QUANTUM BALLISTIC TRANSPORT

IN SEMICONDUCTOR HETEROSTRUCTURES

A Thesis

Submitted to the Faculty

of

Purdue University

- by

Michael James McLennan

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Electrical Engineering

May 1987

To my darling Maria,
for her patience and support
on many a sleepless night.

# ACKNOWLEDGMENTS

# TABLE OF CONTENTS

# LIST OF FIGURES

# ABSTRACT

McLennan, Michael James. M.S.E.E., Purdue University. May 1987. Quantum Ballistic Transport in Semiconductor Heterostructures. Major Professor: Supriyo Datta.

The development of epitaxial growth techniques has sparked a growing interest in an entirely quantum mechanical description of carrier transport. Fabrication methods, such as molecular beam epitaxy (MBE), allow for the growth of ultra-thin layers of differing material compositions. Structures can be designed to exploit the wave-nature of carriers, broadening the possibilities of device design. This thesis represents the first step in the development of a quantum mechanical transport theory. Wave phenomena exhibited by electrons are discussed, and applications in proposed devices are presented. A theory of quantum ballistic transport is developed, emphasizing concerns for a numerical implementation of the analysis. Finally, example calculations are presented, illustrating quantitatively the physics of quantum transport.

# CHAPTER 1
# INTRODUCTION

Wave-particle duality is not a new concept in physics. Until recently, however, the wave nature of carriers has been largely ignored in electrical engineering; a classical, particle description of transport has been adequate. The optical analogy is familiar: A ray description is appropriate, unless the refractive index of a medium varies rapidly, compared to the wavelength of incident radiation. In the past, semiconductor device analysis has avoided using quantum mechanics directly, by incorporating effects of the rapidly varying crystal potential into parameters such as effective mass and energy bands. The remaining, slowly varying potential, due to built-in fields and applied bias, has been treated in a classical framework. The maturity of epitaxial growth techniques, however, has sparked recent interest in developing an entirely quantum mechanical formalism for carrier transport. Techniques such as molecular beam epitaxy (MBE) and metalorganic chemical vapor deposition (MOCVD) permit the growth of ultra-thin layers of differing material composition, with interfaces as sharp as an atomic monolayer. Since the change in bandgap from layer to layer is abrupt on the scale of an electron DeBroglie wavelength (typically 100 Å - 1000 Å), transport properties of such structures must be determined quantum mechanically. The superlattice, a periodic structure with alternating layers of wide bandgap and narrow bandgap materials, is representative of this new class of devices, and has been the focus of intense investigation.

Esaki and Tsu [1,2] proposed the superlattice in 1969, for application in negative differential resistance (NDR) devices. In the Kronig-Penney model, periodicity of the crystal potential gives rise to allowed and forbidden energy bands. Similarly, the periodicity of material layers in a superlattice results in a folding of energy bands, into a reduced Brillouin Zone. The conduction-band is segmented into a series of "minibands," in which electrons are allowed to propagate, and gaps, in which propagation is attenuated. Because carriers are confined to characteristic, "resonant"

energies, the transport mechanism is known as resonant tunneling. For low biases, applied along the growth axis, carriers propagate via resonant tunneling. As bias is increased, resonances in each quantum well become misaligned, and current drops sharply. The resulting negative differential resistance is a useful phenomena, for producing both amplification and oscillation. Although the immaturity of epitaxial growth techniques sabotaged early experiments, in 1972 Esaki et. al. [3] reported NDR in a GaAs/AlGaAs superlattice. Two years later, Chang, Esaki, and Tsu [4] confined attention to a double barrier resonant tunneling device, two periods of a GaAs/AlGaAs superlattice, and reported NDR at temperatures below 77 ° K.

In 1983, Sollner et. al. [5] revived interest, reporting a large NDR region in the current-voltage characteristic of a GaAs/AlGaAs double barrier resonant tunneling device. The reported current peak-to-valley ratio was 6:1 at 25 ° K. Furthermore, current response measured at a driving frequency of 2.5 THz was remarkably similar to the response expected from DC measurements, indicating the potential for high frequency applications. One year later, Sollner et. al. [6] reported the first oscillations generated by a resonant tunneling device, with a power output of 5 $\mu$W at frequencies up to 18 GHz. Shewchuk et. al. reported the first room temperature observation of NDR [7], and suggested experimental guidelines for the stable measurement of NDR regions in the presence of oscillations [8]. Tsuchiya et. al. [9] demonstrated a resonant tunneling device with a number of improvements, also exhibiting NDR at room temperature. Thin (25 Å) barriers of AlAs were used to reduce thermionic emission current, and increase resonant tunneling current. Undoped "spacer" layers of GaAs surrounding the double barrier region were included, to reduce impurity migration into the quantum well. Recently, Ray et. al. [10] reported the first room temperature observation of NDR in a resonant tunneling device fabricated by metalorganic chemical vapor deposition (MOCVD). Lee et. al. [11] reported the first room temperature observation of NDR in a strained-layer GaAs/AlGaAs/InGaAs device. This survey is not comprehensive, but is intended to convey the widespread interest in resonant tunneling devices. A thorough recount of the history of resonant tunneling has been presented by Esaki [12].

The ability to "sculpture" energy bands in one dimension has expanded the possibilities of device design, and has underscored the need for a quantum mechanical description of transport. Although several formalisms

are being developed, a significant amount of work remains. The product of this thesis, an analysis program for ballistic devices, is merely the first step in this enormous task.

## 1.1 Quantum Mechanical Transport Phenomena

### 1.1.1 Coherent Resonant Tunneling

Coherent resonant tunneling is a direct result of the wave nature of electrons. As such, it can be understood with an analogy to wave optics. The optical analogue of a double barrier resonant tunneling device is a Fabry-Perot interferometer, two partially silvered mirrors in perfect alignment. Radiation incident on a Fabry-Perot cavity experiences unity transmission at particular resonant frequencies. Waves multiply reflected between the two mirrors constructively interfere, reducing the overall reflection. At incident frequencies for which the spacing between mirrors is an integral number of half-wavelengths, overall reflection is zero; transmission is unity. However familiar this result may seem in optics, the quantum mechanical effect is striking: Two rectangular potential barriers are completely "invisible" to electrons incident at particular resonant energies.

To justify this result analytically, the transmission coefficient of a resonant tunneling device can be calculated as follows. Consider the double barrier resonant tunneling device, two potential barriers of $Al_xGa_{1-x}As$ surrounding a quantum well of GaAs, shown in figure 1.1. For simplicity, band bending has been neglected, and only the conduction-band profile is shown. Electrons incident from the left contact, and transmitted through the structure, could follow any one of an infinite number of paths, allowing for multiple reflections between the potential barriers. If the Schrödinger equation is solved for each barrier separately, wavefunction amplitudes for transmitted and reflected plane-wave solutions can be determined. Assuming these amplitudes, t and r, represent electrons incident from the left of a barrier, and assuming the amplitudes $t'$ and $r'$ represent a similar solution, for electrons incident from the right, the total transmitted amplitude for each path can be determined. For the first path, electrons are transmitted through the first barrier, traverse the quantum well, and are transmitted through the second barrier; the wavefunction is multiplied by each transmitted amplitude, and acquires phase in the well region. The

Barrier 1    Barrier 2

$E_C$

$\leftarrow a \rightarrow$

Electrons

$t_1 \, e^{ika} \, t_2$

$t_1 \, [r_2 r'_1 e^{2ika}] \, e^{ika} \, t_2$

$t_1 \, [r_2 r'_1 e^{2ika}]^2 \, e^{ika} \, t_2$

$t_1 \, [r_2 r'_1 e^{2ika}]^3 \, e^{ika} \, t_2$
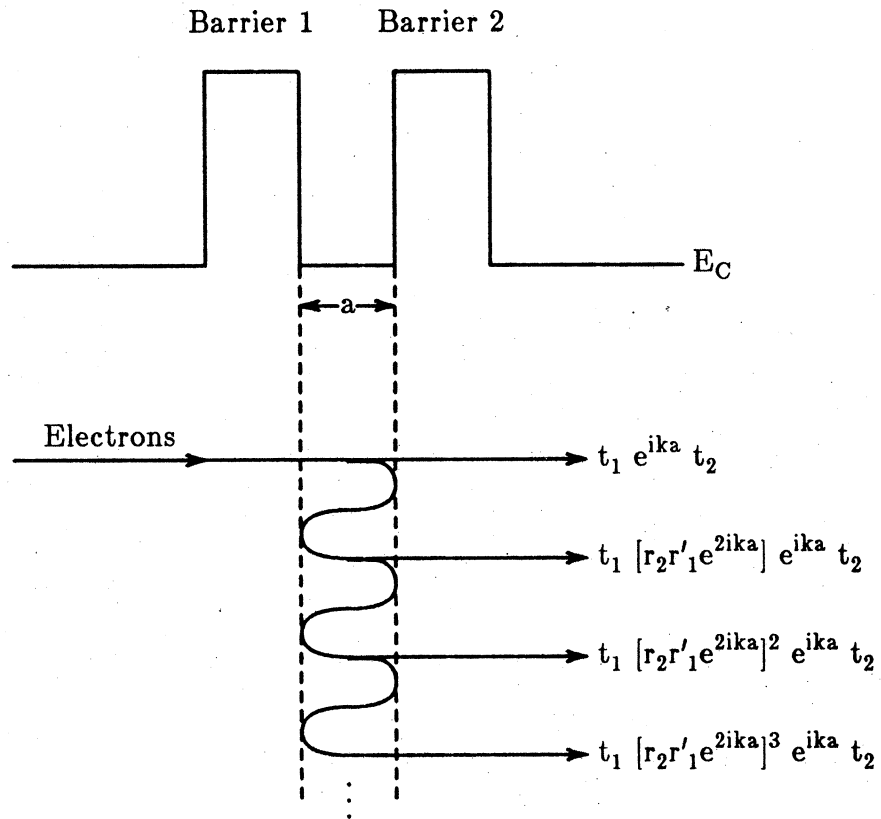
Figure  1.1    Determination  of  transmission  coefficient  by  summing  an infinite series of multiple reflections

transmitted amplitude for each of the multiply reflected paths is derived similarly, assuming scattering is infrequent, so that the phase of an electron remains coherent for the entire time of transit. The total transmitted amplitude can be calculated by summing contributions from every possible path:

$$t_b = t_2 \left[ 1 + r'_1 r_2 e^{i2ka} + (r'_1 r_2 e^{i2ka})^2 + ... \right] e^{ika} t_1$$

$$= \frac{t_2 e^{ika} t_1}{1 - r'_1 r_2 e^{i2ka}}$$

where the subscripts of transmitted and reflected amplitudes denote the barrier encountered. Assume, for simplicity, that the two potential barriers are identical. Since the structure of figure 1.1 is in equilibrium, the transmission coefficient is the squared magnitude of the total transmitted amplitude:

$$T = \frac{T_B^2}{1 + R_B^2 - 2R_B \cos(2ka + 2\phi)} \qquad \text{with} \quad r_1' = | r_1' | \, e^{i\phi} = r_2$$

where

$$T_B = \left| t_1 \right|^2 = \left| t_2 \right|^2$$

$$R_B = \left| r'_1 \right|^2 = \left| r_2 \right|^2$$

For electrons incident with $ka = n\pi - \phi$, for any integer n,

$$T = \frac{T_B^2}{[1 - R_B]^2} = \frac{T_B^2}{[T_B]^2} = 1.$$

This is the result expected: When the spacing between potential barriers is an integer multiple of the incident electron half-wavelength, transmission is unity, in agreement with the analogy of the Fabry-Perot interferometer. Of course, the reflection from each potential barrier could introduce a phase $\phi$, altering the Fabry-Perot condition for constructive interference. The energies of electrons satisfying this criterion are called "resonant" energies, because of the constructive interference of multiple reflections. If the barrier height were taken to infinity, the resonant energies would be bound states of an infinite potential well. As such, resonant energies are considered quasi-bound states of the well region.

Although the analysis presented above provides significant insight into the interpretation of resonant tunneling, several assumptions limit application of the result. Ricco and Azbel [13] discuss the effect, without assuming identical potential barriers in equilibrium. For electrons incident at the resonant energy, neglecting constants, the transmission coefficient is proportional to $T_{min}/T_{max}$, where $T_{min}$ is the smaller of the transmission coefficients of the two barriers, and $T_{max}$ is the larger. For identical barriers, transmission is unity at resonance; however, structures designed to be symmetrical in equilibrium are quite distorted by the application of bias. Away from equilibrium, the maximum possible transmission could be drastically smaller than one. For electrons incident at non-resonant energies, the transmission coefficient is proportional to $T_{min}T_{max}$. Hence, the transmission for resonant tunneling structures is characterized in energy space by narrow spikes, located at each of the resonant energies. An important consideration for current density is the width of these spikes in energy space, since only incident electrons near the resonant energy experience significant transmission. For electrons with an energy slightly off-resonance, the transmission is approximately $T_{min}T_{max}$, indicating that thin barriers, which yield a large transmission, provide a broad resonance width. For identical barriers, only the energy resonance width is controlled by barrier thickness.

Returning to the equilibrium condition, we will now investigate the transport across an arbitrary number of potential barriers. Assume, for simplicity, an infinite superlattice, a device with an infinite number of identical potential barriers and wells. In determining the energy bands for a bulk semiconductor, the periodicity of the atomic core potential results in a folding of the free-electron dispersion relationship into a Brillouin Zone; moreover, energy bands are perturbed from the free-electron condition. Similarly, the periodicity of a superlattice potential results in a folding of energy bands into a reduced Brillouin Zone, since the period of a superlattice (one barrier and one well) is much larger than the period of the crystal potential. Of course, further perturbation of the band structure can be accounted for, with an analysis such as the Kronig-Penney model taking into account the form of the superlattice potential. No real device, however, can have infinite extent. Just as the finite size of a crystal discretizes the Brillouin Zone into a fine mesh of allowed wavevectors, a finite number of superlattice periods further restricts the number of allowed wavevectors. It has been shown [14] that for a finite superlattice of N periods, there are N-1

energies in each band for which propagation is not attenuated (transmission is unity); the range of Bloch wavenumbers is segmented into N intervals, yielding N-1 resonant energies for each energy band.

This discussion is represented schematically in figures 1.2 and 1.3. Energy bands for an infinite superlattice of period L are shown in figure 1.2 for the entire range of unique Bloch wavenumbers. Considering a double barrier resonant tunneling device (a two-period finite superlattice), the range of Bloch wavenumbers is segmented into two intervals, yielding one resonant energy in each of the energy bands. A plot of transmission coefficient versus energy, figure 1.3, shows a single peak at each resonant energy, for each band. For a triple barrier resonant tunneling device, the range of Bloch wavenumbers is segmented into three intervals, yielding two resonant energies in each of the energy bands. Peaks in the transmission coefficient of the double barrier structure appear to be split into doublets, for each energy band. As the number of superlattice periods increases, nearly all energies in each band are allowed; conduction "minibands" arise.

### 1.1.2 The Origin of Negative Differential Resistance

The motivation for the preceeding discussion of resonant tunneling lies in its application, negative differential resistance (NDR). By definition, differential resistance is understood as

$$R_d = \frac{\partial V}{\partial I}.$$

Regions of NDR appear in a current-voltage relationship wherever current decreases as voltage increases.

With the aid of figure 1.4, the origin of NDR can be understood as conservation of transverse momentum, of electrons tunneling into a system of states with reduced dimensionality [15]. Consider a double barrier resonant tunneling device at absolute zero temperature, in equilibrium. Electrons incident from a Fermi sea in the contacts are restricted to energies below the Fermi level. If the first quasi-bound state is above the Fermi level, as shown in the diagram, then there are no electrons with sufficient energy for resonant tunneling. The application of a small bias, however, lowers the quasi-bound state in energy. Electrons with the proper longitudinal momentum are available for tunneling; specifically, consider electrons at the Fermi energy. Assuming the device is free of impurities and
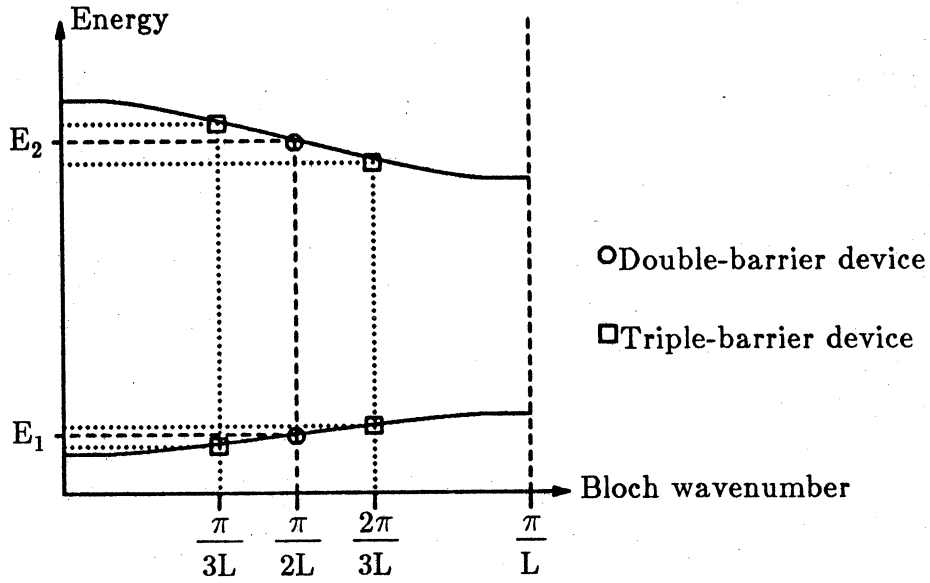
Figure 1.2    Energy bands of a superlattice in the range of unique Bloch wavenumbers
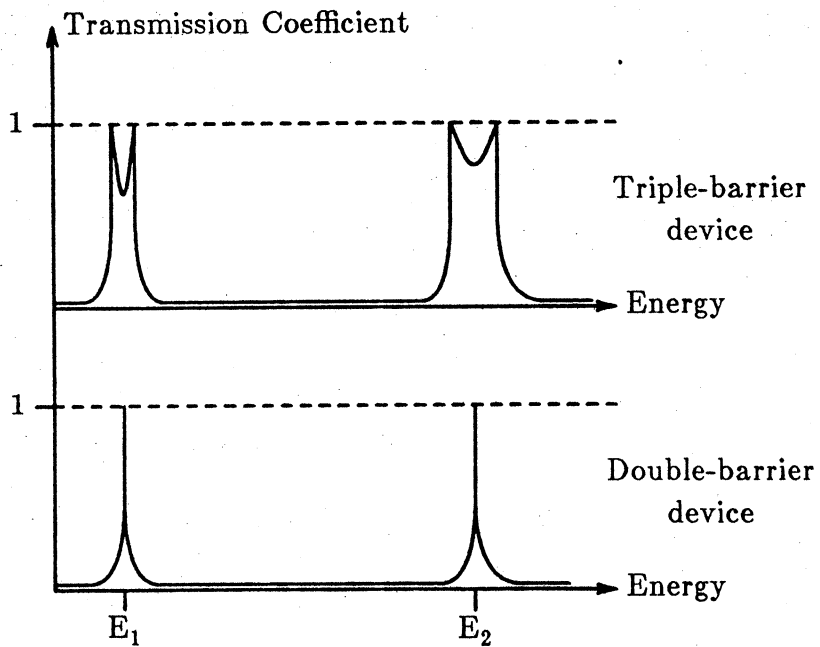


Figure 1.3    Transmission coefficient versus energy of incident electrons for two-period and three-period finite superlattices
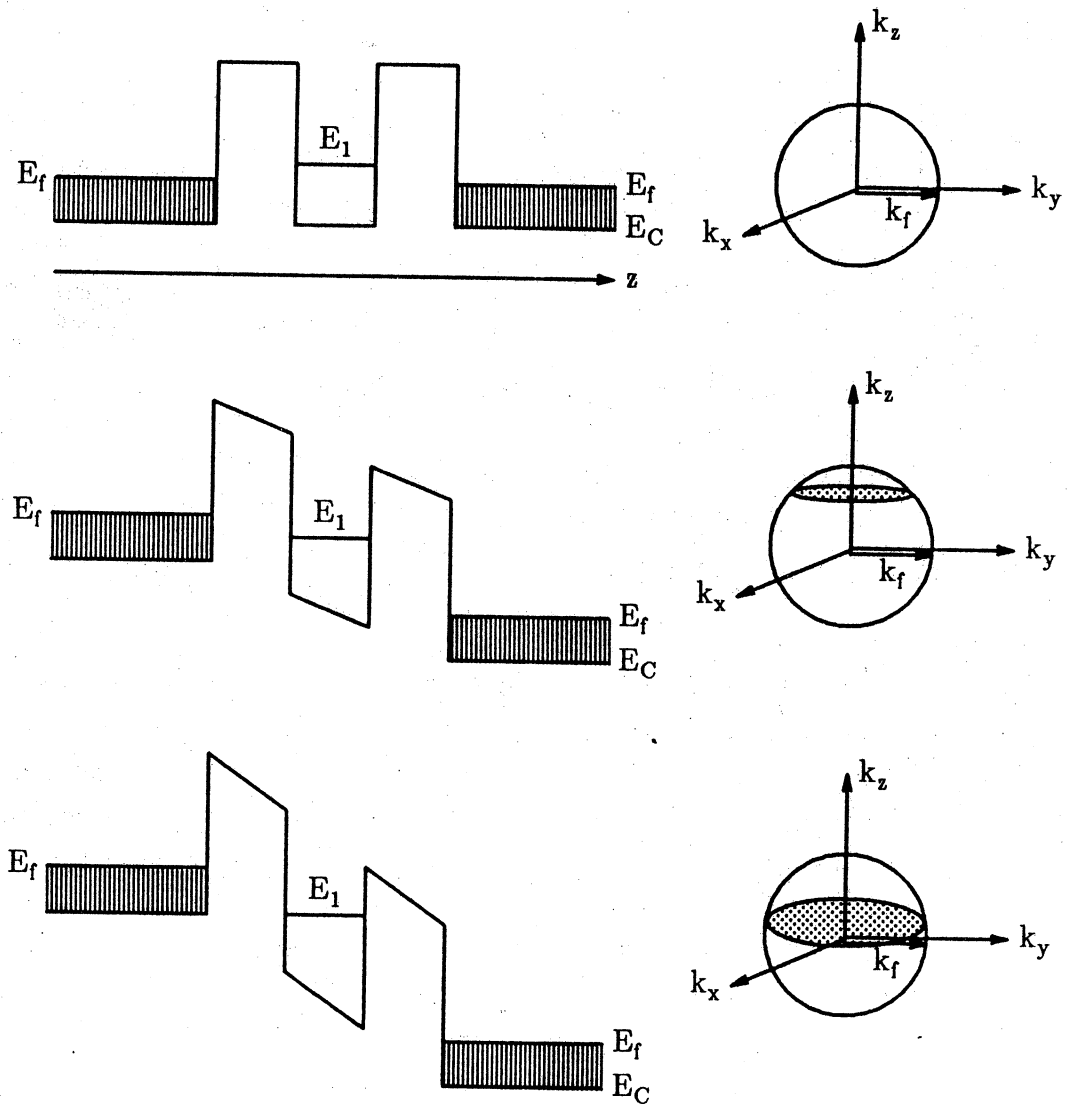
Figure 1.4 Negative differential resistance as a consequence of conservation of transverse momentum

inhomogeneities, energy and transverse momentum ($k_x$, $k_y$) must be conserved. The total number of electrons available for tunneling at the Fermi energy, therefore, corresponds to the shaded disk of the Fermi sphere. Further application of bias continues to lower the required longitudinal momentum, increasing the area of the shaded disk and the associated number of tunneling electrons. This area increases with bias, until the quasi-bound state falls below the conduction-band edge of the supplying contact, and current drops sharply, yielding negative differential resistance.

The dependence of NDR on conservation conditions discussed above was demonstrated experimentally by Morkoc et. al. [16]. A structure was fabricated with a quantum well, bounded on one side by a thin (25 Å) potential barrier, and bounded on the other by a thick (500 Å) barrier. Incident electrons could tunnel only through the thin barrier into the well, to be collected from the well by lateral diffusion. The device contrasts double barrier diodes, in which electrons propagate across the entire structure. Nevertheless, a pronounced NDR region was observed in the current-voltage characteristics at 77 ° K and 300 ° K, verifying the origin of NDR.

### 1.1.3 Coherent Resonant Tunneling *versus* Sequential Tunneling

In the discussion of coherent resonant tunneling in section 1.1.1, electron transport was assumed to be ballistic; collisions were assumed to be infrequent, so that multiply reflected electrons could exit the device before suffering a scattering event. Luryi [15] was the first to question this assumption, proposing an alternative transport mechanism. If, in the course of multiple reflections, an electron suffered a scattering event, the wavefunction phase would be randomized. Therefore, scattering events destroy the coherence needed for constructive interference in the Fabry-Perot effect. Without coherence, the wavefunction amplitude would not build-up within the well, and transmission would not be enhanced; tunneling, however, would still occur in a sequential fashion, and negative differential resistance, dependent only upon conservation considerations, would still be observed.

A question of the dominant transport mechanism, whether coherent resonant tunneling or sequential tunneling, was resolved by Capasso et. al. [17]; the argument presented is as follows. To achieve Fabry-Perot enhancement of transmission, a finite time $\tau_0$ is required for the

wavefunction of a multiply reflected electron to build-up within a well. Stone and Lee [18] proved that if this time constant $\tau_0$ is longer than the scattering time $\tau$, the transmission at resonance will be reduced by a factor proportional to $\tau_0/(\tau_0 + \tau)$; sequential tunneling will dominate. The scattering time $\tau$ is the inverse of the total scattering rate, including elastic processes due to impurities and inhomogeneities, and inelastics processes such as phonon scattering. In terms of the energy of electrons, the intrinsic resonance width $\Gamma_r$, the full width at half-maximum of a transmission resonance, is related to $\hbar/\tau_0$. As discussed earlier, thin barriers provide a broad intrinsic resonance width. A measure of scattering in terms of energy, the collision broadening of electrons, is related to $\hbar/\tau$. If the collision broadening is greater than the intrinsic resonance width, sequential tunneling will dominate. Therefore, conditions favorable for observing coherent resonant tunneling include using structures with thin barriers, to increase the intrinsic resonance width, and performing experiments at low temperatures, to reduce the collision broadening.

## 1.1.4 Effective Mass Filtering

By exploiting the competition between coherent resonant tunneling and sequential tunneling, Capasso et. al. [17] proposed an intriguing application: effective mass filtering. Since the tunneling probability through a potential barrier depends exponentially upon effective mass, heavy holes face a narrower intrinsic resonance width than electrons. In a properly designed superlattice, assuming collision broadening energies for both carriers are nearly equal, the transmission of electrons could be enhanced by coherent resonant tunneling, while heavy holes, relying on sequential tunneling, remain relatively localized in quantum wells. Superlattices can be used to filter the transport of carriers, allowing electrons to propagate and impeding heavy holes.

Applying this effect to photodetectors, Capasso et. al. [19] demonstrated a large photoconductive gain, which could be tuned by altering superlattice design parameters. In photodetectors, gain is determined by the ratio of electron and hole velocities, if the hole lifetime exceeds the hole transit time. If holes are relatively localized, gain is determined by the ratio of the electron lifetime to the electron transit time [17]. These properties are uncontrollable for conventional photodetectors using bulk semiconductors; superlattices, however, can be fabricated to provide a particular electron

transit time, allowing greater freedom in the design of photodetectors.


## 1.2 Proposed Devices

Success in the demonstration of resonant tunneling has instigated a number of proposals [20-24], for devices with a third terminal to control NDR characteristics. One such structure, described by Capasso et. al. [20] and shown in figure 1.5, is a heterojunction bipolar transistor with a single quantum well in the base region. Collector current is determined by the number of minority carriers, injected from the emitter, which traverse the base region. The presence of the resonant tunneling structure in the base restricts transmission of electrons to those with energies near resonance. As the base-emitter potential is manipulated, peaks in the collector current appear when the quasi-bound states of the quantum well align with the energy at which the maximum flux of electrons is injected. It was noted in section 1.1.1 that the application of bias across a double barrier device destroys the symmetry of the two barriers, degrading transmission. Since the controlling bias for this transistor is applied across the base-emitter junction, rather than across the base region itself, transmission peaks for the resonant tunneling structure remain close to unity; resulting peaks in collector current would be pronounced. When operated as shown in figure 1.6, the peaks and valleys in collector current give rise to an output voltage swing between binary levels. In this configuration, an array of transistors could be used to provide binary output for a high-speed, analog-to-digital converter.

Another device, the negative resistance Stark effect transistor (NERSET), was proposed by Bonnefoi et. al. [21]. The structure, presented in figure 1.7, is a double barrier resonant tunneling device, with an extra (base) contact. Because it is shielded by a thick (1000 Å - 1500 Å) potential barrier, base current is negligible. A plot of collector current versus collector-emitter voltage exhibits NDR, as expected for a resonant tunneling device; however, the voltage at which NDR is observed can be shifted, by adjusting the base-emitter voltage. The position of the collector contact is such that the emitter is not completely shielded from electric fields produced by the base.

Finally, Nakata et. al. [22] proposed a triode with a metal-insulator superlattice in the base, acting as an artificial semiconductor. Shown in figure 1.8, this resonant electron transfer triode (RETT) is expected to
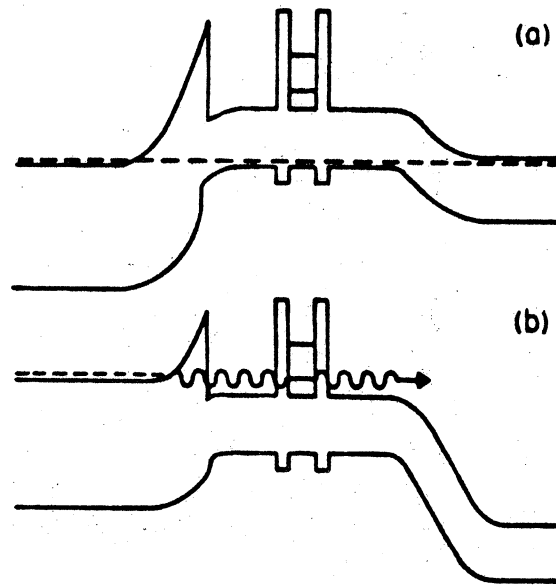
Figure 1.5    Resonant Tunneling Transistor (RTT) proposed by Capasso *et. al.* [20], shown (a) in equilibrium, and (b) with resonant tunneling through the first quasi-bound state
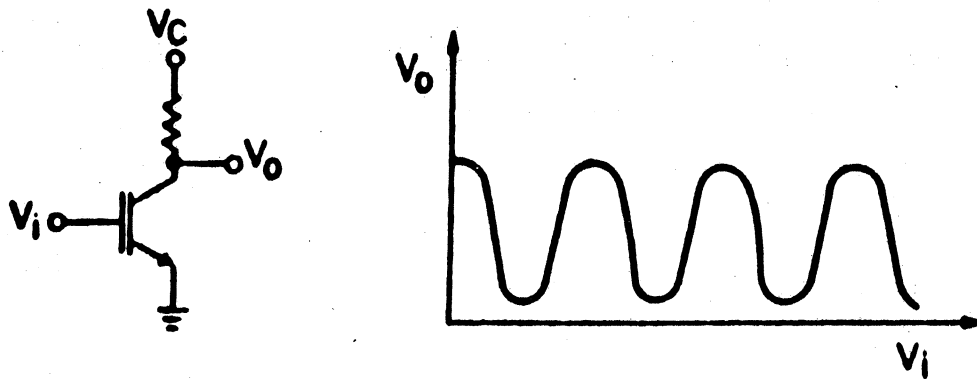


Figure 1.6    Voltage transfer characteristic for the RTT, presented by Capasso *et. al.* [20]
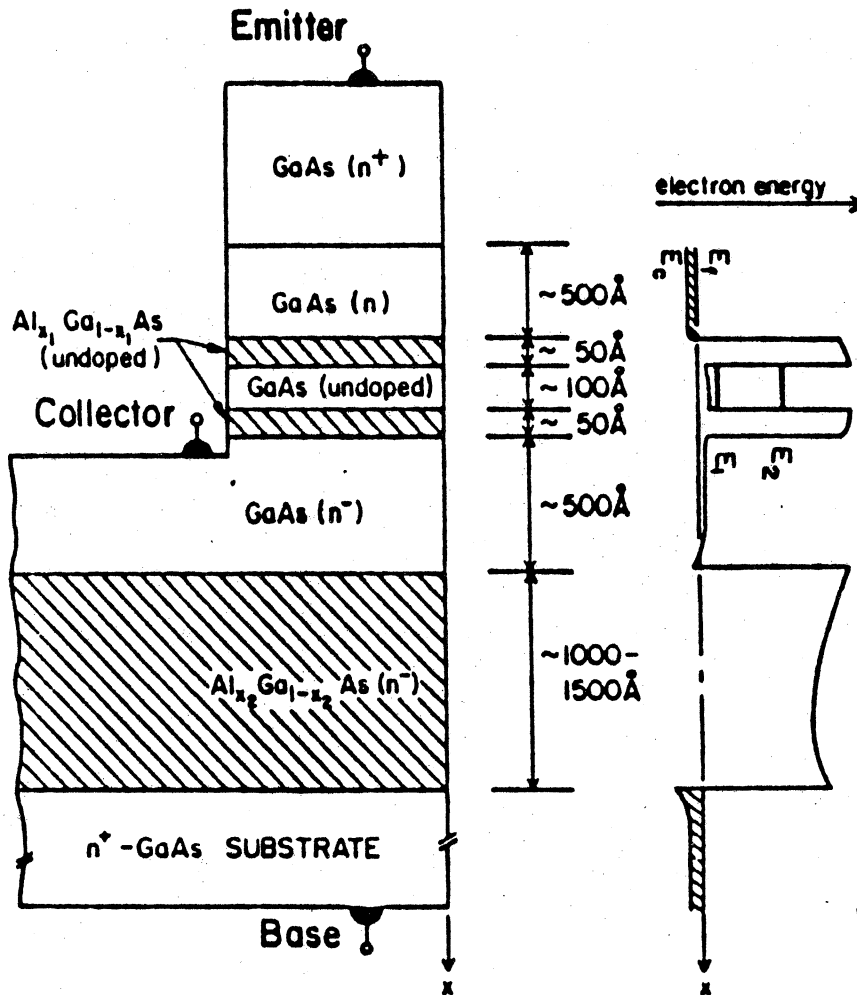
Figure 1.7   Cross section and conduction-band profile of the negative resistance Stark effect transistor (NERSET) proposed by Bonnefoi *et. al.* [21]
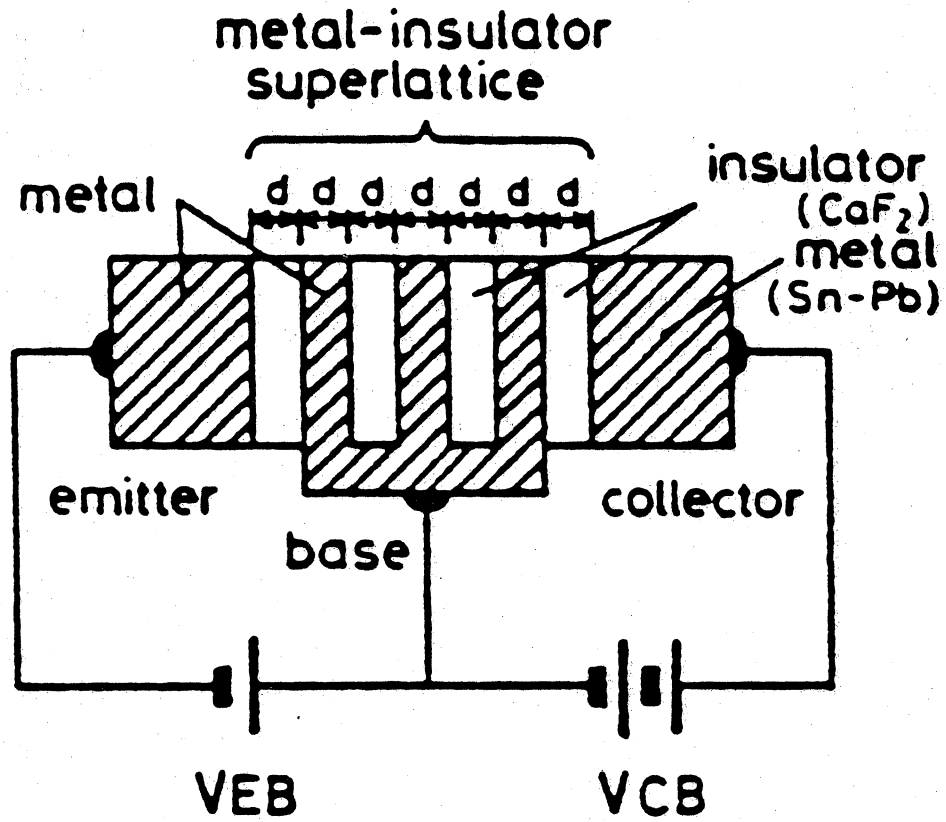
Figure 1.8    Resonant electron transfer triode (RETT) proposed by Nakata
*et. al.* [22]

perform well in high-speed applications because of the low resistivity of metal contacts. An artificial conduction-band in the base region is formed by the periodicity of the metal-insulator superlattice, as discussed in section 1.1.1. When the device is biased such that the emitter Fermi level is aligned with the artificial conduction-band in the base, electrons resonantly tunnel from emitter to collector. Of course, empty states are assumed to be available in the collector; the collector-base junction must be biased accordingly.

Other device proposals exist [23,24]; apart from these, the three presented above are easily understood, and one-dimensional in nature. Descriptions of device operation were intended to be qualitative, emphasizing the need for a quantum mechanical analysis tool, for a variety of one-dimensional applications. The remainder of this thesis is devoted to the development of such a tool.

## 1.3 Overview of Thesis

Chapter 2 provides the analytical foundation necessary for a one-dimensional, quantum mechanical analysis of arbitrary semiconductor heterostructures. Expressions for the evaluation of electron density and current density are presented, and two techniques for the solution of the Schrödinger equation are discussed in detail. The incorporation of space-charge effects, by a self-consistent solution of Schrödinger's and Poisson's equations, is also described. Numerical considerations for the implementation of this analysis are discussed in Chapter 3, highlighting the importance of integration in the calculation of electron and current densities. Example calculations are presented in Chapter 4, and concluding comments are given in Chapter 5. In addition to demonstrating quantum mechanical effects, the qualitative discussion of electron transport in the preceeding sections was presented to offer the insight into transport phenomena that a numerical simulation cannot provide.

# CHAPTER 2
# BALLISTIC ELECTRON TRANSPORT

To reduce a quantum mechanical description of carrier transport to a level feasible for numerical calculations, a number of simplifications must be made. First, the conduction-band profile is assumed to vary in only one dimension; both material composition and doping density are uniform in planes transverse to the direction of electron propagation. Second, participating semiconductors are assumed to be n-type, so that electrons provide the dominant contribution to both carrier and current densities; holes and electron-hole interactions are neglected. Furthermore, electron-electron interactions are assumed to be unimportant, allowing for a single particle approach to transport. Third, an envelope function solution for the electron wavefunction is assumed to apply. Effects of the crystal potential are parameterized by energy bands and an electron effective mass, which change abruptly at material interfaces. Bastard and Brum [25] have demonstrated that this assumption is reasonable, since, for the majority of lattice-matched heterostructures, Bloch wavefunctions differ little from one material to another. The validity of this assumption, however, should be questioned if material thicknesses are reduced to a few atomic sublayers. Finally, devices are assumed to be short, relative to an electron mean-free path, so that scattering events are infrequent, and can be neglected. With regard to the discussion of resonant tunneling in section 1.1.3, this is perhaps a drastic simplification. The dwell time of a multiply reflected electron in a "short" resonant tunneling device could be much longer than the time between collisions. Presently, much work remains for the incorporation of scattering into quantum transport theory.

## 2.1 Formulation of the Analysis

A number of formalisms have been pursued, including the quantum density matrix [26], Wigner function studies [27], and direct solution of the Schrödinger equation. As an analogue of the classical distribution function, the quantum density matrix represents the distribution of electrons in position space, along is main diagonal, and in momentum space, along its cross diagonal. From an equilibrium solution, it evolves in time according to the quantum Liouville equation, to which a collision term may be added for the simulation of scattering. Frensley [26], however, has reported some numerical instabilities when the time of simulation is large, and boundary conditions are difficult to implement. Since the Wigner distribution function can be obtained from the quantum density matrix by a transformation of variables, it is not surprising that some of the same numerical difficulties have been encountered [27]. Although neither of these methods has achieved the popularity of a direct solution of the Schrödinger equation, both are promising for future studies, offering a convenient means of adding collision terms.

Because of the insights provided by both transfer matrix and scatter matrix solutions, the details of these methods are presented below in section 2.2. Both methods, in spite of their differences, merely provide a means of solving a reduced Schrödinger equation. The following approach, common to both methods, was first proposed by Tsu and Esaki [28], and later modified by Vassell *et. al.* [29] to account for a spatially varying effective mass. The physical situation is pictured in figure 2.1: Contacts, assumed to be in local thermodynamic equilibrium, inject plane-wave electrons into an arbitrary, one-dimensional device, with a spectrum of wavevectors $\vec{k}$; the potential profile is static, assuming phonon scattering can be neglected, so that the electron wavefunction obeys the time-independent Schrödinger equation:

$$\frac{-\hbar^2}{2m^*(z)}\nabla_t^2\Psi_{\vec{k}}(\vec{r}) + \frac{\partial}{\partial z}\left[\frac{1}{m^*(z)}\frac{\partial}{\partial z}\Psi_{\vec{k}}(\vec{r})\right] + E_C(z)\Psi_{\vec{k}}(\vec{r}) = E\Psi_{\vec{k}}(\vec{r}) \quad (2.1)$$

where $\nabla_t$ represents the gradient of transverse directions (x and y), and effective mass variations along the z-axis have been taken into account, following Vassell [29]. The conduction-band profile $E_C(z)$ includes the electrostatic potential $\Phi(z)$, as well as the conduction-band offset for different material layers $E_m(z)$:
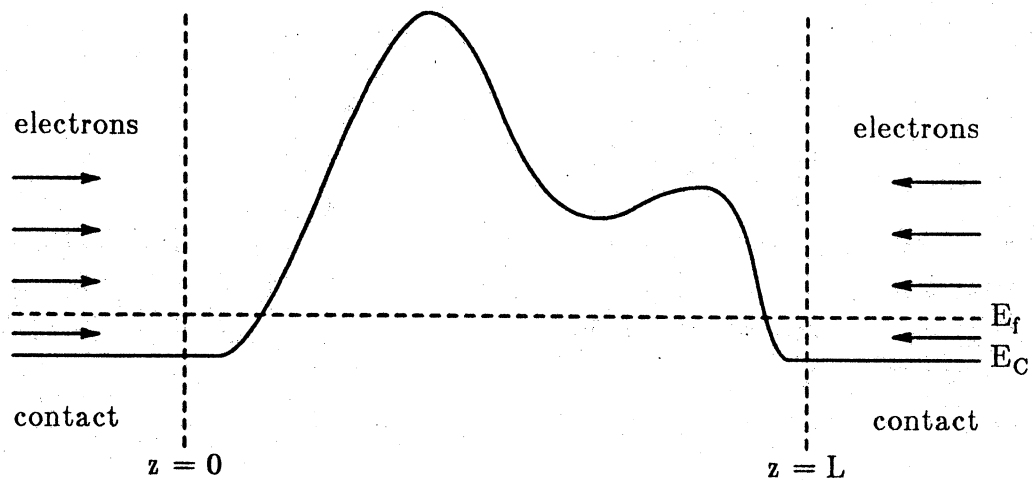
Figure 2.1    Plane-wave electrons, injected from contacts in thermodynamic equilibrium, propagate through an arbitrary potential profile

$$E_C(z) = E_m(z) - q\Phi(z) \tag{2.2}$$

Assuming the device dimensions in the transverse plane are much larger than the confinement along the z-axis, the longitudinal and transverse motions of the electron can be separated, and the wavefunction $\Psi_{\vec{k}}(\vec{r})$ can be written as:

$$\Psi_{\vec{k}}(x,y,z) = \frac{e^{ik_x x} e^{ik_y y}}{\sqrt{\Omega}} \psi_k(z) \tag{2.3}$$

where $\Omega$ is the volume of normalization. Upon substituting this expression into equation 2.1, and simplifying the result, one obtains:

$$\frac{\partial}{\partial z}\left[ \frac{1}{\gamma(z)} \frac{\partial}{\partial z} \psi_k(z) \right] + \frac{2m_c^*}{\hbar^2}[E - \frac{E_t}{\gamma(z)} - E_C(z)]\psi_k(z) = 0 \tag{2.4}$$

where E is the total energy of the injected electron; $E_t$ is the transverse energy $\hbar^2 k_t^2 / 2m_c^*$; $m_c^*$ is the effective mass in the injecting contact, and the function $\gamma(z) = m^*(z)/m_c^*$ represents the spatial variation of effective mass. By somehow approximating the conduction-band profile of an arbitrary device, the reduced Schrödinger equation, equation 2.4, can be solved numerically. Once this is accomplished, any other quantities of interest can be calculated, as described below.

### 2.1.1 Calculation of Electron Density

The electron density can be obtained by summing the magnitudes of electron wavefunctions, injected from the contacts with a spectrum of wavevectors. Consider electrons injected from the left contact (z = 0). The density of electrons propagating from left to right is

$$n^{l-r}(z) = \sum_{\vec{k}} | \Psi_{\vec{k}}(\vec{r}) |^2 f(E_k)$$

$$= \frac{1}{4\pi^3} \int d^3\vec{k} | \psi_k(z) |^2 f(E_k)$$

where

$$f(E_k) = \frac{1}{1 + e^{(E_k - E_f)/k_b T}} \tag{2.6}$$

is the Fermi-Dirac factor, and

$$E_k = \frac{\hbar^2 k^2}{2m_c^*} + E_C(0) \qquad \text{where} \quad k^2 = k_x^2 + k_y^2 + k_z^2$$

with $E_C(0)$ being the conduction-band edge in the injecting contact, which is later taken as a reference point, *i.e.*, $E_C(0) = 0$.

Note that, as shown in equation 2.4, $\psi_k(z)$ depends on the transverse energy $E_t$, complicating the integration. A rigorous estimate of $n(z)$ would then entail calculating $\psi_k(z)$ for all momenta, longitudinal and transverse, and integrating numerically. To avoid this complication and keep the solution tractable, we follow Vassell [29], and replace $E_t$ by its thermal average $(k_BT)$ in the reduced Schrödinger equation. The wavefunction can then be removed from the integration over transverse momentum:

$$n^{l-r}(z) = \int_0^\infty \frac{dk_z}{2\pi} \mid \psi_{k_z,k_BT} \mid^2$$

$$\int_0^\infty \frac{dk_t k_t}{\pi} [1 + \exp[(E_C(0) - E_f + \frac{\hbar^2}{2m_c^*}(k_z^2 + k_t^2))/k_BT]]^{-1} \quad (2.7)$$

and the integration over transverse momentum $k_t$ can be performed exactly:

$$n^{l-r}(z) = \int_0^\infty \frac{dk_z}{2\pi} \mid \psi_{k_z,k_BT} \mid^2 \sigma^{l-r}(k_z) \qquad (2.8)$$

where $\sigma^{l-r}(k_z)$ represents the integration over transverse momentum:

$$\sigma^{l-r}(k_z) = \frac{m_c^* k_BT}{\pi \hbar^2} \ln[1 + \exp[(E_f - E_C(0) - \frac{\hbar^2 k_z^2}{2m_c^*})/k_BT]] \qquad (2.9)$$

Following a similar derivation, the electron density associated with the flow of electrons from right to left can be obtained. The result is identical to equations 2.8 and 2.9, with the replacement of $E_C(0)$ by $E_C(L)$, and with the proper calculation of the wavefunction for injection by the right contact $(z = L)$. The total electron density is the sum of the contributions of two oppositely flowing streams:

$$n(z) = n^{l-r}(z) + n^{r-l}(z) \qquad (2.10)$$

## 2.1.2 Calculation of Current Density

In a manner similar to the calculation of electron density, the total current density can be resolved into two components: one due to electrons impinging from the left contact, and the other due to electrons from the right. For electrons propagating from left to right, the current density is given by

$$J^{l-r} = \sum_{\vec{k}} \frac{-q\hbar}{m_c^*} \, k_z \, T^{l-r}(\vec{k}) \, f(E_{\vec{k}})$$

$$= \frac{-q\hbar}{4\pi^3 m_c^*} \int d^3\vec{k} \, k_z \, T^{l-r}(\vec{k}) \, f(E_{\vec{k}})$$

where $T^{l-r}(\vec{k})$ is the transmission coefficient for electrons propagating from left to right:

$$T^{l-r}(\vec{k}) = \frac{k_z(L)/m^*(L)}{k_z(0)/m_c^*} \, | \, \psi_k^{l-r}(L) \, |^2 = \frac{k_z(L)}{k_z(0)} \frac{1}{\gamma(L)} \, | \, \psi_k^{l-r}(L) \, |^2$$

All other definitions apply, as presented in section 2.1.1. It was previously assumed that the wavefunction exhibits a weak dependence on the transverse energy $E_t$, so that $E_t$ could be taken as its constant thermal average, in the solution of the Schrödinger equation. Since the transmission coefficient is proportional to the magnitude of the wavefunction, this same assumption allows it to be removed from the integration over transverse momentum,

$$J^{l-r} = \frac{-q\hbar}{m_c^* \pi} \int_0^\infty \frac{dk_z}{2\pi} \, k_z \, T^{l-r}(k_z) \int_0^\infty \frac{dk_t}{2\pi} \, k_t \, f(E_k) \tag{2.11}$$

and the transverse integration can be performed exactly, as in section 2.1.1:

$$J^{l-r} = \frac{-q\hbar}{m_c^*} \int_0^\infty \frac{dk_z}{2\pi} \, k_z \, T^{l-r}(k_z) \, \sigma^{l-r}(k_z) \tag{2.12}$$

For electrons incident from the right contact, the solution is similar to equation 2.12, with use of the expressions $T^{r-l}(k_z)$ and $\sigma^{r-l}(k_z)$. As discussed in section 2.1.1, the factors $\sigma(k_z)$ differ only by the conduction-band edge of the injecting contact. Similarly, the roles of left and right must be switched for the transmission coefficient:

$$T^{r-l}(k_z) = \frac{k_z(0)}{k_z(L)} \frac{1}{\gamma(0)} \mid \psi_k^{r-l}(0) \mid^2$$

The total current density is then the difference of the two oppositely flowing streams of electrons:

$$J(z) = J^{l-r}(z) - J^{r-l}(z) \tag{2.13}$$

## 2.2 Solution of the Schrödinger Equation

To complete the analysis, the reduced Schrödinger equation, equation 2.4, must be solved for the entire spectrum of longitudinal wavevectors $k_z$; integrations for electron and charge densities can then be performed. For an arbitrary device, some numerical approximations of the conduction-band profile and effective mass variation must be made. In the following descriptions of transfer matrix and scatter matrix solutions, the device is segmented into a series of small intervals, over which both the conduction-band profile and effective mass are assumed to be constant. This situation is represented schematically in figure 2.2. In an interval of constant potential and constant effective mass, solutions of the Schrödinger equation are known analytically. Hence, the problem is reduced to matching solutions across the device at interval boundaries. An inspection of equation 2.4,

$$\frac{\partial}{\partial z}\left[\frac{1}{\gamma(z)}\frac{\partial}{\partial z}\psi_k(z)\right] + \frac{2m_c^*}{\hbar^2}[E - \frac{E_t}{\gamma(z)} - E_C(z)]\psi_k(z) = 0$$

repeated here for convience, leads to the condition:

$$\lim_{\iota\to 0}\int_{z-\iota}^{z+\iota}\frac{\partial}{\partial\xi}\left[\frac{1}{\gamma(\xi)}\frac{\partial}{\partial\xi}\psi_k(\xi)\right]d\xi = 0$$

Proper solutions of the Schrödinger equation require continuity of $\psi_k(z)$ and $\frac{1}{\gamma(z)}\frac{\partial}{\partial z}\psi_k(z)$, rather than the usual continuity of the wavefunction and its derivative. By representing solutions in a matrix form, these continuity conditions can be assured by cascading transfer or scatter matrices across the device. An application of contact boundary conditions, then, completes the solution.
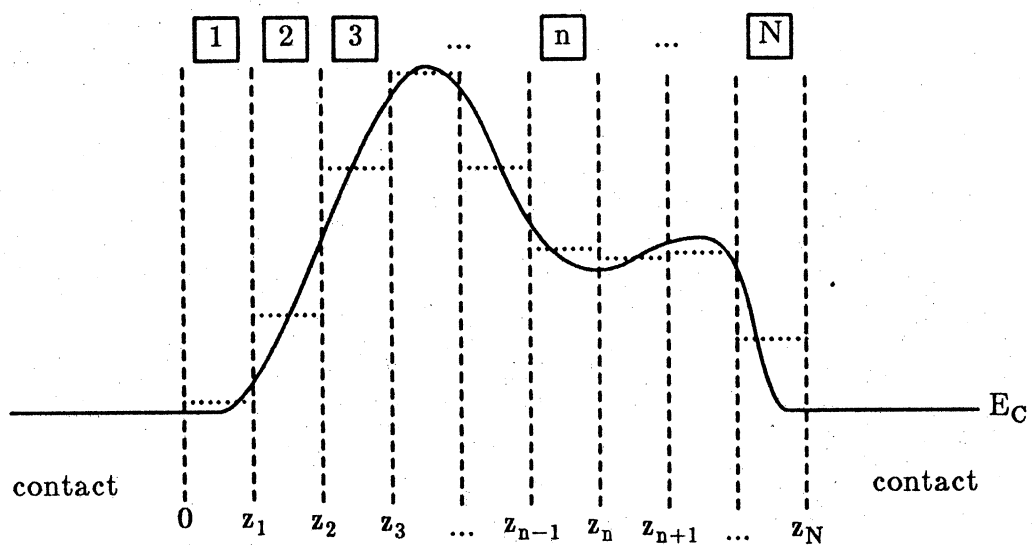
Figure 2.2    An arbitrary device is represented by a series of tiny intervals, over which the conduction-band profile is assumed to be constant

## 2.2.1 Transfer Matrix Technique

### 2.2.1.1 Deriving the Transfer Matrix

Consider a solution of the reduced Schrödinger equation in an interval $n$, shown in figure 2.2. Within the interval, both the effective mass and the conduction-band profile are constants; the Schrödinger equation further reduces to:

$$\frac{\partial^2}{\partial z^2}\psi_n(z) = -\beta_n^2\psi_n(z) \tag{2.14}$$

with the wavevector $\beta_n$, equivalent to $k_z(z_n)$, given by:

$$\beta_n^2 = \gamma_n \frac{2m_c^*}{\hbar^2}\left[E_z + E_t\left(1 - \frac{1}{\gamma_n}\right) - E_{C,n}\right]$$

where $\gamma_n$ and $E_{C,n}$ are taken as their average values at the interval boundaries. The longitudinal energy $E_z = \hbar^2 k_z^2/2m_c^*$ is constant, with the value determined by the incident momentum $k_z$. This second-order partial differential equation can be reduced to two first-order equations. Written in matrix form, equation 2.14 becomes:

$$\frac{\partial}{\partial z}\begin{bmatrix}\psi_n \\ \psi_n'\end{bmatrix} = \begin{bmatrix}0 & 1 \\ -\beta_n^2 & 0\end{bmatrix}\begin{bmatrix}\psi_n \\ \psi_n'\end{bmatrix} \tag{2.15}$$

where the prime notation denotes partial differentiation with respect to $z$. Equation 2.15 is of the form

$$\frac{\partial}{\partial z}\vec{u}(z) = A\vec{u}(z)$$

where $\vec{u}$ represents the column vector, and $A$ is the matrix. Since the matrix $A$ is a constant in the interval, the solution of equation 2.15 is an exponential:

$$\vec{u}(z) = \vec{u}(0)e^{zA} = \vec{u}(0)[I + zA + \frac{z^2}{2!}AA + \frac{z^3}{3!}AAA + \cdots]$$

After carrying out the Taylor series expansion with the matrix $A$ of equation 2.15, the wavefunction and its derivative at the end of an interval can be related to that of the interval start:

$$
\begin{bmatrix} \psi_n \\ \psi_n' \end{bmatrix}_{z=z_n^-} = \begin{bmatrix} \cos(\beta_n \delta_n) & \sin(\beta_n \delta_n)/\beta_n \\ -\beta_n \sin(\beta_n \delta_n) & \cos(\beta_n \delta_n) \end{bmatrix} \begin{bmatrix} \psi_n \\ \psi_n' \end{bmatrix}_{z=z_{n-1}^+} \tag{2.16}
$$

with $\delta_n$ being the width of the interval, $\delta_n = z_n - z_{n-1}$. An important subtlety is that the column vectors shown above are within the same interval $n$, by an infinitesimal amount. What remains is to apply continuity conditions to the wavefunction, and write the result in the form of a transfer matrix. As described earlier, proper solutions of the Schrödinger equation require:

$$
\psi_n(z_{n-1}^+) = \psi_{n-1}(z_{n-1}^-)
$$

$$
\frac{1}{\gamma_n} \frac{\partial}{\partial z} \psi_n(z_{n-1}^+) = \frac{1}{\gamma_{n-1}} \frac{\partial}{\partial z} \psi_{n-1}(z_{n-1}^-)
$$

With reference to figure 2.2, the plus and minus superscripts of $z_n$ represent positions an infinitesimal amount of the positive and negative sides of the interface. From this point forward, the superscript notation is dropped, with an understanding that positions $z_n$ lie within the interval $n$. Including the above continuity conditions, the transfer matrix describing the development of the wavefunction across an interval, is given by:

$$
\begin{bmatrix} \psi_n \\ \psi_n' \end{bmatrix}_{z=z_n} = \begin{bmatrix} \cos(\beta_n \delta_n) & \dfrac{\gamma_n}{\gamma_{n-1}} \dfrac{\sin(\beta_n \delta_n)}{\beta_n} \\ -\beta_n \sin(\beta_n \delta_n) & \dfrac{\gamma_n}{\gamma_{n-1}} \cos(\beta_n \delta_n) \end{bmatrix} \begin{bmatrix} \psi_n \\ \psi_n' \end{bmatrix}_{z=z_{n-1}} \tag{2.17}
$$

If the conduction-band profile were approximated by piece-wise linear segments, rather than constant intervals, the basis of the transfer matrix would be Airy functions. A solution of this type, reported by Lui and Fukuma [30], requires fewer intervals, but matrix elements are more difficult to calculate.

Note that, for an electron tunneling through a classically forbidden region, the wavevector $\beta_n$ would be purely imaginary. In this case, the sine and cosine functions are hyperbolic, and the elements of the transfer matrix could be quite large. In the process of cascading transfer matrices, described below, elements of a composite matrix can become large enough to cause overflow in a computer simulation. The checks needed to detect and avoid overflow are a serious limitation for the transfer matrix technique.

## 2.2.1.2 Applying the Transfer Matrix

The task remaining is to impose boundary conditions at the contacts. In the constant potential of the contacts, the electron wavefunction can be represented by plane-wave solutions traveling in both directions. As shown in figure 2.3, a plane-wave of unit amplitude is injected, and the resulting transmitted and reflected amplitudes must be determined. Assuming contacts are ideally ohmic, electrons reaching a contact are completely absorbed. The wavefunction and its derivative at the end of the device can be related to that at the start by cascading transfer matrices across each interval:

$$\begin{bmatrix} \psi \\ \psi' \end{bmatrix}_{z=z_N} = [M_N]\,[M_{N-1}]\,[M_{N-2}] \cdots [M_1] \begin{bmatrix} \psi \\ \psi' \end{bmatrix}_{z=0}$$

$$\begin{bmatrix} \psi \\ \psi' \end{bmatrix}_{z=z_N} = [M] \begin{bmatrix} \psi \\ \psi' \end{bmatrix}_{z=0}$$

where $[M_i]$ is the transfer matrix for section $i$, and $[M]$ is the overall transfer matrix. Consider electrons injected from the left contact. The plane-wave solutions shown in figure 2.3 can be represented as wavefunction vectors,

$$\begin{bmatrix} \psi \\ \psi' \end{bmatrix}_{z=0} = \begin{bmatrix} 1 + r \\ i\beta_0(1 - r) \end{bmatrix} \qquad \begin{bmatrix} \psi \\ \psi' \end{bmatrix}_{z=z_N} = \begin{bmatrix} t \\ i\beta_N t \end{bmatrix}$$

and the resulting matrix equation,

$$\begin{bmatrix} t \\ i\beta_N t \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} 1 + r \\ i\beta_0(1 - r) \end{bmatrix}$$

can be solved for the transmitted and reflected amplitudes, t and r. Here, the overall transfer matrix $[M]$ has been written showing each of its components. The reflected amplitude, therefore, is given by,

$$r = \frac{\left(\dfrac{\beta_0}{\beta_N}m_{22} - i\beta_0 m_{12}\right) - \left(m_{11} + i\dfrac{m_{21}}{\beta_N}\right)}{\left(\dfrac{\beta_0}{\beta_N}m_{22} - i\beta_0 m_{12}\right) + \left(m_{11} + i\dfrac{m_{21}}{\beta_N}\right)} \qquad (2.18)$$

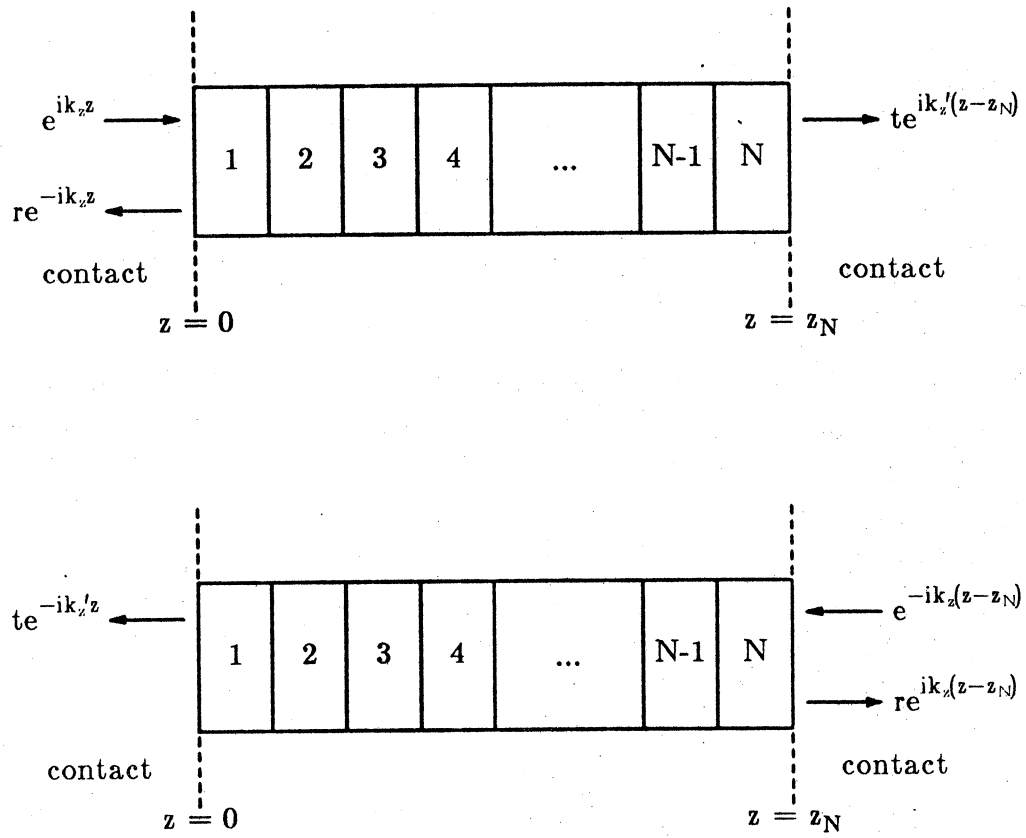Finally, the wavefunction and its derivative can be found at each point in the device:

Figure 2.3    Boundary conditions in the contacts for plane-wave injection

$$\left[\begin{array}{c} \psi \\ \psi' \end{array}\right]_{z=z_n} = [M_n] \, [M_{n-1}] \, \cdots \, [M_1] \left[\begin{array}{c} 1+r \\ i\beta_0(1-r) \end{array}\right] \qquad (2.19)$$

The solution shown above could be equally represented as:

$$\left[\begin{array}{c} \psi \\ \psi' \end{array}\right]_{z=z_n} = \Big[ [M_n] \, [M_{n+1}] \, \cdots \, [M_N] \Big]^{-1} \left[\begin{array}{c} t \\ i\beta_N t \end{array}\right]$$

However, in a numerical simulation, where the transmitted amplitude of tunneling electrons could be small enough to underflow, the incorrect value $t = 0$ would cause the entire wavefunction solution to be zero. Equation 2.19 is therefore the prudent choice for all transfer matrix applications.

### 2.2.2 Scatter Matrix Technique

### 2.2.2.1 Deriving the Scatter Matrix

The technique of scatter matrices is similar to the transfer matrix method, except that the column vectors used to represent the wavefunction are positively and negatively traveling wave amplitudes, rather than the wavefunction and its derivative. In general, scatter matrices relate the incoming and outgoing carrier fluxes at an interface. Neglecting recombination and generation processes, the number of carriers is conserved; consequently, scatter matrices are normally unitary. In the following approach, a scatter matrix is derived, relating incoming and outgoing wave amplitudes. Because wave amplitudes, rather than carrier fluxes, are employed as the basis, the resulting scatter matrix is not unitary. The number of carriers is conserved, however, provided that the proper definition of the transmission coefficient, presented in conjunction with transfer matrices, is used. As assumed in the transfer matrix technique, the conduction-band profile and spatially varying effective mass are broken into a series of intervals, over which both can be taken as constants. Solutions satisfying the reduced Schrödinger equation can be represented by positively and negatively traveling plane-waves, within an interval:

$$\frac{\partial^2}{\partial z^2} \psi_n(z) = -\beta_n^2 \psi_n(z) \qquad (2.20)$$

$$\psi_n(z) = \phi_n^+ e^{i\beta_n z} + \phi_n^- e^{-i\beta_n z} \tag{2.21}$$

where

$$\beta_n^2 = \gamma_n \frac{2m_c^*}{\hbar^2}\left[E_z + E_t\left(1 - \frac{1}{\gamma_n}\right) - E_{C,n}\right]$$

The scatter matrix we hope to obtain is of the form:

$$\begin{bmatrix} \phi_n^- \\ \phi_{n+1}^+ \end{bmatrix} = \begin{bmatrix} R & T' \\ T & R' \end{bmatrix} \begin{bmatrix} \phi_n^+ \\ \phi_{n+1}^- \end{bmatrix} \tag{2.22}$$

Consider the interface between intervals $n$ and $n+1$, shown in figure 2.4. For a plane-wave of unit amplitude injected from the left, the transmitted and reflected amplitudes, t and r, can be determined. For electrons incident from the left of the interval, the wavefunction in each interval is assumed to be:

$$\psi_n = \phi_n^+ e^{i\beta_n z} + \phi_n^- e^{-i\beta_n z}$$

$$= e^{i\beta_n z} + re^{-i\beta_n z}$$

$$\psi_{n+1} = \phi_{n+1}^+ e^{i\beta_{n+1} z} + \phi_{n+1}^- e^{-i\beta_{n+1} z}$$

$$= te^{i\beta_{n+1} z}$$

Reflections from interfaces to the right of interval $n$ are neglected for the present analysis, so that the input $\phi_{n+1}^-$ is zero. However, multiple reflections are considered explicitly in cascading scatter matrices, to obtain a composite matrix for the whole device. In figure 2.4, the position of the origin for an alternative coordinate system $z'$ has been shifted to the interface, to simplify the analysis; all exponential terms, evaluated at the interface, are unity. By demanding continuity of the wavefunction, and its derivative divided by the effective mass parameter, the following relationships are established:

$$1 + r = t$$

$$i\frac{\beta_n}{\gamma_n}(1 - r) = i\frac{\beta_{n+1}}{\gamma_{n+1}}t$$

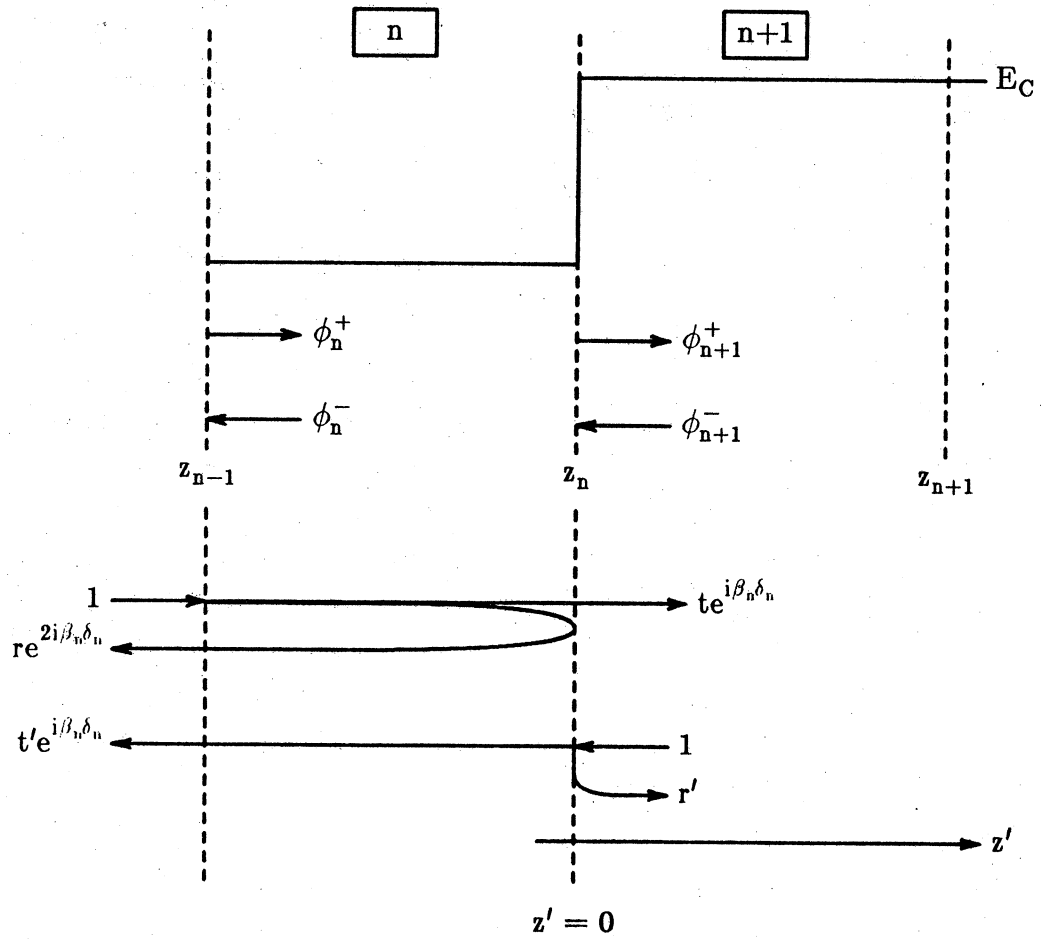from which the transmitted and reflected amplitudes can be determined:

Figure 2.4    Derivation of a scatter matrix by considering transmission and reflection of plane-wave solutions

$$t = \frac{2\beta_n/\gamma_n}{\beta_n/\gamma_n + \beta_{n+1}/\gamma_{n+1}} \qquad r = \frac{\beta_n/\gamma_n - \beta_{n+1}/\gamma_{n+1}}{\beta_n/\gamma_n + \beta_{n+1}/\gamma_{n+1}}$$

Of course, these amplitudes are valid only at the interface. For the scatter matrix to describe the entire interval $n$, the phase acquired by an electron while propagating in the interval must also be taken into account, as shown in figure 2.4. Following a similar derivation, the transmitted and reflected amplitudes, $t'$ and $r'$, for an electron incident from the right of the interface, can also be determined. The result of this analysis is embodied in a scatter matrix, relating the amplitudes of incoming and outgoing wavefunctions:

$$\begin{bmatrix} \phi_n^- \\ \phi_{n+1}^+ \end{bmatrix} = \begin{bmatrix} \dfrac{\kappa_n - \kappa_{n+1}}{\kappa_n + \kappa_{n+1}}e^{i2\beta_n\delta_n} & \dfrac{2\kappa_{n+1}}{\kappa_n + \kappa_{n+1}}e^{i\beta_n\delta_n} \\[3mm] \dfrac{2\kappa_n}{\kappa_n + \kappa_{n+1}}e^{i\beta_n\delta_n} & \dfrac{\kappa_{n+1} - \kappa_n}{\kappa_n + \kappa_{n+1}} \end{bmatrix} \begin{bmatrix} \phi_n^+ \\ \phi_{n+1}^- \end{bmatrix} \qquad (2.23)$$

where

$$\kappa_n = \frac{\beta_n}{\gamma_n}$$

$$\kappa_{n+1} = \frac{\beta_{n+1}}{\gamma_{n+1}} \quad .$$

### 2.2.2.2 Cascading Scatter Matrices

In the transfer matrix technique, presented in section 2.2.1, the method of cascading matrices was simple matrix multiplication; the transfer matrix related the wavefunction and its derivative at one end of an interval to that at the other end. Scatter matrices, however, relate the output wave amplitudes, on opposite sides of an interval, to the input amplitudes, also on opposite sides of an interval. A more complicated method of cascading scatter matrices is therefore required. Keeping in mind the definition of a scatter matrix, represented by equation 2.22, we will derive the composite scatter matrix for a region described by two, individual matrices.

The elements of a scatter matrix (R, T', T, R') represent the amplitudes of plane-wave electrons transmitted across (T, T') or reflected from (R, R') an interval. Within each interval is an interface (i.e., a discontinuity in the

conduction-band or effective mass), causing the wavefunction transmission and reflection. It follows that, in a region of two intervals, an electron could be multiply reflected between two interfaces. By summing the amplitudes of electrons following an infinite number of multiply reflected paths, the overall transmitted and reflected amplitudes can be determined. This analysis is pictured in figures 2.5 and 2.6. In figure 2.5, electrons incident from the left of the region could be transmitted straight through the structure, or could experience an infinite number of multiple reflections before being transmitted. The total transmitted amplitude is the sum of the contributions from each possible path:

$$T = T_2[1 + R_1' R_2 + (R_1' R_2)^2 + \cdots ]T_1$$

where the subscripts indicate the interval in which transmission or reflection occurs. The phase acquired by the wavefunction in the course of multiple reflections, considered explicitly in section 1.1.1, is contained in the definition of scatter matrix elements, and need not be added in this analysis. Since the infinite series of multiply reflected terms is a geometric series, the composite transmitted amplitude can be simplified:

$$T = T_2[1 - R_1' R_2]^{-1}T_1 \tag{2.24}$$

Following a similar derivation, the remaining elements of the composite scatter matrix can be determined, completing the analysis:

$$R = R_1 + T_1' R_2[1 - R_1' R_2]^{-1}T_1 \tag{2.25}$$

$$T' = T_1' [1 + R_2[1 - R_1' R_2]^{-1}R_1' ]T_2' \tag{2.26}$$

$$R' = R_2' + T_2[1 - R_1' R_2]^{-1}R_1' T_2' \tag{2.27}$$

In contrast to transfer matrices, the elements of a scatter matrix remain bounded. Even if the product $R_1' R_2$ were to approach unity, making the inverse of $[1 - R_1' R_2]$ large, transmitted amplitudes $T_1'$ and $T_2$ would approach zero, by particle conservation laws. Since many computers allow small numbers to underflow benignly to zero, a scatter matrix solution can be implemented with relatively little error checking.

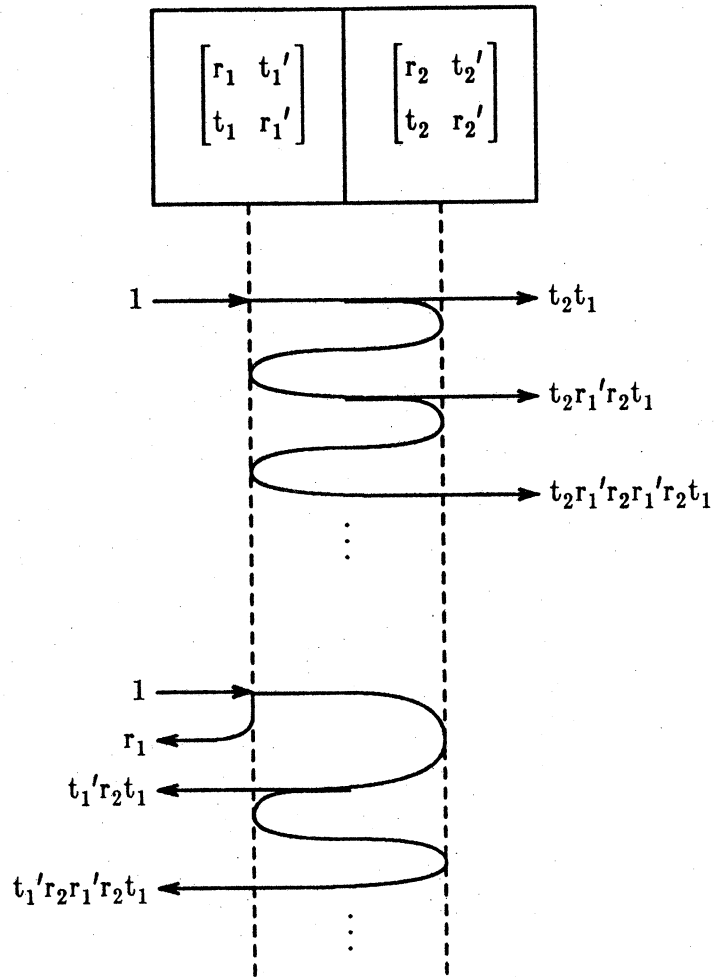Figure 2.5   Derivation of T and R for a composite scatter matrix, by considering multiple reflections of the wavefunction

Figure 2.6   Derivation of T' and R' for a composite scatter matrix, by considering multiple reflections of the wavefunction

### 2.2.2.3 Applying the Scatter Matrix

With the ability to cascade scatter matrices, and with the same contact boundary conditions described for transfer matrices (see figure 2.3), the scatter matrix solution can be completed. Assuming the scatter matrices for each interval are known, we define the composite scatter matrix for an interval $n$ as the cascade of all scatter matrices from interval 1 to interval $n$:

$$\mathbf{S}_1^C = \mathbf{S}_1$$

$$\mathbf{S}_2^C = \left\{ \mathbf{S}_1 , \mathbf{S}_2 \right\}$$

$$\vdots$$

$$\mathbf{S}_n^C = \left\{ \mathbf{S}_{n-1}^C , \mathbf{S}_n \right\}$$

where $\mathbf{S}_i$ is the individual scatter matrix for interval $i$, $\mathbf{S}_n^C$ is the composite scatter matrix for intervals 1 to $n$, and $\{\bullet,\bullet\}$ represents the cascade operation described in section 2.2.2.2. An application of the contact boundary conditions shown in figure 2.3 results in a trivial solution for the transmitted and reflected amplitudes. For a device of N intervals, the composite scatter matrix for the entire structure is $\mathbf{S}_N^C$. Remembering the definition of a scatter matrix, one would expect the overall transmitted and reflected amplitudes, for electrons injected from the left contact, to be $T_N^C$ and $R_N^C$, respectively. This is the result obtained:

$$\begin{bmatrix} r \\ t \end{bmatrix} = \begin{bmatrix} R_N^C & T'^C_N \\ T_N^C & R'^C_N \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

The wavefunction amplitude at the left contact ($z = z_0$) is then $1 + R_N^C$, and that at the right contact ($z = z_N$) is $T_N^C$. Similarly, the wavefunction amplitudes for electrons injected from the right contact are $1 + R_N'^C$ (at $z = z_N$), and $T_N'^C$ (at $z = z_0$).

What remains is to determine the wavefunction at all other points in the device. This is accomplished by transforming the scatter matrix (equation 2.22) into a transmission matrix:

$$\begin{bmatrix} \phi_{n+1}^+ \\ \phi_{n+1}^- \end{bmatrix} = \begin{bmatrix} T - R'\,[T']^{-1}R & R'\,[T']^{-1} \\ -[T']^{-1}R & [T']^{-1} \end{bmatrix} \begin{bmatrix} \phi_n^+ \\ \phi_n^- \end{bmatrix} \qquad (2.28)$$

The transmission matrix relates the amplitudes of positively and negatively traveling waves, on one side of an interval, to that on the other side of an interval; it is a restatement of equation 2.22, with a more convenient vector basis. For electrons injected from the left contact, the composite scatter matrix $S_n^C$ can be transformed into a transmission matrix, relating the wave amplitudes at a position $z = z_n$ to the known amplitudes in the contact. Hence, the wavefunction amplitude can be determined at any point $n$ in the device:

$$\begin{bmatrix} \phi_n^+ \\ \phi_n^- \end{bmatrix} = \begin{bmatrix} T_n^C - R_n'^C\,[T_n'^C]^{-1}R_n^C & R_n'^C\,[T_n'^C]^{-1} \\ -[T_n'^C]^{-1}R_n^C & [T_n'^C]^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ r \end{bmatrix} \qquad (2.29)$$

$$\psi_n = \phi_n^+ + \phi_n^- \qquad (2.30)$$

As mentioned above, $r$ is the overall reflected amplitude, taken as $R_N^C$ from the composite scatter matrix for all $N$ intervals of a device. For electrons injected from the right contact, the procedure is similar; however, scatter matrices must be cascaded in the opposite order, so that the wavefunction at any point in the device can be related to that of the injecting contact.

## 2.3 Incorporating Space-Charge Effects

As explained in section 2.1.1, the calculation of electron density requires a solution of the reduced Schrödinger equation (equation 2.4), for a spectrum of incident wavevectors $k_z$. For both of the solution techniques presented above, the conduction-band profile was assumed to be known. Included in the conduction-band profile, however, is the electrostatic potential $\Phi(z)$ (see equation 2.2), which depends on the electron density, as specified by the Poisson equation:

$$\frac{d}{dz}\left( \epsilon(z)\frac{d}{dz}\Phi(z) \right) = -q[N_D^+(z) - N_A^-(z) - n(z)] \qquad (2.31)$$

where $\epsilon(z)$ is the dielectric constant, $N_D^+(z)$ and $N_A^-(z)$ are respectively the donor and acceptor impurity concentrations, and $n(z)$ is the electron density.

A proper device analysis requires solutions for electron density and electrostatic potential to be self-consistent; therefore, space-charge effects can be incorporated into the analysis by an iterative solution for both electron density and electrostatic potential. First, a guess for the electrostatic potential, such as a classical solution of the problem, is used to determine the electron density. This density is then used in the solution of equation 2.31, for a corrected guess to the electrostatic potential. A modified electron density is then determined from the corrected potential, and the process is continued until convergence is satisfactory. Although space-charge effects were neglected in many early analyses [26-32], recent reports [33-34] have established the importance of a self-consistent solution; the example calculations, presented in Chapter 4, confirm this.

# CHAPTER 3
# NUMERICAL IMPLEMENTATION

Given the theoretical foundation presented in Chapter 2, a numerical simulation of an arbitrary device would seem straightforward; the flowchart of figure 3.1 illustrates a simplified view of the analysis. First, parameters specifying the device--the position-space grid, electrostatic potential, conduction-band discontinuities, effective masses, and dielectric constants--must be read in. Next, by injecting electrons from each contact from a mesh of $k_z$ values, the total electron and current densities can be determined. Finally, if a self-consistent solution is desired, the Poisson equation can be solved for a corrected electrostatic potential, and iteration between solutions for both electron density and electrostatic potential can be performed. In the process of implementing these calculations, however, several difficulties arise. One example, discussed in section 2.2.1.1, is that elements of a transfer matrix can become unbounded for tunneling electrons. The following sections address some details of a numerical implementation, emphasizing serious concerns for the proper integration of the wavefunction magnitude.

## 3.1 Numerical Integration Concerns

The electron and current densities can be determined by integrating the magnitude of the wavefunction, as explained in Chapter 2. Relevant integration formulas, for electrons injected from the left contact, are repeated here, for convenience:

$$n^{l \to r}(z) = \int_0^\infty \frac{dk_z}{2\pi} \mid \psi_{k_z, k_f, T} \mid^2 \sigma^{l \to r}(k_z)$$

$$J^{l \to r} = \frac{-q \hbar}{m_c^*} \int_0^\infty \frac{dk_z}{2\pi} \, k_z \, T^{l \to r}(k_z) \, \sigma^{l \to r}(k_z)$$

START

Read in
device

Compute:
$n^{l-r}$, $J^{l-r}$

Compute:
$n^{r-l}$, $J^{r-l}$

Compute new
potential $\Phi$

$n = n^{l-r} + n^{r-l}$
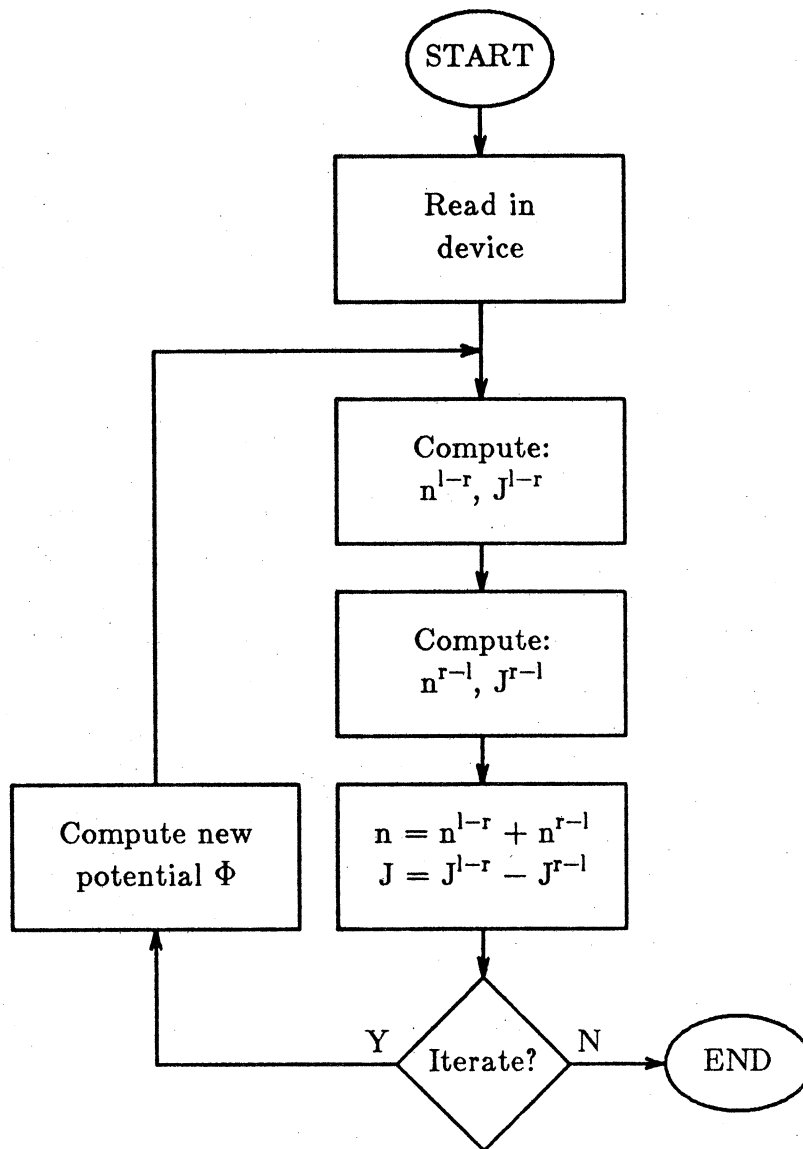$J = J^{l-r} - J^{r-l}$

Y    Iterate?    N

END

Figure 3.1    Simplified flowchart of the analysis

where

$$T^{l-r}(k_z) = \frac{k_z(L)}{k_z(0)} \frac{1}{\gamma(L)} \mid \psi^{l-r}_{k_z,k_BT}(L) \mid^2$$

Since the wavefunction must be determined numerically for a given value of $k_z$, the above integrations involve the specification of some finite mesh of $k_z$ values. A simple approach would be a uniform mesh of wavevectors $k_z$, with an application of the Trapazoidal rule or Simpson's rule, to account for integration. Such an approach, however, would fail miserably in certain devices; specifically, resonant tunneling devices offer the greatest challenge for numerical integration, because resonant peaks of the wavefunction within the well region can be difficult to resolve.

Figure 3.2 underscores the importance of this concern, with a plot of the wavefunction magnitude versus position and $k_z$, for a particular resonant tunneling device. The quantity $\log_{10}[\sigma(k_z) \mid \psi_{k_z,k_BT} \mid^2]$ plotted is that used in the integration for electron density, presented on a logarithmic scale so that all features are shown clearly. It represents the probability of finding an electron, with a particular wavevector, at a particular position, weighted by the transverse integration of Fermi-Dirac statistics. Electrons, injected from the right-hand contact, are largely reflected at low energies, (small $k_z$), setting up patterns of standing waves. At the resonant energy, however, the wavefunction peaks sharply in the well region, and a small ridge of transmitted electrons can be seen extending to the left contact. Because this peak provides the dominant contribution to electron density in the well region, a uniform mesh of wavevectors $k_z$ would have to be sufficiently fine to resolve the resonant peak. Since the peak could be arbitrarily narrow with respect to $k_z$, a criterion for "sufficiently fine" is difficult to establish. Moreover, a fine mesh of wavevectors is required only for values of $k_z$ near the resonance. Segmenting the wavevector space into a series of uniform meshes, fine meshes for wavevectors near a resonance, and coarse meshes elsewhere, might seem to be the solution; however, variations in the electrostatic potential (*e.g.*, by iterating for a self-consistent solution) affect the energy at which resonances occur. Therefore, a proper integration requires the ability to identify and resolve all peaks in the wavefunction magnitude, regardless of device specifics.
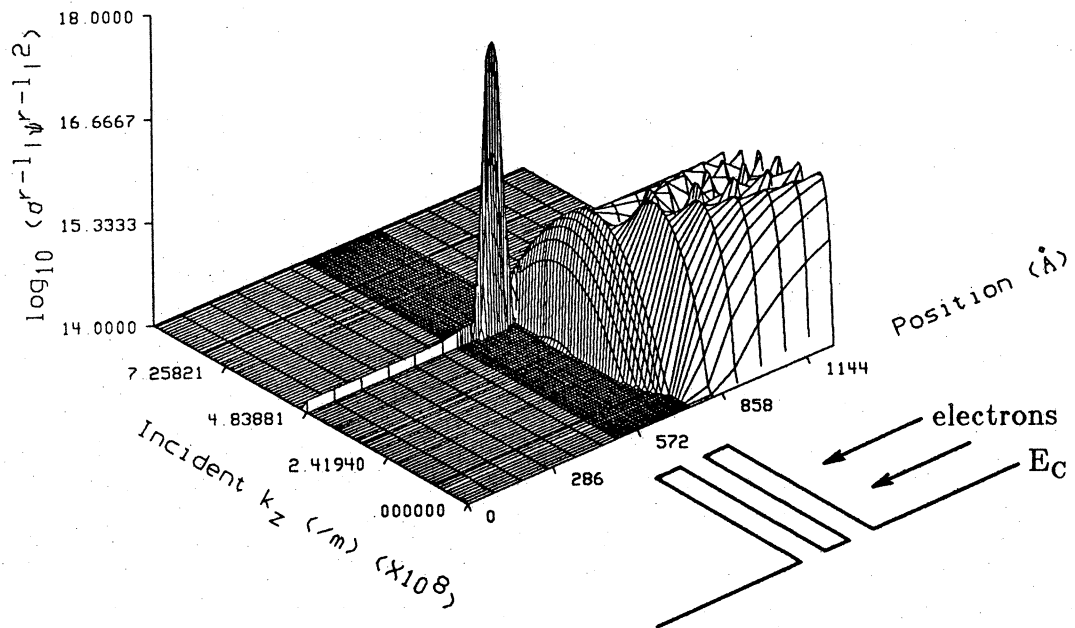
Figure 3.2    The wavefunction magnitude, weighted by the transverse integration $\sigma(k_z)$, for electrons incident from the right contact of a resonant tunneling device

### 3.1.1 Isolating Extrema in Reciprocal Space

In light of the preceeding discussion, it is necessary to determine a method of isolating the incident wavevector $k_z$ for which a maximum occurs in the wavefunction magnitude. Consider the reduced Schrödinger equation, in an interval $n$ of constant potential and constant effective mass:

$$\frac{\partial^2}{\partial z^2}\psi_n + \beta_n^2 \psi_n = 0 \tag{3.1}$$

where the wavevector $\beta_n$, at a position $z = z_n$, is understood as

$$\beta_n^2 = \gamma_n \frac{2m_c^*}{\hbar^2}\left[E_z + E_t\left(1 - \frac{1}{\gamma_n}\right) - E_{C,n}\right] \tag{3.2}$$

Taking the derivative with respect to $k_z$, we obtain:

$$\frac{\partial}{\partial k_z}\frac{\partial^2}{\partial z^2}\psi_n + \beta_n^2 \frac{\partial}{\partial k_z}\psi_n + \psi_n \frac{\partial}{\partial k_z}\beta_n^2 \tag{3.3}$$

Note that the derivative of the wavevector $\beta_n$ can be simplified:

$$\frac{\partial}{\partial k_z}\beta_n^2 = \frac{\partial}{\partial k_z}\left[\gamma_n \frac{2m_c^*}{\hbar^2}\left(\frac{\hbar^2 k_z^2}{2m_c^*} + E_t(1 - 1/\gamma_n) - E_{C,n}\right)\right] \cdot$$

$$\frac{\partial}{\partial k_z}\beta_n^2 = 2\gamma_n k_z \tag{3.4}$$

where the identity $E_z = \hbar^2 k_z^2/2m_c^*$ has been invoked. Upon substituting this result back into equation 3.3, and multiplying through by the complex conjugate of the wavefunction $\psi_n^*$, we obtain:

$$\psi_n^* \frac{\partial}{\partial k_z}\frac{\partial^2}{\partial z^2}\psi_n + \beta_n^2 \psi_n^* \frac{\partial}{\partial k_z}\psi_n + 2\gamma_n k_z \psi_n^* \psi_n = 0 \tag{3.5}$$

A similar procedure can be applied to the complex conjugate of equation 3.1, leading to the result:

$$\psi_n \frac{\partial}{\partial k_z}\frac{\partial^2}{\partial z^2}\psi_n^* + \beta_n^2 \psi_n \frac{\partial}{\partial k_z}\psi_n^* + 2\gamma_n k_z \psi_n \psi_n^* = 0 \tag{3.6}$$

Note that $\beta_n^2$ is a real quantity, and is therefore invariant under complex conjugation. The sum of equations 3.5 and 3.6 can be written as:

$$\psi_n^* \frac{\partial}{\partial k_z} \frac{\partial^2}{\partial z^2} \psi_n + \psi_n \frac{\partial}{\partial k_z} \frac{\partial^2}{\partial z^2} \psi_n^* + \beta_n^2 \frac{\partial}{\partial k_z} \left( \psi_n^* \psi_n \right) + 4\gamma_n k_z \psi_n^* \psi_n = 0 \quad (3.7)$$

The condition for extrema in wavevector space is $\partial(\psi_n^* \psi_n)/\partial k_z = 0$, which cancels one term in the equation above. Assuming the magnitude of the wavefunction is not zero, the following relationship holds when a extremum of the wavefunction magnitude is found:

$$k_z = \frac{1}{4\gamma_n \psi_n^* \psi_n} \left[ \psi_n^* \frac{\partial}{\partial k_z} (\beta_n^2 \psi_n) + \psi_n \frac{\partial}{\partial k_z} (\beta_n^2 \psi_n^*) \right] \quad (3.8)$$

Note that an extremum in wavevector space could be found simultaneously at several points in position space. Hence, an evalution of equation 3.8 should be carried out at all relevant points in position space. For the calculation of current density, the relevant points are at the left and right contacts; for the calculation of electron density, all points in position space could be relevant.

Equation 3.8 will be satisfied exactly by incident wavevectors $k_z$ producing an extremum, in an interval $n$ in position space; for incident wavevectors near an extremum, we define a predicted wavevector $k_p^n$, in an interval $n$:

$$k_p^n = \frac{1}{4\gamma_n \psi_n^* \psi_n} \left[ \psi_n^* \frac{\partial}{\partial k_z} (\beta_n^2 \psi_n) + \psi_n \frac{\partial}{\partial k_z} (\beta_n^2 \psi_n^*) \right] \quad (3.9)$$

When the incident wavevector $k_z$ nears an extremum for a particular interval in position space, the ratio $k_p^n/k_z$ will tend to unity. In the course of stepping through wavevectors $k_z$, the ratio $k_p^n/k_z$ can be evaluated numerically, by approximating the derivative with respect to $k_z$. When this ratio nears unity, the corresponding wavefunction extremum can be resolved. A clear example of this type of calculation is provided by electrons incident on an infinite potential wall from a region of constant potential and constant effective mass. In this example, the total reflection of electrons produces standing waves, of the form:

$$\psi_{k_z}(z) = e^{ik_z z} - e^{-ik_z z}$$

$$\psi^* \psi = 2(1 - \cos 2k_z z)$$

The wavefunction magnitude $\psi^* \psi$ reaches an extremum when $2k_z z = m\pi$, for any integer m. In the framework of the present analysis, the ratio $k_p/k_z$ is:

$$\frac{k_p}{k_z} = \frac{1}{(1 - \cos 2k_z z)}\left[(1 - \cos 2k_z z) + 4z k_z \sin 2k_z z\right] \qquad (3.10)$$

which is unity when $2k_z z = m\pi$, for odd integers m, corresponding to the maxima of the wavefunction magnitude. For the minima, the wavefunction magnitude is zero, so the ratio $k_p/k_z$ becomes unbounded. Note that, for incident wavevectors slightly on either side of a maximum, the ratio $k_p/k_z$ yields values on either side of unity. The wavevector corresponding exactly to the maximum can then be found by a linear interpolation between the two ratios surrounding unity.

Of course, in an application of the above analysis, the steps in wavevector space must be small, relative to the distance between maxima. The remaining task is to define a suitable increment for the incident wavevector, $k_z$. Returning to the example of an infinite potential wall, we find that extrema are separated by $\pi/2z$, in wavevector space; in this case, the increment between incident wavevectors should be a fraction of this separation. In a simulated device, potential steps will have a finite magnitude. For small incident wavevectors, however, a large potential step would appear infinite, and electrons would experience strong reflection. On the other hand, small potential steps would seem insignificant to electrons incident with a large wavevector, and reflection would be neglegible. These two extremes define a range of situations encountered in device analysis. The case of strong reflection, causing peaks in the wavefunction characteristic of standing waves, is the case of interest; with little reflection, the wavefunction magnitude is nearly constant. Therefore, a conservative estimate for the wavevector increment can be obtained by considering the case of an infinite potential wall. At a distance L from the potential wall in position space, extrema are separated in wavevector space by:

$$\Delta k_z = \frac{\pi}{2L} \qquad (3.11)$$

For an arbitrary device, the length an electron must travel before experiencing reflection could be the length of the device, assuming the only reflecting interface is near one of the contacts. If the length L in equation 3.11 is taken as the length of the device, this relationship provides the desired estimate of the wavevector increment. Of course, a better resolution of the wavefunction--at the expense of execution time--could be obtained, by taking a fraction of equation 3.11 as the wavevector increment.

## 3.1.2 Gaussian Quadrature with Legendre Polynomials

Once the maxima of the wavefunction magnitude have been resolved, numerical integration can be performed. Intervals of wavevector space between maxima could be integrated with a number of numerical methods; of these, Gaussian quadrature [35] is the most attractive prospect. Considering the burden of iterating for a self-consistent solution, the number of wavefunction evaluations must be kept to a minimum. Gaussian quadrature allows for an extremely high-order approximation, with relatively few function evaluations. A set of Legendre polynomials is used to interpolate between points of a non-uniform mesh. Each mesh point is a zero of a Legendre polynomial, mapped into the space of integration. Specifically, the integral of any function f(x) can be approximated as:

$$\int_{x_1}^{x_2} f(x) \, dx = \frac{x_2 - x_1}{2} \sum_{i=1}^{m} w_i \, f(x_i) \tag{3.12}$$

with the transformation,

$$x_i = \frac{x_2 + x_1}{2} + \frac{x_2 - x_1}{2} \xi_i \tag{3.13}$$

where $\xi_i$ are the zeroes of a Legendre polynomial of order $m$, and $w_i$ are function weights; both $zi_i$ and $w_i$ are determined from a detailed derivation, and results are tabulated in many numerical analysis texts [35]. Using a Legendre polynomial of order $m$ in equation 3.12, the method is exact for all polynomial integrands of degree less than 2m. Estimates of error in this approximation involve high-order derivates of the function f(x). For this reason, practical error estimates are obtained by comparison of integration results for a number of different orders $m$. In general, caution should be used in applying Gaussian quadrature to a function which oscillates rapidly, or which becomes singular at any point. In the interval of wavevector space between maxima, however, the wavefunction is well behaved; Gaussian quadrature can be applied with confidence.

## 3.2 Iteration For a Self-Consistent Solution

For any arbitrary device, there is no guarantee that the process of iteration between solutions for electron density and electrostatic potential will converge. Suppose, as an illustration, an initial guess of the electrostatic potential is taken as zero, throughout a device. In the lightly

doped regions, a net negative charge would appear, after solution for the electron density. Solution of the Poisson equation (equation 2.31) would dictate that the electrostatic potential would have a positive curvature in this region; the corresponding conduction-band profile would have a negative curvature. Hence, the conduction-band profile would float upward in energy, in a region of negative charge. Upon solving again for the electron density, the negative charge previously observed in lightly doped regions would be reduced. If the changes in electrostatic potential and electron density from iteration to iteration are small, the process will converge. To insure that changes are small between iterations, a method of under-relaxed iteration is often used. In this method, only a small fraction of the corrected electrostatic potential is added to the previous solution:

$$\Phi^{(i+1)}(z) = \Phi^{(i)}(z) + \alpha \, \Phi'(z)$$

Here, $\Phi^{(i)}(z)$ is the electrostatic potential, used in an iteration $i$ to compute the electron density; the corrected electrostatic potential $\Phi'(z)$ is the actual solution to the Poisson equation; the relaxation constant $\alpha$ is a fraction chosen to force convergence. Although convergence can be achieved with this method, the process requires a sizable number of iterations. Moreover, a correct choice of $\alpha$ involves considerable guesswork.

An alternate method, used for the calculations in Chapter 4, involves the solution of a non-linear Poisson equation. The computed electron density can be modeled in the Poisson equation with a Boltzmann factor:

$$\frac{d}{dz}\left[\epsilon(z)\frac{d}{dz}\Phi(z)\right] = -q\left[N_D^+ - N_A^- - n_0 e^{q(\Phi(z)-F_n(z))/k_BT}\right] \qquad (3.14)$$

where the parameter $F_n$ acts like a quasi-Fermi level for electrons:

$$F_n(z) = \Phi^{(i)}(z) - \frac{k_BT}{q}\ln\left(\frac{n^{(i)}(z)}{n_0}\right) \qquad (3.15)$$

Given an electrostatic potential $\Phi^{(i)}(z)$ and the electron density $n^{(i)}(z)$ computed from it, the parameter $F_n$ can be calculated at each position-space node, as shown above. A new solution for electrostatic potential $\Phi^{(i+1)}(z)$ is then obtained from equation 3.14. Since the electron density is modeled as a function of the potential, no relaxation criterion is necessary; interactions

between the electron density and the electrostatic potential are approximately accounted for. Although a solution of equation 3.14 is more complicated [36], convergence can be obtained in relatively few iterations.

# CHAPTER 4
## EXAMPLE CALCULATIONS:
## RESONANT TUNNELING DEVICES

The machinery developed in the preceding chapters can be applied to any one-dimensional semiconductor device. Resonant tunneling devices, exhibiting transport properties of a quantum mechanical origin, provide particularly interesting results. Of the many devices studied experimentally [4-11], we have chosen to examine the resonant tunneling diode reported by Ray *et. al.* [10], since its structure emphasizes the importance of a self-consistent calculation. The device, pictured in figure 4.1, was fabricated by metalorganic chemical vapor deposition (MOCVD), with two $Al_{0.45}Ga_{0.55}As$ barriers surrounding a GaAs potential well. Contacts are heavily doped $(2 \times 10^{18} cm^{-3})$ GaAs, separated from the resonant tunneling region by undoped GaAs spacer layers. The inclusion of spacer layers has several advantages. First, the layers tend to reduce the migration of impurities from the contacts to the resonant tunneling region, thereby reducing impurity scattering. Second, a greater degree of symmetry in the conduction-band profile is maintained, since an applied bias is dropped across a longer, undoped region. As Ricco and Azbel pointed out [13], asymmetry in the conduction-band profile degrades the peak in resonant transmission, reducing the resonant tunneling effect. Finally, the presence of spacer layers pronounces the upward shift of the conduction-band profile in the undoped region, reducing the component of thermionic emission current. This upward shift also has a profound influence on the current-voltage relationship, which is investigated below.

To clearly illustrate the effects of space-charge in the analysis, results are contrasted for solutions with and without self-consistency. In the latter approach, hereafter referred to as a "flatband" analysis, space-charge effects are completely neglected, and the electrostatic potential is taken as zero in equilibrium; the application of bias appears in the conduction-band profile as linear voltage drops across lightly doped regions. A self-consistent analysis,
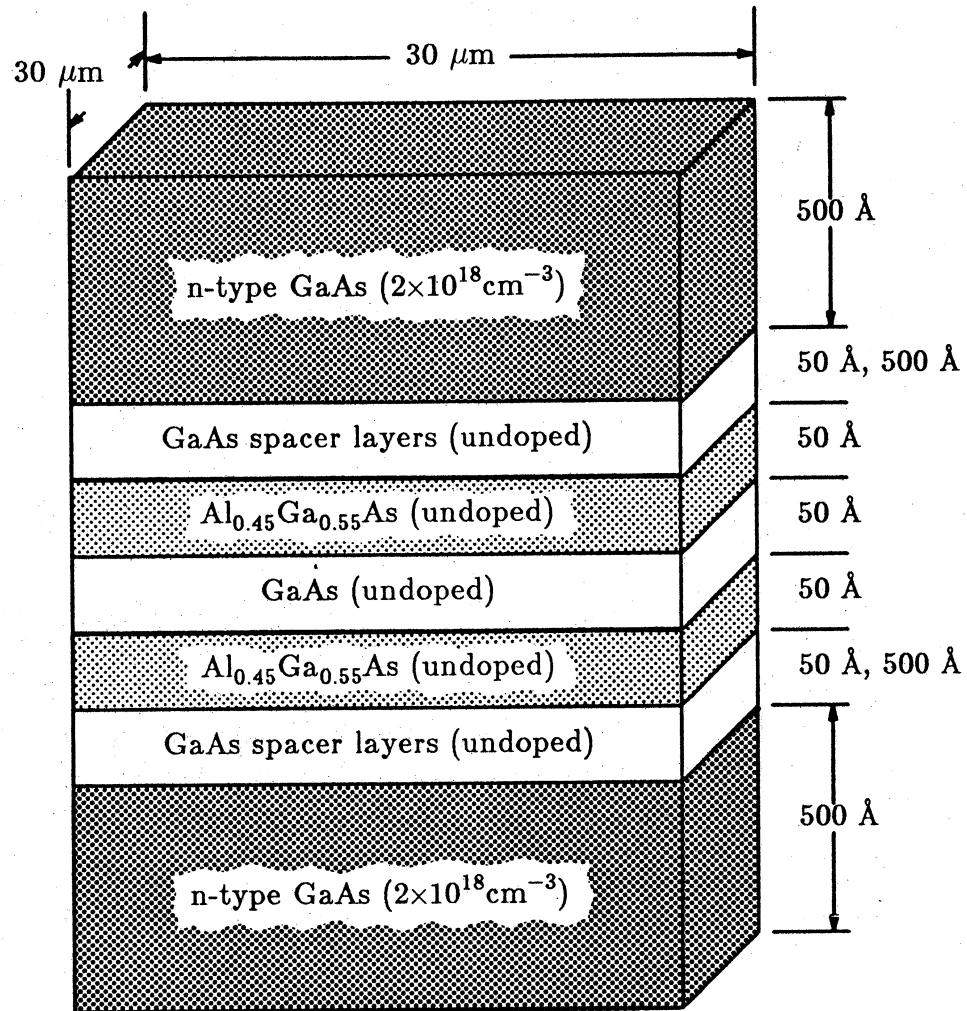
Figure 4.1    Resonant tunneling device fabricated by Ray *et. al.* [10]

however, allows for iteration beyond the flatband solution, accounting for the effects of space-charge. In the following sections, two structures of the type shown in figure 4.1 are simulated at room temperature (300° K)--one with 50 Å spacer layers, and the other with 500 Å spacer layers.

## 4.1 Equilibrium Electrostatics

Figure 4.2 presents the equilibrium conduction-band profiles, for both self-consistent and flatband analyses, for the device of figure 4.1 with 50 Å spacer layers. The consideration of space-charge leads to an upward shift in the undoped region of the conduction-band profile. Band-bending is expected, since the presence of electrons creates a negative charge in the undoped region, causing the conduction-band edge to float upward in energy. A shift of the conduction-band edge, in turn, leads to an upward shift in all quasi-bound state energies. For simplicity, only the first quasi-bound state of each analysis is shown in figure 4.2. As explained in Chapter 1, electrons incident at a resonant energy are multiply reflected between the two potential barriers, creating a significant electron density within the quantum well. This density, pictured in figure 4.3 for both flatband and self-consistent analyses, is smaller for the self-consistent calculation. Since space-charge effects shift all of the quasi-bound states to higher energies, fewer electrons are available to populate the well. Of course, electron density drops rapidly within the potential barriers, as expected classically for a wide bandgap material.

As the length of the spacer layers is increased, the effects of space-charge become more pronounced; from a classical viewpoint, in equilibrium, the Fermi level of intrinsic GaAs must align with that of the heavily doped contacts. Consequently, a larger upward shift in the conduction-band edge can be observed for the device of figure 4.1, with 500 Å spacer layers. Equilibrium conduction-band profiles and electron densities are presented in figures 4.4 and 4.5, respectively. Again, a large electron density is formed within the quantum well, although the consideration of space-charge greatly reduces the peak density. Since band-bending in this case extends well beyond the double barrier region, electron density is decreased within the spacer layers. It is apparent that space-charge effects play a significant role in the equilibrium solution; an even greater emphasis on the importance of self-consistency is established in the following section.
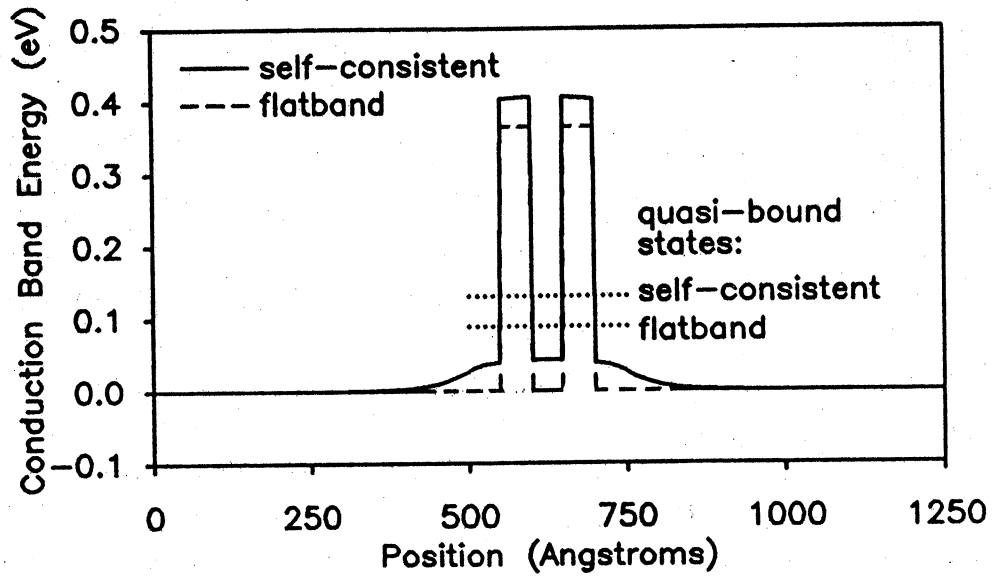
Figure 4.2    Equilibrium conduction-band profiles of the resonant tunneling device shown in figure 4.1, with 50 Å spacer layers
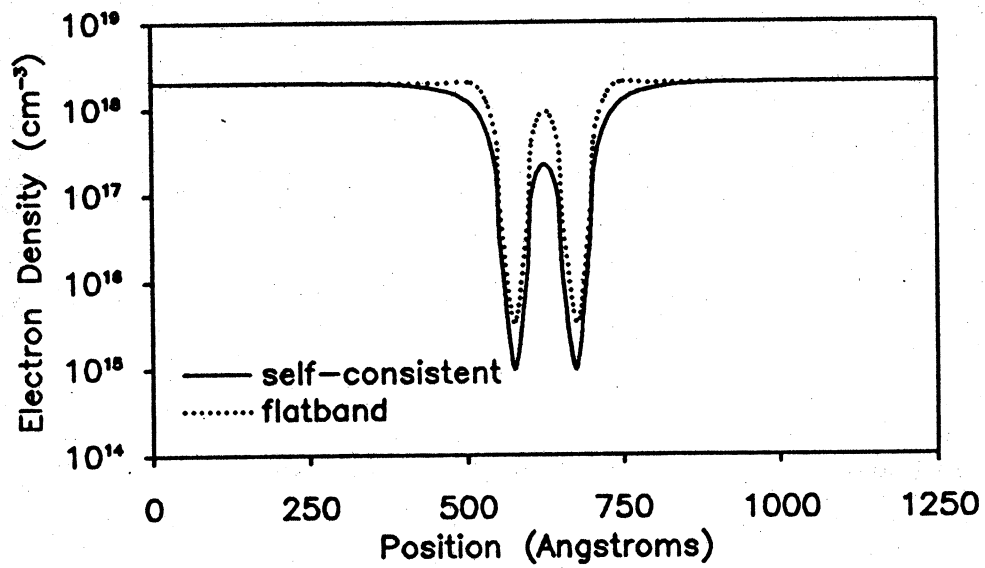


Figure 4.3    Equilibrium electron densities of the resonant tunneling device shown in figure 4.1, with 50 Å spacer layers
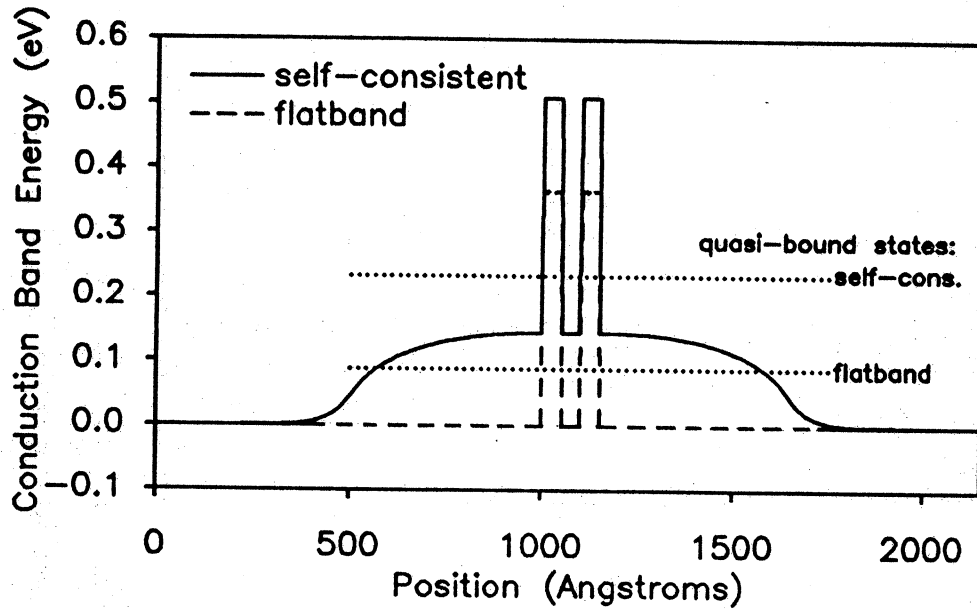
Figure 4.4    Equilibrium conduction-band profiles of the resonant tunneling
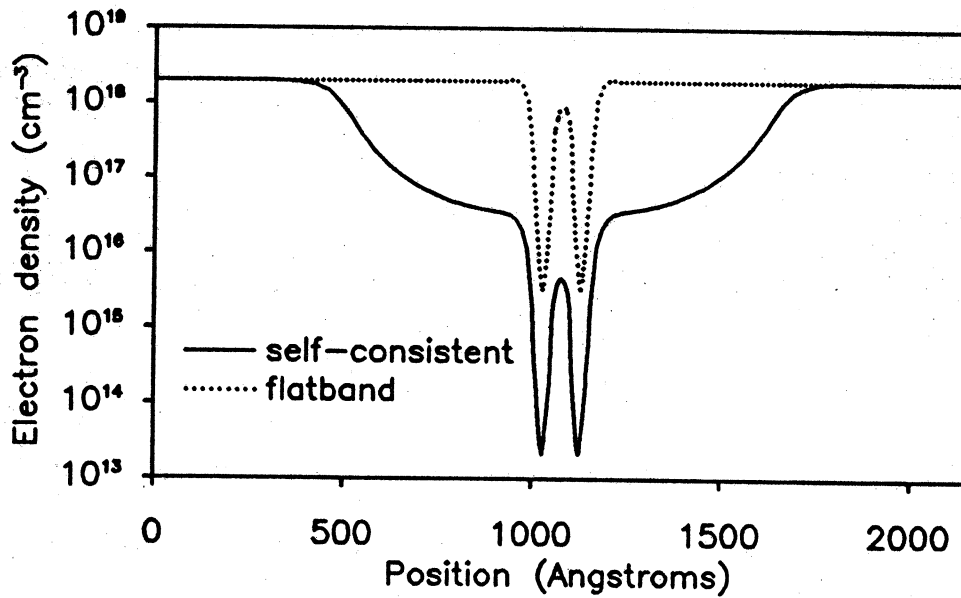device shown in figure 4.1, with 500 Å spacer layers



Figure 4.5    Equilibrium electron densities of the resonant tunneling device
shown in figure 4.1, with 500 Å spacer layers

## 4.2 Current-Voltage Relationships

Under the application of bias, resonant tunneling devices exhibit negative differential resistance (NDR). The origin of this phenomenon was explained in section 1.1.2 as a result of conservation of energy and transverse momentum. As an applied bias is increased, more and more electrons are available for tunneling. Current reaches a maximum, then drops sharply when the quasi-bound state falls below the conduction-band edge of the supplying contact. Therefore, the bias for which NDR is observed depends upon the height of the first quasi-bound state above the conduction-band edge in the supplying contact. Assuming the potential drop across the structure is approximately linear, a bias corresponding to approximately twice this height is required for the onset of NDR:

$$V_{NDR} \simeq \frac{2}{q}(E_1 - E_{C,contact})$$

From the results presented in figure 4.2, NDR should be observed at a bias of $V_{NDR} \simeq 0.262$ V for the self-consistent calculation, and $V_{NDR} \simeq 0.178$ V, for the flatband result. As shown in figure 4.6, a plot of the current-voltage relationships for the device of figure 4.1 with 50 Å spacer layers, these estimates are quite close. This demonstrates one important consideration for self-consistent solutions: Any variation in the quasi-bound state energies appears twice as large in the translation of NDR along the voltage axis.

Returning to figure 4.6, we note that the inclusion of self-consistency has broadened the NDR region. In the self-consistent calculation, current density reaches a maximum when the quasi-bound state level is well above the conduction-band edge in the contact. This is illustrated in figures 4.7 and 4.8, which present plots of the conduction-band profile at biases corresponding to current maxima (points $P$ and $Q$ of figure 4.6). After the maximum current is attained, a larger additional bias is required in the self-consistent case to pull the resonant level below the conduction-band edge in the contact, and reach the point of minimum current. The result is a broadening of the self-consistent NDR region, compared to the flatband solution.

Finally, the peak current of the NDR region is reduced for the self-consistent calculation. For small biases, fewer electrons are available at the higher energies of self-consistent quasi-bound states to be transmitted across the device. Near the onset of NDR, which occurs at a higher bias in the self-consistent case, the transmission coefficient is severely degraded. Current
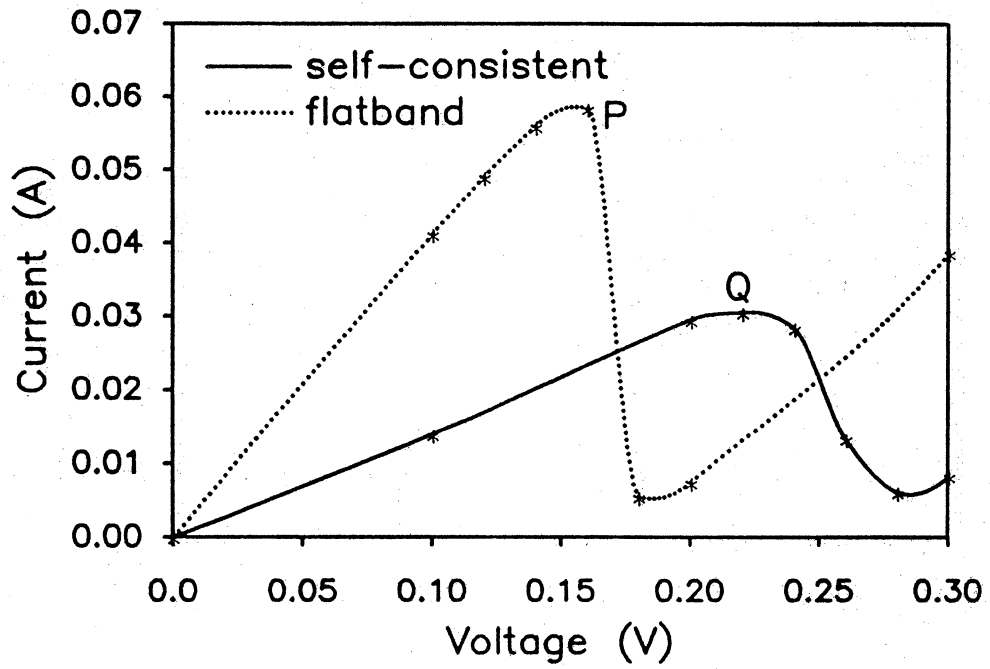
Figure 4.6    Current-voltage relationships for the structure of figure 4.1, with 50 Å spacer layers
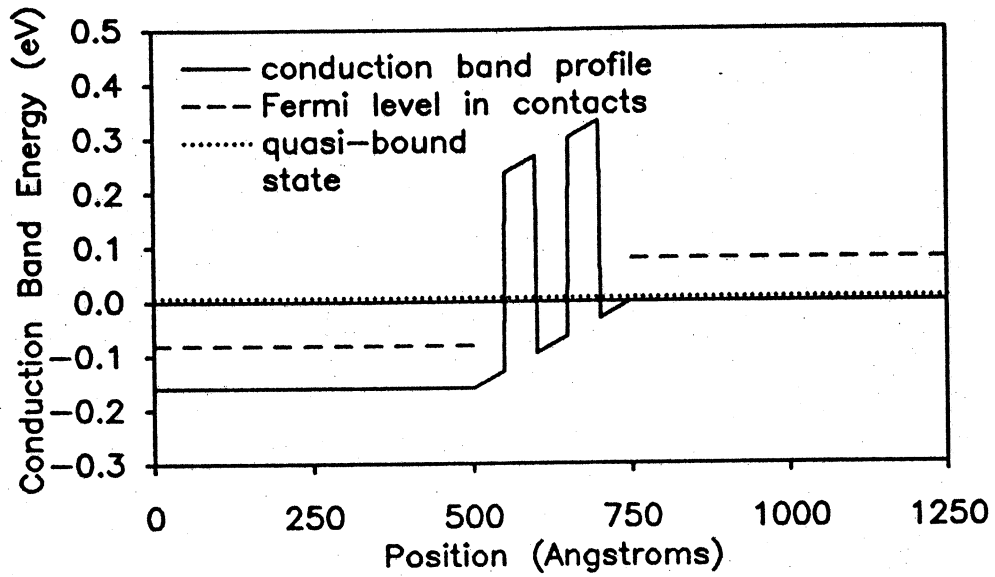
Figure 4.7    Conduction-band profile corresponding to a bias of maximum current for a flatband analysis (point $P$ of figure 4.6)
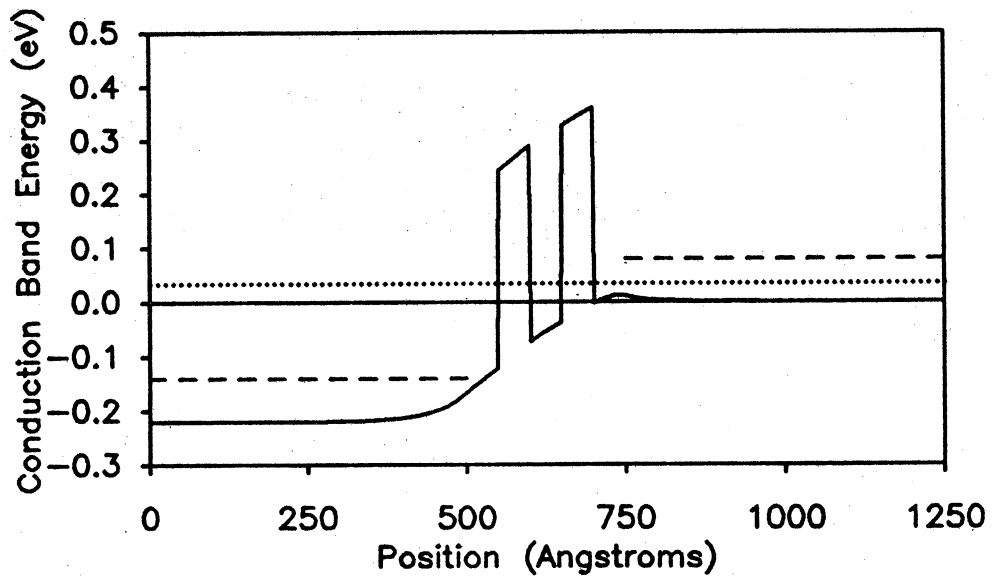


Figure 4.8    Conduction-band profile corresponding to a bias of maximum current for a self-consistent analysis (point $Q$ of figure 4.6)

is maximized when the product of a degrading transmission peak and an increasing flux of incident electrons is maximized. As mentioned above, this occurs when the quasi-bound state energy is well above the conduction-band edge in the contact; therefore, current reaches a smaller maximum value in a self-consistent calculation, relative to the flatband result.

Although the incorporation of space-charge effects brings quantum transport theory one step closer to experiment, meaningful comparison must await a more precise knowledge of device parameters. Experimental unknowns, such as nominal doping densities, interface and bulk charges, and contact resistances, can significantly affect modeling efforts. The preceding discussion, in particular, highlights the sensitivity of the current-voltage relationship with respect to space-charge. Numerical simulations can still be performed, however, demonstrating much of the physics of one-dimensional devices.

# CHAPTER 5
# SUMMARY AND CONCLUSIONS

## 5.1 Summary

A quantum mechanical description of carrier transport in one dimension has been presented. Electrons, injected from contacts in thermodynamic equilibrium, propagate through an arbitrary device in accordance with the Schrödinger equation. Two solution techniques, the methods of cascading transfer matrices and scatter matrices, have been described to calculate the electron wavefunction. Once the wavefunction has been determined, macroscopic quantities of interest, such as electron density and current density, can be calculated by summing contributions from all electrons. To insure accurate integrations of the wavefunction magnitude, a method of isolating extrema in wavevector space was developed. The determination of electron and current densities can be applied to any conduction-band profile. A self-consistent solution, however, requires iteration between calculations for electron density and the electrostatic potential. Although the analysis is computationally demanding, self-consistency is vital for a correct simulation. Example calculations presented in Chapter 4 underscore its importance.

## 5.2 Conclusions

Much of the physics of quantum transport can be demonstrated with the analysis presented; the simulation of resonant tunneling devices provided in Chapter 4 is but one example. Two assumptions made to simplify the analysis, however, limit its application. First, the magnitude of the wavefunction was assumed to be weakly dependent on the transverse energy. The integration over transverse momentum could then be performed analytically. If the effective mass is constant in a device, this assumption is not needed; the wavefunction in a homostructure is independent of the transverse energy, if motion is constrained in one dimension only. For material systems such as GaAs/AlGaAs, the variation of effective mass is

small, and the assumption is reasonable. For devices composed of widely different materials, however, the integration over transverse momentum should be performed numerically, greatly increasing the execution time of a simulation. Second, transport was assumed to be ballistic, so that scattering events are infrequent and unimportant. Although this assumption could be reasonable for short device geometries, or for experiments performed at extremely low temperature, scattering in general is an integral part of carrier transport. In the present analysis, for example, only electronic states at energies above the conduction-band edge of a contact can be populated. The introduction of scattering events would allow carriers to decay into lower states. Moreover, the performance of many devices is deeply affected by scattering processes. Negative differential resistance in resonant tunneling devices is degraded when multiply reflecting electrons lose coherence; effective mass filtering relies on the broadening of carrier energy due to collisions, to effectively localize holes. Although a description of scattering is desirable, it is currently a topic of much research, and a proper treatment must be deferred to future analyses. Despite these two limiting assumptions, the transport theory developed in this thesis represents a first step in the understanding of quantum phenomena.

# LIST OF REFERENCES

## LIST OF REFERENCES

[1] Esaki, L., and Tsu, R., "Superlattice and Negative Conductivity in Semiconductors," IBM Research Note, RC-2418, 1969.

[2] Esaki, L., and Tsu, R., "Superlattice and Negative Differential Conductivity in Semiconductors," IBM Journal of Research and Development, pp. 61-65, 1970.

[3] Esaki, L., Chang, L. L., Howard, W. E., and Rideout, V. L., "Transport Properties of a GaAs-GaAlAs Superlattice," Proceedings of the 11th International Conference on Physics in Semiconductors, Warsaw, Poland, pp. 431-436, 1972.

[4] Chang, L. L., Esaki, L., and Tsu, R., "Resonant Tunneling in Semiconductor Double Barriers," Applied Physics Letters, 24(12), pp. 593-595, 1974.

[5] Sollner, T. C. L. G., Goodhue, W. D., Tannenwald, P. E., Parker, C. D., and Peck, D. D., "Resonant Tunneling Through Quantum Wells at Frequencies Up to 2.5 THz," Applied Physics Letters, 43(6), pp. 588-590, 1983.

[6] Sollner, T. C. L. G., Tannenwald, P. E., Peck, D. D., and Goodhue, W. D., "Quantum Well Oscillators," Applied Physics Letters, 45(12), pp. 1319-1321, 1984.

[7] Shewchuk, T. J., Chapin, P. C., Coleman, P. D., Kopp, W., Fischer, R., and Morkoc, H., "Resonant Tunneling Oscillations in a $GaAs-Al_xGa_{1-x}As$ Heterostructure at Room Temperature," Applied Physics Letters, 46(5), pp. 508-510, 1985.

[8] Shewchuk, T. J., Gering, J. M., Chapin, P. C., Coleman, P. D., Kopp, W., Peng, C. K., and Morkoc, H., "Stable and Unstable Current-Voltage Measurements of a Resonant Tunneling Heterostructure Oscillator," Applied Physics Letters, 47(9), pp. 986-988, 1985.

[9] Tsuchiya, Masahiro, Sakaki, Hiroyuki, and Yoshino, Junji, "Room Temperature Observation of Differential Negative Resistance in an AlAs/GaAs/AlAs Resonant Tunneling Diode," Japanese Journal of Applied Physics, 24(6), pp. L466-L468, 1985.

[10] Ray, S., Ruden, P., Sokolov, V., Kolbas, R., Boonstra, T., and Williams, J., "Resonant Tunneling Transport at 300 K in GaAs-AlGaAs Quantum Wells Grown by Metalorganic Chemical Vapor Deposition," Applied Physics Letters, 48(24), pp. 1666-1668, 1986.

[11] Lee, G. S., Hsieh, K. Y., and Kolbas, R. M., "Room-Temperature Negative Differential Resistance in Strained-Layer GaAs-AlGaAs-InGaAs Quantum Well Heterostructures," Applied Physics Letters, 49(22), pp. 1528-1530, 1986.

[12] Esaki, L., "A Bird's-Eye View on the Evolution of Semiconductor Superlattices and Quantum Wells," IEEE Journal of Quantum Electronics, QE-22(9), pp. 1611-1624, 1986.

[13] Ricco, B., and Azbel, M. Ya., "Physics of Resonant Tunneling. The One-Dimensional Double-Barrier Case," Physical Review B, 29(4), pp. 1970-1981, 1984.

[14] Vezzetti, David J., and Cahay, Marc M., "Transmission Resonances in Finite, Repeated Structures," Journal of Physics D: Applied Physics, 19, pp. L53-L55, 1986.

[15] Luryi, Serge, "Frequency Limit of Double-Barrier Resonant-Tunneling Oscillators," Applied Physics Letters, 47(5), pp. 490-492, 1985.

[16] Morkoc, H., Chen, J., Reddy, U. K., Henderson, T., and Luryi, S., "Observation of a Negative Differential Resistance Due to Tunneling Through a Single Barrier into a Quantum Well," Applied Physics Letters, 49(2), pp. 70-72, 1986.

[17] Capasso, F., Mohammed, K., and Cho, A. Y., "Resonant Tunneling Through Double Barriers, Perpendicular Quantum Transport Phenomena in Superlattices, and Their Device Applications," IEEE Journal of Quantum Electronics, QE-22(9), pp. 1853-1869, 1986.

[18] Stone, A. Douglas, and Lee, P. A., "Effect of Inelastic Processes on Resonant Tunneling in One Dimension," Physical Review Letters, 54(11), pp. 1196-1199, 1985.

[19] Capasso, F., Mohammed, K., and Cho, A. Y., "Quantum Photoconductive Gain by Effective Mass Filtering and Negative Conductance in Superlattice PN Junctions," Physica, 134B, pp. 487-493, 1985.

[20] Capasso, F., and Kiehl, R. A., "Resonant Tunneling Transistor with Quantum Well Base and High-Energy Injection: A New Negative Differential Resistance Device," Journal of Applied Physics, 58(3), pp. 1366-1368, 1985.

[21] Bonnefoi, A. R., Chow, D. H., and McGill, T. C., "Inverted Base-Collector Tunnel Transistors," Applied Physics Letters, 47(8), pp. 888-890, 1985.

[22] Nakata, Y., Asada, M., and Suematsu, Y., "Analysis of Novel Resonant Electron Transfer Triode Device Using Metal-Insulator Superlattice for High Speed Response," IEEE Journal of Quantum Electronics, QE-22(9), pp. 1880-1886, 1986.

[23] Luryi, S., and Capasso, F., "Resonant Tunneling of Two-Dimensional Electrons Through a Quantum Wire: A Negative Transconductance Device," Applied Physics Letters, 47(12), pp. 1347-1349, 1985.

[24] Bonnefoi, A. R., McGill, T. C., and Burnham, R. D., "Resonant Tunneling Transistors with Controllable Negative Differential Resistances," IEEE Electron Device Letters, EDL-6(12), pp. 636-638, 1985.

[25] Bastard, G., and Brum, J. A., "Electronic States in Semiconductor Heterostructures," IEEE Journal of Quantum Electronics, QE-22(9), pp. 1625-1644, 1986.

[26] Frensley, William R., "Simulation of Resonant-Tunneling Heterostructure Devices," Journal of Vacuum Science Technology B, 3(4), pp. 1261-1266, 1985.

[27] Kluksdahl, N., Potz, W., Ravaioli, U., and Ferry, D. K., "Wigner Function Study of a Double Quantum Barrier Resonant Tunneling Diode," preprint.

[28] Tsu, R., and Esaki, L., "Tunneling in a Finite Superlattice," Applied Physics Letters, 22(11), pp. 562-564, 1973.

[29] Vassell, M. O., Lee, Johnson, and Lockwood, H. F., "Multibarrier Tunneling in $Ga_{1-x}Al_xAs/GaAs$ Heterostructures," Journal of Applied Physics, 54(9), pp. 5206-5213, 1983.

[30] Lui, Wayne W., and Fukuma, Masao, "Exact Solution of the Schrodinger Equation Across an Arbitrary One-Dimensional Piecewise-Linear Potential Barrier," Journal of Applied Physics, 60(5), pp. 1555-1559, 1986.

[31] Jogai, B., and Wang, K. L., "Dependence of Tunneling Current on Structural Variations of Superlattice Devices," Applied Physics Letters, 46(2), pp. 167-168, 1985.

[32] Jogai, B., and Wang, K. L., "Free-Electron Density and Transit Time in a Finite Superlattice," Journal of Applied Physics, 59(8), pp. 2968-2970, 1986.

[33] Ohnishi, H., Inata, T., Muto, S., Yokoyama, N., and Shibatomi, A., "Self-Consistent Analysis of Resonant Tunneling Current," Applied Physics Letters, 49(19), pp. 1248-1250, 1986.

[34] Cahay, M., McLennan, M., Datta, S., and Lundstrom, M. S., "Importance of Space-Charge Effects in Resonant Tunneling Devices," Applied Physics Letters, 50(10), pp. 612-614, 1987.

[35] Hornbeck, Robert W., Numerical Methods, pp. 154-159, Quantum Publishers, Inc., New York, New York, 1975.

[36] Lundstrom, M. S., Datta, S., Schuelke, R. J., Bandyopadhyey, S., and Sorlie, P. H., "Physics and Modeling of Heterostructure Semiconductor Devices," Technical Report: TR-EE 84-35, Purdue University, 1984.