Superlattices and Microstructures, Vol. 28, No. 4, 2000
doi:10.1006/spmi.2000.0920
Available online at http://www.idealibrary.com on IDEAL

Nanoscale device modeling: the Green's function method

SUPRIYO DATTA†
School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285, U.S.A.

(Received 24 July 2000)

The non-equilibrium Green's function (NEGF) formalism provides a sound conceptual basis for the devlopment of atomic-level quantum mechanical simulators that will be needed for nanoscale devices of the future. However, this formalism is based on concepts that are unfamiliar to most device physicists and chemists and as such remains relatively obscure. In this paper we try to achieve two objectives: (1) explain the central concepts that define the 'language' of quantum transport, and (2) illustrate the NEGF formalism with simple examples that interested readers can easily duplicate on their PCs. These examples all involve a short n^{++}–n^{+}–n^{++} resistor whose physics is easily understood. However, the basic formulation is quite general and can even be applied to something as different as a nanotube or a molecular wire, once a suitable Hamiltonian has been identified. These examples also underscore the importance of performing self-consistent calculations that include the Poisson equation. The I–V characteristics of nanoscale structures is determined by an interesting interplay between twentieth century physics (quantum transport) and nineteenth century physics (electrostatics) and there is a tendency to emphasize one or the other depending on one's background. However, it is important to do justice to both aspects in order to derive real insights.

Key words: non-equilibrium Green's function, Keldysh or Kadanoff–Baym formalism, self-energy, nanoscale device modeling, quantum transport

1. Introduction

MOS transistors with channel lengths as small as 10 nm are now being actively studied both theoretically and experimentally [1]. At the same time recent demonstrations of molecular switching make molecular electronic devices seem a little closer to reality [2]. It is clear that quantitative simulation tools for this new generation of devices will require atomic-level quantum mechanical models. The non-equilibrium Green function (NEGF) formalism (sometimes referred to as the Keldysh or the Kadanoff–Baym formalism) provides a sound conceptual basis for the devlopment of this new class of simulators. 1D quantum devices like tunneling and resonant tunneling diodes have been modeled quantitatively using NEMO [3] which is based on the NEGF formalism. Although the transport issues in MOS transistors or molecular electronics are completely different, the NEGF formalism should provide a suitable conceptual framework for their analysis as well. However, this formalism is based on concepts that are unfamiliar to most device physicists and chemists

†E-mail: datta@purdue.edu

0749–6036/00/100253 + 26 $35.00/0

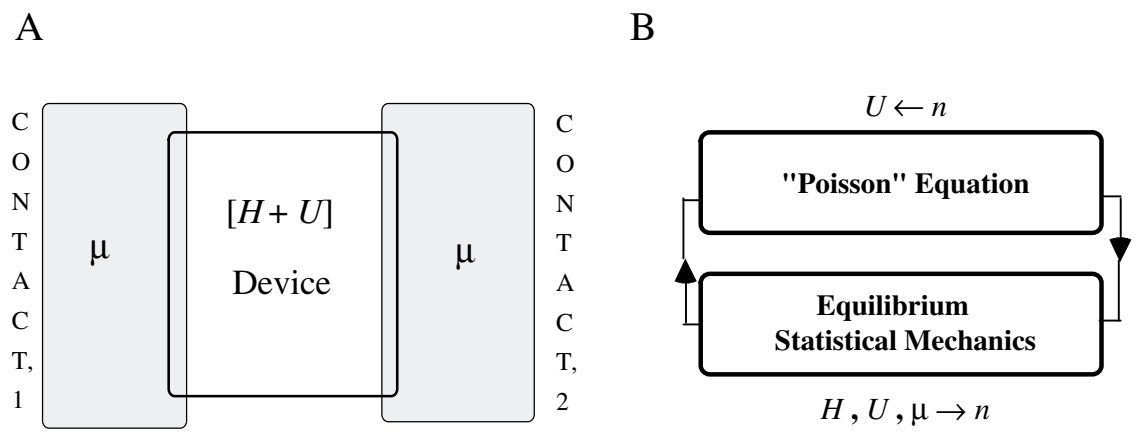

Fig. 1. A, A device in equilibrium; B, self-consistent procedure for the analysis of electronic devices in equilibrium. 'Poisson' is written within quotes as a reminder that it may need to be supplemented with a exchange-correlation potential.

and as such remains relatively obscure despite the obvious value of a fundamentally sound approach on which practical simulation tools for nanoscale devices can be based. In this paper we try to achieve two objectives: (1) explain the central concepts that define the 'language' of quantum transport, and (2) illustrate the NEGF formalism with simple examples that interested readers can easily duplicate on their PCs. The numerical results presented here (Figs 5, 9, 13) were all obtained on a laptop computer and the author will be glad to share his MATLAB programs, typically 40 lines long, with interested readers. These examples all involve a short n^{++}–n^{+}–n^{++} resistor whose physics is easily understood. Their primary purpose is to illustrate how the NEGF formalism is applied to a real device and leads to physically sensible results. The basic formulation is quite general and can even be applied to something as different as a nanotube or a molecular wire.

Most device physicists are familiar with the Schrödinger–Poisson solver. So let us start by recapitulating how the Schrödinger–Poisson solver works for a device in equilibrium (Fig. 1A). The first step is to identify a suitable Hamiltonian, H, that provides an adequate description of the isolated device. For example, if the device operation involves only the electrons in a parabolic conduction band then we could use the effective mass Hamiltonian $H \equiv -(\hbar^2/2m)\nabla^2$. This is what we will use for our illustrative examples in this paper, but the basic formulation could just as well be used with more complicated Hamiltonians like the sp^3s^* Hamiltonian commonly used to provide an accurate description of the valence band or say the 6-31G* Hamiltonian used for molecular conductors. When the device is connected to the contacts there is some charge transferred into or out of the device, which gives rise to a potential, $U(r)$, that has to be calculated self-consistently. The Schrödinger–Poisson solver (Fig. 1B) iterates between the *Poisson equation* which gives us the potential $U(r)$ for a given electron density $n(r)$ relative to that required for local charge neutrality (which is equal to the ionized donor density, $N_D(r)$, in an n-type semiconductor)

$$\nabla \cdot (\varepsilon \nabla U) = q^2[N_D - n] \tag{1.1}$$

and the law of *equilibrium statistical mechanics* which tells us that the electron density $n(r)$ for a given potential profile $U(r)$ is obtained from

$$n(r) = \sum_{\alpha} |\Psi_\alpha(r)|^2 \quad f_0(\varepsilon_\alpha - \mu) \tag{1.2}$$

by filling up the eigenstates $\Psi_\alpha(r)$ of the Schrödinger equation

$$[H + U]\Psi_\alpha(r) = \varepsilon_\alpha \Psi_\alpha(r) \tag{1.3}$$

according to the Fermi function

$$f_0(E-\mu) \equiv (1+\exp[(E-\mu)/k_BT])^{-1} \tag{1.4}$$

μ being the Fermi level. This is the basic approach that has been widely used to model MOS capacitors. Some authors [4] have supplemented the Poisson potential $U(r)$ with an exchange-correlation potential $U_{xc}(r)$ which accounts for the 'hole' that surrounds an individual electron in a conductive medium. This is quite common among quantum chemists [5] who have developed fairly sophisticated approaches for determining $U_{xc}(r)$. In this paper we will not address this issue and simply use the Poisson (or Hartree) potential, with the reminder that it may need to be supplemented with a suitable exchange-correlation potential to account for electron–electron interactions. This self-consistent field approach should provide an adequate equilibrium model for nanoscale devices, unless they happen to be in the 'Coulomb blockade' regime. Let me briefly elaborate on what this means.

An electronic state localized in a sphere of radius R has a single-electron charging energy of approximately $q^2/4\pi\varepsilon R$ which is ~25 meV if $R = 5$ nm and $\varepsilon = 10\varepsilon_0$. If this charging energy exceeds both the thermal energy k_BT and the level broadening due to the connection to the surroundings, then one could be in a regime dominated by single-electron charging effects that is not described well by the self-consistent field method even at equilibrium. One well-known example of this is the fact that donor or acceptor levels are occupied according to a modified Fermi function (ν: level degeneracy)

$$f(E-\mu) \equiv \left(1+\frac{1}{\nu}\exp[(E-\mu)/k_BT]\right)^{-1}$$

rather than the actual Fermi function (cf. eqn (1.4)). This elementary result, familiar to every device scientist, does NOT ordinarily follow from the NEGF formalism without special effort because impurity levels are both localized and weakly coupled. Similar issues could arise in nanoscale devices with weak coupling to contacts (such as the floating gate in a flash memory device) and such devices may require treatments that go beyond the standard NEGF formalism to take single-electron charging effects into account. In this paper we will not discuss this 'Coulomb blockade' regime any further and assume that the energy levels are sufficiently delocalized that electron–electron interactions can be modeled with an appropriate self-consistent field. However, it is important to remember that the NEGF, with all its impressive sophistication, does not automatically include 'everything'.

The problem we wish to address in this paper is that of a device connected to two contacts with two different Fermi levels μ_1 and μ_2 (Fig. 2). What is the electron density, n? We can no longer use eqn (1.2) since there are two different Fermi levels. It would seem that the energy levels in the device would be occupied with a probability $f_\alpha$which has a value intermediate between the source Fermi function $f_0(\varepsilon_\alpha - \mu_1)$ and the drain Fermi function $f_0(\varepsilon_\alpha - \mu_2)$:

$$n(r) = \sum_\alpha \Psi_\alpha(r)\Psi_\alpha^*(r)\, f_\alpha. \tag{1.5}$$

However, the general answer is more complicated than that. Different states can be occupied in a correlated manner described by a density matrix, $\rho_{\alpha\beta}$:

$$n(r) = \sum_{\alpha,\beta} \Psi_\alpha(r)\psi_\beta^*(r)\rho_{\alpha\beta}. \tag{1.6}$$

The central issue in non-equilibrium statistical mechanics is to determine the *density matrix* $\rho_{\alpha\beta}$; once found, all quantities of interest (charge, current, energy current etc) can be calculated. Since this concept is unfamiliar to most device physicists, let us elaborate a little further.

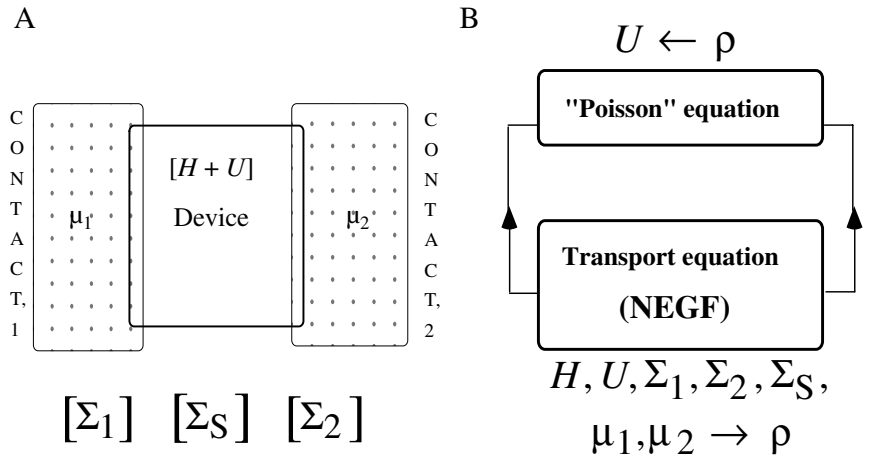

Fig. 2. A, A device driven out of equilibrium by two contacts with different Fermi levels μ_1 and μ_2; B, self-consistent procedure for determining the density matrix ρ from which all quantities of interest (electron density, current etc) can be calculated.

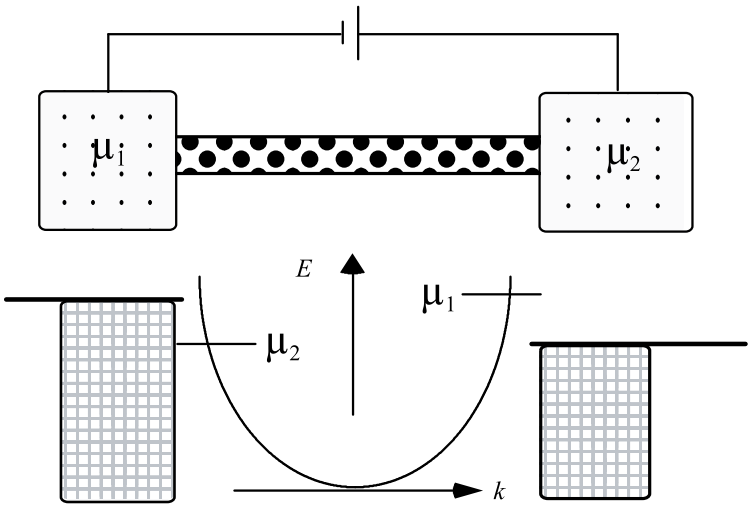

Fig. 3. A ballistic conductor connected to two contacts with different Fermi levels μ_1 and μ_2.

1.1. The density matrix and the current operator

The distinction between an occupation probability f_α and a density matrix $\rho_{\alpha\beta}$ can be appreciated by considering a simple 1D ballistic conductor connected to two contacts with Fermi levels μ_1 and μ_2 (Fig. 3). One of the celebrated results of mesoscopic physics [6–8] is that the conductance of such a ballistic conductor is quantized:

$$g = \frac{I}{[\mu_1 - \mu_2]/-q} = \frac{2q^2}{h}. \tag{1.7}$$

Equation (1.7) is derived simply as follows. The eigenstates of a 1D conductor with periodic boundary conditions can be written as (L: length of conductor)

$$\Psi_{+\mathrm{k}}(x) = e^{+ikx}/\sqrt{L}, \qquad \Psi_{-\mathrm{k}}(x) = e^{-ikx}/\sqrt{L} \tag{1.8}$$

with

$$\varepsilon_{+\mathrm{k}} = \varepsilon_{-\mathrm{k}} = E_c + (\hbar^2 k^2/2m). \tag{1.9}$$

How are these states occupied? Given the complicated nature of the interfaces at the two ends the answer is not obvious but the large body of experimental and theoretical work on point contacts in semiconductors since 1988 has established quite clearly that the +k states are occupied primarily by electrons coming from the left contact while the −k states are occupied primarily by electrons coming from the right contact. Consequently the occupation factors for the +k and −k states are given approximately by the Fermi functions for the left and right contacts respectively:

$$f_{+\mathrm{k}} = f_0(\varepsilon_k - \mu_1), \qquad f_{-\mathrm{k}} = f_0(\varepsilon_k - \mu_2). \tag{1.10}$$

Noting that the probability current carried by these plane wave eigenstates (eqn (1.8)) is given by

$$J_{+\mathrm{k}}(x) = (\hbar k/mL), \qquad J_{-\mathrm{k}}(x) = (-\hbar k/mL) \tag{1.11}$$

we obtain the net current as

$$\begin{aligned} I &= 2\ (\text{for spin}) * (-q) \sum_{k>0} J_{+\mathrm{k}} f_0(\varepsilon_k - \mu_1) + J_{-\mathrm{k}} f_0(\varepsilon_k - \mu_2) \\ &= (-2q) \sum_k \frac{\hbar k}{mL} [\, f_0(\varepsilon_k - \mu_1) - f_0(\varepsilon_k - \mu_2)]. \end{aligned}$$

Converting the summation into an integral using periodic boundary conditions with the usual prescription $\sum_k \to \int dkL/2\pi$ [9],

$$I = \frac{-2q}{h} \int_0^{+\infty} d\varepsilon_k [\, f_0(\varepsilon_k - \mu_1) - f_0(\varepsilon_k - \mu_2)] = \frac{-2q}{h} [\mu_1 - \mu_2]$$

from which eqn (1.7) for the conductance follows readily.

Our purpose in outlining this textbook derivation of the quantized conductance of ballistic conductors is to illustrate an important conceptual point [10]. The form of the eigenstates given by eqn (1.8) is not unique. The '+k' and '−k' states have the same energy so that any linear combination of the two is also an eigenstate. We could just as well have written the eigenstates as

$$\Psi_{c,k}(x) = \sqrt{2/L} \cos kx \qquad \text{and} \qquad \Psi_{s,k}(x) = \sqrt{2/L} \sin kx. \tag{1.12}$$

However, no matter how these states are occupied individually, the net current would be zero, since both the sine and cosine states are equal superpositions of '+k' and '−k' states and carry zero-current individually ($J_{c,k} = J_{s,k} = 0$):

$$1 = 2\ (\text{for spin}) \sum_{k>0} J_{c,k} f_{c,k} + J_{s,k} f_{s,k} = 0. \qquad \text{which is incorrect}$$

Where is the fallacy in this reasoning? The answer is that if we choose to use the cosine and sine states, we cannot calculate the current from occupation probabilities since the current operator is not diagonal in this representation. To see what this means, we first note that in the plane wave representation the current operator J_{op} is given by

$$[J_{op}]_{pw} = \begin{bmatrix} +\mathbf{k} & -\mathbf{k} \\ \hbar k/mL & 0 \\ 0 & -\hbar k/mL \end{bmatrix}. \tag{1.13a}$$

Next we transform the current operator into the cosine and sine representation through a unitary transformation using the transformation matrix [V] whose columns represent the old basis ('pw' : +k, −k) in terms of the new basis ('cs' : c, s):

$$[V] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ +i & -i \end{bmatrix}.$$

We obtain

$$[J_{op}]_{cs} = [V][J_{op}]_{pw}[V]^{+} = \begin{bmatrix} 0 & -i\hbar k/mL \\ +i\hbar k/mL & 0 \end{bmatrix}. \tag{1.13b}$$

The diagonal elements in this representation are indeed zero indicating that neither state carries any current by itself. But if they are occupied in a correlated manner as reflected in the off-diagonal elements of the density matrix, then there could be a net current. In the plane wave representation, the density matrix is diagonal with the diagonal elements given by the occupation probabilities f_α:

$$[\rho]_{pw} = \begin{bmatrix} f_{+\mathrm{k}} & 0 \\ 0 & f_{-\mathrm{k}} \end{bmatrix} \tag{1.14a}$$

so that in the cosine–sine representation

$$\begin{aligned} [\rho]_{cs} &= [V][\rho]_{pw}[V]^{+} \\ &= \frac{1}{2} \begin{bmatrix} f_{+\mathrm{k}} + f_{-\mathrm{k}} & -i(f_{+\mathrm{k}} - f_{-\mathrm{k}}) \\ i(f_{+\mathrm{k}} - f_k) & f_{+\mathrm{k}} + f_{-\mathrm{k}} \end{bmatrix}. \end{aligned} \tag{1.14b}$$

The current is given by $I = 2$ (for spin) $\times (-q) \times$ Trace (ρJ_{op}) and we get the same result in either representation ('pw' or 'cs'). This is only to be expected since the trace is invariant under a unitary transformation and thus remains the same in any representation. However, the point to note is that the current in the 'cs' representation arises from the *off-diagonal* elements of the density matrix and the current operator, rather than the diagonal elements. These elements do not have an intuitive physical meaning, unlike the diagonal elements which are simply the occupation factors f_α. As long as the current is carried by the diagonal terms we can use a semiclassical Boltzmann-like picture. However, if the 'action' is in the off-diagonal elements then we need a more general quantum framework. For nanoscale devices, it is important to have the flexibility to use arbitrary representations since one may not know *a priori* which representation will diagonalize the density matrix. The key problem then is to *find the density matrix ρ in some suitable representation.*

One last point before we proceed. We have mentioned above that the electron density can be calculated from the density matrix using eqn (1.6). We could regard eqn (1.6) as a special case of a unitary transformation into a real space representation:

$$[\rho]_{\text{realspace}} = [V][\rho][V]^{+}. \tag{1.15}$$

The transformation matrix $[V]$ is obtained from the amplitudes of the wavefunctions Ψ_α at points 'r' in real space:

$$[V]_{r,\alpha} = \Psi_\alpha(r)\sqrt{\Omega} \tag{1.16}$$

where Ω is the volume of an individual cell. The factor of $\sqrt{\Omega}$ comes from the process of discretization of the real space coordinate (as we do in the method of finite differences) which is conceptually convenient since it makes the transformation matrix finite-sized. In a discrete representation the normalized wavefunction in eqn (1.8) would be written as

$$\Psi_{+\mathrm{k}}(x) = e^{+ikx}/\sqrt{N} \qquad \text{instead of} \qquad \Psi_{+\mathrm{k}}(x) = e^{+ikx}/\sqrt{L}$$

N being the number of points in a discrete lattice of length L. A factor of $\sqrt{L/N} = \sqrt{a}$ (a: length of the

unit cell) is needed to convert from the continuous to the discrete representation. Substituting eqn (1.16) into eqn (1.15),

$$\rho(r, r') = \Omega \sum_{\alpha,\beta} \Psi_\alpha(r)\Psi^*_\beta(r')\rho_{\alpha\beta} \rightarrow \Omega n(r) = [\rho(r, r')]_{r'=r}. \tag{1.17}$$

The electron density $n(r)$ (see eqn (1.6)) is just the diagonal element ($r' = r$) of the density matrix in real space (divided by the volume of an individual cell). In other words, it is not only that we can calculate the electron density from the density matrix: a more powerful viewpoint is that the density matrix ***is*** the electron density (within a constant factor Ω). Like all quantum mechanical concepts it can be expressed in different representations. In some representation (like the plane wave representation for the ballistic conductor) the density matrix is diagonal; in other representations it is not diagonal. But in any representation the diagonal element tells us the number of electrons occupying a particular basis state in that representation. If we use the real space representation, then the diagonal elements tell us the number of electrons at different points in real space, which is the electron density $n(r)$ times Ω.

Outline: In quantum transport theory, the density matrix is the central quantity from which all quantities of interest can be obtained. For example, the electron density $n(r)$ is obtained from the diagonal elements in the real space representation, while the current is obtained from $I = (-q)$ Trace (ρJ_{op}). The problem then is to find the density matrix in a chosen representation. For this it is not enough just to know the details of the device through $(H + U)$; we also need to know how the device is coupled to the two contacts and the scattering processes that are effective within the device. This information is contained in the self-energy functions Σ_1, Σ_2 and Σ_S (Fig. 2). Given all of this information $\left(H, U, \Sigma_1, \Sigma_2, \Sigma_S, \mu_1 \text{ and } \mu_2\right)$, the NEGF formalism provides clear well-defined relations that can be used to calculate the density matrix from which the electron density and current can be obtained. We will not derive any of the equations, since rigorous derivations based on the second quantized formalism are available in the literature [11]. Instead we will try (1) to motivate the basic equations by showing that they make perfect sense if we use a representation in which the relevant quantities are diagonal and (2) to illustrate these equations with simple MATLAB-based examples that interested readers can conveniently duplicate on a PC. An important ingredient in these calculations is that the transport equation is solved self-consistently with the Poisson equation. The I–V characteristics of nanoscale structures is determined by an interesting interplay of transport physics with electrostatics and it is important to do justice to both aspects if we are to derive real insights.

The numerical examples we present are all centered around a short n^{++}–n^{+}–n^{++} resistor whose physics is easily understood. Our primary purpose is to show that the NEGF formalism leads to physically sensible results for this simple device. However, the basic formulation is quite general and can even be applied to something as different as a molecular wire or a nanotube [12]. For any device the first step is to choose a suitable representation which can be used to write down the quantities H, U, Σ_1, Σ_2, Σ_S (see Fig. 2) in the form of matrices. These are the matrices that contain all the physics of the problem at hand. Given these matrices, the procedure for calculating the density matrix (and hence the electron density and current) is the same for every device, be it a molecule or a nanotube or an n^{++}–n^{+}–n^{++} resistor.

We will start by describing the specific choice of representation that we will be using for our examples (Section 2). We will then discuss the procedure for calculating the equilibrium potential profile and electron density, first in terms of wavefunctions and then in terms of Green's functions in order to illustrate their relationship (Section 3). Next we will discuss devices driven out-of-equilibrium by an applied bias, neglecting any scattering processes inside the device ($\Sigma_s = 0$, Section 4). Finally we will discuss approaches that can be used to incorporate scattering processes into the model ($\Sigma_S \neq 0$, Section 5).

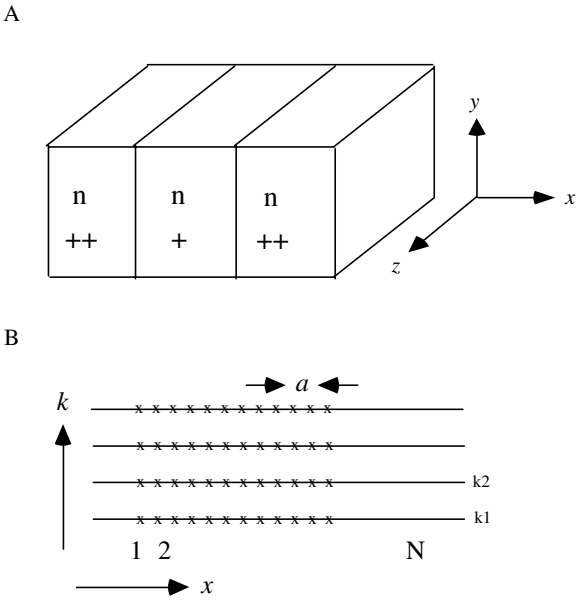

Fig. 4. A, 1D device with a large (effectively infinite) cross-section; B, we use the eigenstate representation for the transverse dimensions (y–z) but a discrete real space lattice for the longitudinal direction.

2. Choice of representation

In formulating a theory of quantum transport we have a choice of what representation to use and the optimum choice depends on the problem at hand. A representation based on eigenstates is often convenient for analytical calculations since the Hamiltonian is diagonal. On the other hand, a real space representation is intuitively more appealing since most of us 'live' in real space. In dealing with 1D devices (Fig. 4A), we find it convenient to use the eigenstate representation for the transverse dimensions (y–z) but a discrete real space lattice for the longitudinal direction. We can separate the overall Hamiltonian H into a longitudinal part H_L and a transverse part H_T ($H + U = H_L + H_T$):

$$H_L \equiv E_c - \frac{\hbar^2}{2m}\frac{d^2}{dx^2} + U(x) \tag{2.1}$$

$$H_T \equiv -\frac{\hbar^2}{2m}\left(\frac{d^2}{dy^2} + \frac{d^2}{dz^2}\right) + U_t(y, z). \tag{2.2}$$

For devices with a large (effectively infinite) cross-section, it is common to ignore the transverse confining potential $U_t(y, z)$ and use periodic boundary conditions in that direction since the real boundary conditions are believed to have minimal effect on the observed properties. The transverse eigenstates are then given by plane waves (S: transverse cross-sectional area)

$$\chi_{\boldsymbol{k}}(\boldsymbol{\rho}) = \exp(i\,\boldsymbol{k}\cdot\boldsymbol{\rho})/\sqrt{S} \tag{2.3}$$

$$H_T\,\chi_{\boldsymbol{k}} = \varepsilon_{\boldsymbol{k}}\chi_{\boldsymbol{k}} \qquad \text{with} \qquad \varepsilon_{\boldsymbol{k}} = \hbar^2 k^2/2m, \tag{2.4}$$

where $\boldsymbol{k}$ and $\boldsymbol{\rho}$ are both 2D vectors in the y–z plane.

For the longitudinal Hamiltonian we use a discrete lattice in real space. To find the matrix representation for H_L the simplest procedure is to use a finite difference approximation for the second derivative in eqn (2.1):

$$[H_L\phi]_n = -t\phi_{n-1} + (E_c + 2t + U_n)\Phi_n - t\phi_{n+1},$$

where $t \equiv \hbar^2/2ma^2$ and $U_n \equiv U(x_n)$. This means that the matrix representing 'H_L' appears as follows:

$$\mathbf{H_L} = \begin{array}{c|cccccc} & |1\rangle & |2\rangle & \dots & |N-1\rangle & |N\rangle \\ |1\rangle & E_c+2t+U_1 & -t & & 0 & 0 \\ |2\rangle & -t & E_c+2t+U_2 & & 0 & 0 \\ & \vdots & \vdots & & \vdots & \vdots \\ |N-1\rangle & 0 & 0 & & E_c+2t+U_{N-1} & -t \\ |N\rangle & 0 & 0 & & -t & E_c+2t+U_N. \end{array} \tag{2.5}$$

It can be shown that if the lattice spacing 'a' is chosen to be small enough that 't' is greater than the energy range of interest, then the discrete lattice representation (often called the tight-binding model) yields fairly accurate results. For an infinitely long uniform structure ($U_n = 0$) it yields a dispersion relation

$$E = E_c + 2t(1 - \cos ka) \tag{2.6}$$

which reduces to a parabolic band $E = \hbar^2k^2/2m$ for small 'ka'.

The overall basis functions can be labeled as $(\boldsymbol{k}, n)$ as shown in Fig. 4B. The elements of the matrix representing $(H + U)$ can be written as

$$[H_L + H_T]_{n,\boldsymbol{k};n',\boldsymbol{k}'} = ([H_L]_{n,n'} + \varepsilon_{\boldsymbol{k}})\delta_{\boldsymbol{k},\boldsymbol{k}'}. \tag{2.7}$$

The point to note is that since the $\boldsymbol{k}$ are eigenstates, there is no off-diagonal matrix element connecting two different 'transverse modes' $\boldsymbol{k}$ and $\boldsymbol{k}'$. As long as we neglect elastic or inelastic scattering processes that couple different transverse modes, we can think of the transverse modes $\boldsymbol{k}$ as separate 1D devices connected in parallel. Each transverse mode $\boldsymbol{k}$ has an extra transverse energy $\varepsilon_{\boldsymbol{k}} = \hbar^2k^2/2m$ that should be added to the longitudinal energy whenever the total energy is required (for example, in the argument of the Fermi function).

Note that the details of the transverse modes could be very different for other devices. For example, a MOS device is essentially a 2D charge sheet so that we have plane wave eigenstates only in one transverse direction; the other transverse direction usually has a small number of discrete subbands. With molecular conductors it is common to treat each molecule as independent, so that there are no transverse modes to worry about. However, in this paper we will assume a uniform 2D conductor as shown in Fig. 4. For all our examples we will use the following parameters: $m = 0.25m_0$, $m_0 = 9.1 \times 10^{-31}$ Kg, $\varepsilon = 10\varepsilon_0$, $\varepsilon_0 = 8.85 \times 10^{-12}$F m^{-1}, $a = 0.3$ nm, $N_D = 10^{20}$ cm^{-3} in the n^{++} regions, each of which is 4.5 nm long and $N_D = 5 \times 10^{19}$ cm^{-3} in the n^{+} region, which is 21 nm long. We have chosen very high doping densities deliberately so that the screening length is short compared with the length of the device. The device has 100 points along the length of the device so that the size of the matrix $[H_L]$ is 100×100. This is a fairly comfortable size for running on a PC and the results presented here (Figs 5, 8 and 13) were all obtained on a laptop computer. State-of-the-art supercomputers can handle much larger matrices and hence much larger devices.

3. Equilibrium

Once we have chosen a suitable representation, we are ready to calculate equilibrium band diagrams for 1D devices. The equilibrium problem can be done in two ways, one that uses the concept of Green's functions and one that does not, and it is instructive to compare the two. That is what we will do in this section using the n^{++}–n^{+}–n^{++} structure shown in Fig. 4A as an example. This discussion will also help introduce the self-energy functions Σ_1 and Σ_2 which describe the connection of the device to the contacts (see Fig. 2).

As shown in Fig. 1B, equilibrium problems can be handled by solving the Poisson equation self-consistently with the law of equilibrium statistical mechanics which requires all the eigenstates of the device (that is, $H + U$) to be filled up according to the Fermi function. This means that the equilibrium density

matrix for a particular transverse mode '$\boldsymbol{k}$' can be written as (we are using square brackets [..] to denote matrices)

$$[\rho_{\boldsymbol{k}}]_{es} = \begin{bmatrix} f_0(\varepsilon_1 + \varepsilon_{\boldsymbol{k}} - \mu) & 0 & 0 & \cdots \\ 0 & f_0(\varepsilon_2 + \varepsilon_{\boldsymbol{k}} - \mu) & 0 & \cdots \\ 0 & 0 & f_0(\varepsilon_3 + \varepsilon_{\boldsymbol{k}} - \mu) & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{3.1}$$

where the subscript '*es*' indicates that the density matrix is expressed in the eigenstate representation: ε_1, ε_2 etc are the eigenenergies of the longitudinal Hamiltonian H_L, while the $\varepsilon_{\boldsymbol{k}}$ are the eigenenergies of the transverse Hamiltonian, H_T (see eqns (2.1), (2.2)). Since all the transverse modes '$\boldsymbol{k}$' are like independent devices in parallel, we need the density matrix summed (or 'traced') over all '$\boldsymbol{k}$':

$$[\rho]_{es} = \sum_{\boldsymbol{k}} [\rho_{\boldsymbol{k}}]_{es} = \begin{bmatrix} F_0(\varepsilon_1 - \mu) & 0 & 0 & \cdots \\ 0 & F_0(\varepsilon_2 - \mu) & 0 & \cdots \\ 0 & 0 & F_0(\varepsilon_3 - \mu) & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{3.2}$$

where

$$F_0(E - \mu) = \sum_{\boldsymbol{k}} f_0(E + \varepsilon_{\boldsymbol{k}} - \mu) = S\frac{mk_{\mathrm{B}}T}{\pi\hbar^2} \ln\left(1 + \exp\left(\frac{\mu - E}{k_{\mathrm{B}}T}\right)\right). \tag{3.3}$$

Note that we have included the sum over spins as well when evaluating F_0 in eqn (3.3). The k-summed density matrix $[\rho]$ looks just like $[\rho_{\boldsymbol{k}}]$ except that the Fermi function f_0 is replaced by the logarithmic function F_0, and this simple replacement takes care of the transverse modes in the y–z plane. We can otherwise proceed with our calculation in the x-direction as if it were a purely 1D problem.

To obtain the density matrix in real space (whose diagonal elements give us the electron density $n(r)$, see eqn (1.17)), we have to perform a unitary transformation:

$$[\rho] = [V][\rho]_{es}[V]^{+} \tag{3.4}$$

where $[V]$ is a matrix whose columns denote the eigenvectors of H_L at each of the points on the discrete lattice. Once we have set up the matrix representing H_L, following the prescription in eqn (2.5) it takes just three commands in MATLAB to obtain the density matrix $[\rho]$ in real space:

$$\begin{aligned} &[V, D] = eig(HL); \\ &rho = \log(1 + \exp((mu - D)./kT)); \\ &rho = V * (rho) * V'; \end{aligned}$$

The first command calculates a diagonal matrix $[D]$ whose diagonal elements are the eigenvalues of H_L and a matrix $[V]$ whose columns are the corresponding eigenvectors. The second and third commands implement eqns (3.2) and (3.4) respectively. Actually we could achieve the same result with just one command:

$$rho = \mathrm{log}m(1 + \exp((mu - HL)/kT));$$

by noting that the density matrix can be written as (I: identity matrix of the same size as $[H_L]$)

$$\begin{aligned} [\rho_{\boldsymbol{k}}] &= f_0([H_L + (\varepsilon_{\boldsymbol{k}} - \mu)I]) \\ [\rho] &= \sum_{\boldsymbol{k}} \rho_{\boldsymbol{k}} = F_0([H_L - \mu I]). \end{aligned} \tag{3.5}$$

Equation (3.5) expresses the density matrix as a function of the Hamiltonian matrix and is really equivalent to eqns (3.2) and (3.4), since the rule for evaluating a function of a matrix is to write down the function in a representation that diagonalizes the matrix (eqn (3.2)) and then transform it back to the original representation (eqn (3.4)). We probably do not save any time by using one command instead of three, but the real value

of eqn (3.5) is conceptual: it states that the *equilibrium density matrix is simply the Fermi function of the Hamiltonian matrix* in any representation. This is a powerful concept that we will make further use of.

3.1. Periodic boundary conditions

We solve eqn (3.5) self-consistently with the Poisson equation

$$\frac{d^2U}{dx^2} = \frac{q^2}{\varepsilon}[N_D - n] \tag{3.6}$$

using the standard Newton–Raphson technique (see for example, Appendix C, Ref. [3]) to obtain the equilibrium potential profile $U(x)$ and electron density as shown in Fig. 5. The point to note is that if we use the matrix $[H_L]$ given in eqn (2.5) we obtain the dashed profile for the electron density which goes to zero at the ends of the device. This is because of the particular boundary condition that is implied by our use of $[H_L]$. When we use the finite difference method to write the Schrödinger equation on a discrete lattice we obtain

$$E\Psi_1 = -t\Psi_0 + (E_c + 2t + U_1)\Psi_1 - t\Psi_2 \tag{3.7a}$$

at the left end of the lattice (Fig. 4B). The problem is that we want to get rid of Ψ_0, in order to truncate $[H_L]$ to a finite size. We have the same problem at the right end

$$E\Psi_N = -t\Psi_{N-1} + (E_c + 2t + U_N)\Psi_N - t\Psi_{N+1} \tag{3.7b}$$

where we would like to get rid of Ψ_{N+1}. If we simply truncate the matrix, we are in effect setting $\Psi_0 = \Psi_{N+1} = 0$ which makes the calculated electron density go to zero at the ends. This would be an appropriate boundary condition if we had an infinite potential wall at the ends. However, what we actually have is an open boundary and this is better described by periodic boundary conditions which effectively wrap the right end around and connect it to the left end by setting $H_L(1, N) = H_L(N, 1) = -t$. The electron density then approaches the constant bulk value near the ends as we would expect. However, it is important to note that we are getting rid of end effects by artificially wrapping the device into a ring. We are not really doing justice to the open boundary that we have in the real device. The self-energy method that we will describe later in this section allows us to do that. But before we can describe this method, we need to discuss the Green's function approach for calculating the density matrix.

3.2. Green's function

Let us start from eqn (3.5) and rewrite it in the form

$$\begin{aligned}[\rho_{\mathbf{k}}] &= \int_{-\infty}^{+\infty} dE\, f_0(E + \varepsilon_{\mathbf{k}} - \mu) \quad \delta([EI - H_L]) \\ [\rho] &= \int_{-\infty}^{+\infty} dE\, F_0(E - \mu) \quad \delta([EI - H_L]).\end{aligned} \tag{3.8}$$

Using the standard expression for the delta function (0^+: positive infinitesimal)

$$2\pi\delta(x) = \mathrm{Lim}_{\varepsilon\to 0^+}\left(\frac{2\varepsilon}{x^2 + \varepsilon^2}\right) = \frac{i}{x + i0^+} - \frac{i}{x - i0^+}$$

we can write

$$\delta(EI - H_L) = \frac{i}{2\pi}([(E + i0^+)I - H_L]^{-1} - [(E - i0^+)I - H_L]^{-1}). \tag{3.9}$$

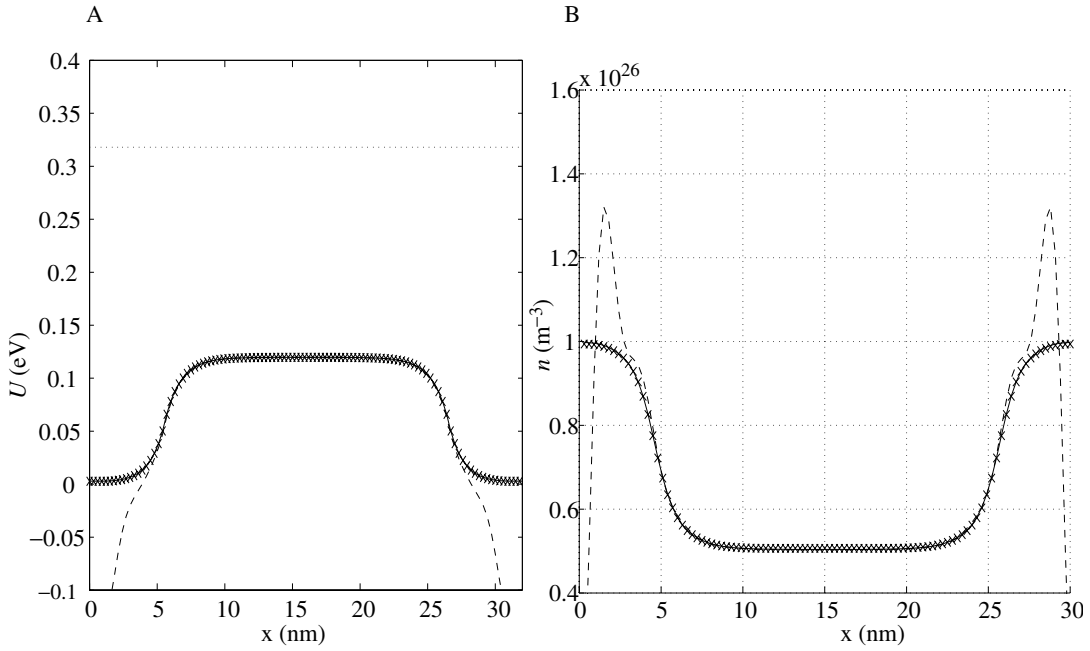

Fig. 5. Self-energy method (crosses); periodic boundary conditions (solid curves); infinite wall boundary conditions (dashed curves). A, Equilibrium potential profile and; B, electron density in a n^+–n–n^+ structure calculated by solving eqns (3.4) and (3.5) self-consistently. The parameters used are listed at the end of Section 2. The dotted line in A indicates the equilibrium Fermi level.

Equations (3.8) and (3.9) can be rewritten in the form

$$[\rho_{\boldsymbol{k}}] = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dE\, f_0(E+\varepsilon_{\boldsymbol{k}}-\mu) \quad [A(E)]$$
$$[\rho] = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dE\, F_0(E-\mu) \quad [A(E)] \tag{3.10}$$

where $[A(E)]$ is known as the *spectral function*

$$[A(E)] = i([G(E)] - [G(E)]^+) \tag{3.11}$$

$[G(E)]$ being the retarded *Green's function* defined as

$$[G(E)] = [(E+i0^+)I - H_L]^{-1}. \tag{3.12}$$

One can see from eqn (3.10) that the spectral function $[A(E)]/2\pi$ can be interpreted as the available density of states which are filled up according to the Fermi function to obtain the electron density. Indeed the diagonal elements of $[A(E)]/2\pi$ in the real space representation give us the local density of states at different points in space (a quantity that can be measured with scanning probe microscopy).

Equation (3.10) represents the Green's function version of eqn (3.5). One might wonder what we have gained by introducing an unnecessary integration over the energy coordinate, E. What makes this extra complication worthwhile is the convenience it affords in the treatment of open systems. Indeed if our interest was limited to closed systems there would be little reason to use Green's functions. But for open systems the Green's function method allows us to focus on the device of interest and replace the effect of all external contacts and baths with self-energy functions $\Sigma_{1,2,S}$ (see Fig. 2A) which are matrices of the same size as the

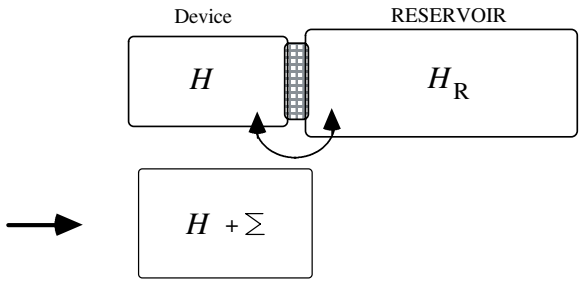

Fig. 6. The interaction of a device with a reservoir can be represented by a self-energy matrix $\sum$.

device Hamiltonian, even though the contacts themselves are much larger entities. This is one of the seminal concepts of many-body physics that we will now discuss.

3.3. Self-energy

The concept of self-energy is used in many-body physics to describe electron–electron and electron–phonon interactions. In the present context, however, we are using this concept to describe something much simpler, namely, the effect of a semi-infinite contact. But the principle is the same. In general, we have a 'device' connected to a large reservoir and the overall Hamiltonian matrix has the form (see Fig. 6)

$$\begin{bmatrix} H & \tau \\ \tau^+ & H_R \end{bmatrix}$$

where the dimension of H_R is huge compared to that of H. The overall Green function has the form

$$\begin{bmatrix} G & G_{DR} \\ G_{RD} & G_R \end{bmatrix} = \begin{bmatrix} (E+i0^+)I - H & -\tau \\ -\tau^+ & (E+i0^+)I - H_R \end{bmatrix}^{-1}.$$

We are only interested in G (and not in G_R or G_{DR} or G_{RD}), because we only care about the details inside the device and not inside the reservoir. It is straightforward to show that (see p. 146, Ref. [6])

$$G = \Big[(E+i0^+)I - H - \Sigma\Big]^{-1} \approx \Big[EI - H - \Sigma\Big]^{-1} \tag{3.13}$$

where

$$\Sigma = \tau g_R \tau^+ \qquad \text{and} \qquad g_R = [(E+i0^+)I - H_R]^{-1}. \tag{3.14}$$

This shows that the effect of the coupling to the reservoir can be accounted for by adding a self-energy matrix Σ to the Hamiltonian H (Fig. 6). This is a very general concept that allows us to eliminate the huge reservoir and work solely within the device subspace whose dimensions are much smaller. Note that Σ is not necessarily an infinitesimal quantity (unlike 0^+); it can be finite with a value defined by the coupling to the reservoir. We will discuss the physical meaning of Σ further at the end of this section.

We could use eqn (3.14) in general to calculate the self-energy for arbitrary reservoirs and coupling matrices τ. It may seem that we have not gained much since we need to invert a huge matrix to obtain g_R which we need to evaluate the self-energy from eqn (3.14)

$$\Sigma(m,n) = \sum_{\mu,\nu} \tau(m,\mu) g_R(\mu,\nu) \tau^+(\nu,n).$$

The indices m, n refer to points within the device while μ, ν refer to points inside the reservoir. However,

the coupling matrix τ couples the points within the device to a small number of points on the surface of the reservoir, so that we only need $g_R(\mu, \nu)$ for points (μ, ν) that are on the surface. This surface Green's function can often be calculated analytically assuming a given model for the reservoir.

For the simple 1D problem at hand, the self-energy can be obtained from fairly elementary arguments without worrying about surface Green's functions. The self-energy matrix $\Sigma_1(E)$ that accounts for the semi-infinite lead on the left is given by ($t \equiv \hbar^2/2ma^2$ as defined earlier before eqn (2.5))

$$\Sigma_1(E) = \begin{array}{c|cccccc} & |1\rangle & |2\rangle & \cdots & |N-1\rangle & |N\rangle \\ |1\rangle & -t\exp(ik_1a) & 0 & & 0 & 0 \\ |2\rangle & 0 & 0 & & 0 & 0 \\ & \vdots & \vdots & & \vdots & \vdots \\ |N-1\rangle & 0 & 0 & & 0 & 0 \\ |N\rangle & 0 & 0 & & 0 & 0 \end{array} \tag{3.15a}$$

where $E = E_c + U_1 + 2t(1 - \cos k_1 a)$.

In other words all we need is to add a term $-t\exp(ik_1a)$ to $H_L(1, 1)$ and we have accounted for the semi-infinite lead exactly, as far as calculating the Green's function is concerned. We can derive this result using an elementary argument. We stated earlier (see eqn (3.7a)) that the basic question at the boundary is how to eliminate Ψ_0 from the equation

$$E\Psi_1 = -t\Psi_0 + (E_c + 2t + U_1)\Psi_1 - t\Psi_2. \qquad \text{(same as eqn (3.7a))}$$

With infinite wall boundary conditions we set Ψ_0 equal to zero while with periodic boundary conditions we set it equal to Ψ_N. In the self-energy method we assume that we only have outgoing (not incoming) waves at the ends. The fact that an actual device has incoming waves as well from the contacts is irrelevant when calculating G.G is the retarded Green's function representing the response of the system to an impulse excitation within the device: $\big[EI - H - \Sigma\big]G = I$, and hence the appropriate boundary condition for G is that we only have *outgoing* waves at the ends. This means that when calculating G we can write

$$\Psi_0 = \Psi_1 \exp[ik_1a]$$

so that eqn (3.7a) becomes

$$E\Psi_1 = -t\exp[ik_1a]\Psi_1 + (2t + U_1)\Psi_1 - t\Psi_2$$

showing that we can take care of the open boundary condition simply by adding a term $-t\exp[ik_1a]$ to point 1, as stated above. Similarly the self-energy matrix $\Sigma_2(E)$ that accounts for the semi-infinite lead on the right has only one non-zero term at point N which is given by

$$\Sigma_2(N, N; E) = -t\exp(ik_2a) \qquad \text{where} \qquad E = E_c + U_N + 2t(1 - \cos k_2 a). \tag{3.15b}$$

The Green's function is obtained from

$$G(E) = \Big[EI - H_L - \Sigma_1 - \Sigma_2\Big]^{-1} \tag{3.16}$$

where the self-energy functions $\Sigma_1(E)$ and $\Sigma_2(E)$ account for the open boundary conditions exactly. The spectral function $A(E)$ is then obtained from eqn (3.11) from which the electron density is obtained using eqn (3.10). As we can see from Fig. 5 the results agree quite well with those obtained directly using periodic boundary conditions. The self-energy method is computationally more intensive, since it requires an integration over energy and looking at Fig. 5 it is not clear that the extra effort is worthwhile. However, it should be noted that the periodic boundary conditions merely get rid of end effects through the artifact of wrapping the device into a ring while the self-energy method treats the open boundary condition exactly. An open system

has a continuous energy spectrum, while a ring has a discrete energy spectrum. The electron density is obtained by integrating over energy and is relatively unaffected by the discretization of the spectrum at least at room temperature. But the difference would be apparent, if we were to look at the density of states, that is, the spectral function. The full power of the self-energy method becomes apparent when we model a device under bias—a problem that cannot be handled with periodic boundary conditions.

3.4. Broadening

It might appear that the self-energy method is just another method for handling boundary effects. With infinite wall boundary conditions we set $\Psi_0 = \Psi_{N+1} = 0$; with periodic boundary conditions we set $\Psi_0 = \Psi_N$; in the self-energy method we set $\Psi_0 = \Psi_1 \exp[ik_1 a]$ and $\Psi_{N+1} = \Psi_N \exp[ik_2 a]$. However, there are two factors that distinguish Σ_1 and Σ_2 from ordinary Hamiltonians. Firstly, they are energy dependent. Secondly, they are not Hermitian. We can write

$$\begin{aligned} H_L + \Sigma_1 + \Sigma_2 &= \left(H_L + \frac{\Sigma_1 + \Sigma_1^+}{2} + \frac{\Sigma_2 + \Sigma_2^+}{2} \right) + \left(\frac{\Sigma_1 - \Sigma_1^+}{2} + \frac{\Sigma_2 - \Sigma_2^+}{2} \right) \\ &= \hat{H}_L - i\Gamma_1/2 - i\Gamma_2/2 \end{aligned}$$

where

$$\hat{H}_L \equiv H_L + \frac{\Sigma_1 + \Sigma_1^+}{2} + \frac{\Sigma_2 + \Sigma_2^+}{2}$$

and

$$\Gamma_1 \equiv i\left[\Sigma_1 - \Sigma_1^+ \right], \qquad \Gamma_2 \equiv i\left[\Sigma_2 - \Sigma_2^+ \right].$$

The point we want to make is that the self-energy terms have two effects. One is to change the Hamiltonian from H_L to $\hat{H}_L$ which changes the eigenstates and their energies. But more importantly, it introduces an imaginary part to the energy determined by the 'broadening' functions Γ_1 and Γ_2. The former represents a minor quantitative change; the latter represents a qualitative change with conceptual implications.

One way to understand the meaning of these functions is to imagine a representation which diagonalizes $\hat{H}_L$. This representation will not necessarily diagonalize Γ_1 and Γ_2—indeed interesting quantum interference effects often arise from the non-diagonal elements of Γ_1 and Γ_2. But if Γ_1 and Γ_2 are also simultaneously diagonalized then the eigenenergies of $\left(H_L + \Sigma_1 + \Sigma_2\right)$ will be given by

$$\varepsilon - i(\gamma_1 + \gamma_2)/2$$

where ε, γ_1 and γ_2 are the corresponding diagonal elements of $\hat{H}_L$, Γ_1 and Γ_2 respectively. This could be viewed as a broadening of the energy level from a delta function $\delta(E - \varepsilon)$ into a line of the form

$$\frac{\gamma_1 + \gamma_2}{(E - \varepsilon)^2 + ((\gamma_1 + \gamma_2)/2)^2}$$

which could have a non-Lorentzian shape since γ_1 and γ_2 are in general energy dependent.

The imaginary part of the energy implies that the wavefunction and the associated probability decays with time which can be written in the form (neglecting any energy dependence of γ_1 and γ_2)

$$\begin{aligned} \Psi &\sim \exp[-i\varepsilon t/\hbar] \exp[-\gamma_1 t/2\hbar] \exp[-\gamma_2 t/2\hbar] \\ |\Psi|^2 &\sim \exp[-\gamma_1 t/\hbar] \exp[-\gamma_2 t/\hbar]. \end{aligned} \tag{3.17}$$

An electron initially placed in that state will escape into the left and right leads with time constants $\hbar/\gamma_1$ and $\hbar/\gamma_2$ respectively. The quantities $\gamma_1/\hbar$ and $\gamma_2/\hbar$ thus represent the rates at which an electron initially in a

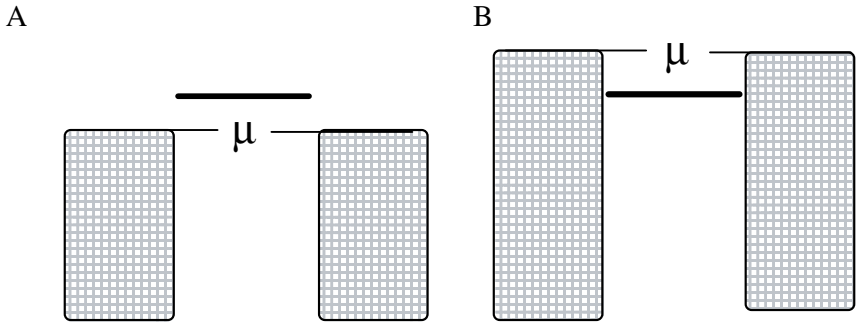

Fig. 7. A discrete level is coupled to a reservoir with a Fermi level μ. The broadening is the same regardless of whether the reservoir states are empty or filled. A, Level coupled to empty states; B, level coupled to filled states.

particular state will escape into the left and right states respectively. For example, we have seen that a 1D lead gives rise to a self-energy that is purely diagonal in real space representation (see eqn (3.15a))

$$\Sigma(1,1) = -t\exp(ika) \rightarrow \Gamma(1,1) = 2t\sin(ka) = \hbar v/a$$

which is quite reasonable since we expect the rate of escape from a lattice site of size 'a' to be v/a.

3.5. The exclusion principle

One final comment before we move on. In calculating the broadening of a level due to the connection to the reservoir, it might seem that the result should depend on whether the reservoir is occupied or not. Consider a discrete level coupled to a reservoir with a Fermi level μ (Fig. 7). One could argue that the broadening would be larger if the reservoir states corresponding to the discrete level are empty (case 'a') than if they are filled (case 'b'). After all, an electron placed on this level would be unable to escape in case 'b' since it would be blocked by the Pauli principle and hence the level should not be broadened. From this point of view, the broadening of a level should be given by

$$i\Sigma^{>} \equiv \Sigma^{\text{out}} = \Gamma(1 - f_0(\varepsilon - \mu)) \qquad \text{(electron escape rate).} \tag{3.18a}$$

However, this argument is not correct, since a hole placed on this level would escape into the reservoir in case 'b' but would be blocked in case 'a'. This means that the broadening would be

$$-i\Sigma^{<} \equiv \Sigma^{\text{in}} = \Gamma f_0(\varepsilon - \mu) \qquad \text{(hole escape rate or electron entry rate)} \tag{3.18b}$$

if we were describing the propagation of holes instead of electrons. But the correct point of view [13] is that electrons and holes are all described by the same self-energy and hence the same broadening which is given by the sum of the electron escape and entry rates:

$$\Gamma = \Sigma^{\text{out}} + \Sigma^{\text{in}} = i\left(\Sigma^{>} - \Sigma^{<}\right). \tag{3.19}$$

The broadening is thus the same irrespective of whether the reservoir is filled or empty.

4. Coherent transport

We have seen in the last section that the equilibrium density matrix is obtained by filling up the available density of states (or spectral function $[A]$) according to the Fermi function (see eqn (3.10)). The next problem

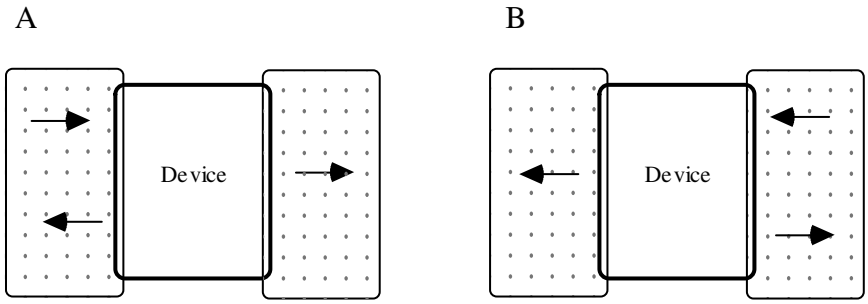

Fig. 8. The eigenstates of a composite contact–device–contact structure can be divided into two groups associated with incident waves from the A, left contact and; B, the right contact. If we neglect scattering processes under bias, then under bias the 'left' eigenstates in A remain in equilibrium with contact 1 (μ_1) and the 'right' eigenstates in B remain in equilibrium with contact 2 (μ_2).

is to find the density matrix if the device is connected to two contacts with different Fermi levels μ_1 and μ_2 (see Fig. 3), and hence different Fermi functions. The solution in general is quite involved: non-equilibrium statistical mechanics is a far more complex subject than equilibrium statistical mechanics. However, the answer is relatively simple, if we neglect scattering processes within the device (that is, if we assume transport to be coherent). This turns out to be a fairly accurate assumption for many ultrashort devices like resonant tunneling diodes. The eigenstates of the composite contact–device–contact structure can then be divided into two groups associated with waves incident from the left and right contacts respectively (see Fig. 8). When a bias is applied, these 'left' eigenstates and 'right' eigenstates remain in equilibrium with the left contact and the right contact respectively. The ballistic conductor is a relatively simple example of this principle where the left eigenstates are the '+k' states and the right eigenstaes are the '−k' states (see Fig. 3).

This simple observation (some might call it an ansatz) leads to an enormous simplification and is at the heart of the transmission formalism that is widely used in mesoscopic physics [6–8]. It allows us to treat a non-equilibrium problem using equilibrium statistical mechanics. At equilibrium, we fill up the full spectral function $[A]$ according to the Fermi function. Under bias, we fill up part of it (the left spectral function $[A_1]$) according to the Fermi function in the left contact and part of it (the right spectral function $[A_2]$) according to the Fermi function in the right contact. The density matrix is given by (cf. eqn (3.10))

$$\rho_{\boldsymbol{k}} = \int \frac{dE}{2\pi} [\, f_0(E + \varepsilon_{\boldsymbol{k}} - \mu_1) A_1 + \, f_0(E + \varepsilon_{\boldsymbol{k}} - \mu_2) A_2]$$

so that

$$\rho = \sum_{\boldsymbol{k}} \rho_{\boldsymbol{k}} = \int \frac{dE}{2\pi} [F_0(E - \mu_1) A_1 + F_0(E - \mu_2) A_2]. \tag{4.1}$$

The Green's function formalism provides a simple way to separate the total spectral function $[A]$ into a left spectral function $[A_1]$ and a right spectral function $[A_2]$:

$$A_1 = G\Gamma_1 G^+, \qquad A_2 = G\Gamma_2 G^+ \tag{4.2}$$

where

$$G = \left[EI - H_L - \Sigma_1 - \Sigma_2 \right]^{-1}, \tag{4.3}$$

$$\Gamma_{1,2} = i \left[\Sigma_{1,2} - \Sigma_{1,2}^+ \right]. \tag{4.4}$$

We can prove that the total spectral function is indeed equal to the sum of the left and right spectral functions:

$$A \equiv i[G - G^+] = A_1 + A_2 = G\Gamma_1 G^+ + G\Gamma_2 G^+ \tag{4.5}$$

by writing eqn (4.3) as

$$G^{-1} = EI - H_L - \Sigma_1 - \Sigma_2 \qquad \text{and} \qquad [G^+]^{-1} = EI - H_L - \Sigma_1^+ - \Sigma_2^+$$

so that

$$G^{-1} - [G^+]^{-1} = i\Gamma_1 + i\Gamma_2.$$

Premultiplying by G and postmultiplying by G^+, we can prove eqn (4.5).

In the last section we discussed how the equilibrium potential profile can be calculated by solving eqn (3.10) (or equivalently eqn (3.5)) self-consistently with the Poisson equation (see eqn (3.6)). Under non-equilibrium conditions we can solve eqn (4.1) self-consistently with the Poisson equation using much the same procedure. The self-consistent potential profile and electron density are shown in Figs 9A,C respectively, for a bias of 0.25 V. Note that the electron density hardly changes under bias as we would expect in such a conductive medium. The potential profile $U(x)$ adjusts in such a way that the resulting electron density is virtually the same before and after the bias is applied. This requires $U(x)$ at the left end to be pulled down relative to the equilibrium value, since the bias causes the right spectral function A_2, which is filled according to μ_2, to be partially emptied. For this reason, when solving the Poisson equation, it is inconvenient to fix U at the ends; a better approach is to impose zero-field conditions at the ends of the device and let U float to whatever value it chooses to [14].

Note that the potential U at the left of the device (see Fig. 9A) is actually pulled down by the applied bias. Deep inside the contact, both +k and −k states will be equally occupied and U must change back to the equilibrium value. This transition is not shown in the figure, but the point is that a significant part ($\sim$0.10 V) of the applied bias of 0.25 V is dropped inside the contact and not inside the device. One could associate this external drop with the ideal contact resistance that leads to a non-zero resistance for ballistic conductors [6–8]. This drop is often obscured in the presence of a large barrier, but is quite apparent in the present example because the barrier is only 100 meV. Such effects are likely to be important for ballistic devices, even with semiclassical models. It is interesting note that $U(x)$ under bias is relatively flat inside the middle n+ region, unlike what we are used to in MOS transistors. This unusual profile results from a combination of two factors: (1) lack of scattering which eliminates any 'voltage drop' inside the device, and (2) high electron density which screens out the end effects within a short distance making the flat potential profile obvious. As we will see in the next section, the potential profile looks more like an ordinary resistor when we introduce a little scattering into the model.

4.1. Current

Once we have the density matrix from eqn (4.1), we not only have the electron density from its diagonal elements (see eqn (1.17)) but also the current from the relation

$$I = (-q)\text{Trace}(\rho J_{op}) \tag{4.6}$$

where J_{op} is the current operator that we discussed in the introduction (see eqns (1.13a) and (1.13b)). Since we are interested in the current in the x-direction, the current operator is $-(i\hbar mL)\,\partial/\partial x$ and can be written as ($N = L/a =$ the number of points on the lattice)

$$[J_{op}] = (t/\hbar N)\begin{bmatrix} 0 & -i & 0 & \cdots \\ +i & 0 & -i & \cdots \\ 0 & +i & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{4.7}$$

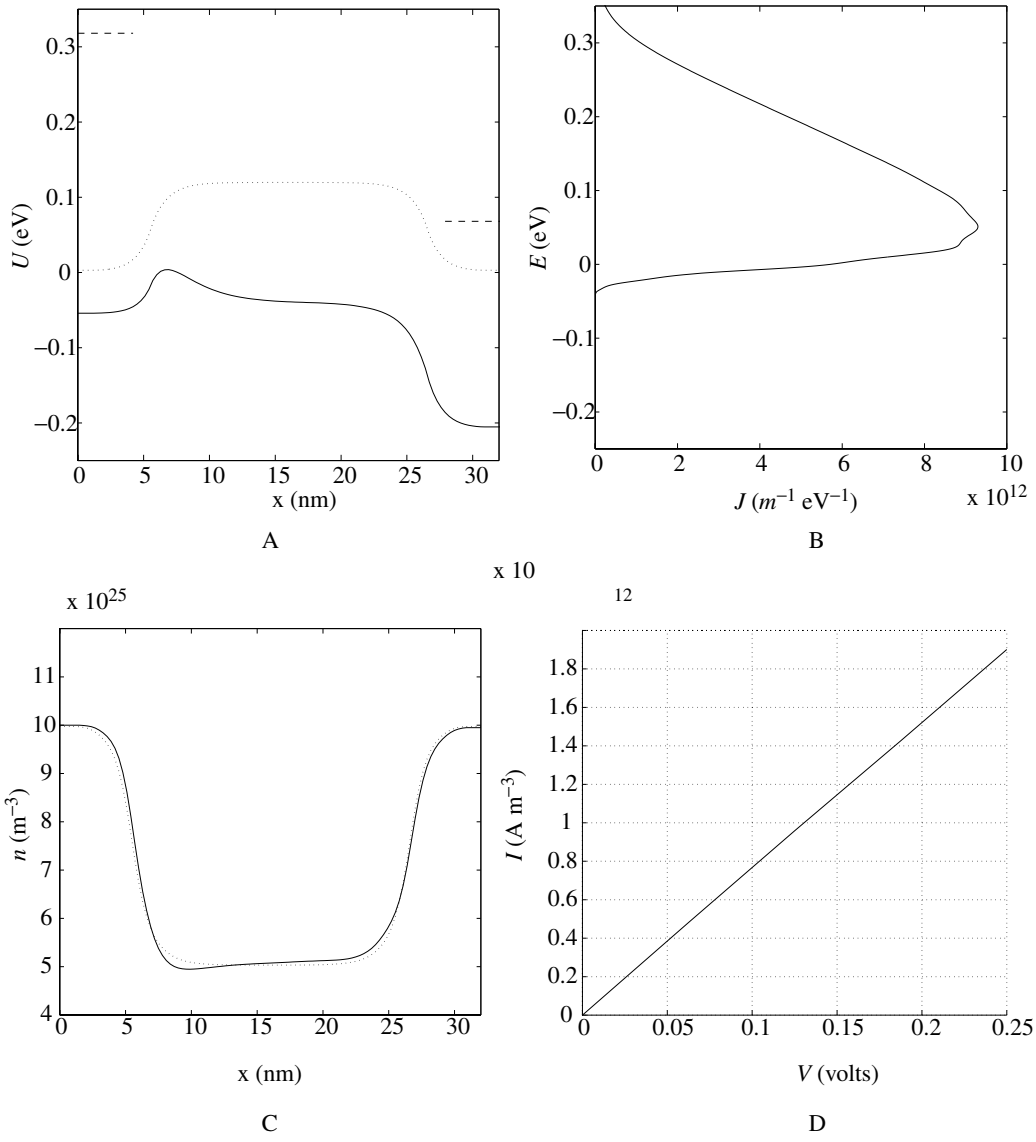

Fig. 9. A, Potential profile; B, energy spectrum of the current density ($J = \tilde{I}(E)/S$) and; C, electron density in a n^+–n–n^+ structure calculated under 0.25 V bias (solid) by solving eqns (4.6) and (3.5) self-consistently using the same parameters as in Fig. 5. Also shown for comparison are the equilibrium profiles in the absence of applied bias (dotted). Dashed curves in A indicate the quasi-Fermi levels in the contacts; D, The current versus voltage characteristic which is linear. The current level is unphysically high due to the high electron density and lack of scattering assumed in this model, as discussed in the text.

in the finite difference representation on a discrete lattice. Note that this operator is the same for all the transverse modes $\boldsymbol{k}$ in our parabolic band model, though in general it may not be so [15]. Note that in the real space representation, the diagonal elements of the Hermitian matrix $(\rho J_{op} + J_{op}\rho)$ are all equal since the dc current must be the same at all 'x'; taking the trace is thus the same as multiplying one of the diagonal elements by the total number of points N.

Figure 9D shows the I–V characteristics for this structure which as expected is quite linear. The current

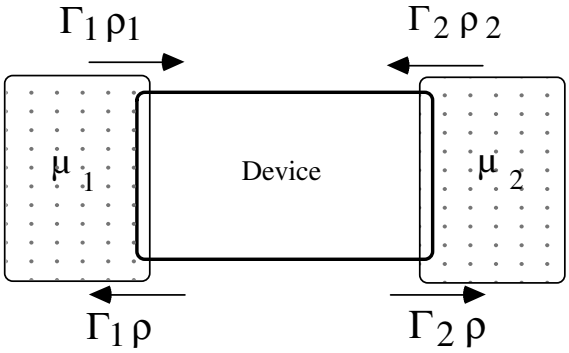

Fig. 10. The current can also be obtained by taking the difference between the influx and outflux at either interface.

level is unphysically high due to the high electron density and lack of scattering assumed in this model. We have chosen very high doping densities deliberately so that the screening length is short compared to the length of the device. It is straightforward to show (from the drift-diffusion equations) that the current in a resistor with $n = 5 \times 10^{19}$ cm^{-3} which is one mean free path long is $\sim 2 \times 10^8$A cm^{-2} for an applied voltage of 0.25 V. The quantum calculations presented here also yield similar current levels which is of course far too large for any real device. We could have chosen a lower doping density and obtained more reasonable current values, but the reduced screening would cause end effects to obscure the flat potential profile within the resistor, since it is only ~20 nm long.

It is instructive to look at the energy spectrum of the current (at a particular bias) calculated from the energy-resolved density matrix, $\tilde{\rho}(E)$

$$\tilde{I}(E) = (-q)\text{Trace}[\tilde{\rho}(E)J_{op}] \qquad \rightarrow I = \int dE\,\tilde{I}(E) \tag{4.8}$$

$$2\pi[\tilde{\rho}(E)] = F_1[A_1(E)] + F_2[A_2(E)] \qquad \rightarrow \rho = \int dE\,\tilde{\rho}(E) \tag{4.9}$$

where

$$F_1 \equiv F_0(E - \mu_1) \qquad \text{and} \qquad F_2 \equiv F_0(E - \mu_2). \tag{4.10}$$

Figure 9C shows the energy spectrum of the current, $\tilde{I}(E)$ from which it is apparent that the current flows above the barrier in the energy range between μ_1 and μ_2, as we would expect. There is a net current only in the energy range where the Fermi functions in the two contacts differ significantly. Other energies remain essentially in equilibrium: they contribute to the electron density, but not to the current.

4.2. An alternative current expression

An alternative expression for the current can be obtained from a rate equation point of view (see Fig. 10) by writing the outflux from the device into contact 1 as

$$I_{\text{out}} = (-q/\hbar) \int dE\ \text{trace}(\Gamma_1\tilde{\rho}) \tag{4.11}$$

which can be understood by noting that the density matrix ρ is like the electron density while $\Gamma_1/\hbar$ represents the rate at which electrons escape into the contact. The influx from the contact into the device can be written by equating it to the outflux we would have if the device were in equilibrium with that contact:

$$I_{\text{in}} = (-q/\hbar) \int dE\ \text{trace}(\Gamma_1\tilde{\rho}_1) \tag{4.12}$$

where

$$2\pi[\tilde{\rho}_1(\mathrm{E})] \equiv \mathrm{F}_1[\mathrm{A}_1(\mathrm{E})] + \mathrm{F}_1[A_2(\mathrm{E})] \tag{4.13}$$

represents the density matrix we would have if F_2 were equal to F_1. The net current is given by $I = I_{\mathrm{in}} - I_{\mathrm{out}}$. Making use of eqns (4.9), (4.11)–(4.13) we obtain

$$I = (-q/h) \int dE \ \mathrm{trace}(\Gamma_1 A_2)[F_1 - F_2]. \tag{4.14}$$

We could go through a similar argument regarding the influx and outflux at the other interface to obtain an equivalent expression for the current:

$$I = (-q/h) \int dE \ \mathrm{trace}(\Gamma_2 A_1)[F_1 - F_2]. \tag{4.15}$$

Equations (4.14), (4.15) provide alternative expressions either of which can be used to calculate the terminal currents without explicitly calculating the density matrix. However, the current operator approach (see eqn (4.8)) allows one to calculate the current flow pattern inside the device (which is trivial for 1D examples, but could be more interesting in higher dimensions or when dissipation is included) rather than just the terminal currents.

4.3. Relation to the transmission formalism

An interesting aspect of eqns (4.14) and (4.15) is that the expression for the current has exactly the same form that is used in the transmission formalism

$$I = (-q/h) \int dE T(E)(F_1 - F_2). \tag{4.16}$$

The function $T(E)$ is typically interpreted as the probability that an electron will transmit from the left to the right contact. Equation (4.16) is often used to calculate the current in tunneling and resonant tunneling devices. Comparing eqn (4.16) with eqns (4.14), (4.15) it is clear that

$$\begin{aligned} T(E) &= \mathrm{trace}(\Gamma_1 A_2) = \mathrm{trace}(\Gamma_2 A_1) \\ &= \mathrm{trace}(\Gamma_1 G \Gamma_2 G^+) = \mathrm{trace}(\Gamma_2 G \Gamma_1 G^+). \end{aligned} \tag{4.17}$$

The NEGF formalism, applied to a coherent device, can thus be viewed simply as a convenient method for evaluating the transmission probability. The basic physics is identical. The real power of the NEGF formalism lies in providing a clear prescription for including scattering processes, as we will discuss next.

5. Non-coherent transport

As we mentioned in the introduction, scattering processes enter the NEGF formalism through the self-energy function Σ_S, which we have ignored so far. For resonant tunneling devices, it is usually fairly accurate to neglect Σ_S, but this may not be adequate for MOSFETs, even 20 nm ones. What complicates this aspect of the problem is that Σ_S is dependent on the density matrix and has to be calculated self-consistently somewhat like the potential $U(x)$. However, $U(x)$ is related to the electron density through the laws of electrostatics while the relation of Σ_S to the density matrix $\tilde{\rho}$ can be considerably more complicated depending on the nature of the scattering mechanism and the level of approximation used. The NEGF formalism provides clear prescriptions for calculating Σ_S for every scattering mechanism that we can think of and thus can be used to investigate the effect of different scattering processes from first principles. Using simple isotropic models for scattering we have shown that the NEGF formalism indeed provides physically reasonable descriptions of energy dissipation in nanoscale devices [16] in good agreement with semiclassical models. However, we will

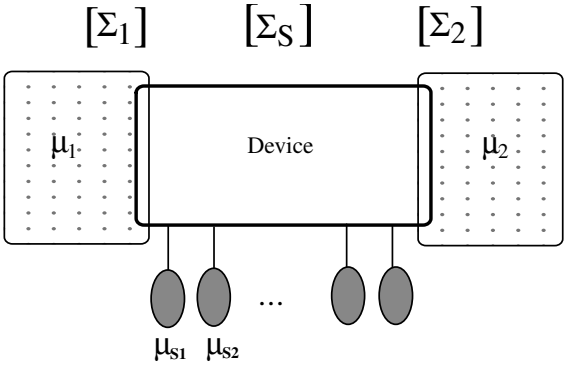

Fig. 11. Dissipative processes can be included with a phenomenological model that is equivalent to adding a separate contact to each lattice site and then adjusting its Fermi level so as to ensure current conservation throughout the device.

not go into any of these models in this paper, which is far too long already. Instead we will present results obtained from a 'toy' model that captures some of the important features of dissipative transport.

Suppose we view the scattering process as just another contact described by Σ_S, no different from the actual contacts described by Σ_1 and Σ_2. We can then simply extend eqns (4.1)–(4.4) to include a third contact

$$G = \left[EI - H - \Sigma_1 - \Sigma_2 - \Sigma_S \right]^{-1} \tag{5.1}$$

$$\Gamma_{1,2,S} = i\left[\Sigma_{1,2,S} - \Sigma^{+}_{1,2,S} \right] \tag{5.2}$$

$$A_1 = G\Gamma_1 G^{+}, \qquad A_2 = G\Gamma_2 G^{+}, \qquad A_S = G\Gamma_S G^{+} \tag{5.3}$$

$$2\pi[\tilde{\rho}(E)] = F_1[A_1] + F_2[A_2] + F_S[A_S]. \tag{5.4}$$

We could set

$$\Sigma_S = -i \begin{bmatrix} \eta_1 & 0 & \cdots & \cdots \\ 0 & \eta_2 & 0 & \cdots \\ 0 & 0 & \eta_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{5.5}$$

where the η are phenomenological parameters related to the scattering time τ by the relation $\tau = \hbar/2\eta$ (see eqn (3.16)). However, the problem is that unlike the real contacts, the 'scattering contact' does not have a well-defined Fermi level μ_S from which we can calculate F_S. If we use eqn (5.4) with F_S calculated from a single μ_S we would effectively be shorting together all the conceptual scattering contacts. A more physically correct model is to let each lattice 'n' site float to a different μ_{Sn} and define an inscattering function

$$\Sigma^{\text{in}}_S = 2 \begin{bmatrix} F_{S1}\eta_1 & 0 & \cdots & \cdots \\ 0 & F_{S2}\eta_2 & 0 & \cdots \\ 0 & 0 & F_{S3}\eta_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{5.6}$$

which is then used to calculate the density matrix from a modified version of eqn (5.4):

$$2\pi[\tilde{\rho}(E)] = F_1 A_1 + F_2 A_2 + G\Sigma^{\text{in}}_S G^{+}. \tag{5.7}$$

How do we know what μ_{Sn} to use for the scattering contacts? If we make a reasonable guess as shown in

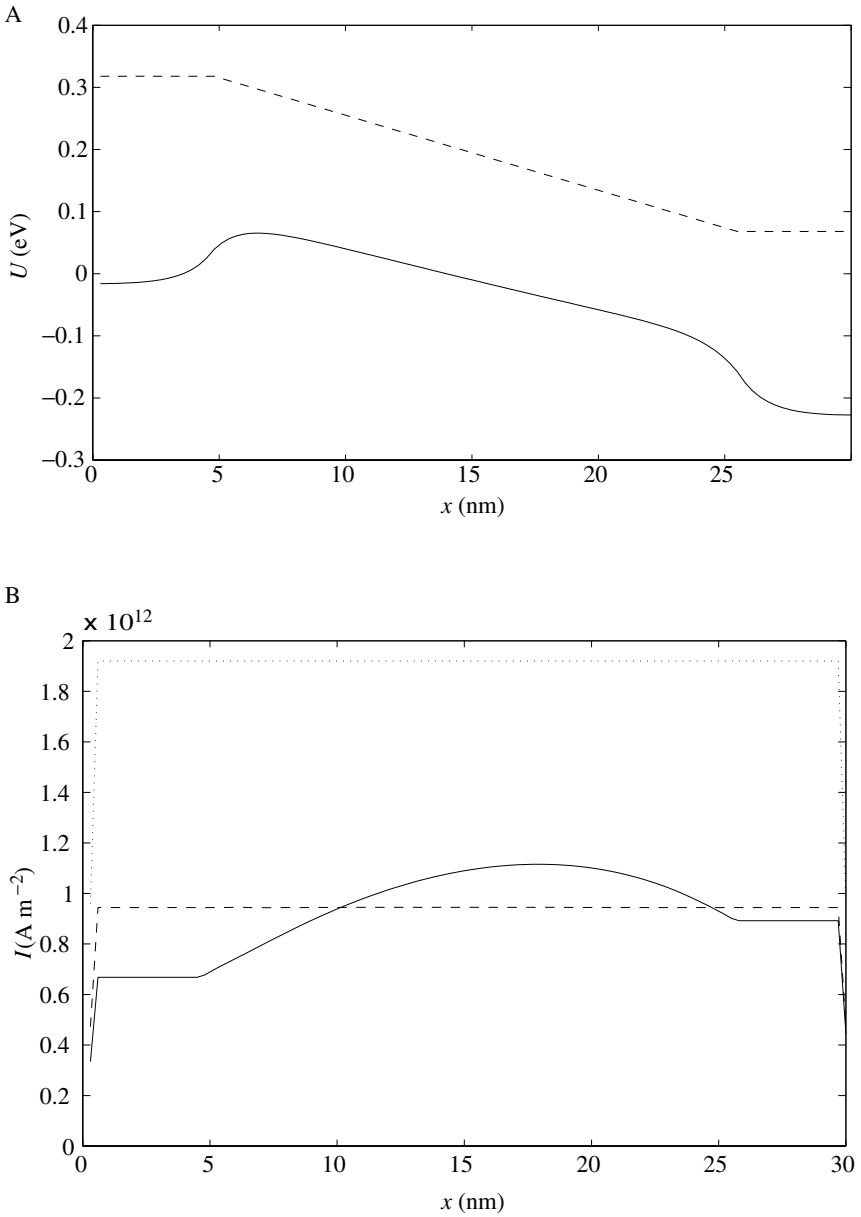

Fig. 12. The same device as in Fig. 9 including scattering processes as described in the text with $\eta = 25$ meV. Assuming a reasonable profile for $\mu_S(x)$, as shown by the dashed curve in A, we obtain the solid profile for the current density across the device. By adjusting μ_{Sn}, the current can be made nearly constant across the device as shown by the dashed line. Also shown for comparison in B is the current without scattering (dotted line).

Fig. 12A, we will find that the current will not be the same everywhere inside the device. We could interpret this lack of current conservation as an inflow or outflow of current at the scattering contacts. But scattering processes lead to an exchange of energy without an exchange of particles, so that we need to ensure that the

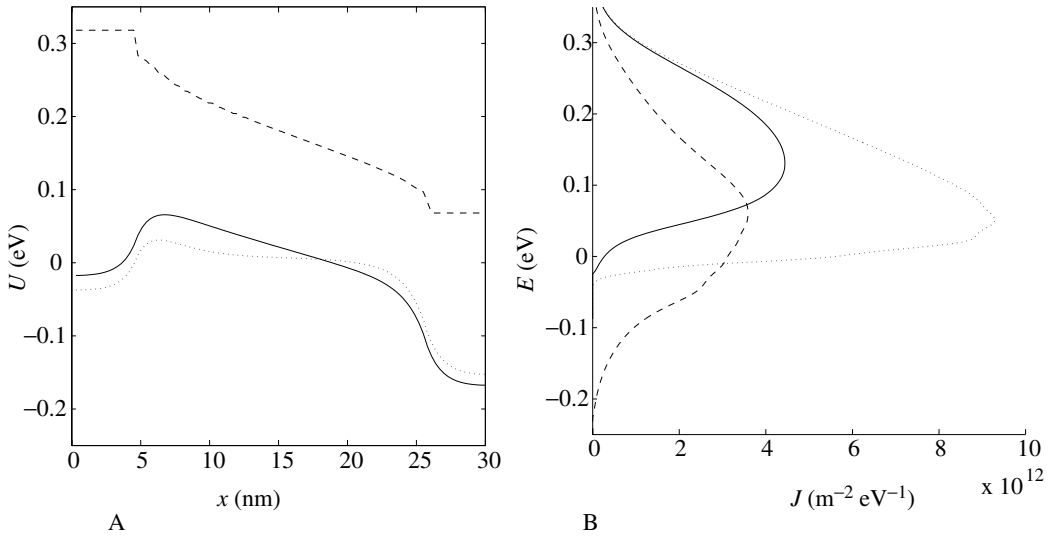

Fig. 13. A, Self-consistent potential profile, $U(x)$ and $\mu_S(x)$ corresponding to the constant current condition shown by the dashed line in Fig. 12; B, energy spectrum of the current ($J = \tilde{I}(E)/S$) at left end of the device (solid curve) and at right end of the device (dashed curve). Also shown for comparison is the potential profile $U(x)$ and the energy spectrum calculated without scattering (dotted curve), which is the same at every 'x'.

current at the scattering contacts is zero. This can be done by adjusting the μ_{Sn} self-consistently so as to reduce dI/dx to zero.

Figure 12 shows the potential profile and current density for the same device as in Fig. 9 including scattering processes through this method with $\eta = 25$ meV in the n^+ region and zero in the n^{++} regions. Assuming a reasonable profile for μ_{Sn} as shown by the dashed curve in (A) we obtain the solid profile for the current density across the device, indicating that current is being injected by the scattering contacts in the left half of the device and extracted by those in the right half. To make the current constant we need to lower μ_S in the left half and raise it in the right half. With proper adjustment, the current can be made nearly constant across the device as shown by the dashed line.

Figure 13 shows the potential profile $U(x)$ and the $\mu_S(x)$ profile corresponding to this constant current condition, along with the energy spectrum of the current. The results look quite reasonable. The potential drops linearly across the device instead of remaining flat since we have now effectively introduced a resistance. The fraction of the voltage dropped inside the device is increased relative to the ballistic case (compare the solid and dotted curves in Fig. 13A). The energy spectrum of the current now moves downwards as we go from contact 1 towards contact 2 as the hot electrons relax their extra energy. By contrast, the coherent transport theory of the last section predicts the same spectrum at every point in the device since there is no mechanism for energy relaxation in the model. this is expected since the divergence of the energy current

$$I_U = \int dE\, E\tilde{I}(E) \tag{5.8}$$

is equal to the power dissipated. In a model with no dissipation, the energy current is spatially constant. But in a model with dissipation, the energy current decreases with distance in agreement with the calculated results shown in Fig. 13B.

This simple phenomenological approach does seem capable of capturing much of the essential physics. Indeed it can be justified in the linear response (or low bias) regime from a microscopic theory if we assume that the scattering is purely isotropic [17]. The basic idea of simulating scattering processes through a floating

contact (originally due to Buttiker [18]) is widely used in mesoscopic physics where much of the attention is usually focused on the linear response regime. A proper microscopic approach, however, requires us to abandon the notion of a Fermi function for the scattering contact altogether and recognize that the inscattering and outscattering functions (cf. eqn (3.19))

$$\Gamma_S = \Sigma_S^{\rm out} + \Sigma_S^{\rm in} = i\left(\Sigma_S^{>} - \Sigma_S^{<}\right) \tag{5.9}$$

are not related to the broadening Γ_S by Fermi functions. Instead they have to be calculated self-consistently from the density matrix, the precise relationship depending on the nature of the scattering process and the level of approximation desired. For example, if we are treating scattering by the emission of phonons with energy $\hbar\omega$ in the first Born approximation then $\Sigma_S^{\rm in}(E) \sim \tilde{\rho}(E+\hbar\omega)$, since electrons with energy $E+\hbar\omega$ get scattered to the energy E. The density matrix at E in turn depends on the inscattering at E through eqn (5.7) and we need a self-consistent calculation of $\Sigma_S^{\rm in}(E)$ and $\tilde{\rho}(E)$. Usually in the NEGF literature, the symbol $-iG^{<}$ is used to denote what we have called $2\pi\tilde{\rho}$, while $+iG^{>}$ is used to denote the empty states $(A-2\pi\tilde{\rho})$ or the hole density matrix and different scattering processes lead to different relationships between $\Sigma_S^{\langle,\rangle}$ on $G^{\langle,\rangle}$ (see for example, Refs [3, 11]). But for a particular microscopic scattering mechanism treated to any order, the precise relationship is clearly laid out in the NEGF formalism, just as in semiclassical theory the scattering rates are clearly known for any process to any order. Indeed, the semiclassical results follow from the NEGF expressions if we assume the eigenstates to be plane waves and use the plane wave representation.

From a practical point of view the real difficulty with including scattering processes is that strictly speaking, different transverse modes, $\boldsymbol{k}$ and different energies, E are not fully independent any more. The inscattering at one $(\boldsymbol{k}, E)$ depends on the density matrix at other $(\boldsymbol{k}', E')$. To make the problem tractable, it is likely that reasonable physically motivated approximations will be needed that are geared towards specific devices. However, the value of the NEGF formalism lies in providing a correct physically sound model that can be used as a starting point for making the necessary approximations.

6. Summary and outlook

As stated in the introduction, we have tried to achieve two objectives: (1) to explain the central concepts that define the 'language' of quantum transport, such as density matrix and self-energy and (2) to illustrate the non-equilibrium Green's function (NEGF) formalism with simple examples that interested readers can easily duplicate on their PCs. The numerical results presented here (Figs 5, 9, 13) were all obtained on a laptop computer and the author will be glad to share his MATLAB programs, typically 40 lines long. These examples all involve a short n^{++}–n^{+}–n^{++} resistor whose physics is easily understood, although the basic formulation is quite general and can be applied to something as different as a nanotube or a molecular wire. Our primary purpose is to illustrate some of the unusual issues that arise in the simulation of short ballistic devices and to show that the NEGF formalism leads to physically sensible results for this simple device. We show that the self-consistent potential profile inside a ballistic conductor (with large cross-section) tends to be flat in the interior of the conductor (see Fig. 9), indicating that the 'voltage drop' is primarily at the ends. Also, a significant fraction of the voltage is dropped inside the contacts and can be associated with the ideal contact resistance well known in mesoscopic physics. However, when we introduce scattering into the model, the potential acquires a non-zero slope inside the conductor as we might expect for an ordinary resistor (see Fig. 13) and the role of the contact resistance is reduced. These examples also underscore the importance of performing self-consistent calculations that include the 'Poisson' equation, augmented with an additional exchange-correlation potential as needed. The I–V characteristics of nanoscale structures is determined by an interesting interplay between twentieth century physics (quantum transport) and nineteenth century physics (electrostatics) and there is a tendency to emphasize one or the other depending on one's

background. But it is important to do justice to both aspects if we are to derive real insights. We believe that self-consistent solutions of the NEGF and 'Poisson' equations should be able to capture the essential physics of most nanoscale devices with the possible exception of those in the 'Coulomb blockade' regime, as discussed in the introduction. However, much work will be needed in the coming years to identify suitable Hamiltonians and scattering models for specific devices that are both accurate and tractable.

Acknowledgements—This work was supported by the National Science Foundation (grant number 9809520-ECS) and the Semiconductor Research Corporation (contract number 99-NJ-724).

References

[1] See for example articles in, IEEE Transactions on Electron Devices, Special Issue on Computational Electronics: New Challenges and Directions, edited by M. S. Lundstrom, R. W. Dutton, D. K. Ferry, and K. Hess (2000).
[2] See for example, C. P. Collier, E. W. Wong, M. Belohradsky, F. M. Raymo, J. F. Stoddart, P. J. Kuekes, R. S. Williams, and J. R. Heath, Science **285**, 391 (1999); J. Chen, M. A. Reed, A. M. Rawlett, and J. M. Tour, Science **286**, 1550 (1999).
[3] See for example, R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, J. Appl. Phys. **81**, 7845 (1997) and references therein.
[4] See for example, D. Vasileska, D. K. Schroder, and D. K. Ferry, IEEE Trans. Electron Devices **44**, 584 (1997).
[5] R. G. Parr and W. Yang, *Density Functional Theory of Atoms and Molecules* (Oxford University Press, 1989).
[6] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, UK, 1997).
[7] Y. Imry, *An Introduction to Mesoscopic Physics* (Oxford University Press, UK, 1997).
[8] D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures* (Cambridge University Press, 1997).
[9] See for example, R. F. Pierret, *Advanced Semiconductor Fundamentals*, Modular Series on Solid State Devices, (Addison-Wesley) Vol. VI, p. 57.
[10] The author is indebted to A. W. Overhauser for suggesting this example.
[11] See for example, G. D. Mahan, Phys. Rep. **145**, 251 (1987); P. Danielewicz, Ann. Phys. **152**, 239 (1984) and references therein. These references deal with bulk conductors and do not explicitly treat the contacts. The use of self-energy functions to represent the contacts can be traced back to C. Caroli, R. Combescot, P. Nozieres, and D. Saint-James, J. Phys. C **5**, 21 (1972) and has been discussed in Chapter 3 of Ref. [6].
[12] See for example, W. Tian, S. Datta, S. Hong, R. Reifenberger, J. I. Henderson, and C. P. Kubiak, J. Chem. Phys. **109**, 2874 (1997) and references therein; M. P. Anantram and T. R. Govindan, Phys. Rev. **B58**, 4882 (1998).
[13] L. P. Kadanoff and G. Baym, *Quantum Statistical Mechanics* (Benjamin/Cummings, 1962); see the discussion in Section 4.4.
[14] Z. Ren and M. S. Lundstrom, (private communication).
[15] For the simple parabolic bands we are discussing here, it is straightforward to identify J_{op}, but this may take considerably more work if we are using more sophisticated models, like the sp^3s^* Hamiltonian.
[16] R. Lake and S. Datta, Phys. Rev. **B46**, 4757 (1992).
[17] M. J. McLennan, Y. Lee, and S. Datta, Phys. Rev. **B43**, 13846 (1991).
[18] M. Buttiker, Phys. Rev. **B33**, 3020 (1986).