1 Prologue: an atomistic view of electrical resistance

Let me start with a brief explanation since this is not a typical "prologue." For one it is too long, indeed as long as the average chapter. The reason for this is that I have a very broad objective in mind, namely to review *all* the relevant concepts needed to understand current flow through a very small object that has only one energy level in the energy range of interest. Remarkably enough, this can be done without invoking any significant background in quantum mechanics. What requires serious quantum mechanics is to understand where the energy levels come from and to describe large conductors with multiple energy levels. Before we get lost in these details (and we have the whole book for it!) it is useful to understand the factors that influence the current–voltage relation of a really small object.

This "bottom-up" view is different from the standard "top-down" approach to electrical resistance. 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. I believe the reason for this "top-down" approach is historical. Till 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 a battery can be connected. No one knew how to attach contact pads to a small molecule till 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 simple, except for unusual things like the Kondo effect that are seen only for a special range of parameters. Of course, it is quite likely that many 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

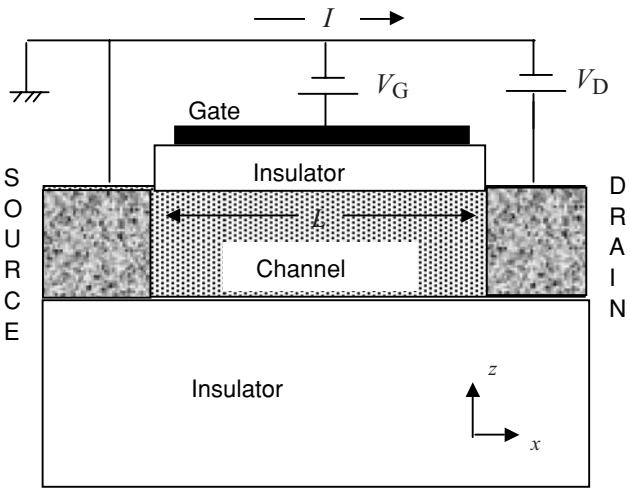

Fig. 1.1 Sketch of a nanoscale field effect transistor. The insulator should be thick enough to ensure that no current flows into the gate terminal, but thin enough to ensure that the gate voltage can control the electron density in the channel.

word" since the traditional top-down approach tends to obscure the simple physics of very small conductors.

The generic structure I will often use is a simple version of a "nanotransistor" consisting of a semiconducting channel separated by an insulator layer (typically silicon dioxide) from the metallic gate (Fig. 1.1). 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. The voltage V_G on the gate is used to control the electron density in the channel and hence its resistance. 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 ~10 μm in 1960 to ~0.1 μm in 2000, allowing circuit designers to pack $(100)^2 = 10\,000$ times more transistors (and hence that much more computing power) into a chip of given surface area. This increase in packing density is at the heart of the computer revolution. How much longer can the downscaling continue? No one really knows. However, one thing seems certain. Regardless of what form future electronic devices take, *we will have to learn how to model and describe the electronic properties of device structures that are engineered on an atomic scale.* The examples I will use in this book may or may not be important twenty years from now. But the problem of current flow touches on some of the deepest issues of physics related to the nature of "friction" on a microscopic scale and the emergence of irreversibility from reversible laws. The concepts we will discuss represent key fundamental concepts of quantum mechanics and non-equilibrium

statistical mechanics that should be relevant to the analysis and design of nanoscale devices for many years into the future.

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 1.1). 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 1.2). 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, 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 channel with one level (in the energy range of interest), which is a fundamental constant related to the charge on an electron and Planck's constant:

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

Actually small channels typically have two levels (one for up spin and one for down spin) at the same energy ("degenerate" levels) making the maximum conductance equal to $2G_0$. We 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 (Section 1.3). In Section 1.4, I will explain the important role played by charging and electrostatics in determining the shape of the current–voltage (I–V) characteristics, and how this aspect is coupled with the equations for quantum transport. Once this aspect has been incorporated we have all the basic physics needed to describe a one-level channel that is coupled "well" to the contacts. But if the channel is weakly coupled, there is some additional physics that I will discuss in Section 1.5. Finally, in Section 1.6, I will explain how the one-level description is extended to larger devices with multiple energy levels, eventually leading to Ohm's law. It is this extension to larger devices that requires the advanced concepts of quantum statistical mechanics that constitute the subject matter of the rest of this book.

1.1 Energy level diagram

Figure 1.1.1 shows the typical current–voltage characteristics for a well-designed transistor of the type shown in Fig. 1.1 having a width of 1 μm in the y-direction perpendicular to the plane of the paper. At low gate voltages, the transistor is in its off state, and very little current flows in response to a drain voltage V_D. Beyond a certain gate voltage, called the threshold voltage V_T, the transistor is turned on and the ON-current increases with increasing gate voltage V_G. For a fixed gate voltage, the current I increases at first with drain voltage, but it then tends to level off and saturate at a value referred to as the

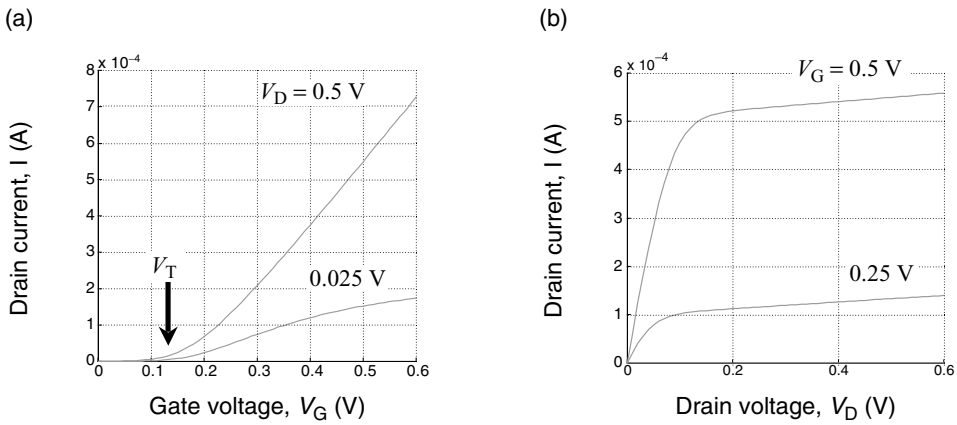

Fig. 1.1.1 (a) Drain current I as a function of the gate voltage V_G for different values of the drain voltage V_D. (b) Drain current as a function of the drain voltage for different values of the gate voltage.

ON-current. Let us start by trying to understand why the current increases when the gate voltage exceeds V_T (Fig. 1.1.1a).

The first step in understanding the operation of any inhomogeneous device structure (like the generic one shown in Fig. 1.1) 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 Fig. 1.1.2. 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 (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 like 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. But in the real world subtle considerations are needed to relate the measured energies to those we use and we will discuss some of these issues in Chapter 2.

We will assume that the large contact regions (labeled source and drain in Fig. 1.1) have a continuous distribution of states. This is true if the contacts are metallic, but not

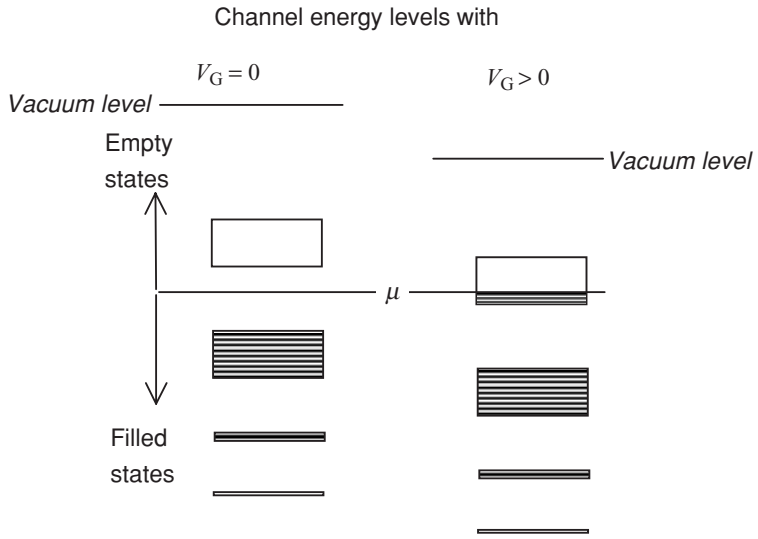

Fig. 1.1.2 Allowed energy levels that can be occupied by electrons in the active region of a device like the channel in Fig. 1.1. A positive gate voltage V_G moves the energy levels down while the electrochemical potential μ is fixed by the source and drain contacts, which are assumed to be in equilibrium with each other ($V_D = 0$).

exactly true of semiconducting contacts, and interesting effects like a decrease in the current with an increase in the voltage (sometimes referred to as negative differential resistance (NDR)) can arise as a result (see Exercise E.1.4); however, we will ignore this possibility in our discussion. The allowed states are occupied up to some energy μ (called the electrochemical potential) which too 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 in 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:

$$f_0(E-\mu) = \frac{1}{1+\exp[(E-\mu)/k_BT]} \tag{1.1.1}$$

Energy levels far below μ are always full so that $f_0 = 1$, while energy levels far above μ are always empty with $f_0 = 0$. Energy levels within a few k_BT of μ are occasionally empty and occasionally full so that the average number of electrons lies

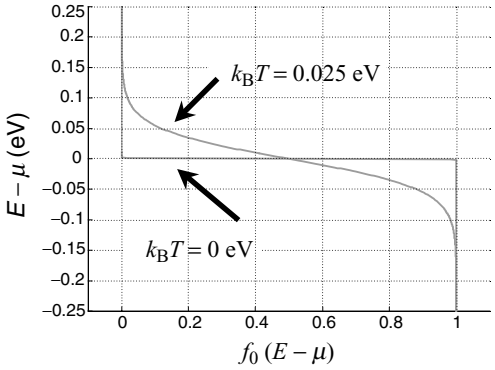

Fig. 1.1.3 The Fermi function (Eq. (1.1.1)) describing the number of electrons occupying a state with an energy E if it is in equilibrium with a large contact ("reservoir") having an electrochemical potential μ.

between 0 and 1: $0 \leq f_0 \leq 1$ (Fig. 1.1.3). Note that this number cannot exceed one because the exclusion principle forbids more than one electron per level.

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 Fig. 1.1.2. 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 the energy difference between the equilibrium electrochemical potential μ and the lowest available empty state (Fig. 1.1.2) 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 Fig. 1.1.4, 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 Fig. 1.1.2. 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 (Fig. 1.1.2) or p-type (Fig. 1.1.4). This point should become clearer in Section 1.2 when we discuss why current flows in response to a voltage applied across the source and drain contacts.

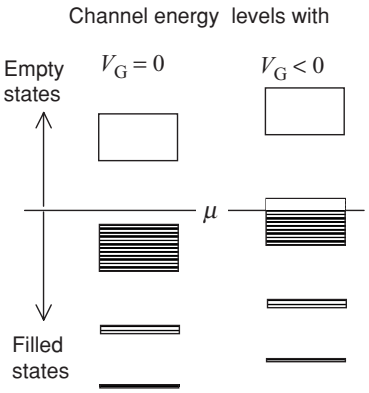

Fig. 1.1.4 Example of p-type or hole conduction. A negative gate voltage ($V_G < 0$) reduces the number of electrons in the channel. Nevertheless the channel will become more conductive once the electrochemical potential μ is driven into the filled band since conduction depends on the availability of states around $E = \mu$ and not on the total number of electrons.

Figures 1.1.2 and 1.1.4 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.

1.2 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 consider 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

$$\mu_1 - \mu_2 = qV_D \tag{1.2.1}$$

giving rise to two different Fermi functions:

$$f_1(E) \equiv \frac{1}{1 + \exp[(E - \mu_1)/k_B T]} = f_0(E - \mu_1) \tag{1.2.2a}$$

$$f_2(E) \equiv \frac{1}{1 + \exp[(E - \mu_2)/k_B T]} = f_0(E - \mu_2) \tag{1.2.2b}$$

Each contact seeks to bring the channel into equilibrium with itself. The source keeps pumping electrons into it, hoping to establish equilibrium. But equilibrium is never

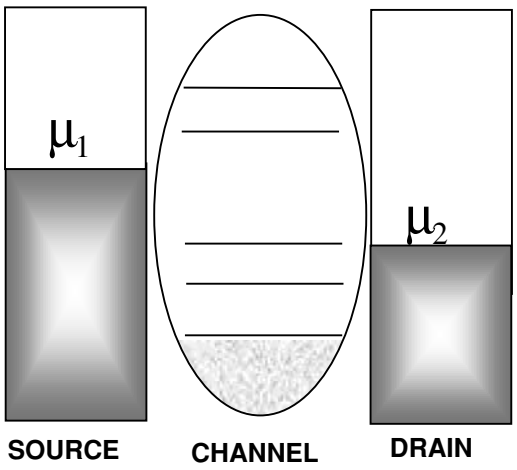

Fig. 1.2.1 A positive voltage V_D applied to the drain with respect to the source lowers the electrochemical potential at the drain: $\mu_2 = \mu_1 - qV_D$. Source and drain contacts now attempt to impose different Fermi distributions as shown, and the channel goes into a state intermediate between the two.

achieved as the drain keeps pulling electrons out in its bid to establish equilibrium with itself. The channel is thus forced into a balancing act between two reservoirs with different agendas and this sends it into a non-equilibrium state intermediate between what the source would like to see and what the drain would like to see (Fig. 1.2.1).

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 (Fig. 1.2.2). 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 Eq. (1.2.2). The average number of electrons N at steady state will be something intermediate between $f_1(\varepsilon)$ and $f_2(\varepsilon)$. There is a net flux I_1 across the left junction that is proportional to $(f_1 - N)$, *dropping the argument* ε for clarity:

$$I_1 = \frac{q\gamma_1}{\hbar}(f_1 - N) \tag{1.2.3a}$$

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

$$I_2 = \frac{q\gamma_2}{\hbar}(f_2 - N) \tag{1.2.3b}$$

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

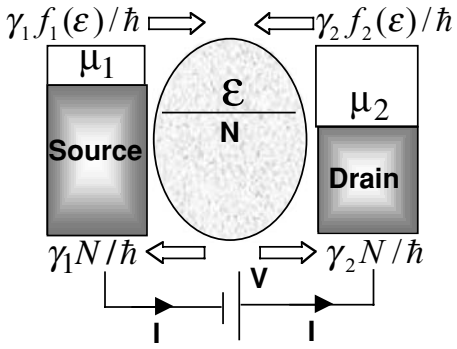

Fig. 1.2.2 Flux of electrons into and out of a one-level channel at the source and drain ends: simple rate equation picture.

dimension per second, so that 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.

Current in a one-level model: At steady state there is no net flux into or out of the channel, $I_1 + I_2 = 0$, so that from Eqs. (1.2.3a, b) we obtain the reasonable result

$$N = \frac{\gamma_1 f_1 + \gamma_2 f_2}{\gamma_1 + \gamma_2} \tag{1.2.4}$$

that is, the occupation N is a weighted average of what contacts 1 and 2 would like to see. Substituting this result into Eq. (1.2.3a) or (1.2.3b) we obtain an expression for the steady-state current:

$$I = I_1 = -I_2 = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}[f_1(\varepsilon) - f_2(\varepsilon)] \tag{1.2.5}$$

This is the current per spin. We should multiply it by two 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 $k_B T$ of the potentials μ_1 and μ_2 that we have $f_1(\varepsilon) \neq f_2(\varepsilon)$ and a current flows. Current flow is thus the 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 (Fig. 1.2.2). 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, Fig. 1.1.2) or filled (p-type, Fig. 1.1.4) 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: Eqs. (1.2.3a, 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. I realize that I run the risk of confusing "satisfied" readers who may want to skip the rest of this section.

The right-hand sides of Eqs. (1.2.3a, b) can be interpreted as the difference between the influx and the outflux from the source and drain respectively (see Fig. 1.2.2). For example, consider the source. The outflux of $\gamma_1 N/\hbar$ is easy to justify 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 influx $\gamma_1 f_1/\hbar$ is harder to justify 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 that the influx would equal the outflux. But the outflux 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 influx remains unchanged since it depends only on the condition in the contacts which remains unchanged (note that the outflux does change giving a net current that we have calculated above).

"Pauli blocking"? Advanced readers may disagree with the statement I just made, namely that the influx "depends only on the condition in the contacts." Shouldn't the influx 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

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

instead of

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

as we have indicated in Fig. 1.2.2. It is easy to see that the net current given by the difference between inflow and outflow is the same in either case, so that 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