, "Gated resonant tunneling devices," *Electron. Lett.*, **27**, 134 (1991); P. Guéret, N. Blanc, R. Germann, and H. Rothuizen, "Confinement and single-electron tunneling in Schottky-gated, laterally squeezed double-barrier quantum-well heterostructures," *Phys. Rev. Lett.*, **68**, 1896 (1992).

[36] V. R. Kolagunta, D. B. Janes, G. L. Chen, K. Webb, M. R. Melloch, and C. Youtsey, "Self-aligned sidewall-gated resonant tunneling transistors," *Appl. Phys. Lett.*, **69**, 374 (1996).

[37] S. Luryi and F. Capasso, "Resonant tunneling of two dimensional electrons through a quantum wire," *Appl. Phys. Lett.*, **47**, 1347 (1985); *erratum ibid.*, **48**, 1693 (1986).

[38] A. Zaslavsky, K.R. Milkove, Y.H. Lee, K.K. Chan, F. Stern, D.A. Grützmacher, S.A. Rishton, C. Stanis, and T.O. Sedgwick, "Fabrication of three-terminal resonant tunneling devices in silicon-based material," *Appl. Phys. Lett.*, **64**, 1699 (1994).

[39] A. Zaslavsky, D.C. Tsui, M. Santos, and M. Shayegan, "Resonant tunneling of two-dimensional electrons into one-dimensional subbands of a quantum wire," *Appl. Phys. Lett.*, **58**, 1440 (1991).

[40] L. N. Pfeiffer, K. W. West, H. L. Stormer, J. P. Eisenstein, K. W. Baldwin, D. Gershoni, and J. Spector, "Formation of a high-quality two-dimensional electron gas on cleaved GaAs," *Appl. Phys. Lett.*, **56**, 1697 (1990).

[41] Ç. Kurdak, D. C. Tsui, S. Parihar, M. B. Santos, H. Manoharan, S. A. Lyon, and M. Shayegan, "Surface resonant tunneling transistor: a new negative transconductance device," *Appl. Phys. Lett.*, **64**, 610 (1994).

[42] F. Capasso, K. Mohammed, and A. Y. Cho, "Resonant tunneling through double barriers, perpendicular quantum transport phenomena in superlattices, and their device applications," *IEEE J. Quantum Electron.*, **QE-22**, 1853 (1986).

[43] A. C. Seabaugh, Y.-C. Kao, and H.-T. Yuan, "Nine-state resonant tunneling diode memory," *IEEE Electron Dev. Lett.*, **13**, 479 (1992).

[44] For a complete discussion see G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Chapter I, Wiley, New York, 1988.

[45] L. V. Keldysh, "Effect of ultrasound on the electron spectrum of a crystal," *Fiz. Tverd. Tela*, **4**, 2265 (1962) [*Sov. Phys. Solid. State*, **4**, 1658 (1963)].

[46] L. Esaki and R. Tsu, "Superlattice and negative differential conductivity in semiconductors," *IBM J. Res. Develop.*, **14**, 61 (1970).

[47] L. Esaki and L. L. Chang, "New transport phenomenon in a semiconductor 'superlattice'," *Phys. Rev. Lett.*, **33**, 495 (1974);
K. K. Choi, B. F. Levine, R. J. Malik, J. Walker, and C. G. Bethea, "Periodic negative conductance by sequential resonant tunneling through an expanding high-field superlattice domain," *Phys. Rev. B*, **35**, 4172 (1987).

[48] A. Sibille, J. F. Palmier, H. Wang, and F. Mollot, "Observation of Esaki-Tsu negative differential velocity in GaAs/AlAs superlattices," *Phys. Rev. Lett.*, **64**, 52 (1990);
H. T. Grahn, K. von Klitzing, K. Ploog, and G. Döhler, "Electrical transport in narrow-miniband semiconductor superlattices," *Phys. Rev. B*, **43**, 12094 (1991).

[49] H. M. James, "Electronic states in perturbed periodic systems," *Phys. Rev.*, **76**, 1611 (1949).

[50] The extent of Wannier-Stark wavefunctions over a finite number of SL periods allows for

their

observation by photocurrent measurements, see E. E. Mendez, F. Agulló-Rueda, and J. M. Hong, "Stark localization in GaAs-GaAlAs superlattices under an electric field," *Phys. Rev. Lett.*, **60**, 2426 (1988).

[51] R. Kazarinov and R. Suris, "Possibility of the amplification of electromagnetic waves in a semiconductor with a superlattice," *Fiz. Tekh. Poluprovodn.*, **5**, 797 (1971) [*Sov. Phys. Semicond.*, **5**, 707 (1971).

[52] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, "Quantum cascade laser," *Science*, **264**, 533 (1994).

[53] H. C. Liu and G. C. Aers, "Resonant tunneling through one-, two-, and three-dimensionally confined quantum wells," *J. Appl. Phys.*, **65**, 4908 (1989).

[54] B. Su, V. J. Goldman, and J. E. Cunningham, "Observation of single-electron charging in double-barrier heterostructures," *Science*, **255**, 313 (1992); "Single-electron tunneling in nanometer-scale double-barrier heterostructure devices," *Phys. Rev. B*, **46**, 7644 (1992).

[55] T. Schmidt, M. Tewordt, R. H. Blick, R. J. Haug, D. Pfannkuche, K. von Klitzing, A. Förster, and H. Luth, "Quantum-dot ground states in a magnetic field studied by

single-electron tunneling spectroscopy on double-barrier heterostructures," *Phys. Rev. B*, **51**, 5570 (1995).

[56] An extensive discussion is available in H. Grabert and M. H. Devoret, eds., *Single Charge Tunneling: Coulomb Blockade Phenomena in Nanostructures*, Plenum Press, New York, 1992. In particular, the chapter by D. V. Averin and K. K. Likharev is devoted to device applications.

[57] T. A. Fulton and G. J. Dolan, "Observation of single-electron charging effects in small junctions," *Phys. Rev. Lett.*, **59**, 109 (1987).

[58] G. Zimmerli, R. L. Kautz, and J. M. Martinis, "Voltage gain in the single-electron transistor," *Appl. Phys. Lett.*, **61**, 2616 (1992).

[59] L. P. Kouwenhoven, A. T. Johnson, N. C. van der Vaart, C. J. Harmans, and C. T. Foxon, "Quantized current in a quantum-dot turnstile using oscillating tunnel barriers," *Phys. Rev. Lett.*, **67**, 1626 (1991).

[60] H. Pothier, P. Lafarge, C. Urbina, D. Esteve, and M. H. Devoret, "Single-electron pump based on charging effects," *Europhys. Lett.*, **17**, 249 (1992);
J. M. Martinis, M. Nahum, and H. D. Jensen, "Metrological accuracy of the electron pump," *Phys. Rev. Lett.*, **72**, 904 (1994).

[61] C. A. Mead, "Tunnel-emission amplifiers," *Proc. IRE*, **48**, 359 (1960).

[62] K. Seeger, *Semiconductor Physics*, 2nd ed., Springer-Verlag, Berlin, 1982;
E. Schöll, "Theory of oscillatory instabilities in parallel and perpendicular transport in heterostructures," in: N. Balkan, B. K. Ridley, and A. J. Vickers, eds., *Negative Differential Resistance and Instabilities in 2D Semiconductors*, Plenum Publishing, New York, 1993, pp. 37-51.

[63] P. J. Price, "Monte Carlo calculation of electron transport in solids," in: *Semiconductors and Semimetals*, vol. 14, Academic Press, New York, 1979, pp. 249-308;
C. Moglestue, *Monte Carlo Simulation of Semiconductor Devices*, Chapman and Hall, New York, 1993.

[64] A thorough discussion of real-space transfer effects is available in a review article by Z. S. Gribnikov, K. Hess, and G. A. Kosinovsky, "Nonlocal and nonlinear transport in semiconductors: real-space transfer effects," *J. Appl. Phys.*, **77**, 1337 (1995).

[65] M. Heiblum, M. I. Nathan, D. Thomas, and C. M. Knoedler, "Direct observation of

ballistic transport in gallium arsenide," *Phys. Rev. Lett.*, **55**, 2200 (1985);
M. Heiblum, I. Anderson, and C. M. Knoedler, "DC performance of ballistic tunneling hot-electron-transfer amplifiers," *Appl. Phys. Lett.*, **49**, 207 (1986).

[66] K. Seo, M. Heiblum, C. M. Knoedler, J. Oh, J. Pamulapati, and P. Bhattacharya, "High-gain pseudomorphic InGaAs base ballistic hot-electron devices," *IEEE Electron Dev. Lett.*, **10**, 73 (1989).

[67] T. S. Moise, A. C. Seabaugh, E. A. Beam III, J. N. Randall, "Room-temperature operation of a resonant-tunneling hot-electron transistor based integrated circuit," *IEEE Electron. Dev. Lett.*, **14**, 441 (1993).

[68] A. A. Grinberg and S. Luryi, "Electron tranmission across interface of different one-dimensional crystals," *Phys. Rev. B*, **39**, 7466 (1989).

[69] J. R. Hayes, A. F. J. Levi, and W. Wiegmann, "Hot electron spectroscopy," *Electron. Lett.*, **20**, 851 (1984);
A. F. J. Levi, J. R. Hayes, P. M. Platzmann, and W. Wiegmann, "Injected hot-electron transport in GaAs," *Phys. Rev. Lett.*, **55**, 2071 (1985).

[70] A. Palevski, M. Heiblum, C. P. Umbach, C. M. Knoedler, A. N. Broers, and R. H. Koch, "Lateral tunneling, ballistic transport, and spectroscopy of a two-dimensional electron gas," *Phys. Rev. Lett.*, **62**, 1776 (1989);
A. Palevski, C. P. Umbach, and M. Heiblum, "A high gain lateral hot-electron device," *Appl. Phys. Lett.*, **55**, 1421 (1989).

[71] J. Spector, H. L. Stormer, K. W. Baldwin, L. N. Pfeiffer, and K. W. West, "Ballistic electron transport beyond 100 μm in 2D electron systems," *Surf. Sci.*, **228**, 283 (1990);
A. Yacoby, U. Sivan, C. P. Umbach, and J. M. Hong, "Hot ballistic transport and phonon emission in a two-dimensional electron gas," *Phys. Rev. Lett.*, **66**, 1938 (1991).

[72] For example, see J. Song, B. W.-P. Hong, C. J. Palmstrom, B. P. van der Gaag, and K. B. Chough, "Ultra-high-speed InP/InGaAs heterojunction bipolar transistors," *IEEE Electron Dev. Lett.*, **15**, 94 (1994).

[73] A. A. Grinberg and S. Luryi, "Coherent transistor," *IEEE Trans. Electron. Dev.*, **40**, 1512 (1993).

[74] Z. S. Gribnikov, "Negative differential conductivity in a multilayer heterostructure," *Fiz. Tekh. Poluprovodn.*, **6**, 1380 (1972) [*Sov. Phys. Semicond.*, **6**, 1204 (1973)].

[75] K. Hess, H. Morkoç, H. Shichijo, and B. G. Streetman, "Negative differential resistance through real-space electron transfer," *Appl. Phys. Lett.*, **35**, 469 (1979).

[76] M. Keever, H. Shichijo, K. Hess, S. Banerjee, L. Witkowski, H. Morkoç, and B. G. Streetman, "Measurements of hot-electron conduction and real-space transfer in GaAs/$Al_xGa_{1-x}As$ heterojunction layers," *Appl. Phys. Lett.*, **38**, 36 (1981).

[77] N. Z. Vagidov, Z. S. Gribnikov, and V. M. Ivastchenko, "Modeling of electron transport in real space in GaAs/$Al_xGa_{1-x}As$ heterostructures (with low and high values of *x*)," *Fiz. Tekh. Poluprovodn.*, **24**, 1087 (1990) [*Sov. Phys. Semicond.*, **24**, 684 (1990)].

[78] A. Kastalsky and S. Luryi, "Novel real-space hot-electron transfer devices," *IEEE Electron. Dev. Lett.*, **4**, 334 (1983);
S. Luryi, A. Kastalsky, A. C. Gossard, and R. H. Hendel, "Charge injection transistor based on real-space hot-electron transfer," *IEEE Trans. Electron. Dev.*, **31**, 832 (1984).

[79] A. Kastalsky, R. Bhat, W. K. Chan, and M. Koza, "Negative-resistance field-effect transistor grown by organometallic chemical vapor deposition," *Solid State Electron.*, **29**, 1073 (1986).

[80] P. M. Mensz, P. A. Garbinski, A. Y. Cho, D. L. Sivco, and S. Luryi, "High transconductance and large peak-to-valley ratio of negative differential conductance in three-terminal InGaAs/InAlAs real-space transfer devices," *Appl. Phys. Lett.*, **57**, 2558 (1990).

[81] M. R. Hueschen, N. Moll, and A. Fischer-Colbrie, "Improved microwave performance in transistors based on real-space electron transfer," *Appl. Phys. Lett.*, **57**, 386 (1990);
G. L. Belenky, P. A. Garbinski, S. Luryi, P. R. Smith, A. Y. Cho, R. A. Hamm, and D. L. Sivco, "Microwave performance of top-collector charge injection transistors on InP substrates," *Semicond. Sci. Technol.*, **9**, 1215 (1994).

[82] C. L. Wu, W. C. Hsu, M. S. Tsai, and H. M. Shieh, "Very strong negative differential resistance real-space transfer transistor using a mulitple δ-doping GaAs/InGaAs pseudomorphic heterostructure," *Appl. Phys. Lett.*, **66**, 739 (1995).

[83] M. Mastrapasqua, C. A. King, P. R. Smith, and M. R. Pinto, "Charge injection transistor and logic elements in Si/$Si_{1-x}Ge_x$ heterostructures," in: S. Luryi, J. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology*, Kluwer, Dordrecht, 1996.

[84] I. G. Kizilyalli and K. Hess, "Physics of real-space transfer transistors," *J. Appl. Phys.*, **65**,

2005 (1989).

[85] A. A. Grinberg, A. Kastalsky, and S. Luryi, "Theory of hot-electron injection in CHINT/NERFET devices," *IEEE Trans. Electron. Dev.*, **34**, 409 (1987).

[86] M. Mastrapasqua, S. Luryi, G. L. Belenky, P. A. Garbinski, A. Y. Cho, and D. L. Sivco, "Multi-terminal light emitting logic device electrically reprogrammable between OR and NAND functions," *IEEE Trans. Electron. Dev.*, **40**, 1371 (1993);
G. L. Belenky, P. A. Garbinski, S. Luryi, M. Mastrapasqua, A. Y. Cho, R. A. Hamm, T. R. Hayes, E. J. Laskowski, D. L. Sivco, and P. Smith, *J. Appl. Phys.*, **73**, 8618 (1993).

[87] N. Yokoyama, K. Imamura, S. Muto, S. Hiyamizu, and H. Nishi, "A new functional resonant-tunneling hot electron transistor (RHET)," *Japan. J. Appl. Phys.*, **24**, L-853 (1985);
N. Yokoyama, K. Imamura, H. Ohnishi, T. Mori, S. Muto, and A. Shibatomi, "Resonant-tunneling hot electron transistor (RHET)," *Solid State Electron.*, **31**, 577 (1988).

[88] For a review of resonant tunneling bipolar transistor research see F. Capasso, S. Sen, and F. Beltram, "Quantum-effect devices," in: S. M. Sze, ed., *High-Speed Semiconductor Devices*, Wiley, New York, 1990, pp. 465-520.

[89] S. Sen, F. Capasso, A. Y. Cho, and D. L. Sivco, "Multiple state resonant tunneling bipolar transistor operating at room temperature and its application as a frequency multiplier," *IEEE Electron. Dev. Lett.*, **9**, 533 (1988).

[90] E. R. Brown, J. R. Söderstrom, C. D. Parker, L. J. Mahoney, K. M. Molvar, and T. C. McGill, "Oscillations up to 712 GHz in InAs/AlSb resonant-tunneling diodes," *Appl. Phys. Lett.*, **58**, 2291 (1991).

[91] W. F. Chow, *Principles of Tunnel Diode Circuits*, Wiley, New York, 1964.

[92] J. Shen, G. Kramer, S. Tehrani, H. Goronkin, and R. Tsui, "Static random access memories based on resonant interband tunneling diodes in the InAs/GaSb/AlSb material system," *IEEE Electron. Dev. Lett.*, **16**, 178 (1995).

[93] T. Mori, S. Muto, H. Tamura, and N. Yokoyama, "A static random access memory cell using a double-emitter resonant-tunneling hot electron transistor for gigabit-plus memory applications," *Japan. J. Appl. Phys.*, **33**, 790 (1994).

[94] M. Takatsu, K. Imamura, H. Ohnishi, T. Mori, T. Adachibara, S. Muto, and N. Yokoyama, "Logic circuits using resonant-tunneling hot-electron transistors (RHET's)," *IEEE*

J. Solid-State Circuits, **27**, 1428 (1992);
N. Yokoyama, H. Ohnishi, T. Mori, M. Takatsu, S. Muto, K. Imamura, and A. Shibatomi, "Resonant hot electron transistors," in: J. Shah, ed., *Hot Carriers in Semiconductor Nanostructures: Physics and Applications*, Academic Press, San Diego, 1992, pp. 443-467.

[95] A. C. Seabaugh and M. A. Reed, "Resonant-tunneling transistors," in: N. G. Einspruch and W. R. Frensley, eds., *Heterostructures and Quantum Devices*, San Diego: Academic Press, 1994.

[96] S. Luryi, P. Mensz, M. R. Pinto, P. A. Garbinski, A. Y. Cho, and D. L. Sivco, "Charge injection logic," *Appl. Phys. Lett.*, **57**, 1787 (1990);
K. Imamura, M. Takatsu, T. Mori, Y. Bamba, S. Muto, and N. Yokoyama, "Proposal and demonstration of multi-emitter HBTs," *Electron. Lett.*, **30**, 459 (1994).

[97] J. Faist, F. Capasso, C. Sirtori, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, "Vertical transition quantum cascade laser with Bragg confined excited state," *Appl. Phys. Lett.*, **66**, 538 (1995).

[98] C. Sirtori, J. Faist, F. Capasso, D. L. Sivco, and A. Y. Cho, "Narrowing of the intersubband absorption spectrum by localization of continuum resonances in a strong electric field," *Appl. Phys. Lett.*, **62**, 1931 (1993).

[99] J. Faist, F. Capasso, C. Sirtori, D. L. Sivco, J. N. Baillargeon, A. L. Hutchinson, and A. Y. Cho, "High power mid-infrared ($\lambda \sim 5$ μm) quantum cascade lasers operating above room temperature," *Appl. Phys. Lett.*, **68**, 3680 (1996).

[100] V. B. Gorfinkel, S. Luryi, and B. Gelmont, "Theory of gain spectra for quantum cascade lasers and temperature dependence of their characteristics at low and moderate carrier concentrations," *IEEE J. Quantum Electron.*, **32**, xxx (1996).

[101] B. Gelmont, V. B. Gorfinkel, and S. Luryi, "Theory of the spectral line shape and gain in quantum wells with intersubband transitions," *Appl. Phys. Lett.*, **68**, 2171 (1996).

[102] D. C. Tsui, H. L. Störmer, and A. C. Gossard, "Two-dimensional magnetotransport in the extreme quantum limit," *Phys. Rev. Lett.*, **48**, 1559 (1982).

[103] L. V. Iogansen, "Errors in papers on resonant electron tunneling in finite superlattices," *Pis'ma Zh. Tekh. Fiz.*, **13**, 1143 (1987) [*Sov. Tech. Phys. Lett.*, **13**, 478 (1987)].

[104] S. Luryi, A. A. Grinberg, and V. B. Gorfinkel, "Heterostructure bipolar transistor with enhanced forward diffusion of minority carriers," *Appl. Phys. Lett.*, **63**, 1537 (1993).