Quantum Transport

William R. Frensley

October 13, 1998

[Ch. 9 of *Heterostructures and Quantum Devices*, W. R. Frensley and N. G. Einspruch editors, A volume of *VLSI Electronics: Microstructure Science*. (Academic Press, San Diego) Publication date: March 25, 1994]

Contents

I Introduction

The progress of heterostructure technology has permitted the fabrication and study of devices which inherently employ quantum-mechanical effects in their operation [1, 3, 2, 4]. Their existence and technological

potential have created the need for effective theoretical tools to help understand, describe, predict and optimize the performance of quantum devices. The fundamental property of heterostructures, that the energy-band structure (or electron dispersion relation) varies with position, requires some modifications to the standard textbook treatments of such things as scattering amplitudes. The present chapter describes the theoretical tools appropriate for quantum-scale heterostructure devices, and attempts to place the significant approaches in an appropriate context.

1 Open Systems

A general feature of electron devices is that they are of use only when connected to a circuit, and to be so connected any device must possess at least two terminals, contacts, or leads. As a consequence, every device is a open system with respect to electron flow [5]. This is the overriding fact that determines which theoretical models and techniques may be appropriately applied to the study of quantum devices. For example, the quantum mechanics of pure, normalizable states, such as those employed in atomic physics, does not contribute significantly to an understanding of devices, because such states describe closed systems. To understand devices, one must consider the unnormalizable scattering states, and/or describe the state of the device in terms of statistically mixed states, which casts the problem in terms of quantum kinetic theory.

As a practical matter, a device is of use only when its state is driven far from thermodynamic equilibrium by the action of the external circuit. The nonequilibrium state is characterized by the conduction of significant current through the device and/or the appearance of a nonnegligible voltage drop across the device.

In classical transport theory, the openness of the device is addressed by the definition of appropriate boundary conditions for the differential (or integro-differential) transport equations. Such boundary conditions are formulated so as to approximate the behavior of the physical contacts to the device, typically Ohmic or Schottky contacts [6]. In the traditional treatments of quantum transport theories, the role of boundary conditions is often taken for granted, as the models are constructed upon an unbounded spatial domain. The proper formulation and interpretation of the boundary conditions remains an issue, however, and will be examined in the present work.

It should be understood that, unless otherwise specified, all models to be considered here are based upon a single-band, effective-mass Schrödinger equation. To perform calculations using any of the theories to be discussed and obtain results which may be compared to experimental data, one must numerically evaluate a number of quantities involving fundamental constants. When working with quantum-scale devices, it is convenient to work in a set of metric units chosen to approximate the scale of the phenomena under consideration. Such a set of units is specified in Table 1, and the frequently-used physical constants are given in Table 1 in terms of these units.

II Tunneling Theory

The simplest model of quantum transport in devices is to describe the problem in terms of the scattering of the electron wavefunction by a spatially varying potential. One assumes that this potential is situated between two electron reservoirs, each of which emits particles with an equilibrium distribution into the scattering region. The reservoirs will, in general, have different chemical potentials, their difference representing an

Table 1: Convenient Units for Quantum Transport Calculations

Quantity	Unit	Value
Energy	electron Volt	1.602189×10^{-19} J
Mass	free electron mass (m_0)	9.10953×10^{-31} kg
Length	nanometer	10^{-9} m
Time	femtosecond	10^{-15} s
Particle density	nm^{-3}	10^{21} cm^{-3}
Current density	q nm^{-2} fs^{-1}	1.602189×10^{10} A cm^{-2}

Table 2: Values of Frequently Used Constants

Quantity	Expression	Value
Reduced Planck constant	$\hbar$	0.658217 eV fs
Kinetic energy factor	$\hbar^2/2m_0$	0.0381001 eV nm^2
Poisson factor	q/ϵ_0	18.095 V nm
Boltzmann constant	k_B	8.61733×10^{-5} eV/K

applied bias voltage. The net flux of electrons passing between the reservoirs constitutes the electrical current conducted by the device. A single-particle Schrödinger equation can only describe a situation in which the electrons move perfectly coherently throughout the device. Any loss of coherence due to inelastic collisions requires a higher-level description. Nevertheless, the solutions of Schrödinger's equation remain one of the fundamental tools available to understand and predict the behavior of quantum-scale devices.

1 General Formulation

Let us consider a one-dimensional, single-band model. In a semiconductor heterostructure, the electron wavefunctions are described most simply by the effective-mass Schrödinger equation:

$$E\psi = -\frac{\hbar^2}{2}\frac{\partial}{\partial x}\frac{1}{m^*(x)}\frac{\partial}{\partial x}\psi + V(x)\psi. \tag{1}$$

The form of the kinetic-energy operator in (1) is the simplest Hermitian form one can use if the materials parameters (such as the effective mass m^*) vary with position [7, 8]. Now, let us assume that all variation of the potential and of the materials parameters are confined to an interval $x_l < x < x_r$, so that outside of this interval the form of Schrödinger's equation is translationally invariant. Thus, in the regions $x < x_l$ and $x > x_r$ (the "asymptotic regions"), the solutions of Schrödinger's equation are superpositions of plane waves, and the energies of these plane waves are described by a well-defined dispersion relation $E(k)$. We will refer to quantities in the asymptotic regions by the subscripts l and r for the left- and right-hand regions, respectively.

Now, for any energy $E > V_l$ and $E > V_r$, there will be two independent solutions to Schrödinger's equation representing electrons incident from the left and the right, respectively. In the asymptotic regions,

these solutions will have the form:

$$\psi_l(x) = \begin{cases} e^{ik_l(x-x_l)} + r_l e^{-ik_l(x-x_l)} & x < x_l, \\ t_r e^{ik_r(x-x_r)} & x > x_r; \end{cases} \tag{2a}$$

$$\psi_r(x) = \begin{cases} t_l e^{-ik_l(x-x_l)} & x < x_l, \\ e^{-ik_r(x-x_r)} + r_r e^{ik_r(x-x_r)} & x > x_r. \end{cases} \tag{2b}$$

In general, $k_l \neq k_r$ because $V_l \neq V_r$. There exist several rigorous relationships between the transmission and reflection amplitudes $t_{l,r}$ and $r_{l,r}$. Invoking Green's identity leads to the current-continuity equations

$$v_l(1 - |r_l|^2) = v_r|t_r|^2, \tag{3a}$$

$$v_r(1 - |r_r|^2) = v_l|t_l|^2, \tag{3b}$$

and an orthogonality condition

$$v_l t_l^* r_l + v_r t_r r_r^* = 0. \tag{3c}$$

One may also invoke time-reversal symmetry to find the relationship between t_l and t_r. Noting that ψ_l^*, say, is a solution of Schrödinger's equation with energy E, it must be possible to write ψ_l^* as a linear combination of ψ_l and ψ_r. With a bit of manipulation, one finds

$$v_l t_l = v_r t_r. \tag{4}$$

In most textbooks, the relations (3–4) are presented with the wavenumbers $k_{l,r}$ in place of the velocities $v_{l,r}$. Such expressions are derived within the assumptions that the dispersion relation (or band structure) $E(k)$ is perfectly parabolic and does not depend upon position. Neither assumption is warranted in semiconductor heterostructures, as electrons in heterostructure devices frequently explore non-parabolic regions of the energy-band structure, and the band structure itself (particularly the effective mass) will vary with semiconductor composition and thus position. The expressions (3–4) are valid for nonparabolic and spatially varying dispersion relations and should thus always be used. The velocity is the electron group velocity, given by

$$v = (1/\hbar) dE/dk, \tag{5}$$

using the dispersion relation appropriate to the given semiconductor material.

One conventionally defines the transmission probability T as the ratio of the transmitted to the incident flux, or

$$T = \frac{v_r}{v_l}|t_r|^2 = \frac{v_l}{v_r}|t_l|^2, \tag{6}$$

so that T is same for both directions of incidence. One can also show, using equations (3–4),

$$|r_l|^2 = |r_r|^2 = R. \tag{7}$$

(Because the reflection probability is measured on the same side of the system as the incident flux, there is no velocity correction.) Also note that, from (3),

$$T + R = 1, \tag{8}$$

as one would expect.

Finally, we investigate the normalization and orthogonality properties of the scattering states. Because we are dealing with a continuum of states over which we must integrate to evaluate any physical observable, a "delta-function" normalization is appropriate. With such a normalization convention, any finite contribution to the inner product, such as the integral over (x_l, x_r), may be neglected in comparison to the integrals over $(-\infty, x_l]$ and $[x_r, \infty)$. Thus,

$$\begin{aligned}
\langle \psi'_l | \psi_l \rangle &= \int_{-\infty}^{0} e^{i(k_l - k'_l)x}\, dx + |r_l|^2 \int_{-\infty}^{0} e^{i(k'_l - k_l)x}\, dx + |t_r|^2 \int_{0}^{\infty} e^{i(k_r - k'_r)x} \Lambda\, dx \\
&= \pi[(1 + |r_l|^2)\delta(k_l - k'_l) + |t_r|^2 \delta(k_r - k'_r)], \qquad (9a) \\
\langle \psi'_r | \psi_l \rangle &= t_l^* r_r \int_{-\infty}^{0} e^{i(k'_l - k_l)x}\, dx + t_r r_r^* \int_{0}^{\infty} e^{i(k_r - k'_r)x}\, dx \\
&= \pi[t_l^* r_l \delta(k_l - k'_l) + t_r r_r^* \delta(k_r - k'_r)]. \qquad (9b)
\end{aligned}$$

A relationship similar to (9a) can be written for $\langle \psi'_r | \psi_r \rangle$. The δ-functions can be rewritten in terms of the energy E using

$$\delta(k_l - k'_l) = \frac{dE}{dk_l} \delta(E - E') = \hbar v_l \delta(E - E'), \qquad (10)$$

and similarly for $\delta(k_r - k'_r)$. We then obtain

$$\begin{aligned}
\langle \psi'_l | \psi_l \rangle &= \pi\hbar\delta(E - E')[v_l(1 + |r_l|^2) + v_r |t_r|^2] \\
&= 2\pi\hbar v_l \delta(E - E'), \qquad (11a) \\
\langle \psi'_r | \psi_r \rangle &= 2\pi\hbar v_r \delta(E - E'), \qquad (11b) \\
\langle \psi'_r | \psi_l \rangle &= \pi\hbar\delta(E - E')(t_l^* r_l v_l + t_r r_r^* v_r) = 0, \qquad (11c)
\end{aligned}$$

from equations (3).

2 Evaluation of the Current Density

To evaluate any physical observables, such as the current density, we must specify how the scattering solutions are statistically weighted in the final result. For the case of a continuous spectrum of states, with δ-function normalization, the derivation of the correct expressions are rather tricky, because we seek expressions for densities of charge, current, energy, etc., rather than total quantities (which are of course infinite in an unbounded system). To illustrate the procedure, let us follow through the derivation of the electron density in a spatially uniform three-dimensional semiconductor in equilibrium. We approximate the conduction band structure by a simple parabolic dispersion relation:

$$E(\mathbf{k}) = E_C + \hbar^2 \mathbf{k} \cdot \mathbf{k} / 2m^*, \qquad (12)$$

where $\mathbf{k}$ is the wavevector. The probability that each state $|\mathbf{k}\rangle$ will be occupied by an electron is given by the Fermi-Dirac distribution function:

$$f_{\mathrm{FD}}[E(k) - E_F] = \left\{ 1 + e^{\beta[E(\mathbf{k}) - E_F]} \right\}^{-1}, \qquad (13)$$

where E_F is the Fermi level or chemical potential and $\beta = 1/k_{\mathrm{B}}T$, T being the absolute temperature. (To avoid confusion with the transmission probability which is also denoted by T, the absolute temperature will always be shown multiplied by Boltzmann's constant k_{B}.) Let us now make an *ad hoc* assumption that the

semiconductor crystal is a cube with each side of length L, and apply periodic boundary conditions. Then the stationary quantum states are plane waves (normalized to unit amplitude) of the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}.$$

Due to the periodic boundary conditions, $\mathbf{k}$ must assume discrete values:

$$\mathbf{k} = (2\pi/L)(n_x\mathbf{e}_x + n_y\mathbf{e}_y + n_z\mathbf{e}_z), \tag{14}$$

where n_x, n_y, and n_z are integers. The total number of electrons in the crystal N is just the sum over all of the states $|\mathbf{k}\rangle$ of the probability that each state is occupied

$$N = 2\sum_{\mathbf{k}} f_{\mathrm{FD}}[E(\mathbf{k})], \tag{15}$$

where the factor of 2 comes from the two spin states. Now, because L is large, the allowed values of $\mathbf{k}$ are very closely spaced, and the sum over $\mathbf{k}$ can be well approximated by an integral:

$$\sum_{\mathbf{k}} \rightarrow \frac{L^3}{(2\pi)^3}\int d^3\mathbf{k}. \tag{16}$$

We can now write an expression for the density of electrons n:

$$n = \frac{N}{L^3} = 2\int \frac{d^3\mathbf{k}}{(2\pi)^3} f_{\mathrm{FD}}[E(\mathbf{k})]. \tag{17}$$

Notice that the arbitrary crystal dimension L has dropped out of the final expression.

In order to evaluate densities using expressions such as (17) it is usually more convenient to transform the integration variable to E. By expressing $d^3\mathbf{k}$ in spherical coordinates and manipulating the dispersion relation (12) one finds [9]:

$$\frac{d^3\mathbf{k}}{(2\pi)^3} = \frac{m^*\sqrt{2m^*(E-E_C)}}{2\pi^2\hbar^3}dE. \tag{18a}$$

We will also have occasion to use the corresponding expressions for integrals over two- and one-dimensional $\mathbf{k}$ vectors. For the two-dimensional case (still assuming a parabolic dispersion relation):

$$\frac{d^2\mathbf{k}}{(2\pi)^2} = \frac{m^*}{2\pi\hbar^2}\,dE. \tag{18b}$$

For integration over a one-dimensional k, the definition of the group velocity (5) may be used to obtain an expression valid for any dispersion relation:

$$\frac{dk}{2\pi} = \frac{dE}{2\pi\hbar v(E)}. \tag{18c}$$

Inserting (18a) into (17) leads to the usual expression for the electron density in a semiconductor (as discussed in Chapter 1 of the present volume) $n = N_C\mathcal{F}_{1/2}[\beta(E_F - E_C)]$.

The procedure for evaluating a physical observable in an equilibrium system of infinite extent may thus be generalized from the above discussion. The expectation value of the observable quantity is calculated for each state, taking the scattering states to be normalized to unit amplitude. The density of the observable is then determined by inserting this expectation value into the sum in (17) and evaluating the resulting integral, usually using the relations (18). The two most important observables are the electron density $n(x)$ and the

current density j (which is independent of position in one dimension and steady-state). The expectation value of the density for a state ψ is simply

$$n_\psi(x) = \psi^*(x)\psi(x). \tag{19}$$

The expectation value of j is simple, though the operator itself often is not. If the dispersion relation $E(k)$ is not parabolic and independent of position, the form of the operator j is *not* given by the simple textbook expression $J = (q\hbar/m^*)(\partial/\partial x)$. The current density operator is instead whatever remains of the kinetic energy term of the Hamiltonian after the application of Green's identity as in the derivation of (3), and this obviously depends upon the form of the Hamiltonian itself. For unit-incident-amplitude scattering states, however, the result is invariably

$$\langle\psi_l|J|\psi_l\rangle = qv_r|t_r|^2 = qv_lT, \tag{20a}$$

$$\langle\psi_r|J|\psi_r\rangle = -qv_l|t_l|^2 = -qv_rT. \tag{20b}$$

Of course, in equilibrium, these two currents cancel each other (by the principle of detailed balance) and there is no net current flow.

To investigate the transport properties of a quantum system one must generally evaluate the current flow through the system, and this requires that one examine systems that are out of thermal equilibrium. A common situation, in both experimental apparatus and technological systems, is that one has two (or more) physically large regions densely populated with electrons in which the current density is low, coupled by a smaller region through which the current density is much larger. It is convenient to regard the large regions as "electron reservoirs" within which the electrons are all in equilibrium with a constant temperature and Fermi level, and which are so large that the current flow into or out of the smaller "device" represents a negligible perturbation. The reservoirs represent the metallic contacting leads to discrete devices or experimental samples, or the power-supply busses at the system level. Consequently the electrons flowing from from a reservoir into the device occupy that equilibrium distribution which characterizes the reservoir. In a simple one-dimensional system with two reservoirs, the electrons flowing in from the left-hand reservoir have $k > 0$ and those flowing from the right-hand reservoir have $k < 0$.

Within this picture, the current that is injected from the left-hand reservoir is

$$J_l = 2q\sum_{k_\perp}\int_0^\infty \frac{dk_\parallel}{2\pi}\, f_{\rm FD}[E(k_\parallel,k_\perp) - E_{Fl}]v_l(k_\parallel,k_\perp)T(k_\parallel,k_\perp), \tag{21a}$$

and the current injected from the right-hand reservoir is

$$J_r = 2q\sum_{k_\perp}\int_0^{-\infty} \frac{dk_\parallel}{2\pi}\, f_{\rm FD}[E(k_\parallel,k_\perp) - E_{Fr}]v_r(k_\parallel,k_\perp)T(k_\parallel,k_\perp). \tag{21b}$$

In order to simplify the calculation of J further, we must invoke some special properties of the system. The most useful such property is that symmetry which permits the separation of the spatial variables. The separation of variables is possible if the Hamiltonian can be separated into two parts:

$$H = H_\parallel(x, \partial/\partial x) + H_\perp(y, z, \partial/\partial y, \partial/\partial z). \tag{22}$$

(Here the notation $H_\parallel$ and $H_\perp$ is defined with respect to the direction of current transport.) Then the wavefunction separates into a product of two factors:

$$\psi(\mathbf{r}) = \psi_\parallel(x)\psi_\perp(y,z), \tag{23}$$

and the energy can be separated into a product of two terms:

$$E = E_{\parallel}(k_{\parallel}) + E_{\perp}(k_{\perp}). \tag{24}$$

The expression for the total current density J can now be simplified to

$$J = q \int_{V_0}^{\infty} \frac{dE_{\parallel}}{2\pi\hbar} T(E_{\parallel})[F(E_{\parallel} - E_{Fl}) - F(E_{\parallel} - E_{Fr})], \tag{25}$$

where V_0 is the larger of the two asymptotic potentials (minimum energy for a propagating state) and F is the Fermi-Dirac distribution function summed over the transverse states:

$$F(E) = 2 \sum_{k_{\perp}} \frac{1}{1 + e^{\beta(E+E_{\perp})}}. \tag{26}$$

The form of the sum over $k_{\perp}$ depends upon the spatial configuration of the tunneling system. Note that the velocity factor does not appear in (25) because it was canceled by the density of states.

If the system in question is macroscopically large in its transverse dimensions, the $k_{\perp}$ form a two-dimensional continuum, and $H_{\perp} = \hbar^2 k_{\perp}^2 / 2m_{\perp}^*$. Then using the two-dimensional analog of (16) and (18b) F can be analytically evaluated:

$$F_{2d}(E) = \frac{m_{\perp}^*}{\pi\hbar^2\beta} \ln(1 + e^{-\beta E}). \tag{27}$$

The current density can now be written in the form usually given for the tunneling current [10]:

$$J_{3d} = q \int_{V_0}^{\infty} \frac{dE_{\parallel}}{2\pi\hbar} T(E_{\parallel}) \ln \left\{ \frac{1 + \exp[-\beta(E_{\parallel} - E_{Fl})]}{1 + \exp[-\beta(E_{\parallel} - E_{Fr})]} \right\}. \tag{28}$$

Note that this expression is valid in general with respect to the dispersion relation in the x direction, but requires a parabolic dispersion relation in the transverse directions. The separation of variables leading to (28) is never rigorously valid in a semiconductor heterostructure. The reason for this is that the transverse effective mass $m_{\perp}^*$ will vary with semiconductor composition, which varies in the x direction. In principle, one must do at least a two-dimensional integral (if axial symmetry holds, otherwise a three-dimensional integral) as implied by (21). Nevertheless, (28) is widely used to model the current density in heterostructure devices. The error introduced by assuming separation of variables is probably less severe than that due to the assumption of an infinite coherence length.

If the transverse dimensions are constrained, but separation of variables is still possible, the transverse motion of the electrons consists of a discrete set of standing waves or normal modes. Such systems are referred to as "one-dimensional" systems, quantum wires, or electron waveguides. The symbol $k_{\perp}$ is now interpreted as an index for the discrete transverse modes, and the expression for the current density now becomes

$$J_{1d} = I = 2q \sum_{k_{\perp}} \int_{V_0}^{\infty} \frac{dE_{\parallel}}{2\pi\hbar} T(E_{\parallel}, k_{\perp})[f_{\mathrm{FD}}(E_{\parallel} + E_{\perp} - E_{Fl}) - f_{\mathrm{FD}}(E_{\parallel} + E_{\perp} - E_{Fr})]. \tag{29}$$

3 Evaluation of Scattering-State Wavefunctions

To apply the above results to a given device structure, we need a method to determine the transmission and reflection coefficients for any given energy. These are determined by solving, either explicitly or implicitly,

Schrödinger's equation over the domain $x_l \leq x \leq x_r$. Again, we assume that outside this domain (in the asymptotic regions), the wavefunction consists of a superposition of traveling waves, and we write the general solution

$$\psi(x) = \begin{cases} a_l e^{ik_l(x-x_l)} + b_l e^{-ik_l(x-x_l)} & x < x_l, \\ a_r e^{-ik_r(x-x_r)} + b_r e^{ik_r(x-x_r)} & x > x_r. \end{cases} \tag{30}$$

The task is to find the relation connecting a_l, b_l, a_r, and b_r.

The traditional, "textbook," techniques for solving one-dimensional scattering problems are the Wentzel-Kramers-Brillouin (WKB) approximation and the transmission-matrix scheme [11]. In the WKB approximation, the wavefunction is written as $\psi(x) = e^{i\phi(x)}$, and semi-classical expressions are developed for $\phi(x)$. These require that the potential be slowly varying, and thus are not particularly suited to the analysis of abrupt heterostructures. In the transmission matrix approach the domain is divided into a suitable number of intervals over each of which the potential can be taken to be constant, or perhaps linearly varying. Within each such interval, the wavefunction is expanded in terms of the two independent solutions at the chosen energy (oppositely directed traveling waves if the potential is constant). Then the amplitudes of these waves at the two ends of interval i can be related by a propagation matrix P_i:

$$P_i = \begin{bmatrix} e^{ik_i l_i} & 0 \\ 0 & e^{-ik_i l_i} \end{bmatrix}. \tag{31}$$

The appropriate matching conditions at the boundary between intervals i and i+1 must be derived from the form of the Hamiltonian, and are expressed by a matrix B_i which is typically of the form:

$$B_i = \frac{1}{2} \begin{bmatrix} 1+r & 1-r \\ 1-r & 1+r \end{bmatrix}, \tag{32}$$

where $r = v_i/v_{i+1}$, the velocity ratio. One can then relate the coefficients in the left asymptotic region, incorporated into a vector $\Psi_l = [a_l, b_l]^T$, to those in the right asymptotic region, $\Psi_r = [b_r, a_r]^T$, by a matrix M formed from the product of the appropriate propagation and boundary matrices:

$$\Psi_r = M\Psi_l = P_m B_{m-1} \cdots B_2 P_2 B_1 P_1 \Psi_l. \tag{33}$$

In practical calculations, the transmission matrix approach has proven to be less than satisfactory, because it is prone to arithmetic overflow. In regions where the wavefunction is evanescent, the P matrices contain real elements equal to the attenuation of the region and its inverse. The inverse is likely to be a very large number, and if several evanescent regions are cascaded, the numbers in the matrix will rapidly exceed the dynamic range of floating-point variables. This problem is particularly severe when the transmission matrix scheme is applied to multi-band models, because at any given energy many of the bands will be evanescent [12], but it has also been observed in simple single-band calculations [13].

A much more robust and effective scheme for solving Schrödinger's equation involves expanding the equation on a set of localized basis functions to reduce the differential equation to a set of linear algebraic equations, and directly solving those equations. This approach is numerically faster, more stable, and more readily applied to differing structures. One can take the basis functions to be any set of localized functions: atomic s-orbitals, Wannier orbitals [14], simple finite-element shape functions, *etc.*, but, for one-dimensional models, all such methods produce similar sets of linear equations to be solved. For the present purposes, let us consider the derivation simply in terms of a finite-difference approximation to the differential form of the

effective-mass Schrödinger equation (1), in which the effective mass may vary as a function of x. We assume that the wavefunction is known only on a set of discrete points $x_j = j\Delta$, where Δ is the mesh spacing. The discrete values of the wavefunction will be denoted by $\psi_j = \psi(x_j)$. If the effective mass did not vary, we could simply use the three-point approximation for the second derivative,

$$\frac{d^2\psi}{dx^2} \approx \frac{\psi(x-\Delta) - 2\psi(x) + \psi(x+\Delta)}{\Delta^2}. \tag{34}$$

This reduces the Schrödinger equation to a set of equations of the form:

$$H\psi_j = -s_j\psi_{j-1} + d_j\psi_j - s_{j+1}\psi_{j+1} = E\psi_j, \tag{35}$$

which can be written as a matrix equation with the Hamiltonian being a tridiagonal matrix.

When m^* is allowed to vary, the form of the equations is not changed, but this variation must be taken into account in the values of the matrix elements s_j, d_j. These values may be systematically derived from the variational principle for Schrödinger's equation [15]:

$$\delta L = \delta \int dx\, [(\hbar^2/2m^*)(\nabla\psi^*)\cdot(\nabla\psi) + (V-E)\psi^*\psi] = 0. \tag{36}$$

If we assume that the wavefunction varies in a simple fashion (such as linearly) between mesh points x_j, L can be readily evaluated in terms of the ψ_j and ψ_j^*, and the discrete Schrödinger equation derived by setting $\partial L/\partial\psi_j^* = 0$ for all j. The particular expressions obtained for the matrix elements depends upon the detailed picture we assume for the functional dependence of ψ and m^* within the interval between two adjacent meshpoints. If we assume that the wavefunction varies linearly between ψ_j and ψ_{j+1}, and that the effective mass is constant across this interval with value $m^*_{j+1/2}$, one obtains [16]:

$$d_j = \frac{\hbar^2}{2\Delta^2}\left(\frac{1}{m^*_{j-1/2}} + \frac{1}{m^*_{j+1/2}}\right) + V_j, \tag{37a}$$

$$s_j = \frac{\hbar^2}{2\Delta^2 m^*_{j+1/2}}. \tag{37b}$$

If the effective mass is constant across the mesh interval, one is effectively assuming that the heterojunctions occur on a mesh point, and there is a question of what value of the potential to assign to this point. In general, it is better to assume that the heterojunction occurs halfway between two adjacent mesh points. With this model for the spatial dependence of the effective mass, and still assuming linear variation of the wavefunction, the Hamiltonian matrix elements become:

$$d_j = \frac{\hbar^2}{4\Delta^2}\left(\frac{1}{m^*_{j-1}} + \frac{2}{m^*_j} + \frac{1}{m^*_{j+1}}\right) + V_j, \tag{38a}$$

$$s_j = \frac{\hbar^2}{4\Delta^2}\left(\frac{1}{m^*_{j-1}} + \frac{1}{m^*_j}\right). \tag{38b}$$

These expressions can be improved upon by taking into account the matching condition on the wavefunction implied by the form of the continuous Hamiltonian (1), which is that $\psi(x)$ and $(1/m^*)(d\psi/dx)$ are continuous across a heterojunction. Assuming that the wavefunction is piecewise linear so as to satisfy this condition between meshpoints yields the most accurate discretization [17]:

$$d_j = \frac{\hbar^2}{2\Delta^2}\left(\frac{1}{m^*_{j-1} + m^*_j} + \frac{1}{m^*_j + m^*_{j+1}}\right) + V_j, \tag{39a}$$

$$s_j = \frac{\hbar^2}{2\Delta^2}\frac{1}{m^*_{j-1}+m^*_j}. \tag{39b}$$

If one seeks the eigenstates of a closed system, the tridiagonal Hamiltonian may be readily diagonalized by standard numerical techniques. However, we are here concerned with the problem of the scattering states in an open system. This requires that appropriate boundary conditions be formulated and applied to (35). Lent and Kirkner [18] have demonstrated how to do this in the context of a finite-element electron waveguide calculation, and their approach is called the Quantum Transmitting Boundary Method (QTBM). In the continuous case, one derives the QTBM conditions by evaluating ψ and its derivative ψ' at x_l and x_r using (30). One then solves for the incident amplitudes a_l and a_r in terms of ψ and ψ', and imposes the resulting expressions upon Schrödinger's equation as inhomogeneous boundary conditions. Conditions of this type, in which a linear combination of the function and its derivative are specified, are known as Robbins conditions. They are implicit, in the sense that they must be solved along with the differential equation itself. However, as we shall see below, this presents no problem in a discrete, numerical approach.

In the discrete case, it is simpler to express the QTBM conditions as a linear combination of the values of ψ on two adjacent meshpoints. If the points $j = 1$ and $j = n$ are the limits of the domain in which the potential can vary, we may add boundary points at $j = 0$ and j $= n + 1$. The form of the wavefunction will be taken to be:

$$\psi_j = a_1 z_1^{j-1} + b_1 z_1^{1-j} \quad \text{for} \quad j \le 1, \tag{40a}$$

$$\psi_j = a_n z_n^{n-j} + b_n z_n^{j-n} \quad \text{for} \quad j \ge n, \tag{40b}$$

where z is the propagation factor, and is equal to either $e^{ik\Delta}$ for propagating states or $e^{-\gamma\Delta}$ for evanescent states. The values of z at the boundaries may be directly obtained by solving Schrödinger's equation in the boundary neighborhoods:

$$E = d_1 - s_1(z_1 + z_1^{-1}), \tag{41a}$$

$$E = d_n - s_n(z_n + z_n^{-1}). \tag{41b}$$

The wavefunctions near the boundaries may thus be written:

$$\psi_0 = a_1 z_1^{-1} + b_1 z_1, \tag{42a}$$

$$\psi_1 = a_1 + b_1, \tag{42b}$$

$$\psi_n = a_n + b_n, \tag{42c}$$

$$\psi_{n+1} = a_n z_n^{-1} + b_n z_n. \tag{42d}$$

To obtain the QTBM equations, one solves (42) for a_1 and a_n, obtaining

$$a_1 = \frac{\psi_0 - z_1\psi_1}{z_1^{-1} - z_1} = \alpha_1\psi_0 + \beta_1\psi_1, \tag{43a}$$

$$a_n = \frac{\psi_{n+1} - z_n\psi_n}{z_n^{-1} - z_n} = \alpha_n\psi_{n+1} + \beta_n\psi_n. \tag{43b}$$

Adding (43) to the matrix representation to Schrödinger's equation (35) we obtain the linear system to

be solved:

$$\begin{bmatrix} \alpha_1 & \beta_1 & & & & & \\ -s_1 & d_1 - E & -s_2 & & & & \\ & -s_2 & d_2 - E & -s_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -s_n & d_n - E & -s_{n+1} & \\ & & & & \beta_n & \alpha_n & \end{bmatrix} \begin{bmatrix} \psi_0 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \\ \psi_{n+1} \end{bmatrix} = \begin{bmatrix} a_1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ a_n \end{bmatrix}. \tag{44}$$

To find the left-incident scattering state one would simply set $a_1 = 1$ and $a_n = 0$ (and conversely for the right-incident state) and solve the tridiagonal system for all ψ_j. Because the matrix is tridiagonal, a very fast numerical technique can be applied to its solution [19], though complex arithmetic must be used.

This scheme provides an extremely simple and robust way to evaluate scattering-state wavefunctions. It has been applied to multi-band tight-binding calculations and has solved the stability problems which have severely hampered transmission-matrix calculations of similar problems [20]. It can also be adapted to the problem of locating and characterizing resonant states [21]. The eigenvalues of (44) give the resonant energy (real part) and the resonance width (imaginary part). Because the α and β elements of (44) are energy-dependent, the eigenvalue problem is nonlinear and must be solved by an iterative procedure. Also, the standard theorem stating that the number of eigenvalues equals the dimension of the matrix no longer holds. An example of a calculation using this technique is shown in Figure 1 which shows the resonant states of a double quantum-well structure under bias.

III Near-Equilibrium Transport

The great majority of published work on the subject of quantum transport deals with conditions very near to thermal equilibrium, particularly with very small voltage drop across the transporting system. These conditions are known as the "linear-response regime," because the currents induced are linear in the applied voltage. The reason that such circumstances have received so much attention is not due to the technological importance of the linear-response regime, but is rather due to the difficulty of theoretically describing significant departures from equilibrium. If these departures are negligible, then one may invoke the well-developed machinery of equilibrium statistical physics and simply treat the departure from equilibrium as a small perturbation on the equilibrium state.

One approach to linear response theory is represented by the Kubo formula for the conductivity [22, 23]

$$\sigma_{ij} = \beta \int_0^\infty dt\, e^{i\omega t} \langle J_i(0) J_j(t) \rangle_{\text{eq}}, \tag{45}$$

where the brackets indicate an average over the equilibrium state. This is a form of the fluctuation-dissipation theorem, relating a transport coefficient, which necessarily characterizes a dissipative process, to the fluctuations about the equilibrium state. Another well-known form of the fluctuation-dissipation theorem is the Einstein relation connecting the mobility and the diffusivity in classical transport theory: $\mu = qD/k_{\text{B}}T$. The Kubo formula expresses the conductivity in terms of the autocorrelation of the current density; if one can calculate this autocorrelation from the equations of motion, for example, one can evaluate the frequency-dependent conductivity.

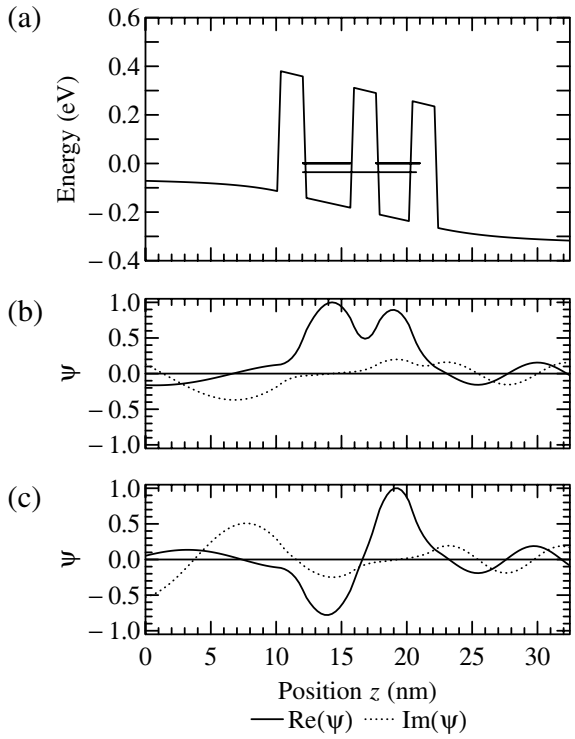

Figure 1: Resonant states of a double quantum-well structure under that voltage bias which brings the resonances into alignment. The resonances were located by finding the eigenvalues of (44). In (a) the conduction-band profile is shown along with an indication of the energy and localization of the resonances. The lower-energy resonant wavefunction is shown in (b) and the higher-energy resonant wavefunction is shown in (c).

1 Conductance — The Landauer Formula

Another general approach to near-equilibrium transport is embodied in the celebrated Landauer formula [24, 25, 26]. This formula expresses the conductance of a system at $T = 0$ in terms of the quantum mechanical transmission coefficients discussed above.

The Landauer formula has become the standard theoretical model by which the results of experiments on electron waveguides [3] and ballistic magnetotransport are interpreted. The presently accepted form of the Landauer formula may be readily derived from the expression for the one-dimensional tunneling current (29). In the limit of absolute zero temperature, the Fermi-Dirac distribution function (13) becomes a step function:

$$\lim_{k_{\rm B}T\to 0} f_{\rm FD}(E - E_F) = \Theta(E_F - E). \tag{46}$$

If a small bias voltage ΔV is applied to the system, $E_{Fr} = E_{Fl} - q\Delta V$. Then

$$f_{\rm FD}(E - E_{Fl}) - f_{\rm FD}(E - E_{Fr}) \approx q\,\Delta V\,\delta(E - E_F),$$

and the integral in (29) may be evaluated to obtain the conductance:

$$g = \frac{\Delta I}{\Delta V} = \frac{q^2}{h}\sum_{k_\perp} T(E_F, k_\perp), \tag{47}$$

where, in the tradition of this field, the factor of 2 due to the spin degeneracy is implied in the sum over the transverse modes (which are now termed "channels"). The constant q^2/h is the "quantum of conductance," and is equal to 39.6 μS, or its inverse is 25.2 kΩ.

One can obtain an alternative form of the Landauer formula by considering a different definition of the voltage drop ΔV [27, 28]. Equation (47) is obtained if one defines the voltage drop by measuring the chemical potentials deep within the respective reservoirs. In experimental terms, this corresponds to a two-terminal measurement. In practice one often employs the Kelvin probe, or a four-terminal arrangement, separating the current conducting from the voltage sensing terminals. In this case it appears that the Landauer formula must be modified because the measured voltage drop will not equal $(E_{Fl} - E_{Fr})/q$. The problem is that, if the transmission coefficient $T \neq 0$ is nonzero, the densities of electrons on either side of the device will not have the same values that they would have in equilibrium. Suppose that the potential for electrons is lower in the right-hand electrode, so the electron flow is from left to right. If the device had a very large energy barrier, so $T \approx 0$, the electron density due to the states with energies E such that $E_{Fr} < E < E_{Fl}$ consists of two equal parts: that due to the incident electrons and that due to the reflected electrons. Now, if $T = 1$, no electrons are reflected, but the density of incident electrons is still the same, so there is only one-half the electron density in the energy range $E_{Fr} < E < E_{Fl}$ in the left-hand lead. Over a small energy range, the Fermi level is proportional to the electron density, and also the electron deficit on the left-hand side is proportional to T, so we have

$$E_{Fl}' = E_{Fl} - \tfrac{1}{2}T(E_{Fl} - E_{Fr}) = E_{Fl} - \tfrac{1}{2}qT\Delta V, \tag{48a}$$

where E_{Fl}' is the quasi-Fermi that would be measured in the left-hand lead. Now the electron density which was missing from the left-hand lead appears in the right-hand lead, raising its quasi-Fermi level. Consequently,

$$E_{Fr}' = E_{Fr} + \tfrac{1}{2}qT\Delta V. \tag{48b}$$

The measured voltage drop is now

$$\Delta V' = (E_{F_l}' - E_{F_r}')/q = (1-T)\Delta V = R\Delta V. \tag{48c}$$

Correcting the Landauer formula (49) for the measured voltage drop leads to

$$g = \frac{\Delta I}{\Delta V'} = \frac{q^2}{h}\frac{T}{R}, \tag{49}$$

which was, in fact, the form originally proposed by Landauer [24].

Which of these two forms, (47) or (49), is appropriate for a given measurement is still a subject of some debate [26]. The question is whether the experimental voltage probes are sufficiently weakly coupled to the transporting system so as to give an unbiased measurement of the local quasi-Fermi level as assumed above. To provide an explicit description of multiprobe experiments, Büttiker derived a formula for the current in each lead in a multiprobe system [29]:

$$I_i = \frac{q}{h}\left[(1-R_{ii})E_{Fi} - \sum_{j\neq i} T_{ij}E_{Fj}\right], \tag{50}$$

where i and j index the leads. This formula is derived assuming that the current in each lead is carried by only one channel. In this connection, one should also note the more general formula for the multi-channel case (assuming the quasi-Fermi level correction), derived by Büttiker *et al.* [29]. If there are N conducting channels to the left and N' conducting channels to the right,

$$g = \frac{2q^2}{\pi\hbar}\frac{\sum_{i=1}^{N'} T_i}{1 + \left[\sum_{i=1}^{N} v_{li}^{-1}R_i\right]\left[\sum_{i=1}^{N} v_{li}^{-1}\right]^{-1} - \left[\sum_{i=1}^{N'} v_{ri}^{-1}T_i\right]\left[\sum_{i=1}^{N'} v_{ri}^{-1}\right]^{-1}}. \tag{51}$$

Because this equation is expressed in terms of the electron velocities, is also valid for non-parabolic energy band structures, as discussed previously.

Despite the attention directed toward linear-response theories, they remain severely limited in the range of physical situations which they address. The Landauer formula is only valid at very low temperatures and very small bias voltages. A finite-temperature form has been derived [28], but it is merely a restatement of the tunneling theory described in Section II. In fact the Landauer formula in its various forms contains no physics beyond that contained in the tunneling theory. In particular, it does not deal with dissipative scattering processes within the transporting system. The Kubo formula, on the other hand, can include such processes if they are incorporated into the evaluation of the current correlation functions. Neither of these approaches is appropriate for a "small-signal analysis" in the engineering sense of this term. Such an analysis studies small departures from a steady-state, but typically far from equilibrium situation (a significant voltage drop occurs). The linear-response theories study small departures from equilibrium, not from a non-equilibrium state.

IV Far-From-Equilibrium Transport

1 Mixed States and Distribution Functions

When a system such as an electron device is driven far from equilibrium by the application of an external voltage, both coherent and incoherent processes will generally occur within the device. Coherent processes

include tunneling and ballistic transport, and incoherent processes include dissipative scattering via phonons, for example. Coherent effects are described by adding complex-valued amplitudes (that is, values of the wavefunction), which is done implicitly in the solution of Schrödinger's equation above. Incoherent effects are described by superposition of real-valued probabilities. An example of such incoherent superposition is the summation of the current density over energies and transverse modes to obtain the total current density, as discussed in Subsection 2. We can formalize the statistical summation procedure described there into a mathematical object known as the density matrix [30, 31]. In terms of the continuum position variable x, the density matrix is actually a complex-valued function of two arguments, and has the general form:

$$\rho(x, x') = \sum_i P_i \psi_i(x) \psi_i^*(x'), \tag{52}$$

where the ψ_i form a complete set of states (*not* necessarily the eigenstates of the Hamiltonian), and the P_i are real-valued probabilities for finding an electron in each state ψ_i. With this definition, the expectation value of any physical observable represented by an operator A is given by:

$$\langle A \rangle = \mathrm{Tr}(A\rho) = \int (A\rho)(x, x)\, dx, \tag{53}$$

where A is taken to operate with respect to the first argument of ρ. Inserting (52) into (53) and rearranging the expression, we get the more familiar form for the expectation value:

$$\langle A \rangle = \sum_i P_i \int \psi_i^*(x) A \psi_i(x)\, dx. \tag{54}$$

In particular, the particle density is given by

$$n(x) = \rho(x, x), \tag{55a}$$

and the current density is

$$J(x) = \frac{q\hbar}{im^*} \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial x'} \right) \rho. \tag{55b}$$

If $E(\mathbf{k})$ is non-parabolic, a more complicated expression is required for the current density.

If the motion of the particles described by the density matrix is purely ballistic (no energy loss) and defined by a Hamiltonian H, the equation describing the evolution of the density matrix may be derived by substituting Schrödinger's equation into (52). The result is the Liouville-von Neumann equation:

$$\frac{\partial \rho}{\partial t} = \frac{1}{i\hbar}[H\rho - \rho H] \equiv \mathcal{L}_\rho \rho, \tag{56}$$

where $\mathcal{L}_\rho$ is a linear operator which operates upon the density matrix and is called the Liouville operator. (Since it operates upon ρ, which is itself a quantum-mechanical operator, $\mathcal{L}_\rho$ is technically a superoperator [5].) The Liouville equation acts upon the density matrix by evolving the wavefunctions, but does not change the probabilities P_i. This is a characteristic of ballistic, or conservative, motion. Irreversible, or dissipative, processes involve transitions between quantum states, and are described by operators which modify the probabilities P_i. Such operators are discussed below.

In the classical systems, the quantity which describes the state of the system corresponding to ρ is the phase-space distribution function $f_\mathrm{c}(r, p)$ where r is now the position and p is the momentum. The classical Liouville equation is

$$\frac{\partial f_\mathrm{c}(r, p, t)}{\partial t} = -v \frac{\partial f_\mathrm{c}(r, p, t)}{\partial r} + \frac{\partial V}{\partial r} \frac{\partial f_\mathrm{c}(r, p, t)}{\partial p} \equiv \mathcal{L}_\mathrm{c}, \tag{57}$$

where v is the velocity and V is the potential in which the particles are moving. The particle and current densities are obtainable from the classical distribution function by

$$n(r,t) = \int dp\, f_{\rm c}(r,p,t), \tag{58a}$$

$$J(r,t) = \int dp\, v(p) f_{\rm c}(r,p,t). \tag{58b}$$

The Liouville equation, in either the classical (57) or quantum (56) context, describes only ideal, conservative motion. Within the scope of these equations, particles can only oscillate within the system, unless one applies boundary conditions which permit particles to escape from it. The form of the equations (for closed systems) does not describe an approach to a steady-state, neither equilibrium nor non-equilibrium. The reason for this involves the eigenvalue spectrum of $\mathcal{L}_\rho$ and $\mathcal{L}_{\rm c}$. The solutions of (56) will consist of a linear combination of terms with time dependence $e^{-i\omega_i t/\hbar}$, where $-i\omega_i$ are the eigenvalues of $\mathcal{L}_\rho$. The Liouville operator [as defined in (56), including the imaginary factor] is anti-Hermitian, so the frequencies ω_i are purely real. Thus the transformation which maps the state of the system at some initial time into some later time is a unitary linear transformation, and we will call the behavior described by such equations "unitary time evolution." Devices of course usually approach a steady state after a sufficiently long time. To describe this behavior, we must incorporate irreversible processes into the equations.

2 Irreversible Processes and Master Equations

Irreversible or energy-dissipating processes always involve transitions between quantum states. Such processes are described, at the simplest level, by master or rate equations [32]. The operators which generate the time-evolution in such equations are of a very different form from that of the Liouville operator.

If the state of a system is described by an array of probabilities or occupation factors P_i for a particle to occupy a (stationary) quantum level i, the time evolution of that system is determined by the rates of transition between the levels i. These rates are usually estimated using the "Fermi golden rule:"

$$W_{ij} = (2\pi/\hbar)|\langle i|H_{\rm int}|j\rangle|^2 \delta(E_i - E_j), \tag{59}$$

where $H_{\rm int}$ is the Hamiltonian describing the interaction that causes the transitions, and W_{ij} is the transition rate from state j to state i. The δ-function ensures energy conservation, but it must be remembered that E_i and E_j are the total energy of each state, including, for example, the energy in an emitted phonon. Thus (59) can describe energy-dissipating processes despite its appearance. If one assumes that these transitions occur independently within any small time interval (the Markov assumption), the transition from state j to state i will produce changes in the corresponding occupation factors:

$$dP_i = -dP_j = W_{ij} P_j\, dt. \tag{60}$$

The occupation of state i increases and that of state j decreases as a result of this particular process, and the amount of change depends only upon the occupation of the initial state. (We neglect here the Pauli exclusion principle, which leads to nonlinear master equations.) If we sum (60) over all possible transition processes, we obtain the master equation:

$$dP_i/dt = \sum_j [W_{ij} P_j(t) - W_{ji} P_i(t)] = MP, \tag{61}$$

where M is the master operator, whose matrix elements are given by

$$M_{ij} = \begin{cases} W_{ij} & i \neq j \\ -\sum_{j \neq i} W_{ij} & i = j \end{cases} . \tag{62}$$

Notice the form of this operator. The off-diagonal elements are all positive and the diagonal elements are all negative, with a magnitude equal to the sum of the off-diagonal elements in the same column. (If one considers an open system, the coupling to external reservoirs can lead to master operators in which the magnitude of the diagonal elements exceeds the sum of the off-diagonal elements [5].) The eigenvalues of an operator of this form will all have real parts less than or equal to zero [5, 32]. Thus the solutions of (61) will consist of a linear combination of terms with a decaying exponential time-dependence, and so will always show a stable approach to some steady state.

The Pauli master equation [33] is the most commonly used model of irreversible processes in simple quantum systems. It can be derived from elementary quantum mechanics plus a Markov assumption. Within these assumptions, the density matrix has the form (52) with the states i being the eigenstates of the system Hamiltonian, and remains of this form at all times. The Pauli master equation is then just (61) with the Fermi golden-rule rates (59). There are a number of conceptual problems with the Pauli equation [33] , not the least of which is that it produces violations of the continuity equation [5]. It is nevertheless employed, either explicitly or implicitly, in almost all semi-classical treatments of electron transport in semiconductors.

Master operators most often occur in the description of stochastic (random) processes, where they describe the average behavior of the system. In such cases there will always be fluctuations (noise) about the solution of the master equation. Diffusion phenomena are the classic example of this. The master operator in the classical diffusion equation $\partial n/\partial t = D\nabla^2 n$ is just the laplacian ∇^2. By examining the form of the finite-difference approximation to the second derivative (34), it is easy to see that this has the form of a master operator (62).

Another case of particular importance (and a source of some confusion) is the gradient operator, which appears in the classical Liouville equation (57) and in the drift term of the drift-diffusion equation ($-v\nabla n$), among many other contexts. The unique property of this operator is that, *depending upon the boundary conditions imposed*, ∇ can be an anti-Hermitian operator (generating unitary time-evolution), or $-v\nabla$ can be a master operator (generating an approach to a steady state). If one applies periodic boundary conditions, the eigenstates of ∇ are of the form e^{ikx}, with real-valued eigenvalues k. The finite-difference approximation appropriate to this situation is the centered-difference form

$$(\partial\phi/\partial x)_j = (\phi_{j+1} - \phi_{j-1})/2\Delta, \tag{63}$$

which (if written in matrix form) is clearly anti-Hermitian. On the other hand, if one applies initial conditions to $-v\nabla$, a single boundary condition should be imposed on the left if $v > 0$ or on the right if $v < 0$. The appropriate discretization in this case is the *upwind difference* [5, 34]

$$\left(-v\frac{\partial n}{\partial x}\right)_j = \frac{|v|}{\Delta}\begin{cases} -n_j + n_{j-1}, & v > 0, \\ -n_j + n_{j+1}, & v < 0, \end{cases} \tag{64}$$

which clearly has the form of a master operator. The upwind difference is known to produce excellent stability in fluid dynamic calculations [34], and the master-operator form is the ultimate explanation of its success.

2.1 The Boltzmann Equation

The Boltzmann equation is the basis for the standard models of electron transport in semiconductors in a semi-classical approximation. It consists of the classical Liouville equation (57) augmented by a master operator of precisely the form (61) to describe collisions between electrons and other particles. The Boltzmann equation is commonly written in the form [35]:

$$\begin{aligned}\frac{\partial f(\mathbf{r},\mathbf{p},t)}{\partial t} = & - \mathbf{v}(\mathbf{p})\cdot\nabla f(\mathbf{r},\mathbf{p},t) - \mathbf{F}\cdot\nabla_{\mathbf{p}} f(\mathbf{r},\mathbf{p},t) \\ & + \sum_{\mathbf{p}'}\left[W(\mathbf{p}',\mathbf{p})f(\mathbf{r},\mathbf{p}',t) - W(\mathbf{p},\mathbf{p}')f(\mathbf{r},\mathbf{p}',t)\right], \end{aligned} \tag{65}$$

where $\mathbf{F}$ is the force on the electron (due to the electric field, for example). $W(\mathbf{p}',\mathbf{p})$ is the scattering rate from $\mathbf{p}$ to $\mathbf{p}'$, as in (59), but now $\mathbf{p}$ is a continuous variable. These scattering rates are conventionally obtained by evaluating the Fermi golden rule for scattering between plane-wave states with wavevectors $\mathbf{k} = \mathbf{p}/\hbar$, and $\mathbf{k}' = \mathbf{p}'/\hbar$. Note that these are the scattering rates between infinitely extended states, but in the Boltzmann equation, each scattering event is assumed to take place at a single point. (If we did not assume this, the continuity equation would be violated.)

The different scattering mechanisms in semiconductors and their rates has been the subject of much theoretical and experimental work. Expressions for these scattering rates can be found in the works of Conwell [36], and Ridley [37]. The evaluation of low-field transport properties of bulk semiconductors is described in detail by Rode [38]

2.2 Numerical Evaluation Methods

The solution of the Boltzmann equation presents a rather difficult problem, because of the large number of variables involved. For a general three-dimensional system there are six arguments of the distribution function (three components each of $\mathbf{r}$ and $\mathbf{p}$). If we were to simply discretize the Boltzmann equation, we would need to represent the distribution function f by an array with six subscripts, corresponding to the six arguments of the continuous function. If we used only ten values for each subscript, the f array would have 10^6 elements, and each evaluation of f would require a corresponding number of operations. Such a discretization of the Boltzmann equation has in fact been investigated by Aubert, Vaissiere, and Nougier [39], for the case of a spatially homogeneous system, so that only the three velocity arguments are required.

By far the most widely used technique for evaluating the Boltzmann description of electron transport has been the Monte Carlo method [40, 41, 42, 43]. With this technique, one does not *solve* the Boltzmann equation directly, but one rather simulates the motion of classical electrons subjected to a combination of free-flight motion and instantaneous random scattering events. The distribution function is then estimated by statistical averages over long times or many particles. The state of the system is represented by the position and velocity vectors of each of a large number of particles. The velocity and position of each particle are integrated over time until a collision is deemed to have occurred (based upon the value of a randomly-chosen value with an appropriate distribution). Other random values then determine the particular scattering mechanism and the velocity of the electron after the scattering. After the scattering, the free-flight motion of the electron is again integrated until the next collision. This procedure is performed for all particles in the chosen ensemble to evaluate the time-evolution of the device. The openness of the device is modeled

by procedures which treat the escape of electrons into and injection of electrons from the contact regions [44, 45].

The Monte Carlo technique permits one to include many of the physical effects that influence electron transport, and to include them at an extremely detailed level. Effects which have been incorporated include the detailed energy-band structure of the semiconductor [46], electron-electron interactions at the level of the self-consistent potential or at a more detailed electron-electron collision description [47], and higher-energy events such as impact ionization [45]. However, there is one very significant constraint imposed by the Monte Carlo procedure: only processes describable by a master operator can be modeled.

2.3 The Wigner Distribution Function

The Wigner distribution function is a mathematical transform of the density matrix which approaches the classical distribution function $f_{\rm c}$ as the system becomes classical (with large dimensions, slowly varying potentials, and/or high temperatures) [5, 48, 49, 50, 51]. This representation of the statistical state has proven to be useful in modeling quantum-effect devices such as the resonant-tunneling diode [52, 53, 54, 55].

To derive the Wigner function from the density matrix $\rho(x,x')$ defined in (52) one rewrites the arguments (x,x') as $r=\frac{1}{2}(x+x')$ and $r'=x-x'$, and then Fourier transforms r' into a momentum variable p. Thus:

$$f_{\rm W}(r,p)=\int_{-\infty}^{\infty}dr'\,\rho\left(r+\tfrac{1}{2}r',r-\tfrac{1}{2}r'\right)e^{-ipr'/\hbar}. \tag{66}$$

Applying the same procedure (which is known as the Wigner-Weyl transformation) to the Liouville-von Neumann equation (56) gives:

$$\frac{\partial f_{\rm W}}{\partial t}=-\frac{p}{m^*}\frac{\partial f_{\rm W}}{\partial r}-\frac{1}{\hbar}\int_{-\infty}^{\infty}\frac{dp'}{2\pi\hbar}V_{\rm W}(r,p-p')f_{\rm W}(r,p'), \tag{67}$$

where the kernel of the potential operator is given by:

$$V_{\rm W}(r,p)=2\int_0^{\infty}dr'\,\sin(pr'/\hbar)\left[V\left(r+\tfrac{1}{2}r'\right)-V\left(r-\tfrac{1}{2}r'\right)\right]. \tag{68}$$

Let us examine the form of these equations. Because (67) is derived from (56) by a mathematical transformation, we would expect that it should also describe unitary time evolution. The condition for unitary evolution is that $\mathcal{L}_{\rm W}$ be an anti-Hermitian operator. The potential operator is anti-Hermitian [because $V_{\rm W}(r,-p)=-V_{\rm W}(r,p)$], and the drift term is anti-Hermitian if periodic boundary conditions are imposed. On the other hand, we have seen that if initial conditions are imposed, the drift term is a master operator, and the equation then describes irreversible time evolution. This is the origin of the usefulness of the Wigner representation for describing electron devices. One applies boundary conditions to $f_{\rm W}$ so as to fix the distribution of electrons entering the domain:

$$f_{\rm W}(x_l,p)|_{p>0}=f_l(p), \tag{69a}$$

$$f_{\rm W}(x_r,p)|_{p<0}=f_r(p), \tag{69b}$$

where $f_l(p)$ and $f_r(p)$ are the distribution functions in the left- and right-hand contacts (reservoirs), respectively. Because these boundary conditions introduce irreversibility into the Liouville equation, one can now evaluate the time-evolution of a device, and observe an approach to steady-state [5]. Inelastic processes

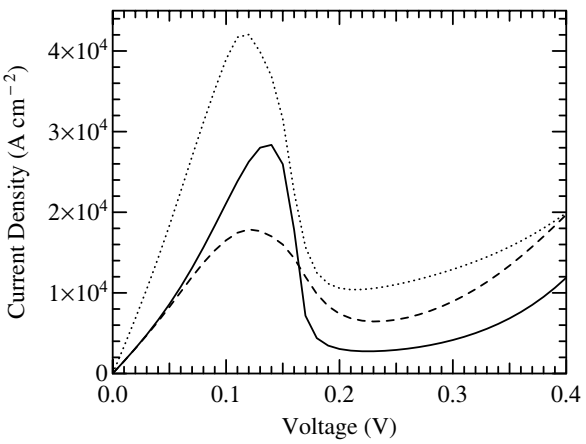

Figure 2: Theoretical $I(V)$ curves for a model resonant-tunneling diode structure . The solid line shows the results of the tunneling theory, equation (28). The dashed line shows the result of evaluating the Wigner function using a local model of the variation of the effective mass [52]. The dotted line shows the result of a Wigner function calculation using the nonlocal effective-mass model of Tsuchiya and co-workers [58]. (Calculations by C. Fernando, University of Texas at Dallas.)

such as phonon scattering may be included in a semi-classical way by adding the Boltzmann collision term [the integral expression in (65)] to the Liouville equation (56) [56, 54], or by even simpler schemes such as the relaxation time approximation [53, 57].

The open-system Wigner function approach has proved to be of use in understanding the behavior of resonant-tunneling diodes. This technique permits evaluation of steady-state behavior in the form of the $I(V)$ curve, and calculations of the large-signal transient response and small-signal ac response [5]. The $I(V)$ curves derived from this model show the expected negative differential-conductance region, but the ratio of the peak to valley currents is always smaller than that obtained from the tunneling theory (28), and is often less than that observed experimentally. Recently, Tsuchiya and co-workers have developed an improved formulation of the Liouville equation (67) which takes the spatial variation of the effective mass into account, and which leads to larger peak-to-valley ratios than the simpler theory [58]. These different formulations are compared in Figure 2.

2.4 Green's Functions

A more sophisticated approach to quantum transport theory is supplied by the Green's-function formulation of many-body theory. This approach had its origins in the development of the theory of quantum electrodynamics in the late 1940's and early 1950's, and inherits this field's emphasis on perturbation expansions described in diagrammatic form. The non-equilibrium Green's-function theory was formulated by Kadanoff and Baym [59] and by Keldysh [60], was elaborated by Langreth [61], and is described in the text by Mahan [62]. The problems and promise of applying this approach to electron devices has been discussed by Jauho [63]. In particular, most of the development of the Green's function approach has assumed uniform electric fields, which is not adequate for the description of quantum devices. Among the more recent work in this

area which addresses problems beyond the uniform field are those of Sols [64] and Rammer [65].

The non-equilibrium Green's functions are defined as expectation values of single-particle creation and annihilation operators, and they describe the state and time evolution of the system. There are four independent functions which appear in the formalism. In the conventional notation $G^<$ gives the distribution of electrons (and reduces to the density matrix or Wigner function in certain limits), $G^>$ gives the distribution of holes, G_r describes time evolution into the future and G_a describes time evolution into the past. These Green's functions are determined by solving a set of Dyson equations (an integral form of Schrödinger's equation) which form a convenient starting point for the development of a perturbation expansion.

Each of the Gs has two position and two time arguments, which can be transformed via the Wigner-Weyl procedure into one each of a position, momentum, time, and energy (or frequency) argument. The presence of the energy dependence (or the two time arguments) distinguishes the Green's function approach from the Wigner function scheme described above. Because the Wigner function measures the state of the device at a particular time, and its evolution is described by a first-order differential equation, it can only comprehend external interactions which occur instantaneously in time. Such behavior is termed "Markovian." The energy dependence of the Green's functions permits a description of processes which are not local in time, or "non-Markovian" processes, because the energy argument provides a way to include convolution integrals over the past history of the system. An example of a process which is non-Markovian is the resonant absorption or emission of a phonon. In order for the energy of the phonon to be well-defined, the interaction must occur over a time greater than the oscillation period of the phonon. A non-Markovian Green's function approach can accurately describe such processes, the Markovian Wigner function approach cannot.

It is fair to say that work on the Green's function approach has produced a great many mathematical formulations and very few explicit calculations of realistic systems. Among the latter is the work of Lake and Datta [66]. They model the resonant-tunneling diode, and include interactions between the electrons and localized phonons. The locality of the interaction removes the momentum argument from the Green's function, and also removes any notion of momentum conservation from the model.

One can identify a general principle here, to the effect that if any of the position, momentum, or energy arguments are missing from the distribution function in a given theory, then a corresponding conservation law is not enforced within that theory. If the position argument is not present, as in the case of the Pauli master equation, then the continuity equation is not enforced. If the momentum argument is absent, as in the approach of Lake and Datta, then conservation of momentum is not enforced. And finally, if the energy argument is absent, as in the Markovian Wigner function theory, then conservation of energy is not enforced. Thus it appears that the only completely satisfactory theory will be a complete Green's function theory which includes all four arguments. Such an approach has not yet been developed into a numerically tractable form.

V Summary

One can model the transport of electrons through quantum devices at a number of different levels of sophistication. The simplest level consists of solving the single-particle Schrödinger equation. Because devices are open systems, the solutions of Schrödinger's equation describing unbounded scattering states are the appropriate basis in which to consider electron transport. Semiconductor heterostructures create some com-

plications in the application of conventional scattering theory, because the electron dispersion relation (band structure) will be non-parabolic and will vary with position. These effects require that the group velocity v be used in most of the fundamental equations of scattering theory where the wavevector k conventionally appears. Extremely robust and efficient numerical techniques have recently been developed which permit evaluation of scattering states, taking into account the these complications.

The near-equilibrium transport properties of a quantum device can be well described with a knowledge of the quantum transmission amplitudes and by invoking the Landauer conductance formula. However, to fully describe the far-from-equilibrium transport on which useful devices depend, one must describe the device in terms of quantum stastical mechanics. Semi-classical formulations, such as the Boltzmann equation and the techniques used to evaluate it, cannot properly deal with quantum interference effects except insofar as they can be described in terms of transition rates. More comprehensive quantum kinetic theories using the Wigner distribution function or non-equilibrium Green's functions have been and are continuing to be developed. To date, however, all of the theories which for which practical computational schemes have been implemented suffer from an inability to deal with one or another of the fundamental conservation laws. Thus there is no one theoretical tool which provides a satisfactory model of all aspects of quantum device behavior.

References

References

[1] E. R. Brown, Chapter 10 of the present volume.

[2] A. C. Seabaugh and M. A. Reed, Chapter 11 of the present volume.

[3] J. H. Davies and G. Timp, Chapter 12 of the present volume.

[4] J. N. Randall, J. H. Luscombe, and R. T. Bate, Chapter 13 of the present volume.

[5] W. R. Frensley, Rev. Mod. Phys. 62, 745 (1990).

[6] S. Selberherr, "Analysis and Simulation of Semiconductor Devices," sect. 5.1. Springer-Verlag, Vienna, 1984.

[7] D. J. BenDaniel and C. B. Duke, Phys. Rev. 152, 683 (1966).

[8] T. Ando, A. B. Fowler and F. Stern, Rev. Mod. Phys. 54, 437 (1982).

[9] , G. Burns, "Solid State Physics," Academic Press, Orlando, 1985, ch. 9.

[10] E. L. Wolf, "Principles of Electron Tunneling Spectroscopy," Oxford University Press, New York, 1985.

[11] E. Merzbacher, "Quantum Mechanics," 2nd ed., John Wiley & Sons, 1970, ch. 6-7.

[12] J. N. Schulman and Y.-C. Chang, Phys. Rev. B 27, 2346 (1983).

[13] J. H. Luscombe, private communication.

[14] D. Z.-Y. Ting, and Y.-C. Chang, Phys. Rev. B 36, 4359 (1987).

[15] J. Mathews and R. L. Walker, "Mathematical Methods of Physics," W. A. Benjamin, New York, 1970, ch. 12.

[16] R. K. Mains, I. Mehdi, and G. I. Haddad, Appl. Phys. Lett. 55, 2631 (1988).

[17] C. Juang, K. J. Kuhn, and R. B. Darling, Phys. Rev. B 41, 12047 (1990).

[18] C. S. Lent and D. J. Kirkner, J. Appl. Phys. 67, 6353 (1990).

[19] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical Recipes in C, The Art of Scientific Computing," pp. 47–48. Cambridge University Press, Cambridge, 1988.

[20] D. Z.-Y. Ting, E. T. Yu, and T. C. McGill, Phys. Rev. B 45, 3583 (1992).

[21] W. R. Frensley, Superlattices and Microstructures 11, 347 (1992).

[22] R. Kubo, J. Phys. Soc. Japan 12, 570 (1957).

[23] R. Kubo, M. Toda, and N. Hashitsume, "Statistical Physics II. Nonequilibrium Statistical Mechanics," Springer-Verlag, Berlin, 1985.

[24] R. Landauer, IBM J. Res. Develop. 1, 233 (1957).

[25] R. Landauer, Phil. Mag. 21, 863 (1970).

[26] A. D. Stone and A. Szafer, IBM J. Res. Develop. 32, 384 (1988).

[27] H.-L. Engquist and P. W. Anderson Phys. Rev. B 24, 1151 (1981).

[28] M. Büttiker, Y. Imry, R. Landauer, and S. Pinhas, Phys. Rev. B 31, 6207 (1985).

[29] M. Büttiker, Phys. Rev. Lett. 57, 1761 (1986).

[30] U. Fano, Rev. Mod. Phys. 29, 74 (1957).

[31] R. P. Feynman, "Statistical Mechanics, A Set of Lectures," W. A. Benjamin, Reading, MA, 1972, ch. 2.

[32] I. Oppenheim, K. E. Shuler, and G. H. Weiss, " Stochastic Processes in Chemical Physics: The Master Equation," MIT Press, Cambridge, Mass., 1977, and reprints included therein.

[33] H. J. Kreuzer, "Nonequilibrium Thermodynamics and its Statistical Foundations," Oxford University Press, Oxford, 1981, ch. 10.

[34] P. J. Roache, "Computational Fluid Dynamics," Hermosa Publishers, Albuquerque, NM, 1976, pp 4–5.

[35] K. Hess, "Advanced Theory of Semiconductor Devices,"Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[36] E. M. Conwell, "High Field Transport in Semiconductors," Supplement 9 to Solid State Physics (F. Seitz and D. Turnbull, eds.). Academic Press, New York, 1967.

[37] B. K. Ridley, "Quantum Processes in Semiconductors," Oxford Univ. Press, New York, 1988.

[38] D. L. Rode, in "Semiconductors and Semimetals, Vol. 10, Transport Phenomena," (R. K. Willardson and A. C. Beer, eds.), pp. 1–89. Academic Press, New York, 1975.

[39] J. P. Aubert, J. C. Vaissiere, and J. P. Nougier, J. Appl. Phys. 56, 1128 (1984).

[40] P. J. Price, in "Semiconductors and Semimetals, Vol. 14, Lasers, Junctions, Transport," (R. K. Willardson and A. C. Beer, eds.), pp. 249–308. Academic Press, New York, 1979.

[41] A. D. Boardman, in "Physics Programs," (A. D. Boardman, ed.), pp. 355–410. John Wiley & Sons, Chichester, 1980.

[42] L. Reggiani, ed., "Hot-Electron Transport in Semiconductors," Topics in Applied Physics, vol. 58. Springer-Verlag, Berlin, 1985.

[43] K. Hess, ed., "Monte Carlo Device Simulation: Full Band and Beyond." Kluwer Academic Publishers, Boston, 1991.

[44] R. W. Hockney and J. W. Eastwood, "Computer Simulation Using Particles," McGraw-Hill, Inc., New York, 1981.

[45] S. E. Laux and M. V. Fischetti, in "Monte Carlo Device Simulation: Full Band and Beyond," (K. Hess, ed.), ch. 1. Kluwer Academic Publishers, Boston, 1991.

[46] H. Shichijo and K. Hess, Phys. Rev. B 23, 4197 (1981).

[47] D. K. Ferry, A. M. Kriman, M.-J. Kann, and R. P. Joshi, in "Monte Carlo Device Simulation: Full Band and Beyond," (K. Hess, ed.), ch. 4. Kluwer Academic Publishers, Boston, 1991.

[48] E. Wigner, Phys. Rev. 40, 749 (1932).

[49] E. J. Heller, J. Chem. Phys. 65, 1289 (1976).

[50] M. V. Berry, Philos. Trans. R. Soc. London 287, 237 (1977).

[51] P. Carruthers and F. Zachariasen, Rev. Mod. Phys. 55, 245 (1983).

[52] W. R. Frensley, Phys. Rev. B 36, 1570 (1987).

[53] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer, Phys. Rev. B 39, 7720 (1989).

[54] R. K. Mains and G. I. Haddad, J. Appl. Phys. 64, 5041 (1988).

[55] K. L. Jensen and F. A. Buot, J. Appl. Phys. 65, 5248 (1989).

[56] W. R. Frensley, Superlattices and Microstructures, 4, 497 (1987).

[57] K. L. Jensen and F. A. Buot, J. Appl. Phys. 67, 7602 (1990).

[58] H. Tsuchiya, M. Ogawa, and T. Miyoshi, IEEE Trans. Electron Devices 38, 1246 (1991).

[59] L. P. Kadanoff and G. Baym, "Quantum Statistical Mechanics." Benjamin-Cummings, Reading, MA, 1962.

[60] L. V. Keldysh, Zh. Eksp. Teor. Fiz. 47, 1515 (1964). [translated in Sov. Phys. JETP 20, 1018 (1965).]

[61] D. C. Langreth, in *Linear and Nonlinear Electron Transport in Solids*, (J. T. Devreese and V. E. van Doren, eds.) p. 3. Plenum, New York, 1976.

[62] G. D. Mahan, "Many-Particle Physics." Plenum Press, New York, 1990.

[63] A.-P. Jauho, Solid-State Electron. 32, 1265 (1989).

[64] F. Sols, Annals of Physics 214, 386 (1992).

[65] J. Rammer, Rev. Mod. Phys. 63, 781 (1991).

[66] R. Lake and S. Datta, Phys. Rev. B 45, 6670 (1992).