united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

ICTP 40th Anniversary

SMR 1564 - 15

SPRING COLLEGE ON SCIENCE AT THE NANOSCALE
(24 May - 11 June 2004)

ELECTRICAL RESISTANCE: AN ATOMISTIC VIEW

Supriyo DATTA
School of Electrical & Computer Engineering, Purdue University
West Lafayette, IN 47907, USA

INSTITUTE OF PHYSICS PUBLISHING NANOTECHNOLOGY
Nanotechnology **15** (2004) S433–S451 PII: S0957-4484(04)77806-X

TUTORIAL

Electrical resistance: an atomistic view

Supriyo Datta

School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

Received 8 March 2004
Published 17 May 2004
Online at stacks.iop.org/Nano/15/S433
DOI: 10.1088/0957-4484/15/7/051

Abstract
This tutorial article presents a 'bottom-up' view of electrical resistance starting from something really small, like a molecule, and then discussing the issues that arise as we move to bigger conductors. Remarkably, no serious quantum mechanics is needed to understand electrical conduction through something really small, except for unusual things like the Kondo effect that are seen only for a special range of parameters. This article starts with energy level diagrams (section 2), shows that the broadening that accompanies coupling limits the conductance to a maximum of q^2/h per level (sections 3, 4), describes how a change in the shape of the self-consistent potential profile can turn a symmetric current–voltage characteristic into a rectifying one (sections 5, 6), shows that many interesting effects in molecular electronics can be understood in terms of a simple model (section 7), introduces the non-equilibrium Green function (NEGF) formalism as a sophisticated version of this simple model with ordinary numbers replaced by appropriate matrices (section 8) and ends with a personal view of unsolved problems in the field of nanoscale electron transport (section 9). Appendix A discusses the Coulomb blockade regime of transport, while appendix B presents a formal derivation of the NEGF equations. MATLAB codes for numerical examples are listed in appendix C. (The appendices are available in the online version only.)

(Some figures in this article are in colour only in the electronic version)

M Supplementary data files are available from the article's abstract page in the online journal; see www.iop.org.

Contents

1. Introduction

It is common to differentiate between two ways of building a small device: a *top-down* approach where we start from something big and chisel out what we want and a *bottom-up* approach where we start from something small such as atoms or molecules and assemble what we want. When it comes to describing electrical resistance, the standard approach could be called a 'top-down' one. We start in college by learning that the conductance, G (inverse of the resistance), of a large macroscopic conductor is directly proportional to its cross-sectional area (A) and inversely proportional to its length (L):

$$G = \sigma A/L \qquad \text{(Ohm's law)}$$

where the conductivity σ is a material property of the conductor. Years later in graduate school we learn about the factors that determine the conductivity and if we stick around long enough we eventually talk about what happens when the conductor is so small that one cannot define its conductivity. In this article I will try to turn this approach around and present a different view of electrical conduction, one that could be called a bottom-up viewpoint [1].

I will try to describe the conductance of something really small, such as a molecule, and then explain the issues that arise as we move to bigger conductors. This is not the way the subject is commonly taught, but I believe the reason is that until recently, no one was sure how to describe the conductance of a really small object, or if it even made sense to talk about the conductance of something really small. To measure the conductance of anything we need to attach two large contact pads to it, across which voltage can be applied. No one knew how to attach contact pads to a small molecule until the late twentieth century, and so no one knew what the conductance of a really small object was. But now that we are able to do so, the answers look fairly clear, and in this article I will try to convey all the essential principles. Remarkably, no serious quantum mechanics is needed to understand electrical conduction through something really small, except for unusual things such as the Kondo effect that are seen only for a special range of parameters. Of course, it is quite likely that new effects will be discovered as we experiment more on small conductors and the description presented here is certainly not intended to be the last word. But I think it should be the 'first word' since the traditional top-down approach tends to obscure the simple physics of very small conductors.

Outline

To model the flow of current, the first step is to draw an equilibrium energy level diagram and locate the electrochemical potential μ (also called the Fermi level or Fermi energy) set by the source and drain contacts (section 2). Current flows when an external device such as a battery maintains the two contacts at different electrochemical potentials μ_1 and μ_2, driving the channel into a non-equilibrium state (section 3). The current through a really small device with only one energy level in the range of interest is easily calculated and, as we might expect, it depends on the quality of the contacts. But what is not obvious (and was not appreciated before the late 1980s) is that there is a maximum conductance for a one-level device which is a fundamental constant related to the charge on an electron, $-q$, and the Planck's constant h:

$$G_0 \equiv q^2/h = 38.7\ \mu\text{S} = (25.8\ \text{k}\Omega)^{-1}. \qquad (1.1)$$

Actually small devices typically have two levels (one for up spin and one for down spin) making the maximum conductance equal to $2G_0$. One can always measure conductances lower than this, if the contacts are bad. But the point is that there is an upper limit to the conductance that can be achieved even with the most perfect of contacts as explained in section 4. We will then discuss how the shape of the current–voltage (I–V) characteristics depends crucially on the electrostatic potential profile which requires a solution of the equations for electrostatics that is self-consistent with those for quantum transport (section 5). Section 6 represents a brief detour, where we discuss the concept of quantum capacitance which can be useful in guessing the electrostatic potential profile without a full self-consistent solution.

Section 7 presents several toy examples to illustrate how the model can be used to understand different current–voltage (I–V) characteristics that are observed for small conductors. This model, despite its simplicity (I use it to introduce an undergraduate course on nanoelectronics), has a rigorous formal foundation. It is really a special case of the non-equilibrium Green function (NEGF) formalism applied to a conductor so small that its electrical conduction can be described in terms of a single energy level. More generally, one needs a Hamiltonian matrix to describe the energy levels and the full NEGF equations can be viewed as a sophisticated version of the simple model with ordinary numbers replaced by appropriate matrices as described in section 8. Finally in section 9 I will conclude by listing what I view as open questions in the field of nanoscale electron transport. Three supplementary appendices are also included. Appendix A describes the multielectron viewpoint needed to describe the new physics (single-electron charging effects) that can arise if a device is coupled weakly to both contacts. Appendix B provides a formal derivation of the NEGF equations for advanced readers using the second-quantized formalism, while appendix C provides a listing of MATLAB codes that can be used to reproduce the numerical examples presented in section 7 and in appendices A, B.

2. Energy level diagram

Consider a simple version of a 'nanotransistor' consisting of a semiconducting channel separated by an insulator layer (typically silicon dioxide) from the metallic gate surrounding the channel (figure 2.1). The voltage V_G on the gate is used to control the electron density in the channel and hence its conductance. The regions marked source and drain are the two contact pads which are assumed to be highly conducting. The resistance of the channel determines the current that flows from the source to the drain when a voltage V_D is applied between them. Such a voltage-controlled resistor is the essence of any field effect transistor (FET) although the details differ from one version to another. The channel length, L, has been progressively reduced from $\sim$10 μm in 1960 to $\sim$0.1 μm in

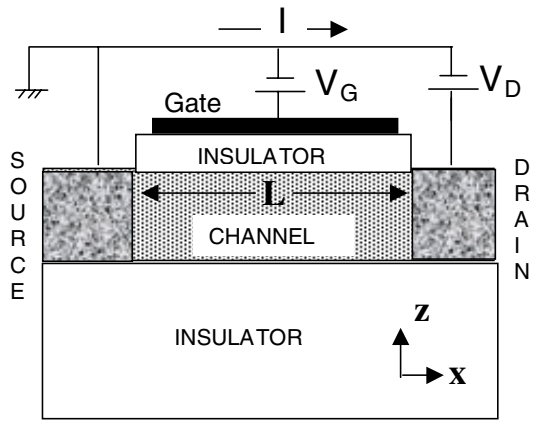


Figure 2.1. A sketch of a nanoscale field effect transistor. The insulator should be thick enough to ensure that no current flows into the gate terminal, but thin enough to ensure that the gate voltage can control the electron density in the channel.

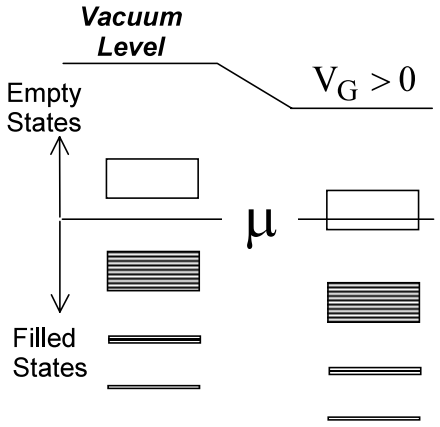


Figure 2.2. Allowed energy levels that can be occupied by electrons in the active region of the device such as the channel in figure 2.1. A positive gate voltage V_G moves the energy levels down while the electrochemical potential μ is fixed by the source and drain contacts which are assumed to be in equilibrium with each other ($V_D = 0$).

2000, allowing circuit designers to pack $(100)^2 = 10\,000$ times more transistors (and hence that much more computing power) into a chip with a given surface area. Laboratory devices have been demonstrated with $L = 0.06$ μm which corresponds to approximately 30 atoms! How do we describe current flow through something this small?

The first step in understanding the operation of any inhomogeneous device structure is to draw an *equilibrium* energy level diagram (sometimes called a 'band diagram') assuming that there is no voltage applied between the source and the drain. Electrons in a semiconductor occupy a set of energy levels that form bands as sketched in figure 2.2. Experimentally, one way to measure the occupied energy levels is to find the minimum energy of a photon required to knock an electron out into vacuum (photoemission or PE experiments). We can describe the process symbolically as

$$S + h\nu \rightarrow S^+ + e^-$$

where 'S' stands for the semiconductor device (or any material for that matter!).

The empty levels, of course, cannot be measured the same way since there is no electron to knock out. We need an inverse photoemission (IPE) experiment where an incident electron is absorbed with the emission of photons:

$$S + e^- \rightarrow S^- + h\nu.$$

Other experiments such as those using optical absorption also provide information regarding energy levels. All these experiments would be equivalent if electrons did not interact with each other and we could knock one electron around without affecting everything else around it. In the real world this is not the case and subtle considerations are needed to relate the measured energies to those we use, but we will not get into this question [2].

We will assume that the large contact regions (labelled source and drain in figure 2.1) have a continuous distribution of states. This is true if the contacts are metallic, but not exactly true of semiconducting contacts and interesting effects, such as a decrease in the current with an increase in the voltage (sometimes referred to as negative differential resistance, NDR), can arise as we will see in section 7 (see also the article by Hersam *et al* [6]). But for the moment let us ignore this possibility and assume the distribution of states to be continuous. They are occupied up to some energy μ (called the electrochemical potential) which also can be located using photoemission measurements. The work function is defined as the minimum energy of a photon needed to knock a photoelectron out of the metal and it tells us how far below the vacuum level μ is located.

Fermi function

If the source and drain regions are coupled to the channel (with V_D held at zero), then electrons will flow in and out of the device bringing them all into equilibrium with a common electrochemical potential, μ, just as two materials in equilibrium acquire a common temperature, T. In this equilibrium state, the average (over time) number of electrons in any energy level is typically not an integer, but is given by the Fermi function:

$$f_0(E - \mu) = \frac{1}{1 + \exp((E - \mu)/k_B T)} \qquad (2.1)$$

which is 1 for energies far below μ and 0 for energies far above μ.

n-type operation

A positive gate voltage V_G applied to the gate lowers the energy levels in the channel. However, the energy levels in the source and drain contacts are unchanged and hence the electrochemical potential μ (which must be the same everywhere) remains unaffected. As a result the energy levels move with respect to μ driving μ into the empty band as shown in figure 2.2. This makes the channel more conductive and turns the transistor ON, since, as we will see in the next section, the current flow under bias depends on the number of energy levels available around $E = \mu$. The threshold gate voltage V_T needed to turn the transistor ON is thus determined by

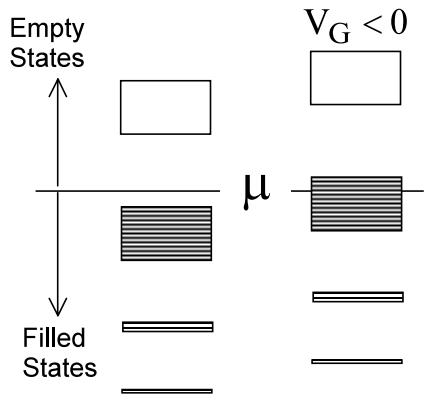


Figure 2.3. An example of p-type or hole conduction. A negative gate voltage ($V_G < 0$) reduces the number of electrons in the channel. Nevertheless the channel will become more conductive once the electrochemical potential μ is driven into the filled band since conduction depends on the availability of states around $E = \mu$ and not on the total number of electrons.

the energy difference between the equilibrium electrochemical potential μ and the lowest available empty state (figure 2.2) or what is called the conduction band edge.

p-type operation

Note that the number of electrons in the channel is not what determines the current flow. A negative gate voltage ($V_G < 0$), for example, reduces the number of electrons in the channel. Nevertheless the channel will become more conductive once the electrochemical potential is driven into the filled band as shown in figure 2.3, due to the availability of states (filled or otherwise) around $E = \mu$.

This is an example of p-type or 'hole' conduction as opposed to the example of n-type or electron conduction shown in figure 2.2. The point is that for current flow to occur states are needed near $E = \mu$, but they need not be empty states. Filled states are just as good and it is not possible to tell from this experiment whether conduction is n-type (figure 2.2) or p-type (figure 2.3). This point should get clearer in the next section when we discuss why current flows in response to a voltage applied across the source and drain contacts.

Figures 2.2 and 2.3 suggest that the same device can be operated as an n-type or a p-type device simply by reversing the polarity of the gate voltage. This is true for short devices if the contacts have a continuous distribution of states as we have assumed. But in general this need not be so: for example, long devices can build up 'depletion layers' near the contacts whose shape can be different for n- and p-type devices.

3. What makes electrons flow?

We have stated that conduction depends on the availability of states around $E = \mu$; it does not matter if they are empty or filled. To understand why, let us ask what makes electrons flow from the source to the drain. The battery lowers the energy levels in the drain contact with respect to the source contact (assuming V_D to be positive) and maintains them at distinct electrochemical potentials separated by qV_D (see figure 3.1):

$$\mu_1 - \mu_2 = qV_D \tag{3.1}$$

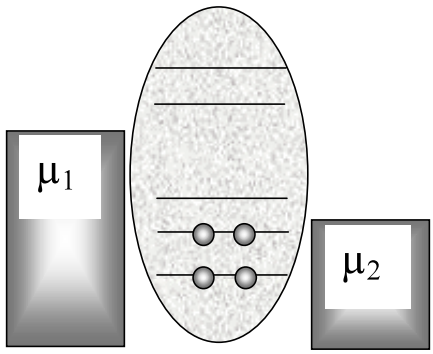


Figure 3.1. A positive voltage V_D applied to the drain with respect to the source lowers the electrochemical potential at the drain: $\mu_2 = \mu_1 - qV_D$. Source and drain contacts now attempt to impose different Fermi distributions as shown and the device goes into a state intermediate between the two.

giving rise to two different Fermi functions:

$$f_1(E) \equiv \frac{1}{1 + \exp((E - \mu_1)/k_B T)} = f_0(E - \mu_1) \tag{3.2a}$$

$$f_2(E) \equiv \frac{1}{1 + \exp((E - \mu_2)/k_B T)} = f_0(E - \mu_2). \tag{3.2b}$$

Each contact seeks to bring the active device into equilibrium with itself. The source keeps pumping electrons into it hoping to establish equilibrium. But equilibrium is never achieved as the drain keeps pulling electrons out in its bid to establish equilibrium with itself. The device is thus forced into a balancing act between two reservoirs with different agendas which sends it into a non-equilibrium state intermediate between what the source would like to see and what the drain would like to see.

Rate equations for a one-level model

This balancing act is easy to see if we consider a simple one-level system, biased such that its energy ε lies between the electrochemical potentials in the two contacts (figure 3.2). Contact 1 would like to see $f_1(\varepsilon)$ electrons, while contact 2 would like to see $f_2(\varepsilon)$ electrons occupying the state where f_1 and f_2 are the source and drain Fermi functions defined in equation (3.1). The average number of electrons N at the steady state will be something intermediate between f_1 and f_2. There is a net flux I_1 across the left junction that is proportional to $f_1 - N$, dropping the argument ε for clarity:

$$I_1 = (-q)\frac{\gamma_1}{\hbar}(f_1 - N) \tag{3.3a}$$

where $-q$ is the charge per electron. Similarly the net flux I_2 across the right junction is proportional to $f_2 - N$ and can be written as

$$I_2 = (-q)\frac{\gamma_2}{\hbar}(f_2 - N). \tag{3.3b}$$

We can interpret the rate constants $\gamma_1/\hbar$ and $\gamma_2/\hbar$ as the rates at which an electron placed initially in the level ε will escape into the source and drain contacts respectively. In principle, we could experimentally measure these quantities which have the dimension per second (γ_1 and γ_2 have the dimension of energy). At the end of this section I will say a few more words about the physics behind these equations. But for the moment, let us work out the consequences.

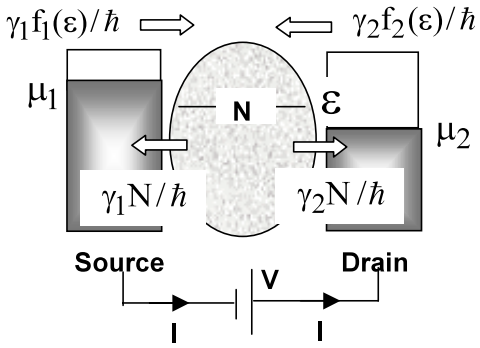


Figure 3.2. The flux of electrons into and out of a one-level device at the source and drain ends: the simple rate equation picture.

Current in a one-level model

At steady state there is no net flux into or out of the device: $I_1 + I_2 = 0$, so from equations (3.2*a*), (3.2*b*) we obtain the reasonable result

$$N = \frac{\gamma_1 f_1 + \gamma_2 f_2}{\gamma_1 + \gamma_2} \tag{3.4}$$

(that the occupation N is a weighted average of what contacts 1 and 2 would like to see). Substituting this result into equations (3.3*a*) or (3.3*b*) we obtain an expression for the steady-state current:

$$I = I_1 = -I_2 = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}[f_1(\varepsilon) - f_2(\varepsilon)]. \tag{3.5}$$

This is the current per spin. We should multiply it by 2 if there are two spin states with the same energy.

This simple result serves to illustrate certain basic facts about the process of current flow. Firstly, no current will flow if $f_1(\varepsilon) = f_2(\varepsilon)$. A level that is way below both electrochemical potentials μ_1 and μ_2 will have $f_1(\varepsilon) = f_2(\varepsilon) = 1$ and will not contribute to the current, just like a level that is way above both potentials μ_1 and μ_2 and has $f_1(\varepsilon) = f_2(\varepsilon) = 0$. It is only when the level lies within a few k_BT of the potentials μ_1 and μ_2 that we have $f_1(\varepsilon) \neq f_2(\varepsilon)$ and a current flows as a result of the '*difference in agenda*' between the contacts. Contact 1 keeps pumping in electrons striving to bring the number up from N to f_1 while contact 2 keeps pulling them out striving to bring it down to f_2. The net effect is a continuous transfer of electrons from contact 1 to 2 corresponding to a current I in the external circuit (figure 3.2). Note that the current is in a direction opposite to that of the flux of electrons, since electrons have negative charge.

It should now be clear why the process of conduction requires the presence of states around $E = \mu$. It does not matter if the states are empty (n-type, figure 2.2) or filled (p-type, figure 2.3) in equilibrium, before a drain voltage is applied. With empty states, electrons are first injected by the negative contact and subsequently collected by the positive contact. With filled states, electrons are first collected by the positive contact and subsequently refilled by the negative contact. Either way, we have current flowing in the external circuit in the same direction.

Inflow/outflow

Equations (3.3*a*), (3.3*b*) look elementary and I seldom hear anyone question them. But they hide many subtle issues that could bother more advanced readers and so I feel obliged to mention these issues briefly at the risk of confusing satisfied readers. The right-hand sides of equations (3.3*a*), (3.3*b*) can be interpreted as the difference between the inflow and the outflow from the source and drain respectively (see figure 3.2). For example, consider the source. The outflow of $\gamma_1 N/\hbar$ is easy to explain since $\gamma_1/\hbar$ represents the rate at which an electron placed initially in the level ε will escape into the source contact. But the inflow $\gamma_1 f_1/\hbar$ is harder to explain since there are many electrons in many states in the contacts, all seeking to fill up one state inside the channel, and it is not obvious how to sum up the inflow from all these states. A convenient approach is to use a thermodynamic argument as follows: if the channel were in equilibrium with the source, there would be no net flux, so the inflow would equal the outflow. But the outflow under equilibrium conditions would equal $\gamma_1 f_1/\hbar$ since N would equal f_1. Under non-equilibrium conditions, N differs from f_1 but the inflow remains unchanged since it depends only on the condition in the contacts which remains unchanged (note that the outflow does change, giving a net current that we have calculated above).

'Pauli blocking'?

Advanced readers may disagree with the statement I have just made, namely that the inflow 'depends only on the condition in the contacts'. Should the inflow not be reduced by the presence of electrons in the channel due to the exclusion principle ('Pauli blocking')? Specifically one could argue that the inflow and outflow (at the source contact) be identified respectively as

$$\gamma_1 f_1(1 - N) \qquad \text{and} \qquad \gamma_1 N(1 - f_1)$$

instead of

$$\gamma_1 f_1 \qquad \text{and} \qquad \gamma_1 N$$

as we have indicated in figure 3.2. It is easy to see that the net current given by the difference between inflow and outflow is the same in either case, so the argument might appear 'academic'. What is not academic, however, is the level broadening that accompanies the process of coupling to the contacts, something we need to include in order to get quantitatively correct results (as we will see in the next section). I have chosen to define inflow and outflow in such a way that the outflow per electron ($\gamma_1 = \gamma_1 N/N$) is equal to the broadening (in addition to their difference being equal to the net current). Whether this broadening (due to the source) is γ_1 or $\gamma_1(1 - f_1)$ or something else is not an academic question. It can be shown that as long as energy relaxing or inelastic interactions are not involved in the inflow/outflow process, the broadening is γ_1 independent of the occupation factor f_1 in the contact.

4. The quantum of conductance

Consider a device with a small voltage applied across it causing a splitting of the source and drain electrochemical potentials (figure 4.1(a)). We can write the current through

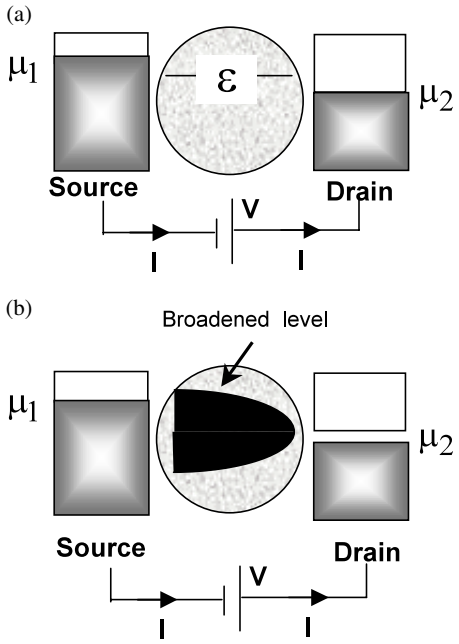


Figure 4.1. (a) A device with a small voltage applied across it causing a splitting of the source and drain electrochemical potentials $\mu_1 > \varepsilon > \mu_2$. (b) The process of coupling to the device inevitably broadens it thereby spreading part of the energy level outside the energy range between μ_1 and μ_2 where current flows.

this device from equation (3.5) and simplify it by assuming that $\mu_1 > \varepsilon > \mu_2$ and the temperature is low enough that $f_1(\varepsilon) \equiv f_0(\varepsilon - \mu_1) \approx 1$ and $f_2(\varepsilon) \equiv f_0(\varepsilon - \mu_2) \approx 0$ (see equation (3.2)):

$$I = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} = \frac{q\gamma_1}{2\hbar} \qquad \text{if } \gamma_2 = \gamma_1. \tag{4.1a}$$

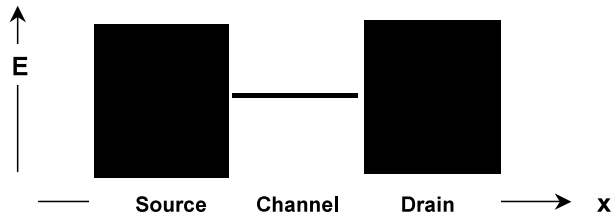
This suggests that we could pump unlimited current through this one-level device by increasing γ_1 ($=\gamma_2$), that is by coupling it more and more strongly to the contacts. However, one of the seminal results of mesoscopic physics is that the maximum conductance of a one-level device is equal to G_0 (see equation (1.1)). What have we missed?

What we have missed is the broadening of the level that inevitably accompanies any process of coupling to it. This causes part of the energy level to spread outside the energy range between μ_1 and μ_2 where current flows. The actual current is then reduced below what we expect from equation (4.1) by a factor $(\mu_1 - \mu_2)/C\gamma_1$ representing the fraction of the level that lies in the window between μ_1 and μ_2, where $C\gamma_1$ is the effective width of the level, C being a numerical constant. Since $\mu_1 - \mu_2 = qV_D$, we see from equation (4.1)

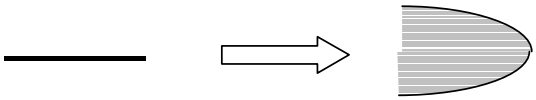
$$I = \frac{q\gamma_1}{2\hbar}\frac{qV_D}{C\gamma_1} \rightarrow G = \frac{I}{V_D} = \frac{q^2}{2C\hbar} \tag{4.1b}$$

that the conductance indeed approaches a constant value independent of the strength of the coupling ($\gamma_1 = \gamma_2$) to the contacts. We will now carry out this calculation a little more quantitatively so as to obtain a better estimate for 'C'.

One way to understand this broadening is to note that, *before* we couple the channel to the source and the drain, the density of states (DOS), $D(E)$, looks something like this (dark indicates a high DOS):



We have one sharp level in the channel and a continuous distribution of states in the source and drain contacts. On coupling, these states 'spill over': the channel 'loses' part of its state as it spreads into the contacts, but it also 'gains' part of the contact states that spread into the channel. Since the loss occurs at a fixed energy while the gain is spread out over a range of energies, the overall effect is to broaden the channel DOS from its initial sharp structure into a more diffuse structure:



There is a 'sum rule' that requires the loss to be exactly offset by the gain, so that integrated over all energy, the level can still hold only one electron. It is common to represent the broadened DOS by a Lorentzian function centred around $E = \varepsilon$ (whose integral over all energy is equal to one):

$$D_\varepsilon(E) = \frac{\gamma/2\pi}{(E-\varepsilon)^2 + (\gamma/2)^2}. \tag{4.2}$$

The initial delta function can be represented as the limiting case of $D_\varepsilon(E)$ as the broadening tends to zero: $\gamma \rightarrow 0$. The broadening γ is proportional to the strength of the coupling as we might expect. Indeed it turns out that $\gamma = \gamma_1 + \gamma_2$, where $\gamma_1/\hbar$ and $\gamma_2/\hbar$ are the escape rates introduced in the last section. This comes out of a full quantum mechanical treatment, but we could rationalize it as a consequence of the 'uncertainty principle' that requires the product of the lifetime ($=\hbar/\gamma$) of a state and its spread in energy (γ) to equal $\hbar$ [3].

Another way to explain the broadening that accompanies the coupling is to note that the coupling to the surroundings makes energy levels acquire a finite lifetime, since an electron inserted into a state with energy $E = \varepsilon$ at time $t = 0$ will gradually escape from that state making its wavefunction look like

$$\exp(-i\varepsilon t/\hbar)\exp(-|t|/2\tau)$$

instead of just

$$\exp(-i\varepsilon t/\hbar).$$

This broadens its Fourier transform from a delta function at $E = \varepsilon$ to the Lorentzian function of width $\gamma = \hbar/\tau$ centred around $E = \varepsilon$ given in equation (4.2). There is thus a simple relationship between the lifetime of a state and its broadening: a lifetime of one picosecond (ps) corresponds to approximately 1.06e−22 J or 0.7 meV. In general the escape of electrons from a level need not follow a simple exponential and the corresponding lineshape need not be Lorentzian. This is usually reflected in an energy-dependent broadening $\gamma(E)$.

The coupling to the contacts thus broadens a single discrete energy level into a continuous density of states given by equation (4.2) and we can include this effect by modifying our expression for the current

$$I = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}[f_1(\varepsilon) - f_2(\varepsilon)] \qquad \text{(same as equation (3.5))}$$

to account for it:

$$I = \frac{q}{\hbar}\int_{-\infty}^{+\infty} \mathrm{d}E\, D_\varepsilon(E)\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}[f_1(E) - f_2(E)]. \qquad (4.3)$$

Equation (4.3) for the current extends our earlier result in equation (3.5) to include the effect of broadening. We could write it in the form

$$I = \frac{q}{h}\int_{-\infty}^{+\infty} \mathrm{d}E\, \bar{T}(E)[f_1(E) - f_2(E)] \qquad (4.4)$$

where the *transmission* $\bar{T}(E)$ is defined as (making use of equation (4.2))

$$\bar{T}(E) \equiv 2\pi D_\varepsilon(E)\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} = \frac{\gamma_1\gamma_2}{(E-\varepsilon)^2 + (\gamma/2)^2}. \qquad (4.5)$$

At low temperatures, we can write

$$f_1(E) - f_2(E) = \begin{cases} 1 & \text{if } \mu_1 > E > \mu_2 \\ 0 & \text{otherwise} \end{cases}$$

so the current is given by

$$I = \frac{q}{h}\int_{\mu_2}^{\mu_1} \mathrm{d}E\, \bar{T}(E).$$

If the bias is small enough that we can assume the density of states and hence the transmission to be constant over the range $\mu_1 > E > \mu_2$, so using equation (4.5) we can write

$$I = \frac{q}{h}[\mu_1 - \mu_2]\frac{\gamma_1\gamma_2}{(\mu-\varepsilon)^2 + ((\gamma_1+\gamma_2)/2)^2}.$$

The maximum current is obtained if the energy level ε coincides with μ, the average of μ_1 and μ_2. Noting that $\mu_1 - \mu_2 = qV_\mathrm{D}$, we can write the maximum conductance as

$$G \equiv \frac{I}{V_\mathrm{D}} = \frac{q^2}{h}\frac{4\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2} = \frac{q^2}{h} \qquad \text{if } \gamma_1 = \gamma_2.$$

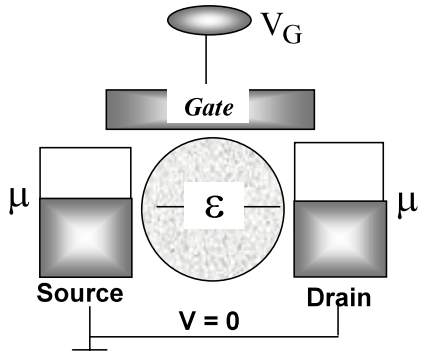
We can also extend the expression for the number of electrons N (see equation (3.4)) to account for the broadened density of states:

$$N = \int_{-\infty}^{+\infty} \mathrm{d}E\, n(E)$$

$$\text{where } n(E) \equiv D_\varepsilon(E)\frac{\gamma_1 f_1(E) + \gamma_2 f_2(E)}{\gamma_1+\gamma_2}. \qquad (4.6)$$

5. Potential profile

Now that we have included the effect of level broadening, there is one other factor that we should include in order to complete our model for a one-level conductor. This has to do with the fact that the voltages applied to the external electrodes (source, drain and gate) change the electrostatic potential in the channel and hence the energy levels. It is easy to see that this can play an important role in determining the shape of the current–voltage characteristics [4]. Consider a one-level device with an equilibrium electrochemical potential μ located slightly above the energy level ε as shown:



When we apply a voltage between the source and drain, the electrochemical potentials separate by qV: $\mu_1 - \mu_2 = qV$. We know that a current flows (at low temperatures) only if the level ε lies between μ_1 and μ_2. Depending on how the energy level ε moves we have different possibilities.

If we ignore the gate we might expect the potential in the channel to be lie halfway between the source and the drain: $\varepsilon \to \varepsilon - (V/2)$, leading to the picture shown in figure 5.1 for positive and negative voltages (note that we are assuming the source potential, relative to which the other potentials are changing, to be held constant). It is apparent that the energy level lies halfway between μ_1 and μ_2 for either bias polarity ($V > 0$ or $V < 0$), leading to a current–voltage characteristic that is symmetric in V.

A different picture emerges, if we assume that the gate is so closely coupled to the channel that the energy level follows the gate potential and is unaffected by the drain voltage or, in other words, ε remains fixed (figure 5.2). In this case the energy level lies between μ_1 and μ_2 for positive bias ($V > 0$) but not for negative bias ($V < 0$), leading to a current–voltage characteristic that can be very asymmetric in V.

The point I wish to make is that the shape of the current–voltage characteristic is affected strongly by the potential profile and even the simplest model needs to account for it. One often hears the question: how do we design a molecule that will rectify? The above example shows that the same molecule could rectify or not rectify depending on how close the gate electrode is located!

So how do we calculate the potential inside the channel? If the channel were an insulator, we could solve Laplace's equation (ε_r: relative permittivity which could be spatially varying)

$$\vec{\nabla}\cdot(\varepsilon_\mathrm{r}\vec{\nabla}V) = 0$$

subject to the boundary conditions that $V = 0$ (source electrode), $V = V_\mathrm{G}$ (gate electrode) and $V = V_\mathrm{D}$ (drain electrode). We could visualize the solution to this equation in terms of the capacitive circuit model shown in figure 5.3, if we treat the channel as a single point ignoring any variation in the potential inside it.

The potential energy in the channel is obtained by multiplying the electrostatic potential, V, by the electronic

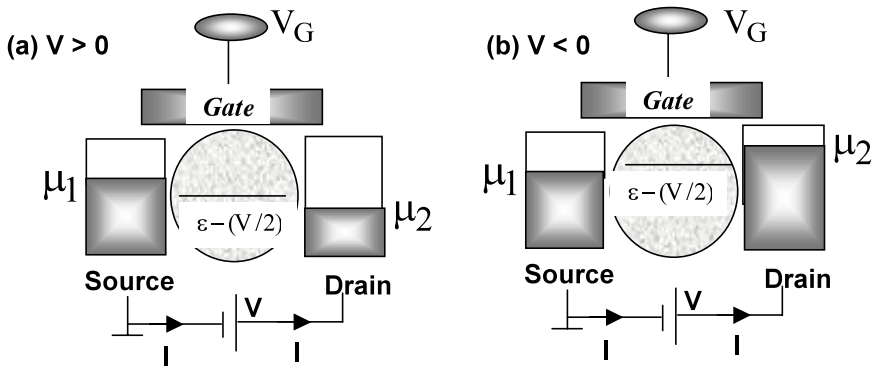


Figure 5.1. If the channel potential lies halfway between the source and drain potentials, significant current will flow for either bias polarity and the current–voltage characteristics will look symmetric.

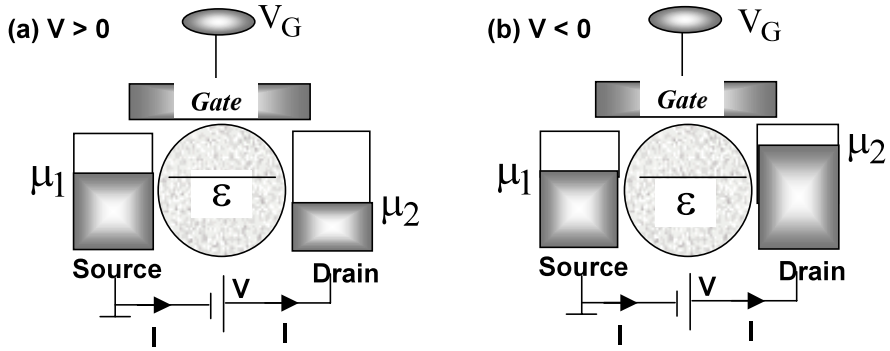


Figure 5.2. If the channel potential is tied to the source and unaffected by the drain potential, significant current will flow (a) for $V > 0$, but not (b) for $V < 0$, making the current–voltage characteristics look rectifying.

charge, $-q$:

$$U_{\mathrm{L}} = \frac{C_{\mathrm{G}}}{C_{\mathrm{E}}}(-qV_{\mathrm{G}}) + \frac{C_{\mathrm{D}}}{C_{\mathrm{E}}}(-qV_{\mathrm{D}}). \qquad (5.1a)$$

Here we have labelled the potential energy with a subscript 'L' as a reminder that it is calculated from the Laplace equation ignoring any change in the electronic charge, which is justified if there are very few electronic states in the energy range around μ_1 and μ_2.

Otherwise there is a change $\Delta\rho$ in the electron density in the channel and we need to solve the Poisson equation

$$\vec{\nabla} \cdot (\varepsilon_{\mathrm{r}} \vec{\nabla} V) = -\Delta\rho/\varepsilon_0$$

for the potential. In terms of our capacitive circuit model, we could write the change in the charge as a sum of the charges on the three capacitors:

$$-q\Delta N = C_{\mathrm{S}}V + C_{\mathrm{G}}(V - V_{\mathrm{G}}) + C_{\mathrm{D}}(V - V_{\mathrm{D}})$$

so the potential energy $U = -qV$ is given by the sum of the Laplace potential and an additional term proportional to the change in the number of electrons:

$$U = U_{\mathrm{L}} + \frac{q^2}{C_{\mathrm{E}}}\Delta N. \qquad (5.1b)$$

The constant $q^2/C_{\mathrm{E}} \equiv U_0$ tells us the change in the potential energy due to *one* extra electron and is called the single-electron charging energy, whose significance we will discuss further in the next section. The *change* ΔN in the number of electrons is calculated with respect to the reference number of electrons, N_0, originally in the channel, corresponding to which its energy level ε is known.

Iterative procedure for self-consistent solution

For a small device, the effect of the potential U is to raise the density of states in energy and can be included in our expressions for the number of electrons, N (equation (4.6)), and the current, I (equation (4.3)), in a straightforward manner:

$$N = \int_{-\infty}^{+\infty} \mathrm{d}E\, D_\varepsilon(E-U)\frac{\gamma_1 f_1(E) + \gamma_2 f_2(E)}{\gamma_1 + \gamma_2} \qquad (5.2)$$

$$I = \frac{q}{\hbar}\int_{-\infty}^{+\infty} \mathrm{d}E\, D_\varepsilon(E-U)\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}[f_1(E) - f_2(E)]. \qquad (5.3)$$

Equation (5.2) has a U appearing on its right-hand side which in turn is a function of N through the electrostatic relation (equation (5.1)). This requires a simultaneous or 'self-consistent' solution of the two equations which is usually carried out using the iterative procedure depicted in figure 5.4. We start with an initial guess for U, calculate N from equation (5.2) with $D_\varepsilon(E)$ given by equation (4.2), calculate an appropriate U from equation (5.1*b*), with U_{L} given by equation (5.1*a*), and compare with our starting guess for U. If this new U is not sufficiently close to our original guess, we

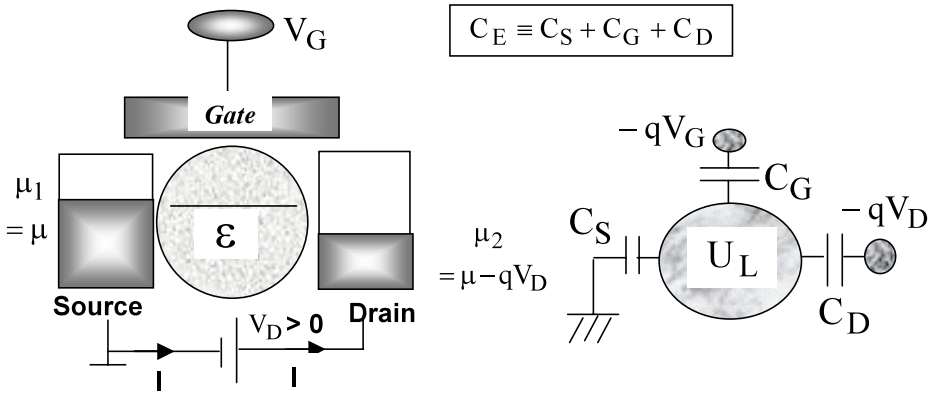


Figure 5.3. A simple capacitive circuit model for the 'Laplace' potential U_L of the active region in response to the external gate and drain voltages, V_G and V_D. The actual potential 'U' can be different from U_L if there is a significant density of electronic states in the energy range around μ_1 and μ_2. The total capacitance is denoted as C_E, where 'E' stands for electrostatic.

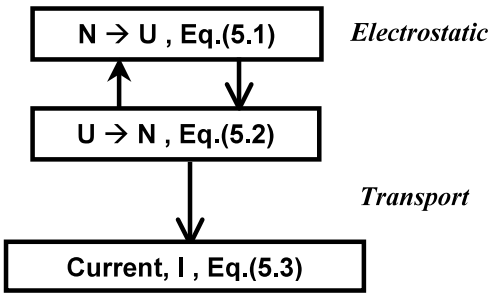


Figure 5.4. The iterative procedure for calculating N and U self-consistently.

revise our guess using a suitable algorithm, say something like

$$U_n = U_o + \alpha[U_c - U_o] \tag{5.4}$$

New guess *Old guess* *Calculated*

where α is a positive number (typically <1) that is adjusted to be as large as possible without causing the solution to diverge (which is manifested as an increase in $U_c - U_o$ from one iteration to the next). The iterative process has to be repeated till we find a U that yields an 'N' that leads to a new U which is sufficiently close (say within a fraction of k_BT) to the original value. Once a converged U has been found, the current can be calculated from equation (5.3).

The self-consistent charging model based on the Poisson equation that we have just discussed represents a good zero-order approximation (sometimes called the Hartree approximation) to the problem of electron–electron interactions, but it is generally recognized that it tends to overestimate the effect. Corrections for the so-called exchange and correlation effects are often added, but the description is still within the one-electron picture which assumes that a typical electron feels some average potential, U, due to the other electrons. Failure of this one-electron picture is known to give rise to profound effects such as magnetism. As we may expect, related effects can manifest themselves in nanoscale transport as well and will continue to be discovered as the field progresses. Such effects are largely outside the scope of this article. However, there is one aspect that is fairly well understood and can affect our picture of current flow even for a simple one-level device putting it in the so-called Coulomb blockade or single-electron charging regime. A proper treatment of this regime requires the multielectron picture described in appendix A.

6. Quantum capacitance

As we have seen, the actual potential U inside the channel plays an important role in determining the shape of the I–V characteristics. Of course, this comes out automatically from the self-consistent calculation described above, but it is important not merely to calculate but also to understand the result. Quantum capacitance is a very useful concept that helps in this understanding [5].

We are performing a simultaneous solution of two relations connecting the potential, U, to the number of electrons, N: an electrostatic relation (equation (5.1)) which is strictly linear and is based on freshman physics, and a transport relation (equation (5.2)) which is non-linear and in general could involve advanced quantum statistical mechanics, although we have tried to keep it fairly simple so far. It is this latter equation that is relatively unfamiliar and one could get some insight by linearizing it around an appropriate point. For example, we could define a potential $U = U_N$, which makes $N = N_0$ and keeps the channel exactly neutral:

$$N_0 = \int_{-\infty}^{+\infty} \mathrm{d}E\, D_\varepsilon(E - U_N)\frac{\gamma_1 f_1(E) + \gamma_2 f_2(E)}{\gamma_1 + \gamma_2}.$$

Any increase in U will raise the energy levels and reduce N, while a decrease in U will lower the levels and increase N. So, for small deviations from the neutral condition, we could write

$$\Delta N \equiv N - N_0 \approx C_Q[U_N - U]/q^2$$
$$\text{where } C_Q \equiv -q^2[\mathrm{d}N/\mathrm{d}U]_{U=U_N} \tag{6.1}$$

is called the quantum capacitance and depends on the density of states around the energy range of interest, as we will show. We can substitute this linearized relation into equation (5.1*b*) to obtain

$$U = U_L + \frac{C_Q}{C_E}[U_N - U] \rightarrow U = \frac{C_E U_L + C_Q U_N}{C_E + C_Q} \tag{6.2}$$

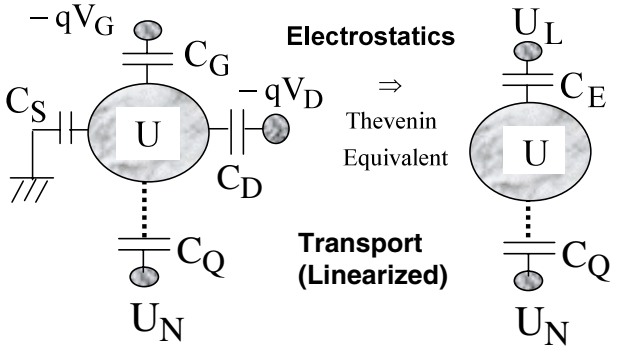


Figure 6.1. Extension of the capacitive network in figure 5.1 to include the quantum capacitance.

showing that the actual channel potential U is intermediate between the Laplace potential, U_L, and the neutral potential, U_N. How close it is to one or the other depends on the relative magnitudes of the electrostatic capacitance, C_E, and the quantum capacitance, C_Q. This is easily visualized in terms of a capacitive network obtained by extending figure 5.1 to include the quantum capacitance, as shown in figure 6.1.

We will now show that a channel with a low density of states in the energy range of interest has a low C_Q making $U = U_L$ as we expect for an insulator. A channel with a high density of states in the energy range of interest has a high C_Q, making $U = U_N$ as we expect for a metal.

Relation between C_Q and the density of states

To establish the connection between the quantum capacitance and the density of states, we rewrite equation (5.2) in the form

$$N = \int_{-\infty}^{+\infty} \mathrm{d}E\, D_\varepsilon(E) \frac{\gamma_1 f_0(E+U-\mu_1) + \gamma_2 f_0(E+U-\mu_2)}{\gamma_1+\gamma_2}$$

and then make use of equation (6.1) for C_Q:

$$\begin{aligned} C_Q &\equiv -q^2[\mathrm{d}N/\mathrm{d}U]_{U=U_N} \\ &= q^2 \int_{-\infty}^{+\infty} \mathrm{d}E\, [D_1(E)F_T(E+U_N-\mu_1) \\ &\quad + D_2(E)F_T(E+U_N-\mu_2)] \end{aligned} \tag{6.3}$$

where

$$D_1(E) \equiv D_\varepsilon(E)\frac{\gamma_1}{\gamma_1+\gamma_2}$$

and

$$D_2(E) \equiv D_\varepsilon(E)\frac{\gamma_2}{\gamma_1+\gamma_2}$$

and we have introduced the thermal broadening function F_T defined as

$$F_T(E) \equiv -\frac{\mathrm{d}f_0}{\mathrm{d}E} = \frac{1}{4k_BT}\,\mathrm{sech}^2\left(\frac{E}{2k_BT}\right). \tag{6.4}$$

Its maximum value is $(1/4k_BT)$ while its width is proportional to k_BT. It is straightforward to show that the area obtained by integrating this function is equal to one, independently of k_BT. This means that at low temperatures $F_T(E)$ becomes very large but very narrow while maintaining a constant area of one and can be idealized as a delta function: $F_T(E) \to \delta(E)$, which allows us to simplify the expression for the quantum capacitance:

$$C_Q \approx q^2[D_1(\mu_1 - U_N) + D_2(\mu_2 - U_N)]. \tag{6.5}$$

This expression, valid at low temperatures, shows that the quantum capacitance depends on the density of states around the electrochemical potentials μ_1 and μ_2, after shifting by the potential U_N.

7. Toy examples

In this section I will first summarize the model that we have developed here and then illustrate it with a few toy examples. We started by calculating the current through a device with a single discrete level (ε) in section 3, and then extended it to include the broadening of the level into a Lorentzian density of states

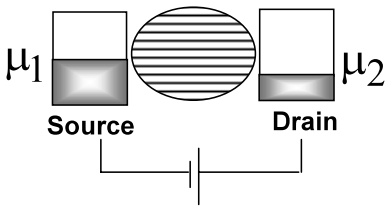
$$D_\varepsilon(E) = 2(\text{for spin}) \times \frac{\gamma/2\pi}{(E-\varepsilon)^2 + (\gamma/2)^2} \qquad \gamma \equiv \gamma_1+\gamma_2 \tag{7.1}$$

in section 4 and the self-consistent potential in section 5:

$$U = U_L + U_0(N - N_0) \tag{7.2}$$

$$\begin{aligned} U_L &= \frac{C_G}{C_E}(-qV_G) + \frac{C_D}{C_E}(-qV_D) \\ U_0 &= q^2/C_E \qquad C_E = C_G + C_S + C_D. \end{aligned} \tag{7.3}$$

The function $D_\varepsilon(E)$ in equation (7.1) is intended to denote the density of states (DOS) obtained by broadening a single discrete level ε. What about a multilevel conductor with many energy levels looking something like this?



If we make the rather cavalier assumption that all levels conduct independently, then we could use exactly the same equations as for the one-level device, replacing the one-level DOS, $D_\varepsilon(E)$, in equation (7.1) with the total DOS, $D(E)$. With this in mind, I will use $D(E)$ instead of $D_\varepsilon(E)$ to denote the density of states and refer to the results summarized below as the *independent level model* rather than the single-level model.

Independent level model: summary

In this model, the number of electrons, N, is given by

$$\begin{aligned} N &= \int_{-\infty}^{+\infty} \mathrm{d}E\, n(E) \\ &\text{where } n(E) = D(E-U)\left(\frac{\gamma_1}{\gamma} f_1(E) + \frac{\gamma_2}{\gamma} f_2(E)\right) \end{aligned} \tag{7.4}$$

while the currents at the two terminals are given by

$$I_1 = \frac{q}{\hbar}\int_{-\infty}^{+\infty} \mathrm{d}E\, \gamma_1[D(E-U)f_1(E) - n(E)] \tag{7.5a}$$

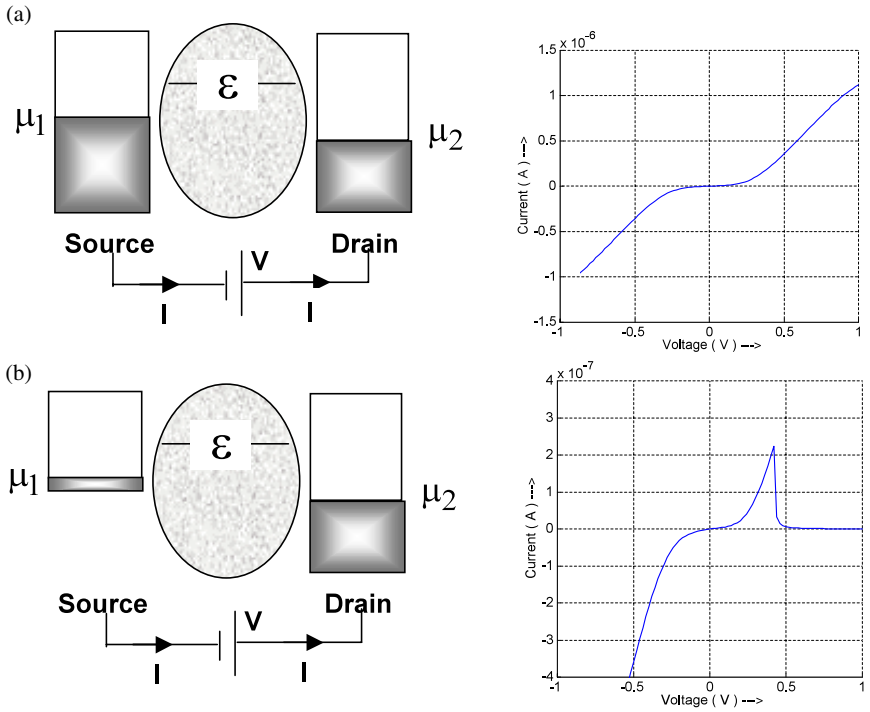


Figure 7.1. Current versus voltage calculated using equations (7.1)–(7.8) with $\mu = 0$, $\varepsilon = 0.2$ eV, $V_G = 0$, $k_B T = 0.025$ eV, $U_0 = 0.25$ eV, $C_D/C_E = 0.5$ and $\gamma_1 = \gamma_2 = 0.005$ eV. The only difference between (a) and (b) is that in (a) γ_1 is independent of energy, while in (b) γ is zero for energies less than zero. In either case γ_2 is assumed to be independent of the energy.

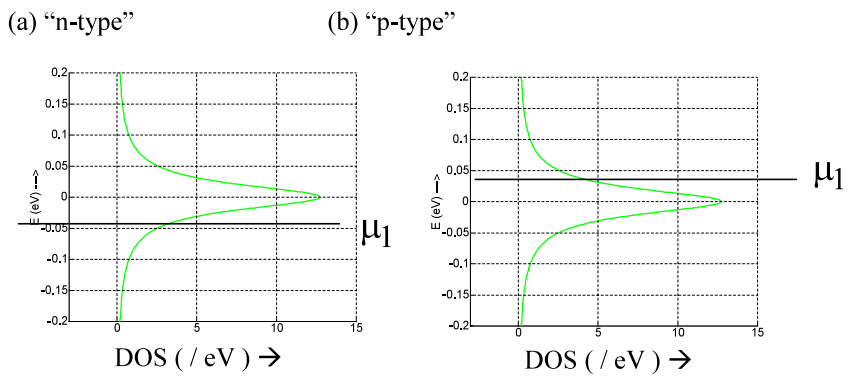


Figure 7.2. We can define (a) n- and (b) p-type conduction depending on whether the electrochemical potential lies on an up slope or a down slope of the DOS.

$$I_2 = \frac{q}{\hbar}\int_{-\infty}^{+\infty} \mathrm{d}E\, \gamma_2[D(E-U)f_2(E) - n(E)]. \qquad (7.5b)$$

At steady state, the sum of the two currents is equated to zero to eliminate $n(E)$:

$$I = \frac{q}{h}\int_{-\infty}^{+\infty} \mathrm{d}E\, \bar{T}(E-U)[f_1(E) - f_2(E)]$$
$$\text{where } \bar{T}(E) = D(E)2\pi\gamma_1\gamma_2/\gamma \qquad (7.6)$$

is called the *transmission*, a concept that plays a central role in the transmission formalism widely used in mesoscopic physics [9]. Note that the Fermi functions f_1 and f_2 are given by

$$f_1(E) = f_0(E-\mu_1) \qquad f_2(E) = f_0(E-\mu_2)$$
$$\text{where } f_0(E) \equiv (1+\exp(E/k_B T))^{-1} \qquad (7.7)$$

where the electrochemical potentials in the source and drain contacts are given by

$$\mu_1 = \mu \qquad \mu_2 = \mu - qV_D \qquad (7.8)$$

where μ is the equilibrium electrochemical potential.

7.1. Negative differential resistance (NDR)

To see how the model works, consider first a one-level device with a broadened DOS given by equation (7.1) with parameters as listed in figure 7.1. As we might expect, the current increases once the applied drain voltage is large enough that the energy level comes within the energy window between μ_1 and μ_2. The current then increases towards a maximum value of $(2q/\hbar)\gamma_1\gamma_2/(\gamma_1+\gamma_2)$ over a voltage range $\sim(\gamma_1+\gamma_2+k_B T)C_E/C_D$ as shown in figure 7.1(a). Here we have assumed

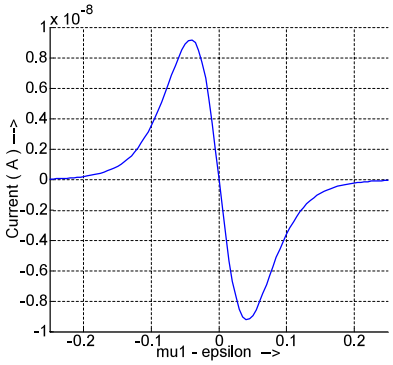


Figure 7.3. The thermoelectric current reverses direction from p-type ($\mu_1 < 0$) to n-type ($\mu_1 > 0$) samples. $\gamma_1 = \gamma_2 = 0.005$ eV, $k_B T_1 = 0.026$ eV and $k_B T_2 = 0.025$ eV.

the broadening due to the two contacts γ_1 and γ_2 to be constants equal to 0.005 eV.

Now suppose γ_1 is equal to 0.005 eV for $E > 0$, but is *zero* for $E < 0$ (γ_2 is still independent of energy and equal to 0.005 eV). The current–voltage characteristics now show negative differential resistance (NDR), that is, a drop in the current with an increase in the voltage, in one direction of applied voltage but not the other, as shown in figure 7.1(b). This simple model may be relevant to the experiment described in [6] though the nature and location of the molecular energy levels remain to be established quantitatively.

7.2. Thermoelectric effect

We have discussed the current that flows when a voltage is applied between the two contacts. In this case the current depends on the density of states near the Fermi energy and it does not matter whether the equilibrium Fermi energy μ_1 lies at the (a) lower end (n-type) or at the (b) upper end (p-type) of the density of states (see figure 7.2).

However, if we simply heat up one contact relative to the other so that $T_1 > T_2$ (with no applied voltage), a thermoelectric current will flow to which the direction will be different in case (a) and in case (b). To see this we could calculate the current from our model with $U = 0$ (there is no need to perform a self-consistent solution), $V_D = 0$ and $V_G = 0$, and with

$$f_1(E) \equiv \frac{1}{1 + \exp\left(\frac{E-\mu_1}{k_B T_1}\right)}$$

and

$$f_2(E) \equiv \frac{1}{1 + \exp\left(\frac{E-\mu_1}{k_B T_2}\right)}.$$

As shown in figure 7.3 the direction of the current is different for n- and p-type samples. This is of course a well-known result for bulk solids where hot point probes are routinely used to identify the type of conduction. But the point I am trying to make is that it is true even for ballistic samples and can be described by the elementary model described here [7].

7.3. Nanotransistor

As another example of the independent level model, let us model a nanotransistor [8] by writing the DOS as (see figure 7.4; W: width in the y-direction)

$$D(E) = m_c W L/\pi\hbar^2 \vartheta(E - E_c) \tag{7.9}$$

making use of the well-known result that the DOS per unit area in a large 2D conductor described by an electron effective mass is equal to $m_c/\pi\hbar^2$, for energies greater than the energy E_c of the conduction band edge. The escape rates can be written down assuming that electrons are removed by the contact with a velocity v_R:

$$\gamma_1 = \gamma_2 = \hbar v_R/L. \tag{7.10}$$

The current–voltage relations shown in figure 7.5 were obtained using these model parameters: $E_c = 0$, $\mu_1 = -0.2$ eV, $m_c = 0.25$ m, $C_G = 2\varepsilon_r\varepsilon_0 WL/t$, $C_S = C_D = 0.05\ C_G$, $W = 1\ \mu$m, $L = 10$ nm, insulator thickness $t = 1.5$ nm, $v_R = 10^7$ cm s^{-1}. At high drain voltages (V_D) the current saturates when μ_2 drops below E_c since there are no additional states to contribute to the current. Note that the gate capacitance C_G is much larger than the other capacitances, which helps to hold the channel potential fixed relative to the source as the drain voltage is increased (see equation (7.3)). Otherwise, the bottom of the channel density of states, E_c, will 'slip down' with respect to μ_1 when the drain voltage is applied, so the current will not saturate. The essential feature of a well-designed transistor is that the gate is much closer to the channel than 'L' allowing it to hold the channel potential constant despite the voltage V_D on the drain.

I should mention that our present model ignores the profile of the potential along the length of the channel, treating it as a little box with a single potential U given by equation (7.2). Nonetheless the results (figure 7.5) are surprisingly close to those of experiments/realistic models, because the current in well-designed nanotransistors is controlled by a small region in the channel near the source whose length can be a small fraction of the actual length L. Luckily we do not need to pin down the precise value of this fraction, since the present model gives the same current independently of L [8].

Ohm's law

It is natural to ask whether the independent level model would lead to Ohm's law if we were to calculate the low bias conductance of a large conductor of length L and cross-sectional area S. Since the current is proportional to the DOS, $D(E)$ (see equation (7.5)), which is proportional to the volume SL of the conductor, it might seem that the conductance $G \sim SL$. However, the coupling to the contacts decreases inversely with the length L of the conductor, since the longer a conductor is, the smaller is its coupling to the contact (see equation (7.10)). While the DOS goes up as the volume, the coupling to the contact goes down as $1/L$, so the conductance

$$G \sim SL/L = S.$$

But Ohm's law tells us that the conductance should scale as S/L; we are predicting that it should scale as 'S'. The reason is that we are really modelling a *ballistic* conductor, where electrons propagate freely, the only resistance arising from the contacts. The conductance of such a conductor is indeed independent of its length. The length dependence of

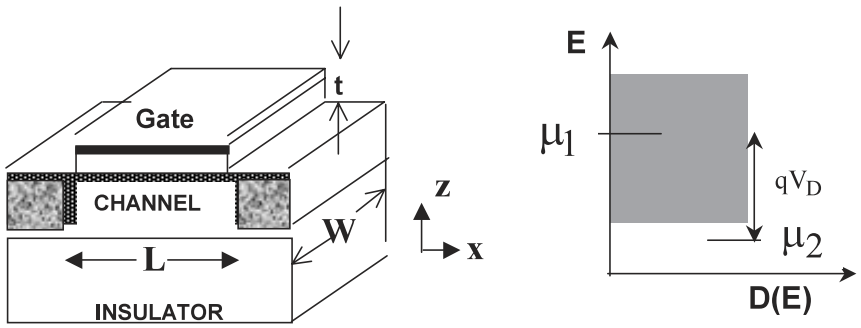


Figure 7.4. A nanotransistor: the physical structure and assumed density of states (DOS) in the channel region.

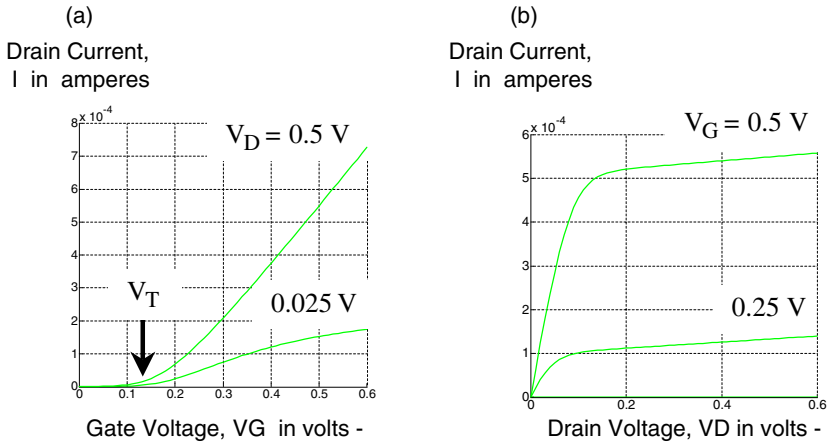


Figure 7.5. (a) The drain current (I) as a function of the gate voltage (V_G) for different values of the drain voltage (V_D); (b) the drain current as a function of the drain voltage for different values of the gate voltage.

the conductance comes from scattering processes within the conductor that are not yet included in our thinking [9].

For example, in a uniform channel the electronic wavefunction is spread out uniformly. But a scatterer in the middle of the channel could split up the wavefunctions into two pieces, one on the left and one on the right, with different energies. One has a small γ_2 while the other has a small γ_1, and so neither conducts very well. This *localization* of wavefunctions would seem to explain why the presence of a scatterer contributes to the resistance, but to get the story quantitatively correct it is in general necessary to go beyond the independent level model to account for interference between multiple paths. This requires a model that treats γ as a *matrix* rather than as simple numbers.

Such 'coherent' scatterers lead to many interesting phenomena, but not to Ohm's law: $R \sim 1/L$. The full story requires us to include phase-breaking scattering processes that cause a change in the state of an external object. For example, if an electron gets deflected by a rigid (that is unchangeable) defect in the lattice, the scattering is said to be coherent. But, if the electron transfers some energy to the atomic lattice causing it to start vibrating, that would constitute a phase-breaking or incoherent process. Purely coherent scatterers can give rise to a measurable resistance R, but cannot give rise to any dissipation, since no energy is removed from the electrons. Indeed there is experimental evidence that the associated Joule heating (I^2R) occurs in the contacts outside the channel, allowing experimentalists to pump a lot more current through a small conductor without burning it up.

Much of the work on small conductors is usually in the coherent limit, but it is clear that including phase-breaking scattering will be important in developing quantitative models. In section 7.4 I will show how this can be done within our simple one-level model. This will lead naturally to the non-equilibrium Green function (NEGF) formalism described in section 8.

7.4. Inelastic spectroscopy

For the purpose of including phase breaking it is useful to recast the equations listed at the beginning of this section in a slightly different form by defining a Green function G:

$$G = \frac{1}{E - \varepsilon - U + (\mathrm{i}\gamma/2)} \qquad \text{where } \gamma = \gamma_1 + \gamma_2 \quad (7.11)$$

such that

$$2\pi D(E) = G(E)\gamma(E)G^*(E) = \mathrm{i}[G - G^*]. \quad (7.12)$$

The electron density can then be written as (cf equation (7.4))

$$2\pi n(E) = G(E)\gamma^{\mathrm{in}}(E)G^*(E) \quad (7.13)$$

in terms of the in-scattering function defined as $\gamma^{\mathrm{in}} = \gamma_1^{\mathrm{in}} + \gamma_2^{\mathrm{in}}$, where

$$\gamma_1^{\mathrm{in}} = \gamma_1 f_1 \qquad \text{and} \qquad \gamma_2^{\mathrm{in}} = \gamma_2 f_2. \quad (7.14a)$$

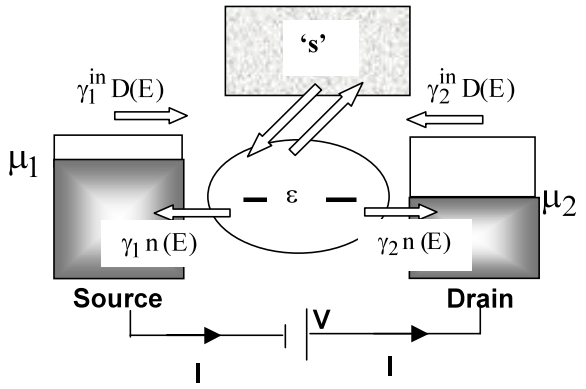


Figure 7.6. Phase-breaking scattering processes can be visualized as a fictitious terminal 's' with its own in-scattering and out-scattering functions.

It is also useful to define an out-scattering function $\gamma^{out} = \gamma_1^{out} + \gamma_2^{out}$, where

$$\gamma_1^{out} = \gamma_1(1-f_1) \qquad \text{and} \qquad \gamma_2^{out} = \gamma_2(1-f_2). \quad (7.14b)$$

Note that

$$\gamma_i = \gamma_i^{out} + \gamma_i^{in}. \quad (7.15)$$

Subtracting equation (7.13) from (7.12) we obtain

$$2\pi p(E) = G(E)\gamma^{out}(E)G^*(E) \quad (7.16)$$

for the *hole density*

$$p(E) = D(E) - n(E) \quad (7.17)$$

obtained by subtracting the electron density from the density of states.

Phase-breaking scattering processes can be visualized as a fictitious terminal 's' with its own in-scattering and out-scattering functions, so

$$\gamma^{in} = \gamma_1^{in} + \gamma_2^{in} + \gamma_s^{in} \quad (7.18a)$$

$$\gamma^{out} = \gamma_1^{out} + \gamma_2^{out} + \gamma_s^{out}. \quad (7.18b)$$

The current (per spin) at any terminal 'i' can be calculated from

$$I_i = (q/\hbar)\int_{-\infty}^{+\infty} dE\, \tilde{I}_i(E) \quad (7.19)$$

with

$$\tilde{I}_i = [\gamma_i^{in} D] - [\gamma_i n]. \quad (7.20)$$

To find γ_s^{in} and γ_s^{out}, one approach is to view the scattering terminal 's' like a real terminal whose electrochemical potential μ_s is adjusted to make the current $I_s = 0$, following the phenomenological approach widely used in mesoscopic physics [9e]. The scattering terminal, however, cannot in general be described by a Fermi function that we can use in equations (7.14a), (7.14b). The NEGF formalism allows us to evaluate γ_s^{in} and γ_s^{out} to any desired approximation from a microscopic theory. In the self-consistent Born approximation,

$$\gamma_s^{in}(E) = \int d(\hbar\omega)\, D^{ph}(\hbar\omega) n(E+\hbar\omega) \quad (7.21a)$$

and

$$\gamma_s^{out}(E) = \int d(\hbar\omega)\, D^{ph}(\hbar\omega) p(E-\hbar\omega) \quad (7.21b)$$

where the 'phonon' spectral function can be written as the sum of an emission term (positive frequencies) and an absorption term (negative frequencies):

$$D^{ph}(\hbar\omega) = \sum_i D_i[(N_i+1)\delta(\hbar\omega - \hbar\omega_i) + N_i\delta(\hbar\omega + \hbar\omega_i)] \quad (7.22)$$

with N_i representing the number of phonons of frequency $\hbar\omega_i$, and D_i its coupling. We assume N_i to be given by the Bose–Einstein factor, but it is conceivable that the phonons could be driven off equilibrium, requiring N_i to be evaluated from a transport equation for the phonons. Low frequency phonons with $\hbar\omega_i$ much smaller than other relevant energy scales can be treated as elastic scatterers with $\hbar\omega_i \sim 0$, $D_i(N_i+1) \approx D_i N_i \equiv D_0^{ph}$. Equations (7.21) then simplify to

$$\gamma_s^{in} = D_0^{ph} n(E) \qquad \text{and} \qquad \gamma_s^{out} = D_0^{ph} p(E)$$
$$\text{so } \gamma_s = \gamma_s^{in} + \gamma_s^{out} = D_0^{ph} D(E). \quad (7.23)$$

Figure 7.7 shows a simple example where the energy level $\varepsilon = 5$ eV lies much above the equilibrium electrochemical potential $\mu = 0$, so current flows by tunnelling. The currents calculated without any phonon scattering (all $D_i = 0$) and with phonon scattering ($D_1 = 0.5$, $\hbar\omega_1 = 0.075$ eV and $D_2 = 0.7$, $\hbar\omega_2 = 0.275$ eV) show no discernible difference. The difference, however, shows up in the conductance dI/dV where there is a discontinuity proportional to D_i when the applied voltage equals the phonon frequency $\hbar\omega_i$. This discontinuity shows up as peaks in d^2I/dV^2 whose location along the voltage axis corresponds to molecular vibration quanta, and this is the basis of the field of inelastic electron tunnelling spectroscopy (IETS) [10].

Note that the above prescription for including inelastic scattering (equations (7.21), (7.22)) is based on the NEGF formalism. This is different from many common theories where exclusion principle factors $(1-f)$ appropriate to the contacts are inserted somewhat intuitively and as such cannot be applied to long devices; by contrast the NEGF prescription can be extended to long devices by replacing numbers with matrices as we will describe in the next section. Indeed as we mentioned in the introduction, what we have described so far can be viewed as a special case of the NEGF formalism applied to a device so small that it is described by a single energy level or a '(1×1) Hamiltonian matrix'. Let us now look at the general formalism.

8. From numbers to matrices: NEGF formalism

The one-level model serves to identify the important concepts underlying the flow of current through a conductor, such as the location of the equilibrium *electrochemical potential* μ relative to the *density of states* $D(E)$, the *broadening* of the level $\gamma_{1,2}$ due to the coupling to contacts 1 and 2 etc. In the general model for a multilevel conductor with 'n' energy levels, all the quantities we have introduced are replaced by a corresponding matrix of size $(n \times n)$:

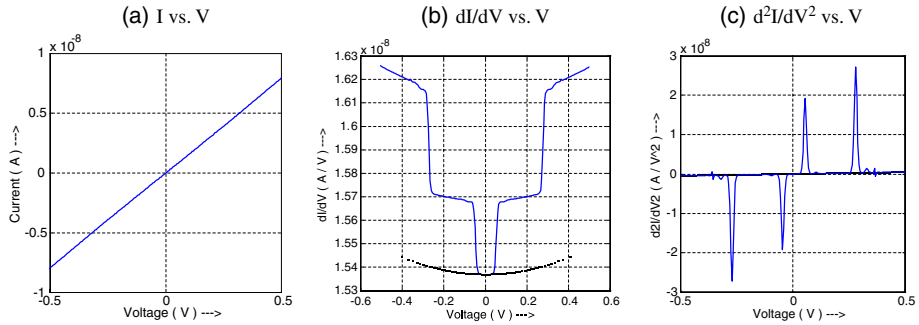


Figure 7.7. (a) Current (I), (b) conductance ($\mathrm{d}I/\mathrm{d}V$) and (c) $\mathrm{d}^2I/\mathrm{d}V^2$ as a function of voltage calculated without phonon scattering (dashed line) and with scattering by phonons (solid curve) with two distinct frequencies having slightly different coupling strengths ($D_1 = 0.5$, $\hbar\omega_1 = 0.075$ eV and $D_2 = 0.7$, $\hbar\omega_2 = 0.275$ eV).

$\varepsilon \to [H]$	*Hamiltonian* matrix
$\gamma_\mathrm{i} \to [\Gamma_\mathrm{i}(E)]$	*Broadening* matrix
$2\pi D(E) \to [A(E)]$	*Spectral function*
$2\pi n(E) \to [G^n(E)]$	*Correlation function*
$2\pi p(E) \to [G^p(E)]$	*Hole correlation function*
$U \to [U]$	*Self-consistent potential* matrix
$N \to [\rho] = \int(\mathrm{d}E/2\pi)[G^n(E)]$	*Density matrix*
$\gamma_\mathrm{i}^\mathrm{in} \to [\Sigma_\mathrm{i}^\mathrm{in}(E)]$	*In-scattering* matrix
$\gamma_\mathrm{i}^\mathrm{out} \to [\Sigma_\mathrm{i}^\mathrm{out}(E)]$	*Out-scattering* matrix

Actually, the effect of the contacts is described by a '*self-energy*' *matrix*, $[\Sigma_{1,2}(E)]$, whose anti-Hermitian part is the broadening matrix: $\Gamma_{1,2} = \mathrm{i}[\Sigma_{1,2} - \Sigma_{1,2}^+]$. The Hermitian part effectively adds to $[H]$, thereby shifting the energy levels—an effect we ignored in the simple model. The Hermitian and anti-Hermitian parts are Hilbert transform pairs. Also, I should mention that I have used

$$G^n(E), \qquad G^p(E), \qquad \Sigma^\mathrm{in}(E), \qquad \Sigma^\mathrm{out}(E)$$

to denote what is usually written in the literature [11, 12] as

$$-\mathrm{i}G^<(E), \qquad +\mathrm{i}G^>(E), \qquad -\mathrm{i}\Sigma^<(E), \qquad +\mathrm{i}\Sigma^>(E),$$

in order to emphasize their physical significance.

The *NEGF equations* for dissipative quantum transport look much like those discussed in section 7.4, but with numbers replaced by matrices:

$$G^n = G\Sigma^\mathrm{in}G^+ \tag{8.1}$$

$$G = [EI - H_0 - U - \Sigma]^{-1} \tag{8.2}$$

$$A = \mathrm{i}[G - G^+] \qquad \Gamma = \mathrm{i}[\Sigma - \Sigma^+] \tag{8.3}$$

where

$$\begin{aligned} \Sigma^\mathrm{in} &= \Sigma_1^\mathrm{in} + \Sigma_2^\mathrm{in} + \Sigma_\mathrm{s}^\mathrm{in} \\ \Sigma &= \Sigma_1 + \Sigma_2 + \Sigma_\mathrm{s}. \end{aligned} \tag{8.4}$$

These equations can be used to calculate the correlation function G^n and hence the density matrix ρ whose diagonal elements give us the electron density:

$$\rho = \int \mathrm{d}E\, G^n(E)/2\pi. \tag{8.5}$$

The current (per spin) at any terminal 'i' can be calculated from

$$I_\mathrm{i} = (q/\hbar)\int_{-\infty}^{+\infty} \mathrm{d}E\, \tilde{I}_\mathrm{i}(E)/2\pi \tag{8.6}$$

with

$$\tilde{I}_\mathrm{i} = \mathrm{Tr}[\Sigma_\mathrm{i}^\mathrm{in}A] - \mathrm{Tr}[\Gamma_\mathrm{i}G^n] \tag{8.7}$$

which is shown in figure 8.1 in terms of an inflow ($\Sigma_\mathrm{i}^\mathrm{in}A$) and an outflow ($\Gamma_\mathrm{i}G^n$). The full time-dependent versions of these equations are derived in sections B.2, B.3 and B.4 from which the steady-state versions stated above are obtained.

Input parameters

To use these equations, we need a channel Hamiltonian $[H_0]$ and the in-scattering $[\Sigma^\mathrm{in}]$ and broadening $[\Gamma]$ functions. For the two contacts, these are related:

$$\Sigma_1^\mathrm{in} = \Gamma_1 f_1 \qquad \text{and} \qquad \Sigma_2^\mathrm{in} = \Gamma_2 f_2 \tag{8.8}$$

and the broadening/self-energy for each contact can be determined from a knowledge of the surface spectral function (a)/surface Green function (g) of the contact and the matrices $[\tau]$ describing the channel contact coupling:

$$\Gamma = \tau a \tau^+ \qquad \text{and} \qquad \Sigma = \tau g \tau^+. \tag{8.9}$$

Finally one needs a model (Hartree–Fock, density functional theory etc) for relating the self-consistent potential U to the density matrix. This aspect of the problem needs further work, since not much of the work in quantum chemistry has been geared towards transport problems.

Scattering contact

The NEGF equations without the 's' contact are often used to analyse small devices and in this form it is identical to the result obtained by Meir and Wingreen (see equation (6) of [12b]). The third 'contact' labelled 's' represents scattering processes, without which we cannot make the transition to Ohm's law. Indeed it is only with the advent of mesoscopic physics in the 1980s that the importance of the contacts (Γ_1 and Γ_2) in interpreting experiments became widely recognized.

Prior to that, it was common to ignore the contacts as minor experimental distractions and try to understand the physics of conduction in terms of the 's' contact, though no one (to my knowledge) thought of scattering as a 'contact'

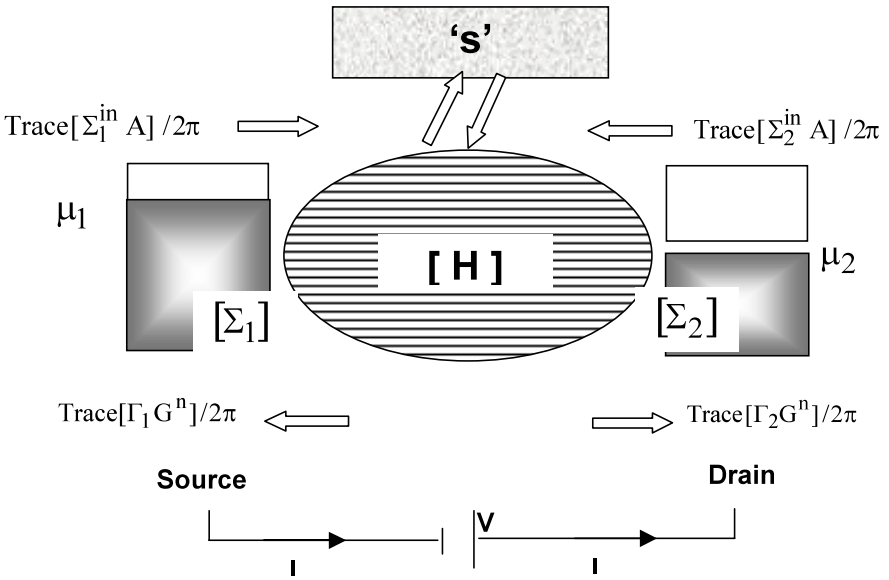


Figure 8.1. (See figure 7.6.) From numbers to matrices: the general matrix model, based on the NEGF formalism. Without the 's contact' this model is equivalent to equation (6) of [12b]. The 's contact' distributed throughout the channel describes incoherent scattering processes [12c]. In general this 'contact' cannot be described by a Fermi function unlike the real contacts.

until Buttiker introduced the idea phenomenologically in the mid-1980s (see [9e]). Subsequently, it was shown [12c] from a microscopic model that incoherent scattering processes in the NEGF method act like a fictitious 'contact' distributed throughout the channel that extracts and reinjects electrons. Like the real contacts, coupling to this 'contact' too can be described by a broadening matrix Γ_s. However, unlike the real contacts, the scattering contact in general cannot be described by a Fermi function so, although the outflow is given by $\mathrm{Tr}[\Gamma_s G^n/2\pi]$, the inflow is more complicated. For the scattering 'terminal', unlike the contacts, there is no simple connection between Σ_s^{in} and Σ_s (or Γ_s). Moreover, these quantities are related to G^n and have to be computed self-consistently. The relevant equations derived in section B.4 can be viewed as the matrix versions of equations (7.21*a*) and (7.21*b*) [17].

Derivation of NEGF equations

The full set of equations are usually derived using the non-equilibrium Green function (NEGF) formalism, also called the Keldysh or the Kadanoff–Baym formalism initiated by the works of Schwinger, Baym, Kadanoff and Keldysh in the 1960s. However, their work was motivated largely by the problem of providing a systematic perturbative treatment of electron–electron interactions, a problem that demands the full power of this formalism. By contrast, we are discussing a much simpler problem, with interactions treated only to lowest order.

Indeed it is quite common to ignore interactions completely (except for the self-consistent potential) assuming 'coherent transport'. The NEGF equations for coherent transport can be derived from a one-electron Schrödinger equation without the advanced formal machinery [12d]. We start by partitioning the Schrödinger equation into three parts, the channel and the source and drain contacts (figure 8.2):

$$i\hbar\frac{d}{dt}\begin{Bmatrix}\Phi_s\\ \psi\\ \Phi_D\end{Bmatrix}=\begin{bmatrix}H_s+i\eta & \tau_S^+ & 0\\ \tau_S & H & \tau_D\\ 0 & \tau_D^+ & H_D+i\eta\end{bmatrix}\begin{Bmatrix}\Phi_s\\ \psi\\ \Phi_D\end{Bmatrix} \tag{8.10}$$

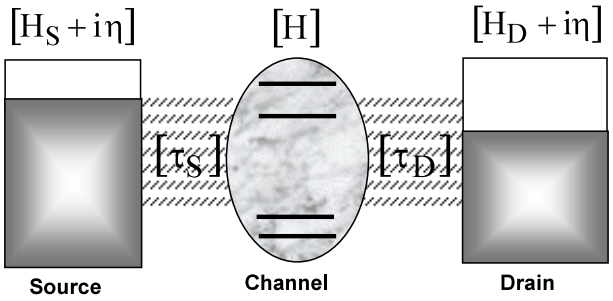


Figure 8.2. A channel connected to two contacts.

with an infinitesimal $i\eta$ added to represent the extraction and injection of electrons from each of the contacts.

It is possible to eliminate the contacts, to write a Schrödinger-like equation for the channel alone:

$$i\hbar\frac{d\psi}{dt}-H\psi-\underbrace{\Sigma\psi}_{\text{Outflow}}=\underbrace{S}_{\text{Inflow}} \tag{8.11}$$

with an additional self-energy term '$\Sigma\psi$' and a source term 'S' that give rise to outflow and inflow respectively. Note that unlike $[H]$, the self-energy $[\Sigma]$ is non-Hermitian and gives rise to an outflow of electrons. The additional terms in equation (8.11) are reminiscent of the frictional term and the noise term added to Newton's law to obtain the Langevin equation

$$m\frac{dv}{dt}+\underbrace{\gamma v}_{\text{Friction}}=\underbrace{F}_{\text{External force}}+\underbrace{N(t)}_{\text{Noise}}$$

describing a Brownian particle [13]. Equivalently, one can move to a collective picture and balance inflow with outflow to obtain the Boltzmann equation. With quantum dynamics too we can express the inflow and outflow in terms of the correlation functions: $G^n\sim\psi\psi^+$, $\Sigma^{in}\sim SS^+$ and relate them

to obtain the NEGF equations (sometimes called the quantum Boltzmann equation).

Beyond the one-electron picture

A proper derivation of the NEGF equations, however, requires us to go beyond this one-electron picture, especially if non-coherent processes are involved. For example, the self-energy term '$\Sigma\psi$' in equation (8.11) represents the outflow of the electrons and it is natural to ask whether Σ (whose imaginary part gives the broadening or the inverse lifetime) should depend on whether the final state (to which outflow occurs) is empty or full. Such exclusion principle factors do not appear as long as purely coherent processes are involved. But they do arise for non-coherent interactions in a non-obvious way that is hard to rationalize from the one-electron picture.

In the one-electron picture, individual electrons are described by a one-electron wavefunction ψ and the electron density is obtained by summing $\psi^*\psi$ from different electrons. A more comprehensive viewpoint describes the electrons in terms of field operators 'c' such that 'c^+c' is the number operator which can take on one of two values '0' or '1' indicating whether a state is empty or full. These 'second-quantized' operators obey differential equations

$$i\hbar \frac{d}{dt}c - Hc - \Sigma c = S \tag{8.12}$$

that look much like the ones describing one-electron wavefunctions (see equation (8.11)). But unlike $\psi^*\psi$ which can take on any value, operators such as c^+c can only take on one of two values '0' or '1', thereby reflecting a particulate aspect that is missing from the Schrödinger equation. This advanced formalism is needed to progress beyond coherent quantum transport to inelastic interactions and onto more subtle many-electron phenomena such as the Kondo effect.

A derivation of equation (8.12) leading to the NEGF equations is provided in appendix B using second quantization for the benefit of advanced readers. However, in this derivation I have not used advanced concepts such as the 'Keldysh contour' which are needed for a systematic treatment of higher order processes. While future works in the field will undoubtedly require us to go beyond the lowest order treatment discussed here, it is not clear whether a higher order perturbative treatment will be useful or whether non-perturbative treatments will be required that describe the transport of composite or dressed particles obtained by appropriate unitary transformations of the bare electron operator 'c'.

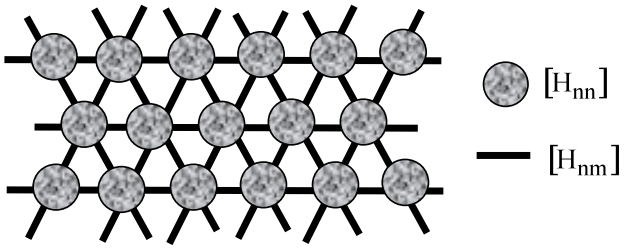
9. Open questions

Let me end by listing what I see as the open questions in the field of nanoscale electronic transport.

Model Hamiltonian

Once the matrices $[H]$ and $[\Sigma]$ are known the NEGF equations provide a well-defined prescription for calculating the current–voltage characteristics. For concrete calculations one needs to adopt a suitable basis such as tight-binding/Huckel/extended Huckel/Gaussian described in the literature [14] in order to write down the matrices $[H]$ and $[\Sigma]$.

We could visualize the Hamiltonian $[H]$ as a network of unit cells described by matrices H_{nn} whose size $(b \times b)$ is determined by the number of basis functions (b) per unit cell. Different unit cells are coupled through the 'bond matrices' $[H_{nm}]$.



The overall size of $[H]$ is $(Nb \times Nb)$, N being the number of unit cells. The self-energy matrix $[\Sigma]$ is also of the same size as $[H]$, although it represents the effect of the infinite contacts. It can be evaluated from a knowledge of the coupling matrices $[\tau_s]$ and $[\tau_D]$ (see figure 8.2) and the surface properties of the contacts, as expressed through its surface Green function (see equation (8.9)). The matrices $[H]$ and $[\Sigma]$ thus provide a kind of intellectual partitioning: $[H]$ expresses the properties of the channel while $[\Sigma]$ depends on the interface with the contacts. In specific problems it may be desirable to borrow $[H]$ and $[\Sigma]$ from two different communities (such as quantum chemists and surface physicists), but the process is made difficult by the fact that they often use different basis functions and self-consistent fields (see below). Much work remains to be done along these lines. Indeed, sometimes it may not even be clear where the channel ends and the contact begins!

Transient transport

Most of the current work to date has been limited to steady-state transport, but it is likely that future experiments will reveal transient effects whose time constants are controlled by the quantum dynamics, rather than circuit or RC effects [18]. The time-dependent NEGF equations [19] should be useful in modelling such phenomena.

Self-consistent field

An important conceptual issue in need of clarification is the treatment of electron–electron interactions. Discovering an appropriate self-consistent field $U(N)$ to replace our simple ansatz (cf equation (5.1*b*))

$$U(N) = q^2[N - N_0]/C_E$$

is arguably one of the central topics in many-electron physics. Quantum chemists have developed sophisticated models for the self-consistent field such as Hartree–Fock (HF) and density functional theory (DFT) in addition to a host of semi-empirical approaches which can all give very different energy level structures. A lot of work has gone into optimizing these models but largely with respect to ground-state calculations and it is not clear what the best choice is for electron transport problems.

One could argue that electron transport involves adding and removing electrons and as such one should be looking at difference between the energies of the $(N \pm 1)$-electron system relative to the ground state of the N-electron system. However,

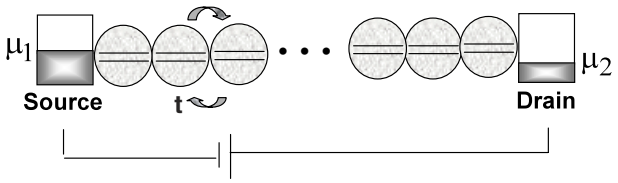


Figure 9.1. A large conductor can be viewed as an array of unit cells. If the conductor is extended in the transverse plane, we should view each unit cell as representing an array of unit cells in the transverse direction.

for large broadening, the wavefunctions are significantly delocalized from the channel into the contacts, so the number of electrons in the channel can change by fractional amounts. The best choice of a self-consistent field for transport problems requires a careful consideration of the degree of delocalization as measured by the relative magnitudes of the broadening and the charging.

Transport regimes

In this context it is useful to distinguish broadly between three different transport regimes for small conductors depending on the degree of delocalization.

Self-consistent field (SCF) regime. If the thermal energy k_BT and/or the broadening γ are comparable to the single-electron charging energy U_0, we can use the SCF method described in this article. However, the optimum choice of the self-consistent potential needs to be clarified.

Coulomb blockade (CB) regime. If U_0 is well in excess of both k_BT and γ, the SCF method is not adequate, at least not the restricted one. More correctly, one could use (if practicable) the multielectron master equation described in appendix A [15].

Intermediate regime. If U_0 is comparable to the larger of k_BT, γ, there is no simple approach: the SCF method does not do justice to the charging, while the master equation does not do justice to the broadening and a different approach is needed to capture the observed physics [16].

With large conductors too we can envision three regimes of transport that evolve out of these three regimes. We could view a large conductor as an array of unit cells as shown in figure 9.1. The inter-unit coupling energy 't' has an effect somewhat (but not exactly) similar to the broadening 'γ' that we have associated with the contacts. If $t \geqslant U_0$, the overall conduction will be in the *SCF* regime and can be treated using the method described here. If $t \ll U_0$, it will be in the *CB* regime and can in principle be treated using the multielectron master equation under certain conditions (specifically if 't' is much less than the level broadening γ_s). On the other hand, large conductors with $\gamma_s \ll t \leqslant U_0$ belong to an intermediate regime that presents major theoretical challenges [20], giving rise to intriguing possibilities. Indeed many believe that the high T_c superconductors (whose microscopic theory is yet to be discovered) consist of unit cells whose coupling is delicately balanced at the borderline of the SCF and the CB regimes.

I believe that the field of nanoelectronics is currently at a very exciting stage where important advances can be expected from both applied and basic points of view. We will continue to acquire a better quantitative understanding of nanoscale devices based on nanowires, nanotubes, molecules and other nanostructured materials. Although many of the observations appear to be described well within the basic self-consistent field model discussed here, much remains to be done in terms of discovering better basis functions for representing the Hamiltonian $[H]$ and self-energy $[\Sigma]$ matrices (see figure 8.1), including inelastic scattering processes and implementing more efficient algorithms for the solution of the quantum transport equations. At the same time we can hope to discover new quantum transport phenomena (both steady state and time dependent) involving strong electron–phonon and electron–electron interactions, which are largely unexplored. A notable exception is the Coulomb blockade arising from strong electron–electron interactions which is fairly well understood. In appendix A, I have tried to provide a brief introduction to this transport regime and relate it to the self-consistent field regime that forms the core of this tutorial. But all this I believe represents the 'tip of the iceberg'. The progress of molecular electronics should lead to greater control of the degree of hybridization between the localized strongly interacting molecular states and the delocalized contact states, thereby allowing a systematic study of different transport regimes. Such a study should reveal many more subtle phenomena involving electrons 'dressed' by a variety of strong interactions that require non-perturbative treatments far beyond those described here.

Acknowledgments

It is a pleasure to thank my colleagues Ron Reifenberger, Magnus Paulsson, Mark Lundstrom and Avik Ghosh for looking at the preliminary versions of this manuscript and providing me with useful feedback.

In keeping with the tutorial spirit of this article, I have only listed a few related references that the reader may find helpful in clarifying the concepts and viewpoint presented here. This is not intended to provide a comprehensive (or even representative) list of the extensive literature in the field of quantum transport.

References

[1] The viewpoint presented here is discussed in more detail in a forthcoming book: Datta S 2004 *Quantum Transport: Atom to Transistor* (Cambridge University Press) based on a graduate course (see http://dynamo.ecn.purdue.edu/~datta)

[2a] For a discussion of the interpretation of one-particle energy levels, see for example, Brus L E 1983 A simple model for the ionization potential, electron affinity, and aqueous redox potentials of small semiconductor crystallites *J. Chem. Phys.* **79** 5566

[2b] More recent references include for example, Bakkers E P A M, Hens Z, Zunger A, Franceschetti A, Kouwenhoven L P, Gurevich L and Vanmaekelbergh D 2001 Shell-tunneling spectroscopy of the single-particle energy levels of insulating quantum dots *Nano Lett.* **1** 551

[2c] Niquet Y M, Delerue C, Allan G and Lannoo M 2002 Interpretation and theory of tunneling experiments on single nanostructures *Phys. Rev.* B **65** 165334

[3] Conductance quantization is usually discussed in terms of $E(k)$ diagrams. I have not seen it discussed for a single level as we do in section 4 by relating the broadening to the lifetime. But the quantization has been related to the 'uncertainty principle': Batra I P 2002 From uncertainty to certainty in quantum conductance of nanowires *Solid State Commun.* **124** 463–7 (An unpublished elaboration of this argument due to M P Anantram is available on request)

[4] The example used to illustrate the importance of the potential profile in determining the current–voltage characteristics in section 5 is related to the experiments discussed in Datta S, Tian W, Hong S, Reifenberger R, Henderson J and Kubiak C P 1997 STM current–voltage characteristics of self-assembled monolayers (SAM's) *Phys. Rev. Lett.* **79** 2530

[5a] The word 'quantum capacitance' was probably first introduced by Luryi S 1988 Quantum capacitance devices *Appl. Phys. Lett.* **52** 501 and has been used by other authors

[5b] See, for example, Katayama Y and Tsui D C 1993 Lumped circuit model of two-dimensional tunneling transistors *Appl. Phys. Lett.* **62** 2563

[6] Guisinger N P, Greene M E, Basu R, Baluch A S and Hersam M C 2004 Room temperature negative differential resistance through individual organic molecules on silicon surfaces *Nano Lett.* **4** 55
See also the article by
Guisinger N P, Basu R, Greene M E, Baluch A S and Hersam M C 2004 *Nanotechnology* **15** S452–8

[7] Paulsson M and Datta S 2003 Thermoelectric effects in molecular electronics *Phys. Rev.* B **67** 241403(R)

[8] The nanotransistor is essentially the same as that described in detail in Rahman A, Guo J, Datta S and Lundstrom M 2003 Theory of ballistic transistors *IEEE Trans. Electron Devices* **50** 1853 (Special Issue on Nanoelectronics)

[9] For more extensive discussions of nanoscale conduction see, for example

[9a] Datta S 1995 *Electronic Transport in Mesoscopic Systems* (Cambridge University Press)

[9b] Imry Y 1997 *Introduction to Mesoscopic Physics* (Oxford University Press)

[9c] Ferry D K and Goodnick S M 1997 *Transport in Nanostructures* (Cambridge University Press)

[9d] Beenakker C W J 1997 Random matrix theory of quantum transport *Rev. Mod. Phys.* **69** 731–808

[9e] Buttiker M 1988 Symmetry of electrical conduction *IBM J. Res. Dev.* **32** 317

[10] See, for example, Wolf E L 1989 *Principles of Electron Tunneling Spectroscopy* (Oxford: Oxford Science Publications)

[11a] For a review of the classic work on the NEGF formalism as applied to infinite homogeneous media, see, for example, Mahan G D 1987 Quantum transport equation for electric and magnetic fields *Phys. Rep.* **145** 251 and references therein

[11b] Recent texts on NEGF formalism include Haug H and Jauho A P 1996 *Quantum Kinetics in Transport and Optics of Semiconductors* (Berlin: Springer) (See also, chapter 8 of [9a])

[12] Many authors have applied the NEGF formalism to problems involving finite structures. The description presented here is based primarily on

[12a] Caroli C, Combescot R, Nozieres P and Saint-James D 1972 A direct calculation of the tunneling current: IV. Electron–phonon interaction effects *J. Phys. C: Solid State Phys.* **5** 21

[12b] Meir Y and Wingreen N S 1992 Landauer formula for the current through an interacting electron region *Phys. Rev. Lett.* **68** 2512

[12c] Datta S 1989 Steady-state quantum kinetic equation *Phys. Rev.* B **40** 5830

[12d] For a derivation of the NEGF equations from a one-particle viewpoint, see [9a], chapters 3, 8. See also a tutorial by Paulson M (*Preprint* cond-mat/0210519)

[13] See, for example, McQuarrie D A 1976 *Statistical Mechanics* (New York: Harper and Row) chapter 20

[14a] NEGF-based models for 1-D semiconductor devices have been extensively developed by the 'NEMO' group and are available for public use. See, for example, Chris Bowen R, Klimeck G, Lake R, Frensley W R and Moise T 1997 Quantitative resonant tunneling diode simulation *J. Appl. Phys.* **81** 3207 (See also http://hpc.jpl.nasa.gov/PEP/gekco/nemo)

[14b] NEGF-based models formalism are also being developed for nanowires, nanotubes and molecules. For a tutorial introduction to the Huckel method for molecular conductors see, Zahid F, Paulsson M and Datta S 2003 *Advanced Semiconductors and Organic Nanotechniques* ed H Morkoc (Amsterdam: Elsevier Science) chapter (Electrical Conduction through Molecules) and references therein (Other review articles by our group are listed on the Website given in [1])

[15] For further reading on the Coulomb blockade regime, see, for example

[15a] Kastner M 1993 Artificial atoms *Phys. Today* **46** 24

[15b] Likharev K 1999 Single-electron devices and their applications *Proc. IEEE* **87** 606

[15c] Beenakker C W J 1991 Theory of Coulomb blockade oscillations in the conductance of a quantum dot *Phys. Rev.* B **44** 1646

[15d] Kouwenhoven L P and McEuen P L 1997 *Nano-Science and Technology* ed G Timp (New York: AIP) chapter 13 (Single Electron Transport through a Quantum Dot)

[15e] Bonet E, Deshmukh M M and Ralph D C 1992 Solving rate equations for electron tunneling via discrete quantum states *Phys. Rev.* B **65** 045317

[16] For further reading about the Kondo resonance observed in this transport regime, see, for example

[16a] Kouwenhoven L and Glazman L 2001 Revival of the Kondo effect *Phys. World* (January) 33

[16b] Fulde P 1991 *Electron Correlations in Molecules and Solids* (Berlin: Springer)

[16c] The Green's function equations described for the Kondo resonance in appendix B.5 are the same as those presented in Meir Y, Wingreen N S and Lee P A 1991 Transport through a strongly interacting electron system: theory of periodic conductance oscillations *Phys. Rev. Lett.* **66** 3048

[17] The lowest order treatment of electron–phonon interaction described in appendix B.4 can be compared to that described in [12a]. If we assume both electron and phonon distributions to be in equilibrium, our results can be shown to reduce to what is known as Migdal's 'Theorem' in the theory of electron–phonon interactions in metals. See for example, section II of Allen P B and Mitrovic B 1982 Theory of Superconducting T_c *Solid State Physics* vol 37, ed H Ehrenreich, F Seitz and D Turnbull (New York: Academic) p 1 (See equation (3.47), p 20)

[18] See for example, Fedorets D, Gorelik L Y, Shekhter R I and Jonson M 2002 Vibrational instability due to coherent tunneling of electrons *Europhys. Lett.* **58** 99

[19] The time-dependent equations in appendix B.3 can be compared to those in Jauho A P, Wingreen N S and Meir Y 1994 Time-dependent transport in interacting and non-interacting resonant tunneling systems *Phys. Rev.* B **50** 5528

[20] See for example, Georges A 2004 Strongly correlated electron materials: Dynamical mean-field theory and electronic structure (*Preprint* cond-mat/0403123)