Lecture 1

The Bottom-up Approach

"Everyone" has a computer these days, and each computer has more than a billion transistors, making transistors more numerous than anything else we could think of. Even the proverbial ants, I am told, have been vastly outnumbered.

There are many types of transistors, but the most common one in use today is the Field Effect Transistor (FET), which is essentially a resistor consisting of a "channel" with two large contacts called the "source" and the "drain" (Fig. 1.1a).

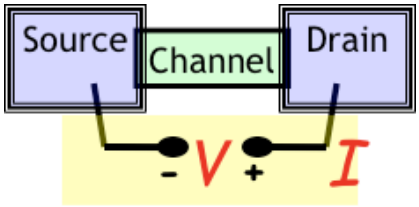


Fig.1.1a.
The Field Effect Transistor (FET) is essentially a resistor consisting of a "channel" with two large contacts called the "source" and the "drain", across which we attach the two terminals of a battery.

The resistance R = Voltage (V) / Current (I) can be switched by several orders of magnitude through the voltage V_G applied to a third terminal called the "gate" (Fig.1.1b) typically from an "OFF" state of ~100 megohms to an "ON" state of ~10 kilohms.

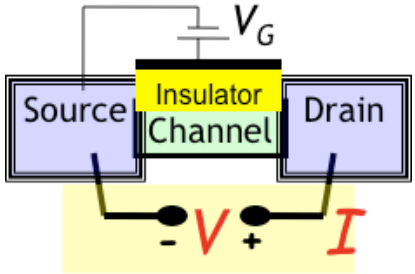


Fig.1.1b.
The resistance $R = V/I$ can be changed by several orders of magnitude through the gate voltage V_G.

Actually, the microelectronics industry uses a complementary pair of transistors such that when one changes from 100M to 10K, the other changes from 10K to 100M. Together they form an inverter whose output is the "inverse" of the input: A low input voltage creates a high output voltage while a high input voltage creates a low output voltage as shown in Fig.1.2.

A billion such switches switching at GHz speeds (that is, once every nanosecond) enable a computer to perform all the amazing feats that we have come to take for granted. Twenty years ago computers were far less powerful, because there were “only” a million of them, switching at a slower rate as well.

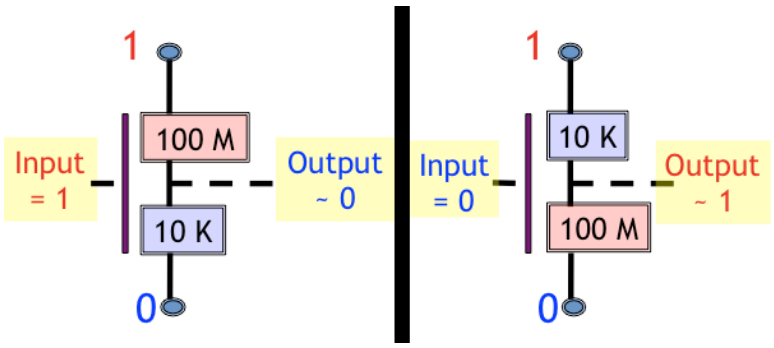


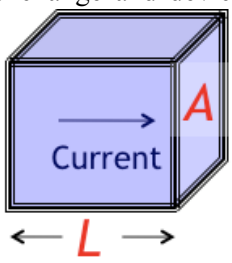
Fig.1.2.
A complementary pair of FET’s form an inverter switch.

Both the increasing number and the speed of transistors are consequences of their ever-shrinking size and it is this continuing miniaturization that has driven the industry from the first four-function calculators of the 1970's to the modern laptops. For example, if each transistor takes up a space of say 10 μm x 10 μm, then we could fit 3000 x 3000 = 9 million of them into a chip of size 3cm x 3cm, since

$$3\,cm\,/\,10\,\mu m \;=\; 3000$$

That is where things stood back in the ancient 1990's. But now that a transistor takes up an area of ~ 1 μm x 1 μm, we can fit 900 million (nearly a billion) of them into the same 3cm x 3cm chip. Where things will go from here remains unclear, since there are major roadblocks to continued miniaturization, the most obvious of which is the difficulty of dissipating the heat that is generated. Any laptop user knows how hot it gets when it is working hard, and it seems difficult to increase the number of switches and/or their speed too much further.

These Lectures, however, are not about the amazing feats of microelectronics or where the field might be headed. They are about a less-appreciated by-product of the microelectronics revolution, namely the deeper understanding of current flow, energy exchange and device operation that it has enabled, based on which we have proposed what we call the bottom-up approach. Let me explain what we mean.



According to Ohm's law, the resistance R is related to the cross-sectional area A and the length L by the relation

$$R \equiv \frac{V}{I} \;=\; \frac{\rho L}{A} \qquad (1.1a)$$

ρ being a geometry-independent property of the material that the channel is made of.

The reciprocal of the resistance is the conductance

$$\frac{I}{V} = \frac{\sigma A}{L} \tag{1.1b}$$

which is written in terms of the reciprocal of the resistivity called the conductivity.

Our conventional view of electronic motion through a solid is that it is "diffusive," which means that the electron takes a random walk from the source to the drain, traveling in one direction for some length of time before getting scattered into some random direction as sketched in Fig.1.3. The mean free path, λ that an electron travels before getting scattered is typically less than a micrometer (also called a micron = 10^{-3} mm, denoted μm) in common semiconductors, but it varies widely with temperature and from one material to another.

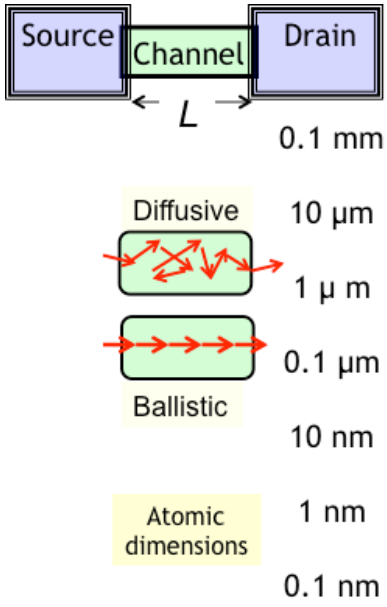


Fig.1.3.
The length of the channel of an FET has progressively shrunk with every new generation of devices ("Moore's Law") and stands today (2010) at ~ 50 nm, which amounts to a few hundred atoms!

Length units:
1 mm = 1000 μm
and 1 μm = 1000 nm

It seems reasonable to ask what would happen if a resistor is shorter than a mean free path so that an electron travels ballistically ("like a bullet") through the channel. Would the resistance still obey Ohm's law? Would it still make sense to talk about its resistance? These questions have intrigued scientists for a long time, but even twenty five years ago one could only speculate about the answers. Today the answers are quite clear and experimentally well established. Even the transistors in commercial laptops now have channel lengths $L \sim 50$ nm, corresponding to a few hundred atoms in length! And in research laboratories people have even measured the resistance of a hydrogen molecule.

It is now clearly established that the resistance of a ballistic conductor can be written in the form

$$R_B = \underbrace{\frac{h}{q^2}}_{\sim 25\ K\Omega} \frac{1}{M} \tag{1.2}$$

where h/q^2 is a fundamental constant and M represents the number of effective channels available for conduction. Note that here we are using the word “channel” not to denote the physical channel in Fig.1.3, but in the sense of parallel paths whose meaning will be clarified in the next few lectures. In future we will refer to M as the number of “modes”.

This result is now fairly well-known, but the common belief is that it applies only to short conductors and belongs in a course on special topics like mesoscopic physics or nanoelectronics. What is not as well-known is that the resistance for both long and short conductors can be written in the form (λ : mean free path)

$$R = \underbrace{\frac{h}{q^2 M}}_{R_B} \left(1 + \frac{L}{\lambda}\right) \tag{1.3}$$

Ballistic and diffusive conductors are not two different worlds, but rather a continuum as the length L is increased. For $L << \lambda$, Eq.(1.3) reduces to the ballistic result in Eq.(1.2), while for $L >> \lambda$, it morphs into Ohm's law in Eq.(1.1). Indeed we could rewrite Eq.(1.3) in the form

$$R = \frac{\rho}{A}(L+\lambda) \tag{1.4}$$

with a new expression

$$\rho = \frac{h}{q^2}\frac{A}{M\lambda} \tag{1.5}$$

that provides a different view of resistivity in terms of the number of modes per unit area and the mean free path.

This is the result we will try to establish in the first few lectures and it illustrates the essence of our bottom-up approach, viewing short conductors not as an aberration but as the starting point to understanding long conductors. For historical reasons, the subject of conduction is always approached top-down, from large complicated conductors down to hydrogen molecules. As long as there was no experimental evidence for what the resistance of a small conductor might be, it made good sense to start from large conductors where the answers were clear. But now that the answers are clear at both ends, a bottom-up view seems called for, at least to complement the top-down view. After all that is how we learn most things, from the simple to the complex: quantum mechanics, for example, starts with the hydrogen atom, not with bulk solids.

But there is a deeper reason why the bottom-up approach can be particularly useful in transport theory and this is the "new perspective" we are seeking to convey in these lectures. One of the major conceptual issues posed by the ballistic resistance R_B in Eq.(1.2), is the question of "where is the heat". Current flow through any resistance R leads to the generation of an amount of heat $VI = I^2R$, commonly known as Joule heating. A ballistic resistance R_B too must generate a heat of I^2R_B.

But how can a ballistic resistor generate heat? Heat generation requires interactions whereby energetic electrons give up their excess energy to the surrounding atoms. A conductor through which electrons zip through without exchanging energy cannot possibly be generating any heat. It is now generally accepted that in such a resistor, all the Joule heat would be dissipated in the contacts as sketched in Fig.1.4. There is experimental evidence that real nanoscale conductors do approach this ideal and a significant fraction of the Joule heat is generated in the contacts.

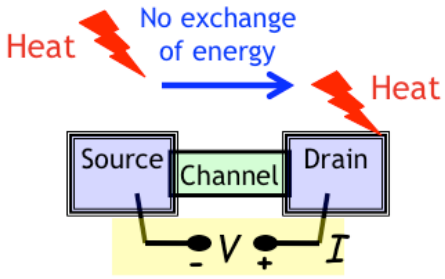


Fig.1.4. **The ideal elastic resistor** with the Joule heat $VI = I^2R$ generated entirely in the contacts as sketched. Many nanoscale conductors are believed to be close to this ideal.

In a sense this seems obvious as my colleague Ashraf often points out. After all a bullet dissipates most of its energy to the object it hits, rather than to the medium it flies through. And yet in the present context, this does seem like a somewhat counter-intuitive result. Clearly the flow of electrons and hence the resistance is determined by the area of the narrow channel that electrons have to squeeze through and not by the large area contacts. But the associated Joule heat occurs in the contacts. And this would be true even if the channel were full of "potholes" that scattered the electrons, as long as the interaction with the electrons is purely ***elastic***, that is does not involve any transfer of energy,

The point we wish to make, however, is that flow or transport always involves two fundamentally different types of processes, namely elastic transfer and heat generation, belonging to two distinct branches of physics. The first involves frictionless mechanics of the type described by Newton's laws or the Schrödinger equation. The second involves the generation of heat described by the laws of thermodynamics. The first is driven by forces or potentials and is reversible. The second is driven by

entropy and is irreversible. Viewed in reverse, such processes look absurd, like heat flowing spontaneously from a cold to a hot surface or an electron accelerating spontaneously by absorbing heat from its surroundings.

Normally the two processes are intertwined and a proper description of current flow in electronic devices requires the advanced methods of non-equilibrium statistical mechanics that integrate mechanics with thermodynamics. Over a century ago Boltzmann taught us how to combine Newtonian mechanics with heat generating or entropy-driven processes

and the resulting Boltzmann transport equation (BTE) is widely accepted as the cornerstone of semiclassical transport theory. The word semiclassical is used because some quantum effects have also been incorporated approximately into the same framework.

A full treatment of quantum transport requires a formal integration of quantum dynamics described by the Schrodinger equation with heat generating processes. This is exactly what is achieved in the non-equilibrium Green function (NEGF) method

originating in the 1960's from the seminal works of Martin and Schwinger (1959), Kadanoff and Baym (1962), Keldysh (1965) and others (see Lecture 19).

The BTE takes many semesters to master and the full NEGF formalism, even longer. Much of this complexity, however, comes from the difficulty of combining mechanics with distributed heat-generating

processes which are a key part of the physics of resistance in long conductors.

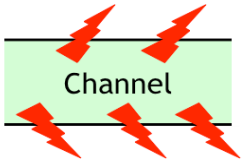


Fig.1.5. Resistance in long conductors primarily arise from distributed heat generating processes along the channel. Prior to 1990, papers dealing with basic transport theory seldom considered the actual physical contacts.

The modern developments in mesoscopic physics and nanoelectronics give us a different perspective, with the ***elastic resistor*** in Fig.1.4 as the starting point. The operation of the elastic resistor can be understood in far more elementary terms because of the clean spatial separation between the mechanical and the heat-generating processes. The former is confined to the channel and the latter to the contacts. As we will see in the next few lectures, the latter is easily taken care of, indeed so easily that it is easy to miss the profound nature of what is being accomplished.

Even quantum transport can be discussed in relatively elementary terms using this viewpoint. My own research has largely been focused in this area developing the NEGF method, but we will get to it only in Part III after we have "set the stage" in Parts I and II using a semiclassical picture.

But does this viewpoint help us understand long conductors? Short conductors may be elastic and conceptually simple, but don't we finally have to deal with distributed heat generation if we want to understand long conductors?

We argue that many properties of long conductors, especially at low bias can be understood in simple terms by viewing them as a series of elastic resistors as sketched in Fig.1.6.

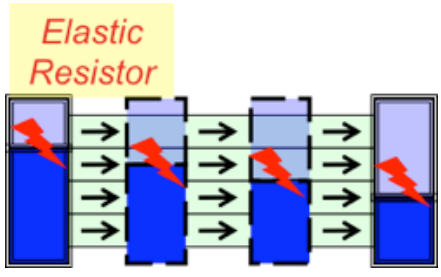


Fig.1.6. Long resistors can be approximately viewed as a series of elastic resistors, as discussed in Section 3.3.

Many well-known results like the conductivity and the thermoelectric coefficients for large conductors, that are commonly obtained from the BTE, can be obtained in a more transparent manner by using this viewpoint, as we will show in the first two parts of these lectures. We will then use this viewpoint in Part III to look at a variety of quantum transport phenomena like resonant tunneling, conductance quantization, the integer quantum Hall effect and spin precession.

In short, the lesson of nanoelectronics we are trying to convey is the utility of the concept of an elastic resistor with its clean separation of mechanics from thermodynamics. The concept was introduced by Rolf Landauer in 1957 and has been widely used in mesoscopic physics ever since the seminal work in the 1980's helped establish its relevance to understanding experiments in short conductors.

What we hope to convey in these lectures is that the concept of an elastic resistor is not just useful for short conductors but provides a fresh new perspective for long conductors as well, that makes a wide variety of devices and phenomena transparent and accessible.

I do not think any of the end results will come as a surprise to the experts. I believe they all follow directly from the BTE or the NEGF and one might well ask whether anything is gained from approximate physical pictures based on elastic resistors. This is a subjective matter

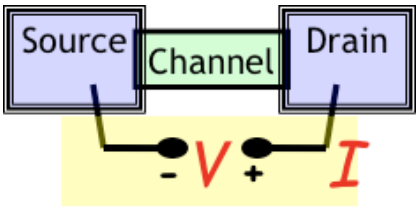
that is not easy to argue. Perhaps Feynman (1963) expressed it best in his Lectures on Physics when he said

> ".. people .. say .. there is nothing which is not contained in the equations .. if I understand them mathematically inside out, I will understand the physics inside out. Only it doesn't work that way. .. A physical understanding is a completely unmathematical, imprecise and inexact thing, but absolutely necessary for a physicist."

I believe the elastic resistor contributes to our physical understanding of the BTE and the NEGF method, without being too "imprecise" or "inexact", and I hope it will facilitate the insights needed to take us to the next level of understanding, discovery and innovation.

Lecture 2

Why Electrons Flow



It is a well-known and well-established fact, namely that when the two terminals of a battery are connected across a conductor, it gives rise to a current due to the flow of electrons across the channel from the source to the drain.

If you ask anyone, novice or expert, what causes electrons to flow, by far the most common answer you will receive is that it is the electric field. However, this answer is incomplete at best. After all even before we connect a battery, there are enormous electric fields around every atom due to the positive nucleus whose effects on the atomic spectra are well-documented. Why is it that these electric fields do not cause electrons to flow, and yet a far smaller field from an external battery does?

The standard answer is that microscopic fields do not cause current to flow, a macroscopic field is needed. This too is not satisfactory, for two reasons. Firstly, there are well-known inhomogeneous conductors like p-n junctions which have large macroscopic fields extending over many micrometers that do not cause any flow of electrons till an external battery is connected.

Secondly, experimentalists are now measuring current flow through conductors that are only a few atoms long with no clear distinction between the microscopic and the macroscopic. This is a result of our progress in nanoelectronics, and it forces us to search for a better answer to the question, "why electrons flow."

2.1 Two Key Concepts

To answer this question, we need two key concepts. First is the ***density of states per unit energy D(E) available for electrons to occupy*** inside the channel (Fig.2.1). For the benefit of experts, I should note that we are adopting what we will call a "point channel model" represented by a single density of states *D(E).* More generally one needs to consider the spatial variation of *D(E),* as we will see in Lecture 8, but there is much that can be understood just from our point channel model.

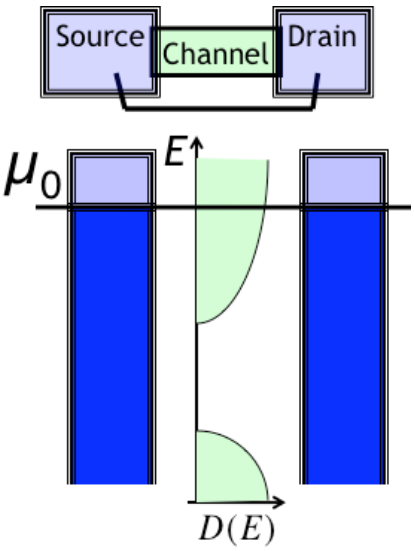


Fig.2.1.
The first step in understanding the operation of any electronic device is to draw the available density of states *D(E)* as a function of energy E, inside the channel and to locate the equilibrium electrochemical potential μ_0 separating the filled from the empty states.

The second key input is the ***location of the electrochemical potential***, μ_0 which at equilibrium is the same everywhere, in the source, the drain and the channel. Roughly speaking (we will make this statement more precise shortly) it is the energy that demarcates the filled states from the

empty ones. All states with energy $E < \mu_0$ are filled while all states with $E > \mu_0$ are empty. For convenience I might occasionally refer to the electrochemical potential as just the ***"potential".***

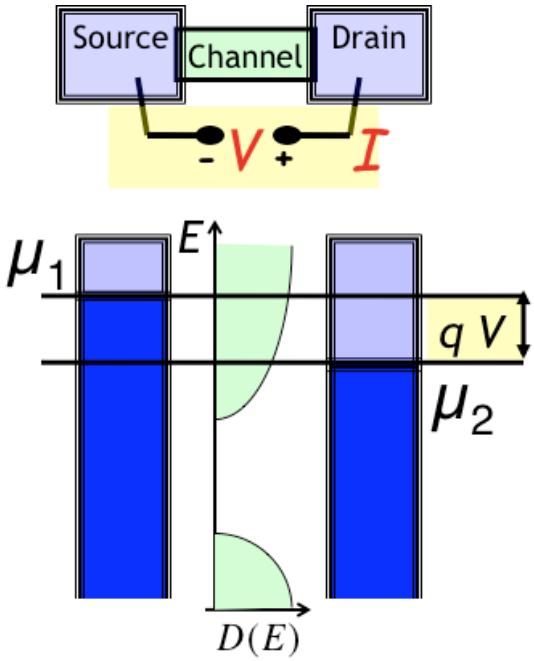


Fig.2.2.
When a voltage is applied across the contacts, it lowers all energy levels at the positive contact (drain in the picture). As a result the electrochemical potentials in the two contacts separate: $\mu_1 - \mu_2 = qV$.

When a battery is connected across the two contacts creating a potential difference V between them, it lowers all energies at the positive terminal (drain) by an amount qV, **-** q being the charge of an electron ($q = 1.6 \times 10^{-19}$ coulombs) making the two electrochemical potentials separate by qV as shown in Fig.2.2:

$$\mu_1 - \mu_2 \quad = \quad qV \tag{2.1}$$

Just as a temperature difference causes heat to flow and a difference in water levels makes water flow, a difference in electrochemical potentials causes electrons to flow. Interestingly, only the states in and around an energy window around μ_1 and μ_2 contribute to the current flow, all the states far above and well below that window playing no part at all. Let us explain why.

2.1.1 Energy Window for Current Flow

Each contact seeks to bring the channel into equilibrium with itself, which roughly means filling up all the states with energies E less than its electrochemical potential μ and emptying all states with energies greater than μ.

Consider the states with energy E that are less than μ_1 but greater than μ_2. Contact 1 wants to fill them up since $E < \mu_1$, but contact 2 wants to empty them since $E > \mu_2$. And so contact 1 keeps filling them up and contact 2 keeps emptying them causing electrons to flow continually from contact 1 to contact 2.

Consider now the states with E greater than both μ_1 and μ_2. Both contacts want these states to remain empty and they simply remain empty with no flow of electrons. Similarly the states with E less than both μ_1 and μ_2 do not cause any flow either. Both contacts like to keep them filled and they just remain filled. There is no flow of electrons outside the window between μ_1 and μ_2, or more correctly outside ± a few kT of this window, as we will discuss shortly.

This last point may seem obvious, but often causes much debate because of the common belief we alluded to earlier, namely that electron flow is caused by the electric field in the channel. If that were true, all the electrons should flow and not just the ones in any specific window determined by the contacts.

2.2 Fermi Function

Let us now make the above statements more precise. We stated that roughly speaking, at equilibrium, all states with energies E below the electrochemical potential μ_0 are filled while all states with $E > \mu_0$ are empty. This is precisely true only at absolute zero temperature. More generally, the transition from completely full to completely empty occurs over an energy range $\sim \pm 2\,kT$ around $E = \mu_0$ where k is the Boltzmann

constant (~ 80 μeV/K) and T is the absolute temperature. Mathematically this transition is described by the Fermi function :

$$f(E) = \frac{1}{\exp\left(\frac{E-\mu}{kT}\right)+1} \tag{2.2}$$

This function is plotted in Fig.2.3 (left panel), though in an unconventional form with the energy axis vertical rather than horizontal. This will allow us to place it alongside the density of states, when trying to understand current flow (see Fig.2.4).

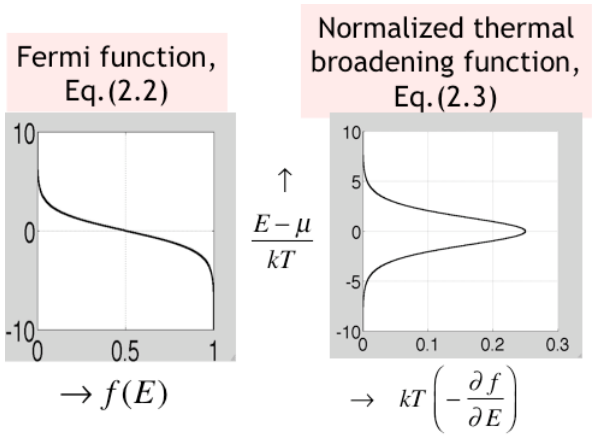


Fig.2.3. Fermi function and the normalized (dimensionless) thermal broadening function.

For readers unfamiliar with the Fermi function, let me note that an extended discussion is needed to do justice to this deep but standard result, and we will discuss it a little further in Lecture 16 when we talk about the key principles of equilibrium statistical mechanics. At this stage it may help to note that what this function (Fig.2.3) basically tells us is that states with low energies are always occupied (*f=1*), while states with high energies are are always empty (*f=0*), something that seems reasonable since we have heard often enough that (1) everything goes to its lowest energy, and (2) electrons obey an exclusion principle that stops

them from all getting into the same state. The additional fact that the Fermi function tells us is that the transition from *f=1* to *f=0* occurs over an energy range of ~ ± *2kT* around μ_0.

2.2.1. Thermal Broadening Function

Also shown in Fig.2.3 is the derivative of the Fermi function, multiplied by *kT* to make it dimensionless:

$$F_T(E,\mu) \;=\; kT\left(-\frac{\partial f}{\partial E}\right) \tag{2.3a}$$

Using Eq.(2.2) it is straightforward to show that

$$F_T(E,\mu) \;=\; \frac{e^x}{(e^x+1)^2}, \quad where \quad x \equiv \frac{E-\mu}{kT} \tag{2.3b}$$

Note:

(1) From Eq.(2.3b) it can be seen that

$$F_T(E,\mu) \;=\; F_T(E-\mu) \;=\; F_T(\mu-E) \tag{2.4a}$$

(2) From Eqs.(2.3b) and (2.2) it can be seen that

$$F_T \;=\; f(1-f) \tag{2.4b}$$

(3) If we integrate F_T over all energy the total area equals *kT:*

$$\int_{-\infty}^{+\infty} dE\, F_T(E,\mu) \;=\; kT \int_{-\infty}^{+\infty} dE\left(-\frac{\partial f}{\partial E}\right)$$

$$=\; kT\left[-f\right]_{-\infty}^{+\infty} \;=\; kT\,(1-0) \;=\; kT \tag{2.4c}$$

so that we can approximately visualize F_T as a rectangular "pulse" centered around $E=\mu$ with a peak value of *1/4* and a width of ~ *4kT*.

2.3 Non-equilibrium: Two Fermi Functions

When a system is in equilibrium the electrons are distributed among the available states according to the Fermi function. But when a system is driven out-of-equilibrium there is no simple rule for determining the distribution of electrons. It depends on the specific problem at hand making non-equilibrium statistical mechanics far richer and less understood than its equilibrium counterpart.

For our specific non-equilibrium problem, we argue that the two contacts are such large systems that they cannot be driven out-of-equilibrium. And so each remains locally in equilibrium with its own electrochemical potential giving rise to two different Fermi functions (Fig.2.4):

$$f_1(E) = \frac{1}{\exp\left(\frac{E-\mu_1}{kT}\right)+1} \tag{2.5a}$$

$$f_2(E) = \frac{1}{\exp\left(\frac{E-\mu_2}{kT}\right)+1} \tag{2.5b}$$

The "little" channel in between does not quite know which Fermi function to follow and as we discussed earlier, the source keeps filling it up while the drain keeps emptying it, resulting in a continuous flow of current.

In summary, what makes electrons flow is the difference in the "agenda" of the two contacts as reflected in their respective Fermi functions, $f_1(E)$ and $f_2(E)$. This is qualitatively true for all conductors, short or long. But for short conductors, the current at any given energy E is quantitatively proportional to

$$I(E) \sim f_1(E) - f_2(E)$$

representing the difference in the probabilities in the two contacts. This quantity goes to zero when E lies way above μ_1, μ_2 since f_1 *and* f_2 are

both zero. It also goes to zero when E lies way below μ_1, μ_2 since f_1 *and* f_2 are both one. Current flow occurs only in the intermediate energy window, as we had argued earlier.

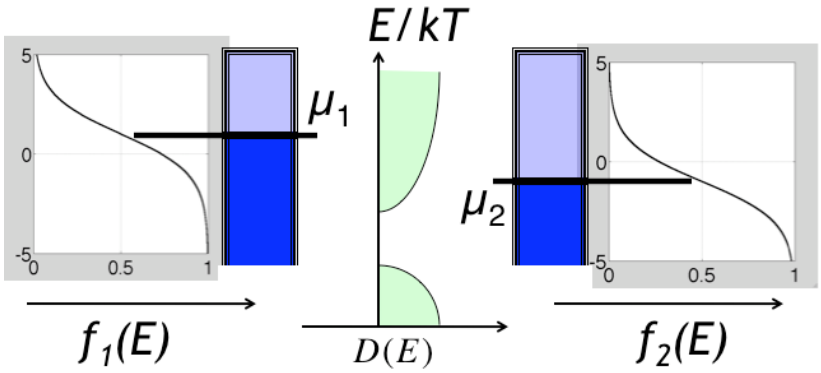


Fig.2.4.
Electrons in the contacts occupy the available states with a probability described by a Fermi function *f(E)* with the appropriate electrochemical potential μ.

2.4 Linear Response

Current-voltage relations are typically not linear, but there is a common approximation that we will frequently use throughout these lectures to extract the "linear response" which refers to the low bias conductance, dI/dV, as $V \rightarrow 0$.

The basic idea can be appreciated by plotting the difference between two Fermi functions, normalized to the applied voltage

$$F(E) = \frac{f_1(E) - f_2(E)}{qV/kT} \tag{2.6}$$

where

$$\mu_1 = \mu_0 + (qV/2)$$

$$\mu_2 = \mu_0 - (qV/2)$$

Fig.2.5 shows that the difference function F gets narrower as the voltage is reduced relative to kT. The interesting point is that as qV is reduced below kT, the function F approaches the thermal broadening function F_T we defined (see Eq.(2.3a)) in Section 2.2:

$$F(E) \rightarrow F_T(E), \quad as \quad qV/kT \rightarrow 0$$

so that from Eq.(2.6)

$$f_1(E) - f_2(E) \approx \frac{qV}{kT} F_T(E,\mu_0) = \left(-\frac{\partial f_0}{\partial E}\right) qV \qquad (2.7)$$

if the applied voltage $\mu_1 - \mu_2 = qV$ is much less than kT.

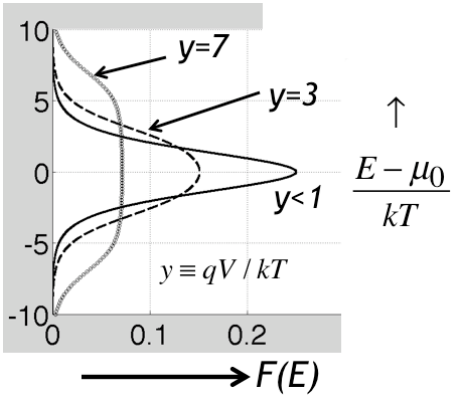


Fig.2.5. *F(E)* from Eq.(2.6) versus *$(E-\mu_0)/kT$* for different values of *$y=qV/kT$*.

The validity of Eq.(2.7) for $qV << kT$ can be checked numerically if you have access to MATLAB or equivalent. For those who like to see a mathematical derivation, Eq.(2.7) can be obtained using the Taylor series expansion described in Appendix A to write

$$f(E) - f_0(E) \approx \left(-\frac{\partial f_0}{\partial E}\right)(\mu - \mu_0) \qquad (2.8)$$

Eq.(2.8) and Eq.(2.7) which follows from it, will be used frequently in these lectures.

2.5. Difference in "Agenda" Drives the Flow

Before moving on, let me quickly reiterate the key point we are trying to make, namely that current is determined by

$$-\frac{\partial f_0(E)}{\partial E} \quad \textit{and not by} \quad f_0(E)$$

The two functions look similar over a limited range of energies

$$-\frac{\partial f_0(E)}{\partial E} \approx \frac{f_0(E)}{kT} \quad \textit{if } E-\mu_0 >> kT$$

So if we are dealing with a so-called "non-degenerate conductor" where we can restrict our attention to a range of energies satisfying this criterion, we may not notice the difference.

But in general these functions look very different (see Fig.2.3) and the experts agree that current depends not on the Fermi function, but on its derivative. However, we are not aware of any elementary treatment that leads to this result.

Freshman physics texts start by treating the force due to an electric electric field F as the driving term and adding a frictional term to Newton's law (τ_m is the so-called "momentum relaxation time")

$$\underbrace{\frac{d(mv)}{dt} = (-qF)}_{Newton's\ Law} - \underbrace{\frac{mv}{\tau_m}}_{Friction}$$

At steady-state ($d/dt = 0$) this gives a non-zero drift velocity, from which one calculates the current. This elementary approach leads to the Drude formula (discussed in Lecture 5) which played a major historical role in our understanding of current flow. But since it treats electric fields as the driving term, it also suggests that the current depends on the total number of electrons. This is commonly explained away by saying that there are mysterious quantum mechanical forces that prevent electrons in full bands from moving and what matters is the number of "free electrons".

But this begs the question of which electrons are free and which are not, a question that becomes more confusing for atomic scale conductors.

It is well-known that the conductivity varies widely, changing by a factor of ~10^{20} going from copper to glass, to mention two materials that are near two ends of the spectrum. But this is not because one has more electrons than the other. The total number of electrons is of the same order of magnitude for all materials from copper to glass.

Whether a conductor is good or bad is determined by the availability of states in an energy window ~ kT around the electrochemical potential μ_0, which can vary widely from one material to another. This is well-known to experts and comes mathematically from the dependence of the conductivity

$$on \quad -\frac{\partial f_0(E)}{\partial E} \quad rather\ than \quad f_0(E)$$

a result that typically requires advanced treatments based on the Boltzmann (Lecture 7) or the Kubo formalism (Lecture 15).

Our bottom-up approach, however, leads us to this result in an elementary way as we have just seen. Current is driven by the difference in the "agenda" of the two contacts which for low bias is proportional to the derivative of the equilibrium Fermi function:

$$f_1(E) - f_2(E) \approx \left(-\frac{\partial f_0}{\partial E}\right) qV$$

There is no need to invoke mysterious forces that stops some electrons from moving, though one could perhaps call f_1 - f_2 a mysterious force, since the Fermi function (Eq.(2.2)) reflects the exclusion principle. In Lecture 11 we will see how this approach is readily extended to describe the flow of phonons which is proportional to $n_1 - n_2$, n being the Bose (not Fermi) function which is appropriate for particles that do not have an exclusion principle.

Lecture 3

The Elastic Resistor

3.1. How an Elastic Resistor Dissipates Heat
3.2. Conductance of an Elastic Resistor
3.3. Why an Elastic Resistor is Relevant

We saw in the last Lecture that the flow of electrons is driven by the difference in the "agenda" of the two contacts as reflected in their respective Fermi functions, $f_1(E)$ and $f_2(E)$. The negative contact with its larger $f(E)$ would like to see more electrons in the channel than the positive contact. And so the positive contact keeps withdrawing electrons from the channel while the negative contact keeps pushing them in.

This is true of all conductors, big and small. But it is generally difficult to express the current as a simple function of $f_1(E)$ and $f_2(E)$, because electrons jump around from one energy to another and the current flow at different energies is all mixed up.

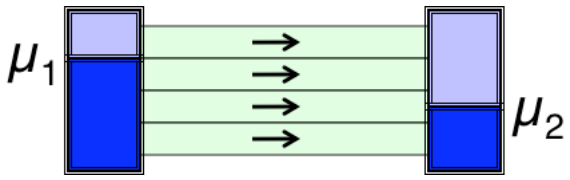


Fig.3.1.
An elastic resistor: Electrons travel along fixed energy channels.

But for the ideal elastic resistor shown in Fig.1.4, the current in an energy range from E to $E+dE$ is decoupled from that in any other energy range, allowing us to write it in the form (Fig.3.1)

$$dI = \frac{1}{q} dE\, G(E)\,(f_1(E) - f_2(E))$$

and integrating it to obtain the total current I. Making use of Eq.(2.7), this leads to an expression for the low bias conductance

$$\frac{I}{V} = \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E} \right) G(E) \tag{3.1}$$

where $(-\partial f_0 / \partial E)$ can be visualized as a rectangular pulse of area equal to one, with a width of $\sim \pm 2kT$ (see Fig.2.3, right panel).

Let me briefly comment on a general point that often causes confusion regarding the direction of the current. As I noted in Lecture 2, because the electronic charge is negative (an unfortunate choice, but something we cannot do anything about) the side with the higher voltage has a lower electrochemical potential. Inside the channel, electrons flow from the higher to the lower electrochemical potential, so that the electron current flows from the source to the drain. The conventional current on the other hand flows from the higher to the lower voltage.

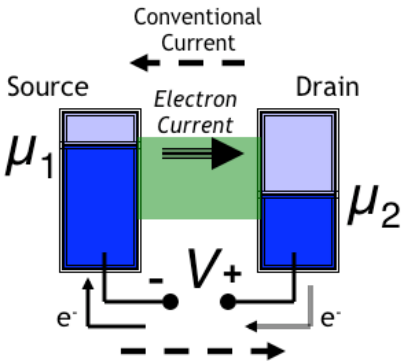


Fig.3.2.
Because an electron carries negative charge, the direction of the electron current is always opposite to that of the conventional current.

Since our discussions will usually involve electron energy levels and the electrochemical potentials describing their occupation, it is also convenient for us to use the electron current instead of the conventional current. For example, in Fig.3.2 it seems natural to say that the current flows from the source to the drain and not the other way around. And

that is what I will try to do consistently throughout these Lectures. In short, we will use the current, *I*, to mean ***electron current.***

Getting back to Eq.(3.1), we note that it tells us that for an elastic resistor, we can define a conductance function *G(E)* whose average over an energy range ~ ± *2kT* around the electrochemical potential μ_0 gives the experimentally measured conductance. At low temperatures, we can simply use the value of *G(E)* at $E = \mu_0$.

This energy-resolved view of conductance represents an enormous simplification that is made possible by the concept of an ***elastic resistor*** which is a very useful idealization that describes short devices very well and provides insights into the operation of long devices as well.

Note that by elastic we do not just mean "ballistic" which implies that the electron goes straight from source to drain, "like a bullet." We also include the possibility that an electron takes a more traditional diffusive path ***as long as it changes only its momentum and not its energy along the way:***

In ***Section 3.2*** we will obtain an expression for the conductance function *G(E)* for an elastic resistor in terms of the density of states *D(E)*.

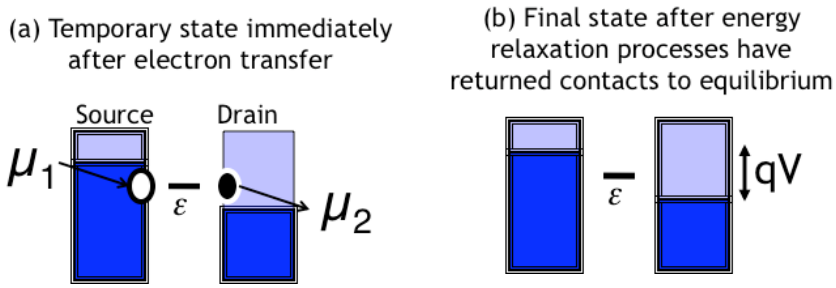
The concept of an elastic resistor is not only useful in understanding nanoscale devices, but it also helps understand transport properties like the conductivity of large resistors by viewing them as multiple elastic resistors in series, as explained in ***Section 3.3***. This is what makes the bottom-up approach so powerful in clarifying transport problems in general.

But before we talk further about the conductance of an elastic resistor, let us address an important conceptual issue. Since current flow (*I*) through

a resistor (R) dissipates a Joule heat of I^2R per second, it seems like a contradiction to talk of an elastic resistor where electrons do not lose energy? The point to note is that while the electron does not lose any energy in the channel of an elastic resistor, it does lose energy both in the source and the drain and that is where the Joule heat gets dissipated. This is a very non-intuitive result that seems to be at least approximately true of nanoscale conductors: ***An elastic resistor has a resistance R determined by the channel, but the corresponding heat I^2R is entirely dissipated outside the channel.***

3.1. How an Elastic Resistor Dissipates Heat

How could this happen? Consider a one level elastic resistor having one sharp level with energy ε. Every time an electron crosses over through the channel, it appears as a "hot electron" on the drain side with an energy ε in excess of the local electrochemical potential μ_2 as shown below:



Energy dissipating processes in the contact quickly make the electron get rid of the excess energy ($\varepsilon - \mu_2$). Similarly at the source end an empty spot (a "hole") is left behind with an energy ε that is much less than the local electrochemical potential μ_1, which gets quickly filled up by electrons dissipating the excess energy ($\mu_1 - \varepsilon$).

In effect, every time an electron crosses over from the source to the drain,

$$\textit{an energy } (\mu_1 - \varepsilon) \textit{ is dissipated in the source}$$

$$\textit{an energy } (\varepsilon - \mu_2) \textit{ is dissipated in the drain}$$

The total energy dissipated is

$$\mu_1 - \mu_2 \quad = \quad qV$$

which is supplied by the external battery that maintains the potential difference μ_1 - μ_2. The overall flow of electrons and heat is summarized in Fig.3.3 below.

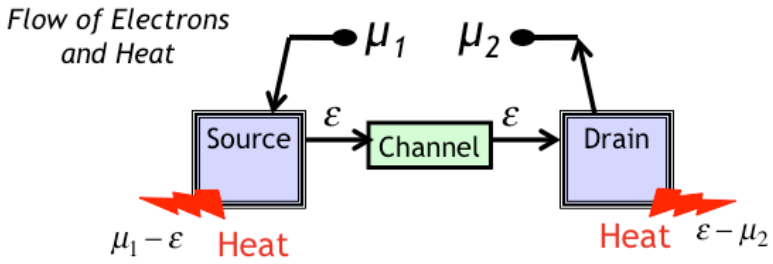


Fig.3.3. Flow of electrons and heat in a one-level elastic resistor having one level with $E = \varepsilon$.

If N electrons cross over in a time t

$$\textit{Dissipated power} \quad = \quad qV * N / t \quad = \quad V * I$$

since

$$\textit{Current} \quad = \quad q * N / t$$

Note that $V*I$ is the same as I^2R and V^2G.

The heat dissipated by an "elastic resistor" thus occurs in the contacts. As we will see next, the detailed mechanism underlying the complicated process of heat transfer in the contacts can be completely bypassed simply by legislating that the contacts are always maintained in equilibrium with a fixed electrochemical potential.

3.2. Conductance of an Elastic Resistor

Consider first the simplest elastic resistor having just one level with energy ε in the energy range of interest through which electrons can squeeze through from the source to the drain. We can write the resulting current as

$$I_{one\ level} = \frac{q}{t}\left(f_1(\varepsilon) - f_2(\varepsilon)\right) \tag{3.2}$$

where t is the time it takes for an electron to transfer from the source to the drain.

We can extend Eq.(3.2) for the current through a one-level resistor to any elastic conductor (Fig.3.1) with an arbitrary density of states *D(E),* noting that all energy channels conduct independently in parallel. We could first write the current in an energy channel between *E and E+dE*

$$dI = dE\,\frac{D(E)}{2}\,\frac{q}{t}\,(f_1(E) - f_2(E))$$

since an energy channel between *E and E+dE* contains *D(E)dE* states, half of which contribute to carrying current from source to drain.

Integrating we obtain an expression for the current through an elastic resistor:

$$I = \frac{1}{q}\int_{-\infty}^{+\infty} dE\,G(E)\,(f_1(E) - f_2(E)) \tag{3.3}$$

where

$$G(E) = \frac{q^2 D(E)}{2t(E)} \tag{3.4}$$

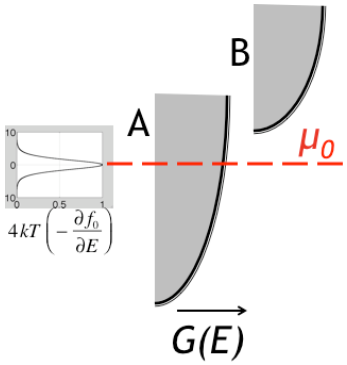
If the applied voltage $\mu_1 - \mu_2 = qV$ is much less than kT, we can use Eq.(2.7) to write

$$I = V \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E} \right) G(E)$$

which yields the expression for conductance stated earlier in Eq.(3.1).

3.2.1. Degenerate and Non-Degenerate Conductors

Eq.(3.1) is valid in general, but depending on the nature of the conductance function *G(E)* and the thermal broadening function $-\partial f_0 / \partial E$, two distinct physical pictures are possible. The first is case A where the conductance function *G(E)* is nearly constant over the width of the broadening function.



We could then pull *G(E)* out of the integral in Eq.(3.1) to write

$$\frac{I}{V} \approx G(E=\mu_0) \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E} \right) = G(E=\mu_0) \qquad (3.5)$$

This relation suggests an operational definition for the conductance function *G(E): It is the conductance measured at low temperatures for a channel with its electrochemical potential μ_0 located at E.*

Case A is a good example of the so-called degenerate conductors. The other extreme is the non-degenerate conductor shown in case B where

the electrochemical potential is located at an energy many *kT*'s below the energy range where the conductance function is non-zero. As a result over the energy range of interest where *G(E)* is non-zero, we have

$$x \equiv \frac{E-\mu_0}{kT} >> 1$$

and it is common to approximate the Fermi function with the Boltzmann function

$$\frac{1}{1+e^{x}} \approx e^{-x}$$

so that

$$\frac{I}{V} \approx \int_{-\infty}^{+\infty} \frac{dE}{kT} G(E)\, e^{-(E-\mu_0)/kT}$$

This non-degenerate limit is commonly used in the semiconductor literature though the actual situation is often intermediate between degenerate and non-degenerate limits.

We will generally use the degenerate limit expressed by Eq.(3.5) writing

$$G = \frac{q^2 D}{2t}$$

with the understanding that the quantities *D* and *t* are evaluated at $E = \mu_0$ and depending on the nature of *G(E)* may need to be averaged over a range of energies using $-\partial f_0 / \partial E$ as a "weighting function" as prescribed by Eq.(3.1).

Eq.(3.4) seems quite intuitive: it says that the conductance is proportional to the product of two factors, namely ***the availability of states (D) and the ease with which electrons can transport through them (1/t)***. This is the key result that we will use in subsequent Lectures.

3.3. Why an Elastic Resistor is Relevant

The elastic resistor model is clearly of great value in understanding nanoscale conductors, but the reader may well wonder how an elastic resistor can capture the physics of real conductors which are surely far from elastic? In long conductors inelastic processes are distributed continuously through the channel, inextricably mixed up with all the elastic processes (Fig.3.4). Doesn't that affect the conductance and other properties we are discussing?

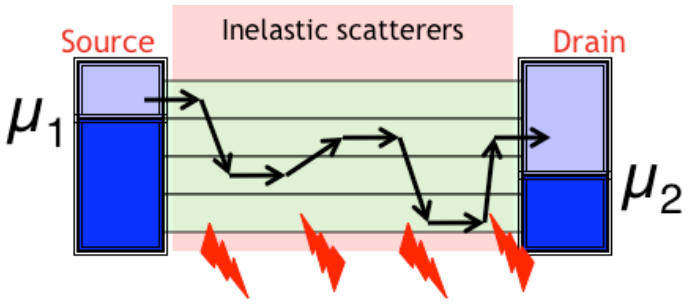


Fig.3.4
Real conductors have inelastic scatterers distributed throughout the channel.

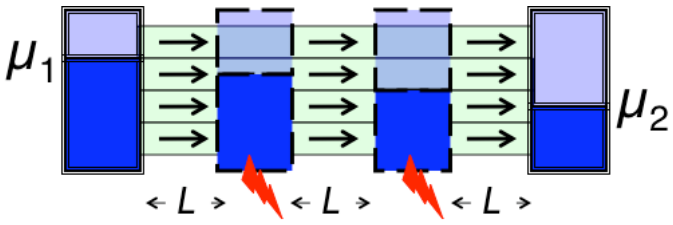


Fig.3.5
A hypothetical series of elastic resistors as an approximation to a real resistor with distributed inelastic scattering as shown in Fig.3.4.

One way to apply the elastic resistor model to a large conductor with distributed inelastic processes is to break up the latter conceptually into a sequence of elastic resistors (Fig.3.5), each much shorter than the physical length L, having a voltage that is only a fraction of the total

voltage V. We could then argue that the total resistance is the sum of the individual resistances.

This splitting of a long resistor into little sections of length shorter than L_{in} (L_{in}: length an electron travels on the average before getting inelastically scattered) also helps answer another question one may raise about the elastic resistor model. We obtained the linear conductance by resorting to a Taylor's series expansion (see Eq.(2.6)). But keeping the first term in the Taylor's series can be justified only for voltages $V < kT/q$, which at room temperature equals 25 mV. But everyday resistors are linear for voltages that are much larger. How do we explain that? The answer is that the elastic resistor model should only be applied to a short length $< L_{in}$ and as long as the voltage dropped over a length L_{in} is less than kT/q we expect the current to be linear with voltage. The terminal voltage can be much larger.

However, this splitting into short resistors needs to be done carefully. A key result we will discuss in the next Lecture is that Ohm's law should be modified

$$\textit{from} \quad \underbrace{R = \frac{\rho}{A} L}_{Eq.(1.1)} \quad \textit{to} \quad \underbrace{R = \frac{\rho}{A}(L + \lambda)}_{Eq.(1.4)}$$

to include an extra fixed resistance $\rho\lambda / A$ that is independent of the length and can be viewed as an interface resistance associated with the channel- contact interfaces. Here λ is a length of the order of a mean free path, so that this modification is primarily important for near ballistic conductors ($L \sim \lambda$) and is negligible for conductors that are many mean free paths long ($L >> \lambda$).

Conceptually, however, this additional resistance is very important if we wish to use the hypothetical structure in Fig.3.5 to understand the real structure in Fig.3.4. The structure in Fig.3.5 has too many interfaces that are not present in the real structure of Fig.3.4 and we have to remember to exclude the resistance coming from these conceptual interfaces.

For example, if each section in Fig.3.5 is of length L having a resistance of

$$R = \frac{\rho(L+\lambda)}{A}$$

then the correct resistance of the real structure in Fig.3.4 of length $3L$ is given by

$$R = \frac{\rho(3L+\lambda)}{A} \quad \textit{and NOT by} \quad R = \frac{\rho(3L+3\lambda)}{A}$$

Clearly we have to be careful to separate the interface resistance from the length dependent part. This is what we will do next.

Lecture 4

Ballistic and Diffusive Transport

4.1. Ballistic and Diffusive Transfer Times
4.2. Channels for Conduction

We saw in the last Lecture that the resistance of an elastic resistor can be written as

$$G = \frac{q^2 D}{2t} \qquad \text{(see Eq.(3.4))}$$

In this Lecture I will first argue that the transfer time t across a resistor of length L for diffusive transport with a mean free path λ can be related to the time t_B for ballistic transport by the relation (***Section 4.1***)

$$t = t_B\left(1 + \frac{L}{\lambda}\right) \tag{4.1}$$

Combining with Eq.(3.4) we obtain

$$G = \frac{G_B \lambda}{L+\lambda} \tag{4.2}$$

where

$$G_B \equiv \frac{q^2 D}{2t_B} \tag{4.3}$$

We could invert Eq.(4.2) to write the new Ohm's law

$$G = \frac{\sigma A}{L+\lambda}$$

where $\sigma A = G_B \lambda$

So far we have only talked about three dimensional resistors with a large cross-sectional area A. Many experiments involve two-dimensional resistors whose cross-section is effectively one-dimensional with a width W, so that the appropriate equations have the form

$$G = \frac{\sigma W}{L + \lambda}$$

where $\sigma W = G_B \lambda$

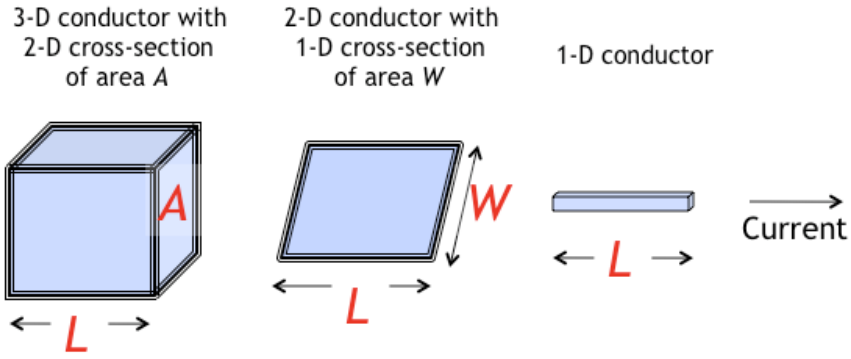


Fig.4.1. 3-D, 2-D and 1-D conductors

Finally we have one-dimensional conductors for which

$$G = \frac{\sigma}{L + \lambda}$$

where $\sigma = G_B \lambda$

We could collect all these results and write them compactly in the form

$$G = \frac{\sigma}{L+\lambda}\{1, \quad W, \quad A\} \tag{4.5}$$

with

$$\sigma = G_B\lambda\left\{1, \quad \frac{1}{W}, \quad \frac{1}{A}\right\} \tag{4.6}$$

where the three items in parenthesis correspond to 1-D, 2-D and 3-D conductors. Note that the conductivity σ has different dimensions in 1-D, 2-D and 3-D, while both G_B and λ have the same dimensions, namely Siemens (S) and meters (m) respectively.

The standard Ohm's law predicts that the resistance will approach zero as the length L is reduced to zero. Of course no one expects it to become zero, but the common belief is that it will approach a value determined by the interface resistance which can be made arbitrarily small with improved contacting technology.

What is now well established experimentally is that even with the best possible contacts, there is a minimum interface resistance determined by the properties of the channel, independent of the contact. The modified Ohm's law in Eq.(4.5) reflects this fact: Even a channel of zero length with perfect contacts has a resistance equal to that of a hypothetical channel of length λ.

But what does it mean to talk about the mean free path λ of a channel of zero length? The answer is that neither ρ nor λ mean anything for a short conductor, but their product $\rho\lambda$ does. The ballistic resistance has a simple meaning that has become clear in the light of modern experiments as we will see in ***Section 4.2.*** It is inversely proportional to the number of channels, *M(E)* available for conduction, which is proportional to, but not the same as, the density of states, *D(E).*

The concept of density of states has been with us since the earliest days of solid state physics. By contrast, the number of channels (or transverse modes) *M(E)* is a more recent concept whose significance was appreciated only after the seminal experiments in the 1980's on ballistic conductors showing conductance quantization.

4.1 Ballistic and Diffusive Transport

Consider how the two quantities in

$$G = \frac{q^2 D}{2t}$$

namely the density of states, D and the transfer time t scale with channel dimensions for large conductors. The first of these is relatively easy to see since we expect the number of states to be additive. A channel twice as big should have twice as many states, so that the density of states $D(E)$ for large conductors should be proportional to the volume ($A*L$).

Regarding the transfer time, t, broadly speaking there are two transport regimes:

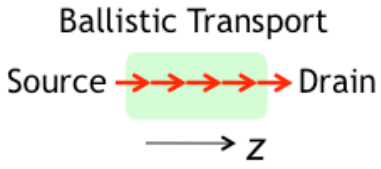
Ballistic regime: Transfer time $t \sim L$

Diffusive regime: Transfer time $t \sim L^2$

Consequently the ballistic conductance is proportional to the area (note that $D \sim A*L$ as discussed above), but ***independent of the length***. This "non-Ohmic" behavior has indeed been observed in short conductors. It is only diffusive conductors that show the "ohmic" behavior $G \sim A/L$.

These two regimes can be understood as follows. In the ballistic regime electrons travel straight from the source to the drain "like a bullet," taking a time

$$t_B = \frac{L}{\bar{u}} \qquad (4.7)$$



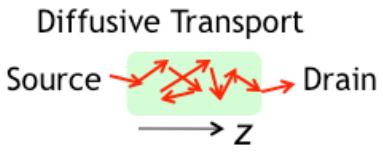
where $\bar{u} = \langle |v_z| \rangle$

is the average velocity of the electrons in the z-direction.

But conductors are typically not short enough for electrons to travel "like bullets." Instead they stumble along, getting scattered randomly by

various defects along the way taking much longer than the ballistic time in Eq.(4.7). We could write

$$t = \frac{L}{\bar{u}} + \frac{L^2}{2\bar{D}}$$



viewing it as a sort of "polynomial expansion" of the transfer time t in powers of L. We could then argue that the lowest term in this expansion must equal the ballistic limit $L/\bar{u}$, while the highest term should equal the diffusive limit well-known from the theory of random walks. This theory (see for example, Berg, 1983) identifies the coefficient $\bar{D}$ as the diffusion constant

$$\bar{D} = \left\langle v_z^2 \tau \right\rangle$$

τ being the mean free time.

We could use Eq.(4.7) to rewrite the expression for the transit time in Eq.(4.8) in the form

$$t = t_B \left(1 + \frac{L\bar{u}}{2\bar{D}} \right)$$

which agrees with Eq.(4.1) if the mean free path is given by

$$\lambda = \frac{2\bar{D}}{\bar{u}}$$

In defining the two constants $\bar{D}, \bar{u}$ we have used the symbol $<\cdots>$ to denote an average over the angular distribution of velocities which yields a different numerical factor depending on the dimensionality of the conductor (see Appendix B).

For ***d = {1, 2, 3}*** dimensions

$$\bar{u} \;=\; \left\langle \left| v_z \right| \right\rangle \;=\; v(E)\left\{1,\ \frac{2}{\pi},\ \frac{1}{2}\right\} \tag{4.8a}$$

and

$$\bar{D} \;=\; \left\langle v_z^2 \tau \right\rangle \;=\; v^2\tau(E)\left\{1,\ \frac{1}{2},\ \frac{1}{3}\right\} \tag{4.8b}$$

so that

$$\lambda \;=\; \frac{2\bar{D}}{\bar{u}} \;=\; v\tau\left\{2,\ \frac{\pi}{2},\ \frac{4}{3}\right\} \tag{4.9}$$

Note that our definition of the ***mean free path*** includes a dimension-dependent numerical factor over and above the standard value of $v\tau$. Couldn't we simply use the standard definition? We could, but then the new Ohm's law would not simply involve replacing L with L plus λ. Instead it would involve L plus a dimension-dependent factor times λ. Instead we have chosen to absorb this factor into the definition of λ.

Interestingly, even in one dimensional conductors the factor is not one, but two. This is because τ is the mean free time after which an electron gets scattered. Assuming the scattering to be isotropic, only half the scattering events will result in an electron traveling towards the drain to head towards the source. The mean free time for backscattering is thus 2τ, making the mean free path $2v\tau$ rather than $v\tau$.

Next we obtain an expression for the ***ballistic conductance*** by combining Eq.(4.3) with Eq.(4.7) to write

$$G_B \;\equiv\; \frac{q^2 D\,\bar{u}}{2L}$$

and then make use of Eq.(4.8a) to write

$$G_B \;\equiv\; \frac{q^2 D\,v}{2L}\left\{1,\ \frac{2}{\pi},\ \frac{1}{2}\right\} \tag{4.10}$$

Finally we can use Eqs.(4.9) and (4.10) in Eq.(4.6) and make use of Eq.(4.8b) to obtain an expression for the ***conductivity***:

$$\sigma = q^2 \bar{D} \frac{D}{L} \left\{ 1, \frac{1}{W}, \frac{1}{A} \right\} \tag{4.11}$$

We have thus obtained expressions for the conductance in the ballistic regime as well as the conductivity in the diffusive regime, starting from our expression

$$G = \frac{q^2 D}{2t}$$

based on the expression for the ballistic and diffusive transfer times

$$t = \frac{L}{\bar{u}} + \frac{L^2}{2\bar{D}}$$

which some readers may not find completely satisfactory. But this approach has the advantage of getting us to the new Ohm's law (Eq.(4.5)) very quickly using simple algebra. In Lecture 6 we will re-derive Eq.(4.5) more directly by solving a differential equation, without invoking the transfer time.

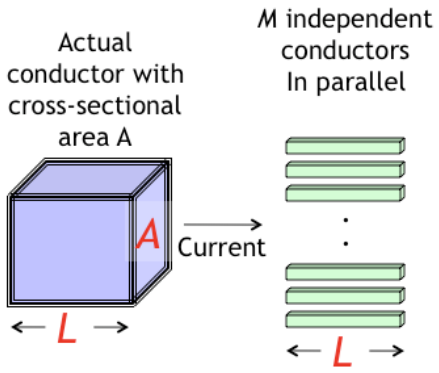
4.2 Channels for Conduction

Eq.(4.10) tells us that the ballistic conductance depends on D/L, the density of states per unit length. Since D is proportional to the volume, the ballistic conductance is expected to be proportional to the cross-sectional area A in 3-D conductors (or the width W in 2-D conductors).

Numerous experiments since the 1980's have shown that for small conductors, the ballistic conductance does not go down linearly with the area A. Rather it goes down in integer multiples of the ***conductance quantum***

$$G_B \equiv \underbrace{\frac{q^2}{h}}_{38\,\mu S} \underbrace{M}_{\text{integer}} \tag{4.12}$$

How can we understand this relation and what does the integer *M* represent? This result cannot come out of our elementary treatment of electrons in classical particle-like terms, since it involves Planck's constant *h*. Some input from quantum mechanics is clearly essential and this will come in Lecture 5 when we evaluate *D(E)*.



For the moment we note that heuristically Eq.(4.8) suggests that we visualize the real conductor as *M* independent channels in parallel whose conductances add up to give Eq.(4.12) for the ballistic conductance.

This suggests that we use Eqs.(4.10) and (4.12) to define a quantity *M(E)*

$$M \equiv \frac{hD\,v}{2L}\left\{1,\ \frac{2}{\pi},\ \frac{1}{2}\right\} \tag{4.13}$$

which should provide us a measure of the number of conducting channels. From Eqs.(4.6) and (4.12) we can write the conductivity in terms of *M* and the mean free path λ :

$$\sigma = \frac{q^2}{h} M\lambda \left\{1,\ \frac{1}{W},\ \frac{1}{A}\right\} \tag{4.14}$$

In the next Lecture we will use a simple model that incorporates the wave nature of electrons to show that for a one-dimensional channel the quantity *M* indeed equals one showing that it has only one channel, while for two- and three-dimensional conductors the quantity *M* represents the number of de Broglie wavelengths that fit into the cross-section, like the modes of a waveguide.

Lecture 5

Conductivity

5.1. E(p) or E(k) Relations
5.2. Counting States
5.3. Drude Formula
5.4. Is Conductivity proportional to Electron Density?
5.5. Quantized Conductance

A common expression for conductivity is the Drude formula relating the conductivity to the electron density *n,* the effective mass *m* and the mean free time

$$\sigma \equiv \frac{1}{\rho} = \frac{q^2 n\tau}{m} \tag{5.1a}$$

This expression is very well-known since even freshman physics texts start by deriving it. It also leads to the widely used concept of mobility

$$\bar{\mu} = \frac{q\tau}{m} \tag{5.1b}$$

such that

$$\sigma = qn\bar{\mu} \tag{5.1c}$$

On the other hand, in Lecture 4 we obtained two equivalent expressions for the conductivity, one as a product of the density of states *D* and the diffusion coefficient $\bar{D}$ (see Eq.(4.11))

$$\sigma(E) = q^2 \bar{D} \frac{D}{L}\left\{1, \frac{1}{W}, \frac{1}{A}\right\} \tag{5.2a}$$

and the other as a product of the number of modes *M* and the mean free path λ :

$$\sigma(E) \;=\; \frac{q^2}{h} M \lambda \left\{ 1, \;\; \frac{1}{W}, \;\; \frac{1}{A} \right\} \tag{5.2b}$$

Note that, like the conductance (see Eq.(3.1)), these expressions for the energy-dependent conductivity also have to be averaged over an energy range of a few kT's around $E=\mu_0$, using the thermal broadening function,

$$\bar{\sigma} \;=\; \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E} \right) \sigma(E) \tag{5.3a}$$

It is this averaged conductivity $\bar{\sigma}$ that should be compared to the Drude conductivity in Eq.(5.1). But for degenerate conductors (see Section 3.2.1) the averaged conductivity $\bar{\sigma}$ is approximately equal to the conductivity at an energy $E = \mu_0$:

$$\bar{\sigma} \;\approx\; \sigma(E = \mu_0) \tag{5.3b}$$

and so we can compare $\sigma(E = \mu_0)$ from Eq.(5.2) to Eq.(5.1).

Although Eq.(5.2b) is not very well-known, the equivalent version in Eq.(5.2a) is a standard result that is derived in many textbooks. However, the usual derivation of Eq.(5.2a) requires advanced concepts like the Boltzmann or the Kubo formalism and so appears much later than Eq.(5.1) in any solid-state physics text. Not surprisingly, most people remember Eq.(5.1) and not Eq.(5.2).

But the point we wish to stress is that while Eq.(5.1) is often very useful, it is a result of limited validity that can be obtained from Eq.(5.2) by making suitable approximations based on a specific model. But when these approximations are not appropriate, we can still use Eq.(5.2) which is ***far more generally applicable***. For example, Eq.(5.2) gives sensible answers even for materials like graphene whose non-parabolic bands make the meaning of mass somewhat unclear, causing considerable confusion when using Eq.(5.1). In general we should really use Eq.(5.2), and not Eq.(5.1), to shape our thinking about conductivity.

There is a fundamental difference between Eq.(5.2) and (5.1). The averaging implied in Eq.(5.3) makes the conductivity a "Fermi surface property", that is one that depends only on the energy levels close to $E=\mu_0$. By contrast, Eq.(5.1) depends on the total electron density n integrated over all energy. But this dependence on the total number is true only in a limited sense.

Experts know that n only represents the density of "free" electrons and have an instinctive feeling for what it means to be free. They know that there are p-type semiconductors which conduct better when they have fewer electrons, but in that case they know that n should be interpreted to mean the number of "holes". For beginners, all this appears confusing and much of this confusion can be avoided by using Eq.(5.2) instead of (5.1).

Interestingly, Eq.(5.2a) was used in a seminal paper to obtain Eq.(3.4)

$$G = \frac{q^2 D}{2t} \qquad \text{(same as Eq.(3.4))}$$

(see Eq.(1) of Thouless (1977)). Instead we have used the concept of an elastic resistor to first obtain Eq.(3.4) from elementary arguments, and then used it to obtain Eq.(5.2a).

Eq.(5.2) stresses that the essential factor determining the conductivity is the density of states around $E=\mu_0$. Materials are known to have conductivities ranging over many orders of magnitude from glass to copper. And the basic fact remains that they all have approximately the same number of electrons. Glass is not an insulator because it is lacking in electrons. It is an insulator because it has a very low density of states or number of modes around $E=\mu_0$.

So when does Eq.(5.2) reduce to (5.1)? Answer: If the electrons are described by a "single band effective mass model" as I will try to show in this Lecture. So far we have kept our discussion general in terms of the density of states, $D(E)$ and the velocity, $v(E)$ without adopting any specific models. These concepts are generally applicable even to

amorphous materials and molecular conductors. A vast amount of literature both in condensed matter physics and in solid state devices, however, is devoted to crystalline solids with a periodic arrangement of atoms because of the major role they have played from both basic and applied points of view.

For such materials, energy levels over a limited range of energies are described by a *E(p)* relation and we will show in this Lecture that irrespective of the specific *E(p)* relation, at any energy E the density of states *D(E),* velocity *v(E)* and momentum *p(E)* are related to the total number of states *N(E)* with energy less than E by the relation (*d*: number of dimensions)

$$D(E)v(E)p(E) = N(E).d \tag{5.4}$$

We can combine this relation with Eq.(5.2a) and make use of Eq.(4.8b), $\bar{D} = v^2\tau / d$, to write

$$\sigma(E) = \frac{q^2\tau(E)}{m(E)} \left\{ \frac{N(E)}{L}, \quad \frac{N(E)}{W\,L}, \quad \frac{N(E)}{A\,L} \right\} \tag{5.5}$$

where we have defined mass as

$$m(E) = \frac{p(E)}{v(E)} \tag{5.6}$$

For parabolic *E(p)* relations, the mass is independent of energy, but in general it could be energy-dependent.

Eq.(5.5) indeed looks like Drude expression (Eq.(5.1a)) if we identify the quantity in parenthesis *{N/L, N/WL, N/AL}* as the electron density, n per unit length, area and volume in 1D, 2D and 3D respectively. At low temperatures, this is easy to justify since the energy averaging in Eq.(5.3) amounts to looking at the value at $E = \mu_0$ and *N(E)* at $E = \mu_0$ represents the total number of electrons (Fig.5.1).

At non-zero temperatures one needs a longer discussion which we will get into later in the Lecture. Indeed as will see, some subtleties are

involved even at zero temperature when dealing with differently shaped density of states.

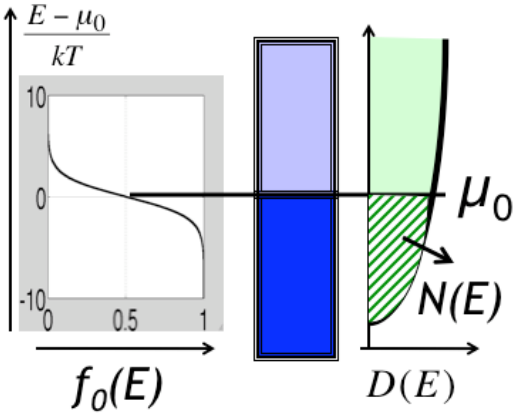


Fig.5.1. Equilibrium Fermi function $f_0(E)$, Density of states D(E) and integrated density of states N(E).

Note, however, that the key to reducing our conductivity expression (Eq.(5.2)) to the Drude-like expression (Eq.(5.5)) is Eq.(5.4) which is an interesting relation for it relates *D(E), v(E)* and *p(E)* at a given energy *E*, to the total number of states *N(E)* obtained by integrating *D(E)*

$$N(E) = \int_{-\infty}^{E} dE\ D(E)$$

How can the integrated value of *D(E)* be uniquely related to the value of quantities like *D(E), v(E)* and *p(E)* at a single energy ? The answer is that this relation holds only as long as the energy levels are given by a single E(p) relation. It may not hold in an energy range with multiple bands of energies or in an amorphous solid not described by an *E(p)* relation. Eq.(5.2) is then not equivalent to Eq.(5.5), and ***it is Eq.(5.2) that can be trusted.***

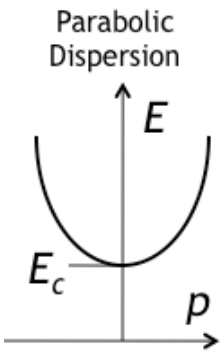
With that long introduction let us now look at how single bands described by an *E(p)* relation leads to Eq.(5.4) and helps us connect our

conductivity expression (Eq.(5.2)) to the Drude formula (Eq.(5.1)). This will also lead to a different interpretation of the quantity *M(E)* introduced in the last Lecture, that will help understand why it is an integer representing the number of channels.

5.1 E(p) or E(k) relations for crystalline solids

The general principle for calculating *D(E)* is to start from the Schrodinger equation treating the electron as a wave confined to the solid. Confined waves (like a guitar string) have resonant "frequencies" and these are basically the allowed energy levels. By counting the number of energy levels in a range *E* to *E+dE,* we obtain the density of states *D(E).*

Although the principle is simple, a first principles implementation is fairly complicated since one needs to start from a Schrodinger equation including the nuclear potential that the electrons feel inside the solid. One of the seminal concepts in solid state physics is the realization that in crystalline solids electrons behave as if they are in vacuum, but with an effective mass different from their natural mass, so that the energy-momentum relation can be written as



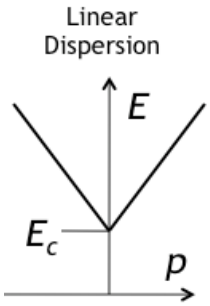
$$E(p) = E_c + \frac{p^2}{2m} \qquad (5.7a)$$

where E_c is a constant. The momentum p is equated to $\hbar k$, providing the link between the energy-momentum relation *E(p)* associated with the particle viewpoint and the dispersion relation *E(k)* associated with the wave viewpoint. Here we will write everything in terms of *p*, but they are easily translated in terms of $k = p/\hbar$.

Eq.(5.7a) is generally referred to as a parabolic dispersion relation and is commonly used in a wide variety of materials from metals like copper to semiconductors like silicon, because it often approximates the actual

E(p) relation fairly well in the energy range of interest. But it is by no means the only possibility. Graphene, a material of great current interest, is described by a linear relation:

$$E = E_c + v_0\, p \qquad (5.7b)$$



where v_0 is a constant. Note that p denotes the magnitude of the momentum and we will assume that the *E(p)* relation is isotropic, which means that it is the same regardless of which direction the momentum vector points.

For any given isotropic *E(p)* relation, the velocity points in the same direction as the momentum, while its magnitude is given by

$$v \equiv \frac{dE}{dp} \qquad (5.8)$$

This is a general relation applicable to arbitrary energy-momentum relations for classical particles. On the other hand, in wave mechanics it is justified as the group velocity for a given dispersion relation *E(k)*.

5.2 Counting states

One great advantage of this principle is that it reduces the complicated problem of electron waves in a solid to that of waves in vacuum, where the allowed energy levels can be determined the same way we find the resonant frequencies of a guitar string: simply by requiring that an integer number of wavelengths fit into the solid. Noting that the de Broglie principle relates the electron wavelength to the Planck's constant divided by its momentum, h / p, we can write

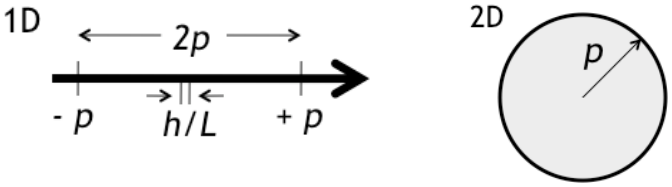
$$\frac{L}{h/p} = Integer \quad \Rightarrow \quad p = Integer * \left(\frac{h}{L}\right)$$

where L is the length of the box. This means that the allowed states are uniformly distributed in p with each state occupying a "space" of

$$\Delta p = \frac{h}{L} \tag{5.9}$$

Let us define a function *N(p)* that tells us the total number of states that have a momentum less than a given value *p*. In *one dimension* this function is written down by dividing the total range of *2p* (from *-p* to *+p*) by the spacing *h/L*:

$$N(p) = \frac{2p}{h/L} = 2L\left(\frac{p}{h}\right) \quad \text{1D}$$



In two dimensions we divide the area of a circle of radius *p* by the spacing *h/L* * *h/W*, *L* and *W* being the dimensions of the two dimensional box.

$$N(p) = \frac{\pi p^2}{(h/L)(h/W)} = \pi WL\left(\frac{p}{h}\right)^2 \quad \text{2D}$$

In three dimensions we divide the volume of a sphere of radius *p* by the spacing *h/L* * *h/W*$_1$ * *h/W*$_2$, *L, W*$_1$ and *W*$_2$ being the dimensions of the three dimensional box. Writing *A = W*$_1$ * *W*$_2$ we have

$$N(p) = \frac{(4\pi/3)p^3}{(h/L)(h^2/A)} = \frac{4\pi}{3}AL\left(\frac{p}{h}\right)^3 \quad \text{3D}$$

We can combine all three results into a single expression for *d = {1, 2, 3} dimensions:*

$$N(p) = \left\{2\frac{L}{h/p},\ \pi\frac{LW}{(h/p)^2},\ \frac{4\pi}{3}\frac{LA}{(h/p)^3}\right\} \tag{5.10}$$

We could use a given *E(p)* relation to turn this function *N(p)* into a function of energy *N(E)* that tells us the total number of states with energy less than *E*.

5.2.1. Density of states, D(E)

This function *N(E)* that we have just obtained must equal the density of states *D(E)* ***integrated*** up to an energy *E*, so that *D(E)* can be obtained from the derivative of *N(E):*

$$N(E) = \int_{-\infty}^{E} dE'\, D(E') \quad \rightarrow \quad D(E) = \frac{dN}{dE}$$

Hence from Eq.(5.10),

$$D(E) \;=\; \frac{dN}{dp}\frac{dp}{dE} \;=\; K_N \frac{dp}{dE}\frac{p^{d-1}d}{h^d}$$

Making use of Eqs.(5.8) and (5.10), we obtain the relation stated earlier

$$D(E)v(E)p(E) = N(E)\,.d \qquad \text{(same as Eq.(5.4))}$$

which is completely general *independent of the actual E(p) relation.*

5.3 Drude formula

As noted earlier, using this relation we can rewrite Eq.(5.2) in the form

$$\sigma(E) \;=\; \frac{q^2\tau(E)}{m(E)}\left\{\frac{N(E)}{L},\;\; \frac{N(E)}{WL},\;\; \frac{N(E)}{AL}\right\} \qquad \text{(same as Eq.(5.5))}$$

with an energy-dependent mass *m(E)* defined in Eq.(5.6)). As we have seen, it is straightforward to connect this relation to the Drude formula (Eq.(5.1)) at low temperatures where the energy averaging in Eq.(5.3) amounts to looking at the value at a single energy $E=\mu_0$. What about non-zero temperatures?

5.3.1. n-type Conductors

Using Eq.(5.5) and assuming m and τ to be energy-independent we have

$$\bar{\sigma} = \frac{q^2\tau}{m}\int_{-\infty}^{+\infty} dE\left(-\frac{\partial f_0}{\partial E}\right)N(E)\left\{\frac{1}{L},\ \frac{1}{WL},\ \frac{1}{AL}\right\} \qquad (5.11)$$

The integral can be carried out "by parts" to yield

$$\int_{-\infty}^{+\infty} dE\left(-\frac{\partial f_0}{\partial E}\right)N(E) = \left[-N(E)f_0(E)\right]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} dE\,\frac{dN}{dE}\,f_0(E)$$

$$= [0-0] + \int_{-\infty}^{+\infty} dE\,D(E)\,f_0(E)$$

$$= \textit{Total Number of Electrons}$$

since $dE\,D(E)\,f_0(E)$ tells us the number of electrons in the energy range from E to $E+dE$. When integrated it gives us the total number of electrons.

Eq.(5.11) then reduces to

$$\bar{\sigma} = \frac{q^2\tau}{m}\,(\textit{Number of electrons})\left\{\frac{1}{\textit{Length}},\ \frac{1}{\textit{Area}},\ \frac{1}{\textit{Volume}}\right\}$$

which is the Drude formula stated in Eq.(5.1a).

5.3.2. p-type conductors

An interesting subtlety is involved when we consider a p-type conductor for which the $E(p)$ relation extends downwards, say something like

$$E(p) = E_v - \frac{p^2}{2m}$$

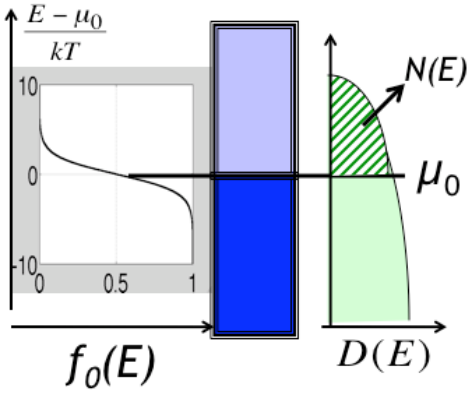


Fig.5.2: Equilibrium Fermi function $f_0(E)$, Density of states $D(E)$ and integrated density of states $N(E)$: p-type conductor.

Instead of $$N(E) = \int_{-\infty}^{E} dE'\, D(E')$$

we now have (see Fig.5.2)

$$N(E) = \int_{E}^{+\infty} dE'\, D(E') \quad \rightarrow \quad D(E) = -\frac{dN}{dE}$$

This is because we defined the function $N(E)$ from $N(p)$ which represents the total number of states with momenta less than p, which means energies greater than E for a p-type dispersion relation.

Now if we carry out the integration by parts as before

$$\int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E}\right) N(E) \;=\; \left[-N(E) f_0(E)\right]_{-\infty}^{+\infty} \;+\; \int_{-\infty}^{+\infty} dE \frac{dN}{dE} f_0(E)$$

we run into a problem because the first term does not vanish at the lower limit where both $N(E)$ and $f_0(E)$ are both non-zero.

We can get around this problem by writing the derivative in terms of $1\text{-}f_0$ instead of f_0:

$$\int_{-\infty}^{+\infty} dE \left(\frac{\partial(1-f_0)}{\partial E} \right) N(E)$$

$$= \left[-N(E)(1-f_0(E)) \right]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} dE \, \frac{dN}{dE} (1-f_0(E))$$

$$= \left[0-0 \right] + \int_{-\infty}^{+\infty} dE \, D(E) (1-f_0(E))$$

$$= \textit{Total Number of "holes"}, P$$

What this means is that with p-type conductors we can use the Drude formula

$$\sigma = q^2 n \tau / m$$

but the *n* now represents the density of empty states or holes. A larger *n* really means fewer electrons.

5.3.3. "Double-ended" density of states

How would we count *n* for a density of states *D(E)* that extends in both directions as shown in Fig.5.3 (left panel). This is representative of graphene, a material of great interest (recognized by the 2010 Nobel prize in physics), whose *E(p)* relation is commonly approximated by

$$E = \pm v_0 \, p \, .$$

People usually come up with clever ways to handle such "double-ended" density of states so that the Drude formula can be used. For example they divide the total density of states into an n-type and a p-type component

$$D(E) = D_n(E) + D_p(E)$$

as shown in Fig.5.3 and the two components are then handled separately, using a prescription that is less than obvious: The conductivity due to the upper half D_n depends on the number of occupied states (electrons),

while that due to the lower half depends on the number of unoccupied states (holes).

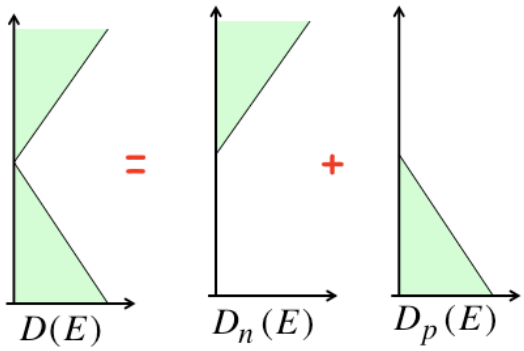


Fig.5.3.
A "double-ended" density of states can be visualized as a sum of an "n-type component" and a "p-type component."

But the point we would like to stress is that there is really no particular reason to insist on using a Drude formula and keep inventing clever ways to make it work. One might just as well use Eq.(5.2) which reflects the correct physics of conduction, ***namely that it takes place in a narrow band of energies around*** μ_0.

5.4. Is conductivity proportional to electron density?

Experimental conductivity measurements are often performed as a function of the electron density and the common expectation based on the Drude formula (Eq.(5.1)) is that conductivity should be proportional to the electron density and any non-linearity must be a consequence of the energy-dependence of the mean free time. What is not often recognized is that for non-parabolic dispersion relations, the mass itself defined as p/v can be energy-dependent and this will affect the conductivity- electron density relation.

First we note that from Eq.(5.10)

$$n(p) \;=\; \left\{ 2\frac{p}{h},\;\; \pi\frac{p^2}{h^2},\;\; \frac{4\pi}{3}\frac{p^3}{h^3} \right\} \qquad (5.12a)$$

where we have defined *n* as *N/L* or *N/WL* or *N/AL* in 1, 2 and 3 dimensions respectively. Writing (*d* = dimensions, *K* = constant)

$$n(p) \;=\; K\, p^d \qquad (5.12b)$$

we have $$\sigma = q^2 \frac{n(p)\tau(p)}{m(p)} \;=\; q^2 K\, p^{d-1}\, v(p)\tau(p) \qquad (5.12c)$$

If we know how the velocity and the mean free time vary with *E* (and hence with *p*) we could eliminate *p* from the expressions for the conductivity, σ and the electron density, *n* to obtain a direct relation between them (for degenerate conductors, as explained in the introduction)

For example, with graphene, $E = \pm v_0 p$, so that the velocity *dE/dp* is a constant (v_0), independent of *p*. If we assume an energy independent mean free time τ, we obtain

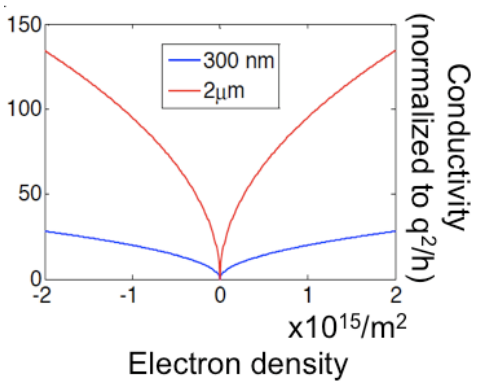
$$\sigma \;=\; \frac{q^2}{h}\lambda\sqrt{\frac{4n}{\pi}} \qquad (5.13)$$

after a little algebra, noting that graphene is two-dimensional and making use of Eq.(4.9) for the mean free path.

To compare with experiments, we need to modify Eq.(5.13) a little to account for the degeneracy factor *g* which denotes the number of equivalent states. For example all non-magnetic materials have two spin states with identical energies, which would make *g=2.* Certain materials also have equivalent "valleys" having identical energy momenta relations so that the *N* we calculate for one valley has to be multiplied by *g* when relating to the experimentally measured electron densities. For graphene, *g = 2*2 = 4.*

Eq.(5.13) applies to a single spin and valley for which the conductivity and the electron density are each *1/g* times the actual, so that

$$\sigma/g = \frac{q^2}{h}\lambda\sqrt{\frac{4n/g}{\pi}} \quad \rightarrow \quad \sigma = \frac{q^2}{h}\lambda\sqrt{\frac{4gn}{\pi}} \tag{5.14}$$



The calculated results from Eq.(5.14) with $\lambda = 2$ μm and with $\lambda =$ 300 nm compares well with the experimental data on graphene reported in Bolotin et al. (2008). Note that the values of the mean free path indicated in the paper are half the values we have used. This is because our definition of mean free path differs from the standard one by a dimension-dependent factor (see Eq.(4.9)).

I should mention, however, that long graphene samples often show a conductivity ~ n and not ~$\sqrt{n}$. This is believed to be because the mean free time and hence the mean free path λ due to charged impurity scattering is ~ E ~$\sqrt{n}$. Eq.(5.14) then predicts a conductivity ~ *n*. It is only for an energy-independent mean free path that Eq.(5.14) predicts a ~ $\sqrt{n}$ dependence of the conductivity and this is only seen in short near ballistic samples for which the mean free path plays no role.

5.5. Quantized Conductance

I noted in the last Lecture 4 that the ballistic conductance is given by

$$G_B \equiv \underbrace{\frac{q^2}{h}}_{38\,\mu S} \underbrace{M}_{\text{int eger}} \qquad \text{(same as Eq.(4.12))}$$

and that experimentally M is found to be an integer in low dimensional conductors at low temperatures. However, in the last lecture we defined M (see Eq.(4.13))

$$M \equiv \frac{hD\,v}{2L}\left\{1, \quad \frac{2}{\pi}, \quad \frac{1}{2}\right\} \tag{5.15}$$

as the product of the density of states and the velocity and it is not at all clear why it should be an integer. Using the *E(p)* relations discussed in this Lecture we will now show that we can interpret *M(p)* in a very different way that helps see its integer nature.

First we make use of Eq.(5.4) to rewrite Eq.(5.15) in the form

$$M = \frac{hN}{2L\,p}\left\{1, \quad \frac{4}{\pi}, \quad \frac{3}{2}\right\} \tag{5.16}$$

where *N(p)* is the total number of states with a momentum that is less than p and we have seen that it is equal to the number of wavelengths that fit into the solid. Making use of Eq.(5.10) for *N(p),* we obtain from Eq.(5.16)

$$M(p) = \left\{1, \quad 2\frac{W}{h/p}, \quad \pi\frac{A}{(h/p)^2}\right\} \tag{5.17}$$

Just as *N(p)* tells us the number of wavelengths that fit into the volume, *M(p) tells us the number that fit into the cross-section* and this result is independent of the actual E(p) relation, since we have not made use of any specific relationship.

Now we are ready to look at the origin of conductance quantization. If we evaluate our expressions for *N(p)* and *M(p)* for a given sample we will in general get a fractional number. However, since these quantities

represent the number of states, we would expect them to be integers and if we obtain say *201.59*, we should take the lower integer *201*.

This point is commonly ignored in large conductors at high temperatures, where experiments do not show this quantization because of the energy averaging over $\mu_0 \pm 2kT$ associated with experimental measurements. For example, if over this energy range, *M(E)* varies from say *201.59* to *311.67*, then it seems acceptable to ignore the fact that it really varies from *201* to *311*.

But in small structures where one or more dimensions is small enough to fit only a few wavelengths the integer nature of *M* is observable and shows up in the quantization of the ballistic conductance. We should then rewrite Eq.(5.17) as

$$M(p) \quad = \quad Int\left\{1, \quad 2\frac{W}{h/p}, \quad \pi\frac{A}{(h/p)^2}\right\} \tag{5.18}$$

where *Int (x)* represents the largest integer less than or equal to *x*.

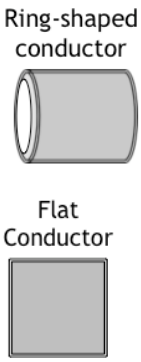
For one dimensional conductors the number of modes is equal to g, which is the number of spins times the number of valleys. Ballistic conductors have a resistance of

$$\frac{h}{q^2 M} \approx \frac{25\ K\Omega}{M}$$

so that the resistance of a 1D ballistic conductor is approximately equal to $25\ K\Omega$ divided by g. This has indeed been observed experimentally. Most metals and semiconductors like GaAs have *g=2*, and the 1D ballistic resistance $\sim 12.5\ K\Omega$. But carbon nanotubes have two valleys as well making g=4 and exhibit a ballistic resistance $\sim 6.25\ K\Omega$.

For two- and three-dimensional conductors, Eq.(5.17) is not quite right, because it is based on the heuristic idea of counting modes by counting the number of wavelengths that fit into the solid (see Eq.(5.5)). Mathematically it can be justified only if we assume periodic boundary conditions, that is if we assume that the cross-section is in the form of a

ring rather than a flat sheet for a 2D conductor. For a 3D conductor it is hard to visualize what periodic boundary conditions might look like though it is easy to impose it mathematically as we have been doing.



Most real conductors do not come in the form of rings, yet periodic boundary conditions are widely used because it is mathematically convenient and people believe that the actual boundary conditions do not really matter. But this is true only if the cross-section is large. For small area conductors the actual boundary conditions do matter and we cannot use Eq.(5.10).

Interestingly a conductor of great current interest has actually been studied in both forms: a ring-shaped form called a carbon nanotube and a flat form called graphene. If the circumference or width is tens of nanometers they have much the same properties, but if it is a few nanometers their properties are observably different including their ballistic resistances.

Lecture 6

Diffusion Equation for Ballistic Transport

6.1. Electrochemical Potentials Out of Equilibrium
6.2. Currents in Terms of Non-Equilibrium Potentials

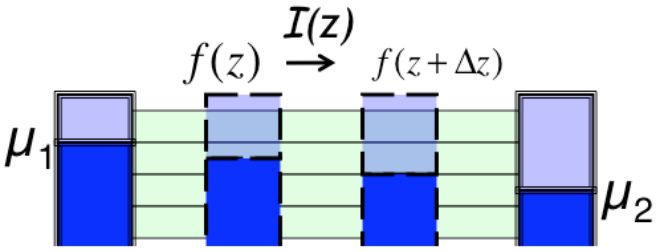
The title of this Lecture may sound contradictory, like the elastic resistor. Doesn't the diffusion equation describe diffusive transport? How can one use it for ballistic transport? An important idea we are trying to get across with our bottom-up approach is the essential unity of these two regimes of transport and hopefully this lecture will help.

The diffusion equation relates the current to the slope of the electrochemical potential $\mu(z)$

$$\frac{I}{A} = -\frac{\sigma}{q}\frac{d\mu}{dz} \tag{6.1a}$$

where σ is the conductivity (Eq.(5.2)) from the last Lecture.

We can obtain this equation by viewing a long conductor as a series of elastic resistors as discussed in Section 3.3:



Using Eq.(3.3) we can write the current *I(z)* in a section of the conductor as

$$I(z) = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, G(E)\,(f(z,E)-f(z+\Delta z,E))$$

From Eq.(4.5) we could write

$$\frac{1}{G(E)} = \rho\frac{\Delta z+\lambda}{A}$$

but the point to note is that part of this resistance represents the interface resistance, which should not be included since there are no actual interfaces except at the very ends. Omitting the interface resistance we can write (Note: $\sigma = 1/\rho$, Eq.(1.1))

$$G(E) = \frac{\sigma A}{\Delta z}$$

Combining this with our usual linear expansion for small potential differences from Eq.(2.7)

$$f(z,E)-f(z+\Delta z,E) \approx \left(-\frac{\partial f_0}{\partial E}\right)\left(\,\mu(z)-\mu(z+\Delta z)\,\right)$$

and defining the conductivity σ as the thermal average of $\sigma(E)$ (Eq.(5.3)), we obtain

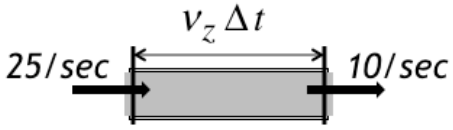
$$I(z) = \frac{1}{q}\frac{\sigma A}{\Delta z}(\mu(z)-\mu(z+\Delta z))$$

letting $\Delta z \rightarrow 0$, we obtain the diffusion equation stated above in Eq.(6.1a).

The diffusion equation is usually combined with a second equation called the continuity equation. For one-dimensional structures (see Fig.6.1), under steady-state conditions, the current must be the same at all *z*:

$$\frac{dI}{dz} = 0 \tag{6.1b}$$

The reason is easy to see. If we have a current of *25* electrons per second entering a section of the conductor and only *10* electrons per second leaving it, then the number of electrons will be building up in this section at the rate of *25-10=15* per second. That is a transient condition, not a steady-state one. Under steady-state conditions the current has to be the same at all points along the z-axis as required by Eq.(6.1b).



The standard approach is to solve Eqs.(6.1a,b) with the boundary conditions

$$\mu(z=0) = \mu_1 \tag{6.2a}$$

$$\mu(z=L) = \mu_2 \tag{6.2b}$$

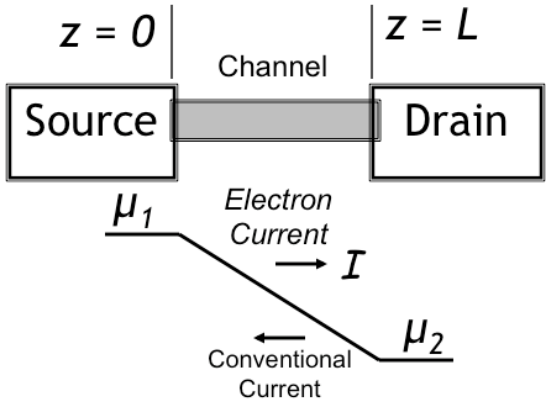


Fig.6.1. Solution to Eqs.(6.1a,b) with the boundary conditions in Eq.(6.2). Note that we are using I to represent the electron current as explained earlier (see Fig.3.2).

It is easy to see that the linear solution sketched in Fig.6.1 meets the boundary conditions in Eq.(6.2) and at the same time satisfies both Eqs.(6.1a,b) since a linear $\mu(z)$ has a constant $d\mu/dz$

$$\frac{d\mu}{dz} = -\frac{\mu_1 - \mu_2}{L}$$

so that from Eq.(6.1a) we have a constant current with $dI/dz = 0$:

$$I = \frac{\sigma A}{q}\frac{\mu_1 - \mu_2}{L}$$

Note that $\mu_1 - \mu_2 = qV$ (Eq.(2.1)), so that

$$I = \frac{\sigma A}{L}V \tag{6.3a}$$

which is the standard Ohm's law and not the generalized one we have been discussing

$$I = \frac{\sigma A}{L+\lambda}V \tag{6.3b}$$

that includes ballistic channels as well.

Can we obtain this result (Eq.(6.3b)) from the diffusion equation (Eqs.(6.1a,b))? Many would say that a whole new approach is needed since quantities like the conductivity or the diffusion coefficient mean nothing for a ballistic channel. The central result I wish to establish in this Lecture is that we can still use Eqs.(6.1a,b) provided we modify the boundary conditions in Eq.(6.2) to reflect the interface resistance that we have been talking about:

$$\mu(z=0) = \mu_1 - \frac{qI\,R_B}{2} \tag{6.4a}$$

$$\mu(z=L) = \mu_2 + \frac{qI\,R_B}{2} \tag{6.4b}$$

R_B being the inverse of the ballistic conductance G_B discussed earlier (see Eqs.(4.6), (4.12)):

$$R_B = \frac{\lambda}{\sigma A} = \frac{h}{q^2 M} \tag{6.5}$$

The new boundary conditions in Eqs.(6.4a,b) can be visualized in terms of lumped resistors $R_B/2$ at the interfaces as shown in Fig.6.2. leading to additional potential drops as shown.

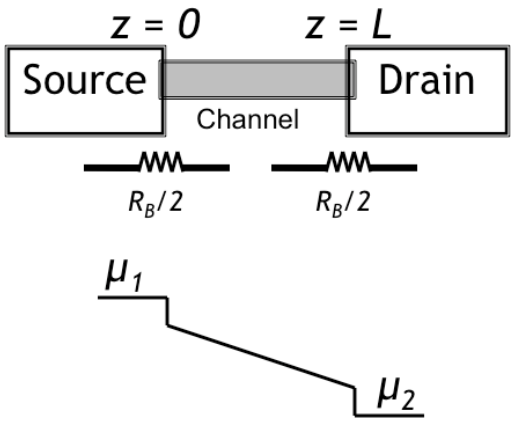


Fig.6.2. Eqs.(6.1a,b) can be used to model both ballistic and diffusive transport provided we modify the boundary conditions in Eq.(6.2) to reflect the two interface resistances, each equal to $R_B/2$.

It is straightforward to see that this ***new boundary condition*** applied to a uniform resistor leads to the new Ohm's law in Eq.(6.3b). Since $\mu(z)$ varies linearly from $z=0$ to $z=L$, the current is obtained from Eq.(6.1a)

$$I = \frac{\sigma A}{q} \frac{\mu(0) - \mu(L)}{L}$$

Using Eqs.(6.4a,b)

$$I = \frac{\sigma A}{q} \left(\frac{\mu_1 - \mu_2}{L} - \frac{qI R_B}{L} \right)$$

Since $\sigma A R_B = \lambda$, $\quad I\left(1+\frac{\lambda}{L}\right) = \frac{\sigma A}{q}\left(\frac{\mu_1-\mu_2}{L}\right)$

Noting that $\mu_1 - \mu_2 = qV$ (Eq.(2.1)), this yields Eq.(6.3b).

But how do we justify this new boundary condition (Eqs.(6.4a,b))? It follows from the new Ohm's law (Eq.(6.3b)) if we assume that the extra resistance $\sigma A / \lambda$ corresponding to $L=0$ is equally divided between the two interfaces.

For a better justification, we need to introduce two different electrochemical potentials μ^+ and μ^- for electrons moving along $+z$ and $-z$ respectively. In previous lectures we talked about electrochemical potentials inside the ***contacts*** which are large regions that always remain close to equilibrium and hence are described by Fermi functions (see Eq.(2.5)) with well-defined electrochemical potentials.

By contrast in this Lecture we are using $\mu(z)$ to represent quantities inside the out-of-equilibrium ***channel***, where it is at best an approximate concept since the electron distribution among the available states need not follow a Fermi function. Even if it does, electronic states carrying current along $+z$ must be occupied differently from those carrying current along $-z$, or else there would be no net current.

This difference in occupation is reflected in different electrochemical potentials μ^+ and μ^- and we will show that the current is proportional to the difference (***Section 6.2***)

$$I = \frac{q}{h} M \left(\mu^+(z) - \mu^-(z)\right) \tag{6.6a}$$

which can also be rewritten in the form

$$I = \frac{1}{qR_B}\left(\mu^+(z) - \mu^-(z)\right) \tag{6.6b}$$

$$= \frac{\sigma A}{q\lambda}\left(\mu^{+}(z)-\mu^{-}(z)\right) \tag{6.6c}$$

using Eq.(4.12). The correct boundary conditions for μ^{+} and μ^{-} are

$$\begin{aligned} \mu^{+}(z=0) &= \mu_1, \\ \mu^{-}(z=L) &= \mu_2 \end{aligned} \tag{6.7}$$

which can be understood by noting that at $z=0$ the electrons moving along $+z$ have just emerged from the left contact and hence have the same distribution and electrochemical potential, μ_1. Similarly at $z=L$ the electrons moving along $-z$ have just emerged from the right contact and thus have the same potential μ_2 (Fig.6.3).

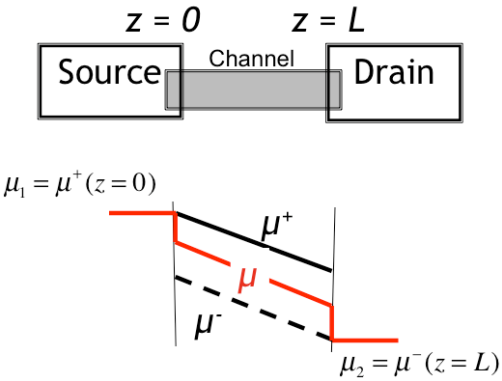


Fig.6.3.
Spatial profile of electrochemical potentials μ^{+}, μ^{-} across a diffusive channel.

In the next Lecture I will show that the current is related to the potentials μ^{+} and μ^{-} by an equation

$$I = -\frac{\sigma A}{q}\frac{d\mu^{+}}{dz} = -\frac{\sigma A}{q}\frac{d\mu^{-}}{dz} \tag{6.8}$$

that looks just like the diffusion equation (Eq.(6.1a)) which applies to the average potential:

$$\mu(z) \;=\; \frac{\mu^{+}(z)+\mu^{-}(z)}{2} \tag{6.9}$$

Eq.(6.8) can be solved with the boundary conditions in Eq.(6.7) to obtain the plot shown in Fig.6.3 for μ^{+}, μ^{-} and their average indeed looks like Fig.6.2 for μ with its discontinuities at the ends.

However, it is not necessary to abandon the traditional diffusion equation (Eq.(6.1a)) in favor of the new diffusion equation (Eq.(6.8)). We can obtain the same results simply by modifying the boundary conditions for $\mu(z)$ as follows:

$$\mu(z=0) \;=\; \left(\frac{\mu^{+}+\mu^{-}}{2}\right)_{z=0} \;=\; \left(\mu^{+}-\frac{\mu^{+}-\mu^{-}}{2}\right)_{z=0}$$

$$\;=\; \mu_1 \;-\; (q\,I\,R_B/2)$$

making use of Eqs.(6.6) and (6.7). Similarly

$$\mu(z=L) \;=\; \left(\mu^{-}+\frac{\mu^{+}-\mu^{-}}{2}\right)_{z=L} \;=\; \mu_2 \;+\; \frac{qI\,R_B}{2}$$

These are exactly the new boundary conditions for the standard diffusion equation that we mentioned earlier (Eqs.(6.4a,b)).

Let me finish up this Lecture by establishing the key result we stated without proof in the above discussion, namely, Eq.(6.6) (Section 6.2). But first let me say a few words about how the non-equilibrium potentials μ^{+} and μ^{-} are defined. (Section 6.1).

6.1. Electrochemical Potentials Out of Equilibrium

As I mentioned earlier, it is conceptually straightforward to talk about electrochemical potentials inside the ***contacts*** which are large regions that always remain close to equilibrium and hence are described by

Fermi functions (see Eq.(2.5)) with well-defined electrochemical potentials. But in an out-of-equilibrium ***channel***, the electron distribution among the available states need not follow a Fermi function.

In general one has to solve a full-fledged transport equation like the semiclassical Boltzmann equation to be introduced in the next Lecture which allows us to calculate the full occupation factors *f(z;E)*. More generally for quantum transport one can use the non-equilibrium Green's function (NEGF) formalism to be introduced in Part three to solve for the quantum version of *f(z;E)*. Can we really represent these distribution functions using electrochemical potentials $\mu^+(z)$ and $\mu^-(z)$?

Interestingly for a perfectly ballistic channel with good contacts, such a representation in terms of $\mu^+(z)$ and $\mu^-(z)$ is exact and not just an approximation. All drainbound electrons (traveling along $+z$, see Fig.6.4) are distributed according to the source contact with $\mu^+ = \mu_1$

$$f^+(z;E) = f_1(E) \equiv \frac{1}{1+\exp\left(\frac{E-\mu_1}{kT}\right)} \tag{6.10a}$$

while all sourcebound electrons (traveling along $-z$) are distributed according to the drain contact with $\mu^- = \mu_2$:

$$f^-(z;E) = f_2(E) \equiv \frac{1}{1+\exp\left(\frac{E-\mu_2}{kT}\right)} \tag{6.10b}$$

This is justified by noting that the drainbound channels from the source are filled only with electrons originating in the source and so these channels remain in equilibrium with the source with a distribution function $f_1(E)$. Similarly the sourcebound channels from the drain are in equilibrium with the drain with a distribution function $f_2(E)$.

Suppose at some energy $f_1(E) = 1$, and $f_2(E) = 0$, so that there are lots of electrons waiting to get out of the source, but none in the drain. We would then expect the drainbound lanes of the electronic highway to be completely full ("bumper-to-bumper traffic"), while the sourcebound lanes would all be empty as shown below in Fig.6.4.

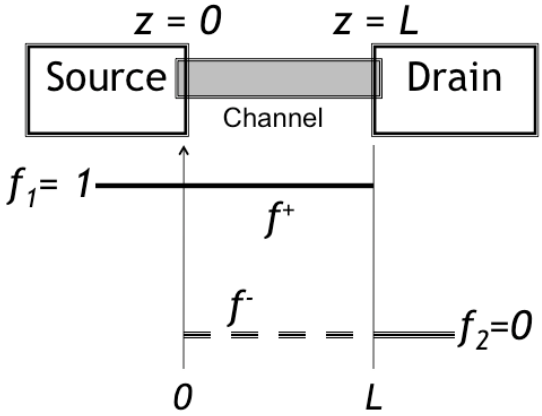


Fig.6.4. Spatial profile of the occupation factors f^+, f^- across a ballistic channel.

Of course this assumes that electrons do not turn around either along the way or at the ends. This means ballistic channels with good contacts where there are so many channels available that electrons can exit smoothly with a very low probability of turning around. If we either have bad contacts or diffusive channels, the solution in Eq.(6.10a,b) wouldn't work. In Lecture 14 on spin valves we will see some consequences of bad contacts, but for the moment let us talk about diffusive channels with good contacts.

Eqs.(6.10a,b) suggest a plausible guess for what we might expect the distributions to look like in a diffusive channel. We assume the same Fermi-like function but with spatially varying electrochemical potentials reflecting the fact that electrons from the drainbound channels continually transfer over to the sourcebound lanes:

$$f^+(z;E) = \frac{1}{1+\exp\left(\dfrac{E-\mu^+(z)}{kT}\right)} \qquad (6.11a)$$

$$f^{-}(z;E) = \frac{1}{1+\exp\left(\frac{E-\mu^{-}(z)}{kT}\right)} \qquad (6.11b)$$

Note that the potentials are in general energy-dependent and could be written as $\mu^{\pm}(z;E)$. In an elastic resistor, every energy is independent and in general each one could exhibit a different spatial variation in the potential if the mean free path is energy-dependent. But for simplicity, we will ignore this point assuming some average energy-independent mean free path.

But if we accept these forms for the occupation factors, then it is straightforward to translate a plot of occupation factors f (like the one in Fig.6.4) into a corresponding plot for the electrochemical potentials by noting that at low bias, the deviation of f from a reference value f_0 is proportional to the deviation of the corresponding μ from the corresponding reference value of μ_0 :

$$f(E)-f_0(E) \approx \left(-\frac{\partial f_0}{\partial E}\right)(\mu-\mu_0) \text{ (same as Eq. (2.8))}$$

This relation, for example, can be used to translate Fig.6.4 into Fig.6.5.

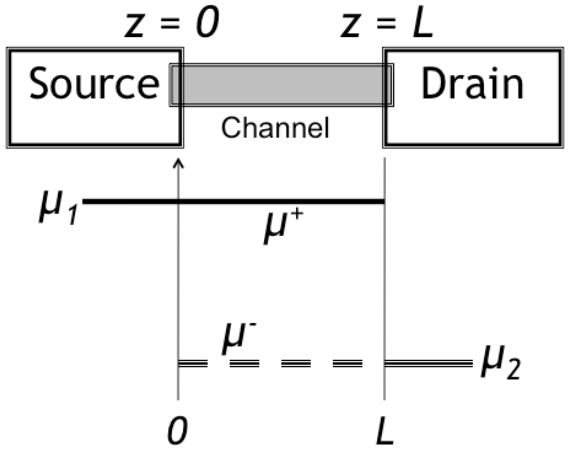


Fig.6.5.
Spatial profile of the electrochemical potentials μ^{+}, μ^{-} across a ballistic channel, obtained from Fig.6.4 by translating f's into μ's using Eq.(2.8).

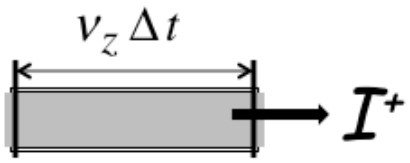
6.2. Current in Terms of Non-Equilibrium Potentials

Usually we talk about the net current I which can be expressed as the difference between the drainbound flux I^+ and the sourcebound flux I^-:

$$I(z) = I^+(z) - I^-(z) \tag{6.12}$$

The current I^+ equals the amount of charge exiting from the right per unit time. In a time Δt, all the charge in a length $v_z \Delta t$ exits, so that

$$I^+ = q * Electrons\ per\ unit\ length * v_z$$



The number of electrons per unit length is equal to half the density of states (since only half the states carry current to the right) per unit length, $D(E) / 2L$, times the fraction f^+ of occupied states, so that

$$I^+(z;E) = q \underbrace{\frac{D(E)}{2L}\bar{u}(E)}_{M(E)/h} f^+(z;E)$$

Here $\bar{u}$ is the average v_z as defined in Eq.(4.7) and making use of the definition of the number of channels M from Eq.(4.13) we have

$$I^+(z;E) = \frac{qM(E)}{h} f^+(z;E) \tag{6.13a}$$

Similarly
$$I^-(z;E) = \frac{qM(E)}{h} f^-(z;E) \tag{6.13b}$$

This allows us to write the current from Eq.(6.12)

$$I(z) = \int_{-\infty}^{+\infty} dE \left(I^{+}(z;E) - I^{-}(z;E) \right)$$

$$= \frac{q}{h} \int_{-\infty}^{+\infty} dE \left(f^{+}(z;E) - f^{-}(z;E) \right) M(E) \tag{6.14}$$

Once again, to get from distribution functions $f^{\pm}$ to electrochemical potentials $\mu^{\pm}$, we make use of the low bias result (Eq.(2.8)) to write

$$f^{+}(z;E) - f^{-}(z;E) = \left(-\frac{\partial f_0}{\partial E} \right) \left(\mu^{+}(z) - \mu^{-}(z) \right) \tag{6.15}$$

so that from Eq.(6.14) we obtain Eq.(6.6a)

$$I(z) = \frac{q}{h} \left(\mu^{+}(z) - \mu^{-}(z) \right) \underbrace{\int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E} \right) M(E)}_{\equiv M} \tag{6.16}$$

provided we identify M with the thermally averaged $M(E)$ as indicated in Eq.(6.16).

Lecture 7

What about Drift?

7.1. Boltzmann Transport Equation, BTE
7.2. Diffusion Equation from BTE
7.3. Equilibrium Fields Do Matter
7.4. The Two Potentials

Interestingly in our Lectures so far we have hardly ever mentioned the electric field, in contrast to most treatments of electronic transport which start by considering the electric field induced force as the driving term. It may seem paradoxical that we could obtain the conductivity without ever mentioning the electric field!

Electric fields are typically visualized as the gradient of an electrostatic potential U/q. By contrast, we have been using the electrochemical potential μ as the basis for our discussions. It is important to recognize the difference between the two "potentials":

$$\underbrace{\mu}_{Electrochemical} = \underbrace{(\mu - U)}_{Chemical} + \underbrace{U}_{Electrostatic} \qquad (7.1)$$

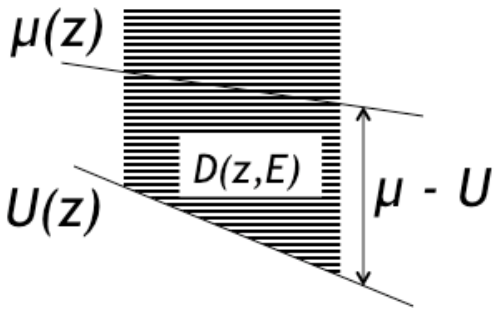


Fig.7.1.
The two potentials: Electrostatic U/q and electrochemical μ/q. $D(z;E)$ denotes the spatially varying density of states.

μ is a measure of the energy upto which the states are filled, while U determines the energy shift of the available states, so that μ - U is a measure of the degree to which the states are filled and hence the number of electrons.

In the last chapter we obtained the equation

$$I/A = -\sigma \frac{d(\mu/q)}{dz} \tag{7.2}$$

But what we really showed was that

$$I/A = -\sigma \frac{d(\mu - U)/q}{dz} \tag{7.3}$$

assuming zero electric field, $dU/dz = 0$. So how do we know what the correct equation is, when we include U?

It would seem that we needed to solve a whole new problem including the effect of the field (= $d(U/q)/dz$) on electrons. However, this is unnecessary because the basic principles of equilibrium statistical mechanics require the current to be zero for a constant μ, just as there can be no heat current if the temperature is constant. Hence the current expression must have the form given in Eq.(7.2) which can be written as the sum of a drift term and a diffusion term

$$I/A = \underbrace{-\sigma \frac{d(\mu - U)/q}{dz}}_{Diffusion} \quad \underbrace{-\sigma \frac{dU/q}{dz}}_{Drift} \tag{7.4}$$

both of which must be described by the same coefficient σ, a requirement that leads to the Einstein relation between drift and diffusion. And that is why we can find σ considering only the diffusion of electrons with $U = 0$, obtain Eq.(7.3) and just replace it with Eq.(7.2) which correctly accounts for "everything." There is really no need work out the drift problem separately. What we called the diffusion equation is

really the ***drift-diffusion equation*** even though we did not consider drift explicitly.

Couldn't we instead have neglected diffusion completely and just gone with the drift term? That way we could stick to the view that current is driven by electric fields and not have to bother with electrochemical potentials. The problem is that if we take this view then one has to invoke mysterious quantum mechanical forces to explain why all electrons are not affected by the field. In our discussion the energy window for transport (F_T, see Fig.2.3) arises naturally from the difference in the "agenda" of the two contacts (see Eqs.(2.7), (2.8))

$$f_1(E) - f_2(E) \;=\; \left(-\frac{\partial f_0}{\partial E}\right)\left(\mu_1 - \mu_2\right)$$

as discussed in Lectures 2, 3. The point is that regardless of which potential we choose to work with, it finally affects transport through the occupation factor, *f*.

In this Lecture we will justify our neglect of drift more explicitly by introducing the Boltzmann Transport Equation (BTE) which is the standard starting point for all discussions of the transport of particles. We too could have used it as the starting point for but we did not do so because it is harder to digest with its multiple independent variables, compared to the ordinary differential equation in Lecture 6, which follows from relatively elementary arguments.

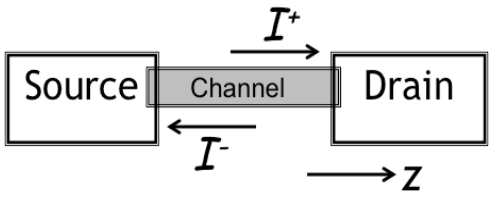
Even in this Lecture we will not really do justice to the BTE. We will introduce it briefly and use it to show that for low bias, the current indeed depends only on $d\mu/dz$ and not on dU/dz, thus putting our discussion of steady-state, low bias transport without electric fields on a firmer footing and identifying possible issues with it.

Note the two qualifying phrases, namely "steady-state" and "low bias." We will show later in this lecture that for time varying transport, the neglect of electric fields can lead to errors, but we will not discuss it further in these lectures. However, even under steady-state conditions,

electric fields can play an important role in determining the full current-voltage characteristics, once we go beyond low bias, as we will discuss in the next Lecture.

7.1 Boltzmann Transport Equation, BTE

In Lecture 6 we introduced electron distribution functions $f^{\pm}$ and electrochemical potentials describing the drainbound and sourcebound currents $I^{\pm}$. Both the drainbound and sourcebound current, however, is composed of electrons traveling at different angles having different z-momemtum p_z, even though they all have the same energy (we are still talking about an elastic resistor) and hence the same total momentum. To include the effect of the electric field we need "momentum-resolved" distribution functions $f^{\pm}(z, p_z, t)$.



The BTE describes the evolution of such "momentum-resolved" distribution functions *f(z,p_z,t)* that tell us the occupation of states with a given momentum p_z and velocity v_z at a location z at time t :

$$\frac{\partial f}{\partial t} + v_z \frac{\partial f}{\partial z} + F_z \frac{\partial f}{\partial p_z} = S_{op} f \tag{7.5}$$

where F_z is the force on the electrons, and $S_{op}f$ symbolically represents the complex scattering processes that continually redistribute electrons among the available velocity states.

The BTE with the right hand side set to zero (that is without scattering processes)

$$\frac{\partial f}{\partial t} + v_z \frac{\partial f}{\partial z} + F_z \frac{\partial f}{\partial p_z} = 0 \qquad (7.6)$$

is completely equivalent to describing a set of particles each with position *z(t)* and momenta $p_z(t)$ that evolve according to the semiclassical laws of motion:

$$v_z \equiv \frac{dz}{dt} = \frac{\partial E}{\partial p_z} \qquad (7.7a)$$

$$F_z \equiv \frac{dp_z}{dt} = -\frac{\partial E}{\partial z} \qquad (7.7b)$$

where $E(z,p_z,t)$ is the total energy.

Eqs.(7.7a,b) describe semiclassical dynamics in single particle terms where the position *z(t)* and momenta $p_z(t)$ for each of the electrons is a dependent variable evolving in time. By contrast, the BTE provides a collective description with all three independent variables z, p_z,t on an equal footing.

To get from Eqs.(7.7) to (7.6) we start by noting that in the absence of scattering, we can write

$$f(z, p_z, t) = f(z - v_z \Delta t, p_z - F_z \Delta t, t - \Delta t)$$

reflecting the fact that any electron with a momentum

$$p_z \text{ at } z \text{ at time } t ,$$

must have had a momentum of

$$p_z - F_z \Delta t \text{ at } z - v_z \Delta t \text{ a little earlier at time } t - \Delta t .$$

Next we expand the right hand side to the first term in a Taylor series to write

$$f(z,p_z,t) \quad = \quad f(z,p_z,t) - \frac{\partial f}{\partial z} v_z \Delta t - \frac{\partial f}{\partial p_z} F_z \Delta t - \frac{\partial f}{\partial t} \Delta t$$

Eq.(7.6) follows readily on canceling out the common terms.

The left hand side of the BTE thus represents an alternative way of expressing the laws of motion. What makes it different from mere mechanics, however, is the stochastic scattering term on the right which makes the distribution function f approach the equilibrium Fermi function when external driving terms are absent. This last point of course is not meant to be obvious. It requires an extended discussion of the scattering operator S_{op} that we talk a little more about in Lecture 16 when we discuss the second law.

For our purpose it suffices to note that a common approximation for the scattering term is the relaxation time approximation (RTA)

$$S_{op} f \quad \cong \quad - \frac{f - f_0}{\tau} \tag{7.8}$$

which assumes that the effect of the scattering processes is proportional to the degree to which a given distribution f differs from the equilibrium distribution f_0.

One comment about why we call this approach ***semiclassical.*** The BTE is ***classical*** in the sense that it is based on a particle view of electrons. But it is not ***fully*** classical, since it typically includes quantum input both in the scattering operator S_{op} and in the form of the energy-momentum relation. For example, graphene is often described by a linear energy-momentum relation

$$\vec{E} = v_0 \vec{p}$$

a result that is usually justified in terms of the bandstructure of the graphene lattice requiring quantum mechanics that Boltzmann did not live to see. But once we accept that, many transport properties of

graphene can be understood in classical particulate terms using the BTE that Boltzmann taught us to use.

7.2 Diffusion equation from BTE

We start by combining the RTA (Eq.(7.8)) with the full BTE (Eq.(7.5)) to obtain for steady-state ($\partial / \partial t = 0$),

$$v_z \frac{\partial f}{\partial z} + F_z \frac{\partial f}{\partial p_z} = -\frac{f - f_0}{\tau} \tag{7.9}$$

In the presence of an electric field we can write the total energy as

$$E(z, p_z) = \varepsilon(p_z) + U(z) \tag{7.10}$$

where $\varepsilon(p_z)$ denotes the energy-momentum relation with $U=0$ and this gets shifted locally by $U(z)$ as sketched in Fig.7.2.

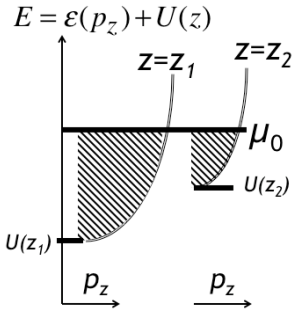


Fig.7.2. The energy momentum relation with $U=0$ is shifted locally by $U(z)$. At equilibrium the electrochemical potential μ_0 is spatially constant.

The first point to note is that the equilibrium distribution with a constant electrochemical potential μ_0

$$f_0(z, p_z) = \frac{1}{\exp\left(\frac{E(z, p_z) - \mu_0}{kT}\right) + 1} \tag{7.11}$$

satisfies the BTE in Eq.(7.9). The right hand side of Eq.(7.9) is obviously zero, but it takes a little differential calculus to see that the left hand side is zero too.

Defining

$$X_0 \equiv E(z,p_z)-\mu_0 = \varepsilon(p_z)+U(z)-\mu_0 \tag{7.12}$$

we have

$$v_z\frac{\partial f_0}{\partial z}+F_z\frac{\partial f_0}{\partial p_z} = \left(\frac{\partial f_0}{\partial X_0}\right)\left(v_z\frac{\partial X_0}{\partial z}+F_z\frac{\partial X_0}{\partial p_z}\right)$$

$$= \left(\frac{\partial f_0}{\partial X_0}\right)\left(v_z\frac{\partial E}{\partial z}+F_z\frac{\partial E}{\partial p_z}\right) = 0$$

making use of Eqa.(7.7a,b).

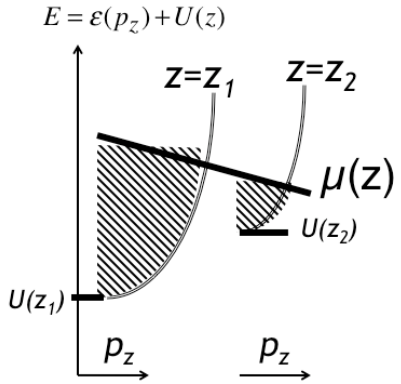


Fig.7.3. Same as Fig.7.2, but the electrochemical potential *μ(z)* varies spatially reflecting a non-equilibrium state.

Out of equilibrium, we assume the distribution function $f(z,p_z)$ to have the same form as Eq.(7.11) but with a spatially varying electrochemical potential $\mu(z)$:

$$f(z,p_z) = \frac{1}{\exp\left(\dfrac{E(z,p_z)-\mu(z)}{kT}\right)+1} \tag{7.13}$$

Using Eq.(7.13), the left hand side of BTE (see Eq.(7.9)) reduces to

$$\left(\frac{\partial f}{\partial X}\right)\left(v_z\frac{\partial X}{\partial z}+F_z\frac{\partial X}{\partial p_z}\right) = \left(\frac{\partial f}{\partial X}\right)\left(-v_z\frac{d\mu}{dz}\right)$$

$$X \equiv E(z,p_z)-\mu(z)$$

where $$= X_0(z,p_z)+\mu_0-\mu(z) \qquad (7.14)$$

We now assume small deviations in $\mu(z)$ from the equilibrium value so that we can write the left hand side as

$$\left(\frac{\partial f}{\partial X}\right)_{X=X_0}\left(-v_z\frac{d\mu}{dz}\right)$$

and use our standard Taylor series expansion (see Eq.(2.8)) to write the right hand side of BTE as

$$-\frac{f-f_0}{\tau} \cong \left(\frac{\partial f}{\partial X}\right)_{X=X_0}\frac{\mu(z)-\mu_0}{\tau}$$

Combining the two sides

$$v_z\frac{d\mu}{dz} = -\frac{\mu(z)-\mu_0}{\tau} \qquad (7.15)$$

We now introduce two separate electrochemical potentials μ^+ and μ^- for the right-moving ($v_z>0$) and left-moving ($v_z<0$) electrons to write

$$\frac{d\mu^+}{dz} = -\frac{\mu^+-\mu_0}{v_z\tau} \quad , \quad \frac{d\mu^-}{dz} = \frac{\mu^--\mu_0}{v_z\tau}$$

Assuming $\mu_0=(\mu^++\mu^-)/2$, we obtain

$$\frac{d\mu^+}{dz} = -\frac{\mu^+-\mu^-}{\lambda} = \frac{d\mu^-}{dz} \qquad (7.16)$$

with $\lambda=2v_z\tau$. Combining with Eq.(6.6a) for the current, we obtain the result (Eq.(6.8)) stated without proof in the last Lecture. Note that we

have included electric fields explicitly and shown that their effect cancels out.

7.3. Equilibrium Fields Do Matter

However, we believe there is an important subtlety worth pointing out. Although the externally applied electric field does not affect the low bias conductance, any inbuilt fields that exist within the conductor under equilibrium conditions can affect its low bias conductance. Let me explain.

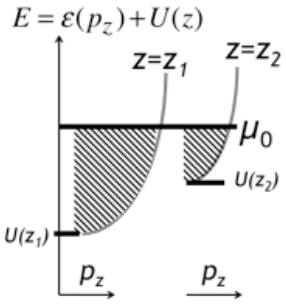
Note that in our treatment above we assumed that under non-equilibrium conditions, the electrochemical potential is a function of z (Eq.(7.13)) and the resulting linearized equation (Eq.(7.15)) does not involve the field $F_z = dU/dz$. However, the field term would not have dropped out so nicely if we were to assume that the electrochemical potential is not just a function of z, but of both z and p_z. Instead of Eq.(7.15) we would then obtain

$$v_z \frac{\partial \mu}{\partial z} + F_z \frac{\partial \mu}{\partial p_z} = - \frac{\mu(z, p_z) - \mu_0}{\tau} \tag{7.17}$$

However, the additional term involving the field F_z does not play a role in determining linear conductivity because it is $\sim V^2$, V being the applied voltage. At equilibrium with $V=0$, $\mu = \mu_0$, so that both derivatives appearing on the left are zero. Under bias, in principle, both could be non-zero and to first order $\sim V$. But the point is that while v_z is a constant, the applied field F_z is also $\sim V$. So while the first term on the left is $\sim V$, the second term is $\sim V^2$.

But this argument would not hold if F_z were not the applied field, but internal inbuilt fields independent of V that are present even at equilibrium. Equilibrium requires a constant μ and NOT a constant U.

The equilibrium condition depicted in Fig.7.2 (also shown here for ease of reference) is quite common in real conductors, with varying *U(z)* corresponding to non-zero fields F_z. Indeed this picture could also represent an interface between dissimilar materials (called "heterostructures") where the discontinuity in band edges is often modeled with effective fields.



The point is that such equilibrium fields can and do affect the low bias conductance. For an ideal homogeneous conductor we do not have such fields. But even then we need to make two contacts in order to measure the resistance. Each such contact represents a heterostructure qualitatively similar to that shown in Fig.7.2 with inbuilt effective (if not real) fields. I think these fields give rise to the interface resistance distinguishing the new Ohm's law from the standard one, but I have not checked.

7.4. The Two Potentials

In these Lectures we will generally focus on steady-state transport involving the injection of electrons from a source and their collection by a drain (Fig.7.4). We have seen that the low bias conductance can be understood in terms of the electrochemical potential μ, without worrying about the electrostatic potential *U*.

However, we would like to briefly consider ac transport through a nanowire far from any contacts where we have a local voltage *V(z,t)* and current *I(z,t)* (Fig.7.5), because this provides a contrasting example where it is important to pay attention to the difference between the two potentials even for low bias, in order to obtain the correct inductance and capacitance.

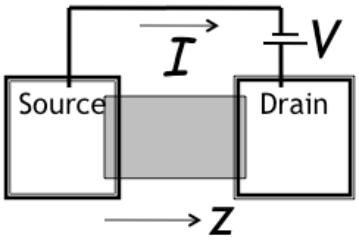


Fig.7.4. So far we have talked of steady-state transport involving the injection of electrons by a source and their collection by a drain contact.

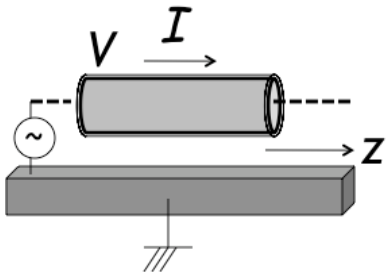


Fig.7.5.
Ac or time varying transport along a nanowire can be described in terms of a voltage *V(z,t)* and a current *I(z,t)*.

For this problem too we start from the BTE with the RTA approximation as in the last section, but we do not set $\partial / \partial t = 0$,

$$\frac{\partial f}{\partial t} + v_z \frac{\partial f}{\partial z} + F_z \frac{\partial f}{\partial p_z} = -\frac{f - f_0}{\tau}$$

and linearize it assuming a distribution of the form (compare Eq.(7.13))

$$f(z, p_z, t) = \frac{1}{\exp\left(\frac{E(z, p_z, t) - \mu(z,t)}{kT}\right) + 1} \tag{7.18}$$

Compared to the steady-state problem (Eq.(7.15)) we now have two extra terms involving the time derivatives of E and μ:

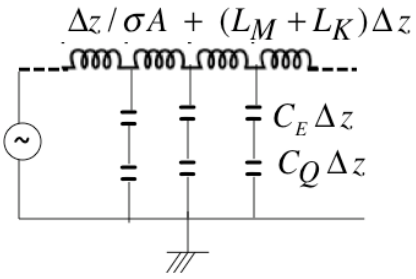
$$\frac{\partial \mu}{\partial t} + v_z \frac{\partial \mu}{\partial z} - \frac{\partial E}{\partial t} = -\frac{\mu(z,t) - \mu_0}{\tau} \tag{7.19}\backslash$$

As we did in the last Section with Eq.(7.15), we can separate Eq.(7.19) into two equations for μ^+ and μ^-, whose sum and difference are identified with voltage and current to obtain a set of equations

$$\frac{\partial(\mu/q)}{\partial z} = -(L_K + L_M)\frac{\partial I}{\partial t} - \frac{I}{\sigma A} \tag{7.20a}$$

$$\frac{\partial(\mu/q)}{\partial t} = -\left(\frac{1}{C_Q} + \frac{1}{C_E}\right)\frac{\partial I}{\partial z} \tag{7.20b}$$

that look just like the transmission line equations with a distributed series inductance and resistance and a shunt capacitance.



The algebra getting from Eq.(7.19) to Eqs.(7.20a,b) is a little long-winded and since time-varying transport is only incidental to our main message we have relegated the details to Appendix D. Those who are really interested can look at the original paper on which this discussion is based (Salahuddin et al., 2005).

But note the two inductors and the two capacitors in series. The ***kinetic inductance*** L_K and the ***quantum capacitance*** C_Q per unit length, arise from transport-related effects

$$L_K = \frac{h}{q^2}\frac{1}{\langle 2Mv_z \rangle} \tag{7.21a}$$

$$C_Q = \frac{q^2}{h}\left\langle \frac{2M}{v_z} \right\rangle \tag{7.21b}$$

while the L_M and the C_E are just the normal ***magnetic inductance*** and the ***electrostatic capacitance*** from the equations of magnetostatics and electrostatics.

The point I wish to make is that the fields enter the expression for the energy $E(z,p_z,t)$ and if we ignore the fields we would miss the $\partial E / \partial t$ term in Eq.(7.19) to obtain

$$\frac{\partial \mu}{\partial t} + v_z \frac{\partial \mu}{\partial z} = - \frac{\mu(z,t) - \mu_0}{\tau}$$

and after working through the algebra obtain instead of Eqs.(7.20a,b)

$$\frac{\partial (\mu / q)}{\partial z} = - L_K \frac{\partial I}{\partial t} - \frac{I}{\sigma A} \tag{7.22a}$$

$$\frac{\partial (\mu / q)}{\partial t} = - \left(\frac{1}{C_Q} \right) \frac{\partial I}{\partial z} \tag{7.22b}$$

Do these equations approximately capture the physics? Not unless we are considering wires with very small cross-sections so that M is a small number making $L_K >> L_M$ and $C_Q << C_E$.

We could recover the correct answer from Eqs.(7.22a,b) by replacing the μ in with μ- U and then using the laws of electromagnetics to replace

$$\frac{\partial U}{\partial t} \quad with \quad \frac{1}{C_E} \frac{\partial I}{\partial z} \qquad \textbf{\textit{and}} \qquad \frac{\partial U}{\partial z} \quad with \quad L_M \frac{\partial I}{\partial t}$$

But these replacements may not be obvious and it is more straightforward to go from Eq.(7.19) to (7.20) as spelt out in Appndix D.

Note that if we specialize to steady-state ($\partial / \partial t = 0$), both Eqs.(7.20) and (7.22) give us back our old diffusion equation (Eq.(6.1)). As we argued earlier, for low bias steady-state transport, the applied electric field can be treated as incidental.

Lecture 8

Electrostatics is Important

In the last Lecture we tried to justify our "field-less" approach to conductivity which comes as a surprise to many since it is commonly believed that currents are driven by electric fields. However, we hasten to add that the field can and does play an important role once we go beyond low bias and our purpose in this lecture is to discuss the role of the electrostatic potential and the corresponding electric field on the current-voltage characteristics beyond low bias.

To illustrate these issues, I will use the nanotransistor, an important device that is at the heart of microelectronics. As we noted at the outset the nanotransistor is essentially a voltage-controlled resistor whose length has shrunk over the years and is now down to a few hundred atoms. But as any expert will tell you, it is not just the low bias resistance, but the entire shape of the current-voltage characteristics of a nanotransistor that determines its utility. And this shape is controlled largely by its electrostatics, making it a perfect example for our purpose.

I should add, however, that this Lecture does not do justice to the nanotransistor as a device. This will be discussed in a separate volume in this series written by Lundstrom, whose model is widely used in the field and forms the basis of our discussion here. We will simply use the nanotransistor to illustrate the role of electrostatics in determining current flow.

We have seen that the elastic transport model characterized by the current formula

$$I = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, G(E)\big(\, f_1(E) - f_2(E)\,\big) \qquad \text{(see Eq.(3.3))}$$

In this Lecture I will use the nanotransistor to illustrate some of the issues that need to be considered at high bias, some of which can be modeled with a simple extension of Eq.(3.3)

$$I = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, G(E-U)\big(\, f_1(E) - f_2(E)\,\big) \qquad (8.1)$$

to include an appropriate choice of the potential U in the channel which is treated as a single point. We call this the *point channel* model to distinguish it from the standard and more elaborate extended channel model which we will introduce at the end of the Lecture.

8.1 The nanotransistor

The nanotransistor is a three-terminal device (Fig.8.1), though ideally no current should flow at the gate terminal whose role is just to control the current. In other words, the current-drain voltage, I- V_D, characteristics are controlled by the gate voltage, V_G (see Fig.8.2). The low bias current and conductance can be understood based on the principles we have already discussed. But currents at high V_D involve important new principles.

The basic principle underlying an FET is straightforward (see Fig.8.3). A positive gate voltage V_G changes the potential in the channel, lowering all the states down in energy, which can be included by replacing Eq.(8.1) with Eq.(8.2) and setting $U = qV_G$.

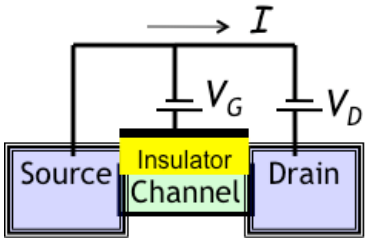


Fig.8.1.
Sketch of a field effect transistor (FET): Channel length, L; Transverse width, W (Perpendicular to page).

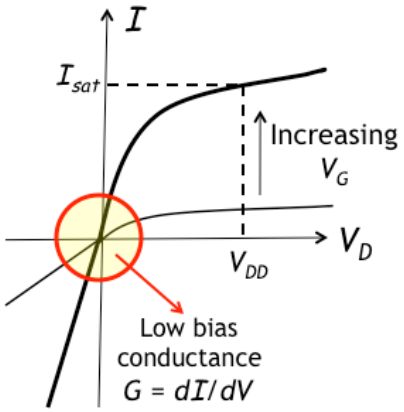


Fig.8.2.
Typical current-voltage, I-V_D characteristic and its variation with V_G for an FET.

For an n-type conductor this increases the number of available states in the energy window of interest around μ_1 and μ_2 as shown. Of course for a p-type conductor (see Fig.7.2) the reverse would be true leading to a complementary FET (see Fig.0.2) whose conductance variation is just the opposite of what we are discussing. But we will focus here on n-type FET's.

We will not discuss the low bias conductance since these involve no new principles. Instead we will focus on the current at high bias, specifically on why the current-voltage, I- V_D characteristic is (1) non-linear, and (2) "rectifying," that is different for positive and negative V_D.

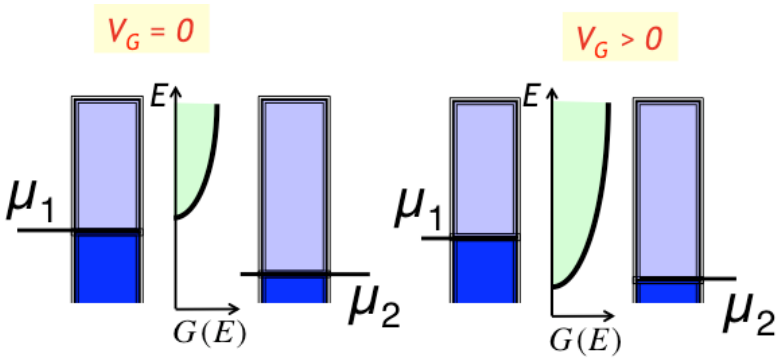


Fig.8.3.
A positive gate voltage V_G increases the current in an FET by moving the states down in energy.

8.2 Why the current saturates

Fig.8.2 shows that as the voltage V_D is increased the current does not continue to increase linearly. Instead it levels off tending to saturate. Why? The reason seems easy enough. Once the electrochemical potential in the drain has been lowered below the band edge the current does not increase any more (Fig.8.4).

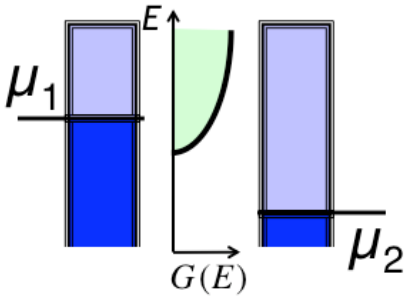


Fig.8.4.
The current saturates once μ_2 drops below the band-edge.

The saturation current can be written from Eq.(8.1)

$$I_{sat} = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, G(E-U)\, f_1(E) \qquad (8.2)$$

by dropping the second term $f_2(E)$ assuming μ_2 is low enough that $f_2(E)$ is zero for all energies where the conductance function is non-zero. In the simplest approximation

$$U^{(1)} = -qV_G$$

The superscript *1* is included to denote that this expression is a little too simple, representing a first step that we will try to improve.

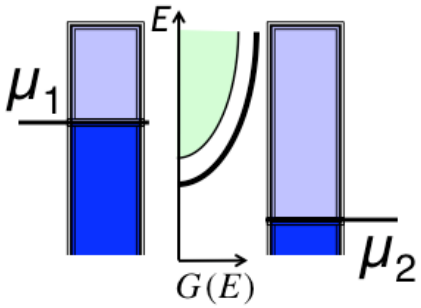


Fig.8.5.
The current does not saturate completely because the states in the channel are also lowered by the drain voltage.

If this were the full story the current would have saturated completely as soon as μ_2 dropped a few kT below the band edge. In practice the current continues to increase with drain voltage as sketched in Fig.8.6. The reason is that when we increase the drain voltage we do not just lower μ_2, but also lower the energy levels inside the channel (Fig.8.5) similar to the way a gate voltage would. The result is that the current keeps increasing as the conductance function $G(E)$ slides down in energy by a fraction α (< 1) of the drain voltage V_D, which we could include in our model by choosing

$$U^{(2)} = \alpha(-qV_D) + \beta(-qV_G) \equiv U_L \qquad (8.3)$$

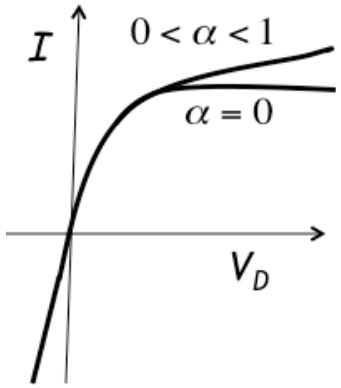


Fig.8.6.
Current in an FET would saturate perfectly if the channel potential were unaffected by the drain voltage.

Indeed the challenge of designing a good transistor is to make α as small as possible so that the channel potential is hardly affected by the drain voltage. If α were zero the current would saturate perfectly as shown in Fig.8.6 and that is really the ideal: a device whose current is determined entirely by V_G and not at all by V_D or in technical terms, a high transconductance but low output conductance. For reasons we will not go into, this makes designing circuits much easier.

To ensure that V_G has far greater control over the channel than V_D it is necessary to make the insulator thickness a small fraction of the channel length. This means that for a channel length of a few hundred atoms we need an insulator that is only a few atoms thick in order to ensure a small α. This thickness has to be precisely controlled since thinner insulators would leak unacceptably. We mentioned earlier that today's laptops have a billion transistors. What is even more amazing is that each has an insulator whose thickness is precisely controlled down to a few atoms!

8.3 Role of charging

There is a second effect that leads to an increase in the saturation current over what we get using Eq.(8.3) in (8.1). Under bias, the occupation of the channel states is less than what it is at equilibrium. This is because at equilibrium both contacts are trying to fill up the channel states, while under bias only the source is trying to fill up the states while the drain is

trying to empty it. Since there are fewer electrons in the channel, it tends to become positively charged and this will lower the states in the channel as shown in Fig.8.5, even for perfect electrostatics (α = 0) resulting in an increase in the current.

This effect can be captured within the point channel model (Eq.(8.1)) by writing the channel potential as

$$U = U_L + U_0(N - N_0) \tag{8.4}$$

where U_L is given by our previous expression in Eq.(8.3). The extra term represents the change in the channel potential due to the change in the number of electrons in the channel, N under non-equilibrium conditions relative to the equilibrium number N_0, U_0 being the change in the channel potential energy per electron.

To use Eq.(8.4), we need expressions for N_0, N. N_0 is the equilibrium number of channel electrons, which can be calculated simply by filling up the density of states, $D(E)$ according to the equilibrium Fermi function $f_0(E)$.

$$N_0 = \int_{-\infty}^{+\infty} dE\ D(E - U)\ f_0(E) \tag{8.5a}$$

while the number of electrons, N in the channel under non-equilibrium conditions is given by

$$N = \int_{-\infty}^{+\infty} dE\ D(E - U)\ \frac{f_1(E) + f_2(E)}{2} \tag{8.5b}$$

assuming that the channel is "equally" connected to both contacts. Note that the calculation is now a little more intricate than what it would be if U_0 were zero. We now have to obtain a solution for U and N that satisfy both Eqs.(8.4) and (8.5) simultaneously through an iterative procedure as shown schematically in Fig.8.7.

Once a self-consistent U has been obtained, the current is calculated from Eq.(8.1), or an equivalent version that is sometimes more convenient numerically and conceptually.

$$I = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, G(E)\,(f_1(E+U)-f_2(E+U)) \qquad (8.6)$$

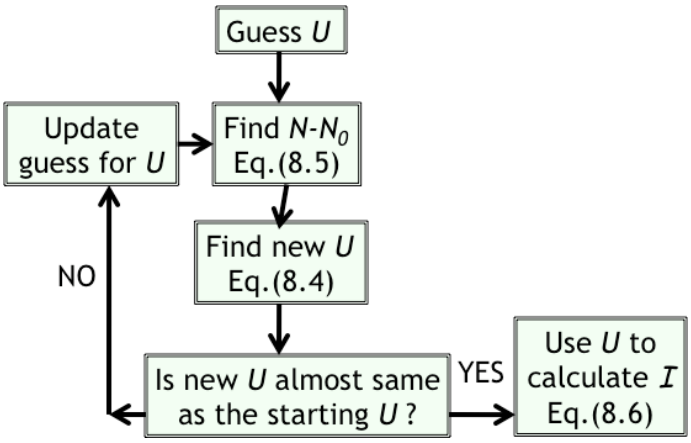


Fig.8.7. Self-consistent procedure for calculating the channel potential U in point channel model.

This simple point channel model often provides good agreement with far more sophisticated models as discussed in Rahman et al. (2003).

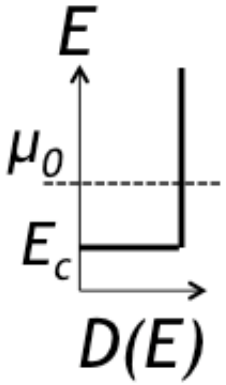


Fig.8.8 shows the current versus voltage characteristic calculated numerically (MATLAB code included in Appendix) assuming a 2-D channel with a parabolic dispersion relation for which the density of states is given by (L: Length, W: Width)

$$D(E) = g\frac{mLW}{2\pi\hbar^2}\,\vartheta(E - E_c)$$

where ϑ represents the unit step function. The numerical results are obtained using *g=2*, *m = 0.2* 9.1e-31 Kg*, $\beta = 1$, $\alpha = 0$ and $U_0 = 0$ *or* ∞ as indicated with *L=1 μm, W=1 μm* assuming ballistic transport, so that

$$G(E) \;=\; \frac{q^2}{h}\,M(E),$$

M(E) being the number of modes given by

$$M(E) = g\frac{2W}{h}\sqrt{2m(E - E_c)}\;\vartheta(E - E_c)$$

The current-voltage characteristics in Fig.8.8 has two distinct parts, the initial linear increase followed by a saturation of the current. Although these results were obtained numerically, both the slope and the saturation current can be calculated analytically, especially if we make the low temperature approximation that the Fermi functions change abruptly from *1* to *0* as the energy *E* crosses the electrochemical potential μ. Indeed we used a *kT* of *5 meV* instead of the usual *25 meV*, so that the numerical results would compare better with simple low temperature estimates.

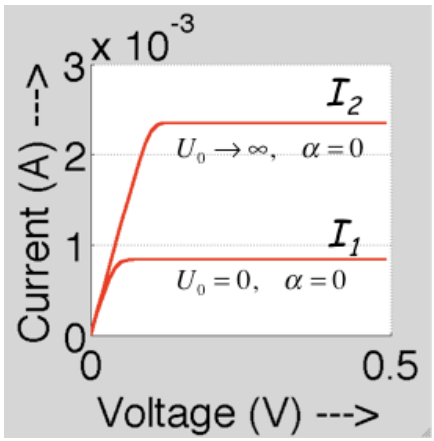
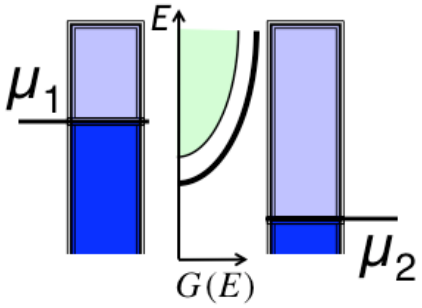


Fig.8.8. Current-voltage characteristics calculated numerically using the self-consistent point channel model shown in Fig.8.7.

There are two key points we wanted to illustrate with this example. Firstly, the initial slope of the current-voltage characteristics is unaffected by the charging energy. This slope defines the low bias conductance that we have been discussing till we came to this Lecture. The fact that it remains unaffected is reassuring and justifies our not bringing up the role of electrostatics earlier.

Secondly, the saturation current is strongly affected by the electrostatics and changes by a factor of ~ 2.8 from a model with zero charging energy to one with a very large charging energy. This is because of the reason mentioned at the beginning of this section. With $U_0 = 0$, the channel states remain fixed and the number of electrons N is equal to $N_0/2$, since $f_1=1$ and $f_2=0$ in the energy range of interest. With very large U_0, to avoid $U_0(N-N_0)$ becoming excessive, N needs to be almost equal to N_0 even though the states are only half-filled. This requires the states to move down as sketched with a corresponding increase in the current.



8.4 "Rectifier" Based on Electrostatics

Let us now look at an example that can be handled using the point channel model just discussed though it does not illustrate any issues affecting the design of nanotransistors. I have chosen this example to illustrate a fundamental point that is often not appreciated, namely that an otherwise symmetric structure could exhibit asymmetric current-voltage characteristics (which we are loosely calling a "rectifier"). In other words, we could have

$$I(+V_D) \neq I(-V_D)$$

for a symmetric structure, simply because of *electrostatic asymmetry.*

Consider a nanotransistor having perfect electrostatics represented by α = 0 (Eq.(8.3)), connected (a) in the standard configuration (Fig.8.9a) and (b) with the gate left floating (Fig.8.9b).

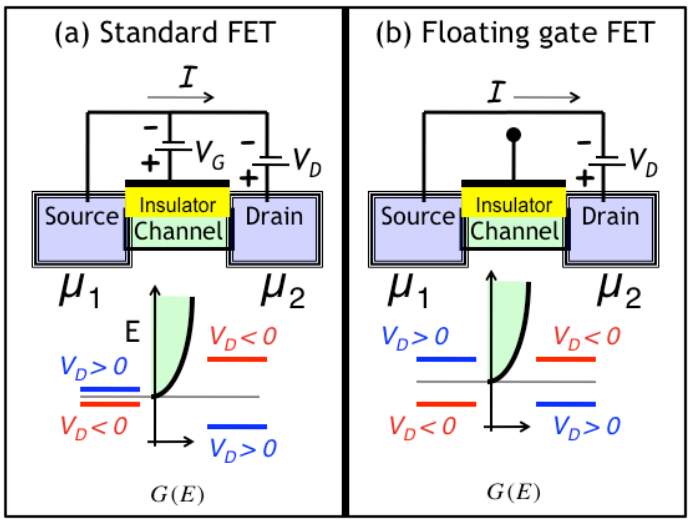


Fig.8.9. (a) Standard FET assuming perfect electrostatics. (b) Floating gate FET

The basic device is assumed physically symmetric, so that one could not tell the difference between the source and drain contacts. This may not be true of real transistors, but that is not important, since we are only trying to make a conceptual point. The configuration in (a) has electrostatic asymmetry, since the gate is held at a fixed potential with respect to the source, but not with respect to the drain. But configuration (b) is symmetric in this respect too, since the gate floats to a potential halfway between the source and the drain.

Fig.8.10 shows the current-voltage characteristics calculated using the model summarized in Fig.8.7 (MATLAB code in Appendix F), for each of the structures shown in (a) and (b). The parameters are the same as those used for the example shown in Fig.8.8, except that the equilibrium electrochemical potential is located exactly at the bottom of the band as shown in Fig.8.9: $\mu_0 = E_c$

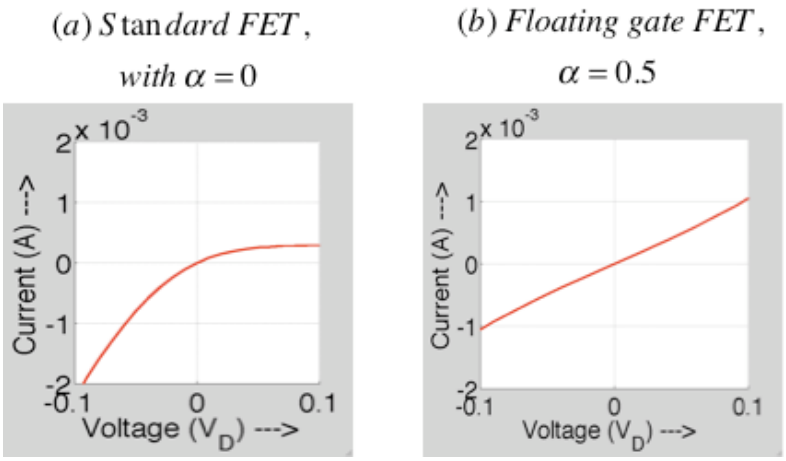


Fig.8.10. Current-voltage characteristics obtained from the point channel model corresponding to the confgurations shown in Fig.8.9.

The standard FET connection corresponds to $\alpha = 0$ assuming perfect electrostatics, while the same physical structure in the floating gate connection corresponds to $\alpha = 0.5$. The former gives a rectifying characteristic, while the latter gives a linear characteristic, often called "Ohmic". The point is that it is not necessary to design an asymmetric channel to get asymmetric *I-V* characteristics. Even the simplest symmetric channel can exhibit non-symmetric $I(V_D)$ characteristic if the electrostatics is asymmetric.

Note also that the linear conductance given by the slope *dI/dV* around V=0 is unaffected by our choice of α and can be predicted without any reference to the electrostatics, even though the overall shape obviously cannot.

8.5 Extended Channel Model

The point channel elastic model that we have described (Eqs.(8.1), (8.2)) integrates our elastic resistor with a simple electrostatic model for the

channel potential U/q, allowing it to capture some of the high bias physics that the pure elastic resistor misses. Let me end this Lecture by noting some of the things it misses.

The point channel model ignores *the electric field in the channel* and assumes that the density of states $D(E)$ stays the same from source to drain. In the real structure, however, the electric field lowers the states at the drain end relative to the source as sketched here. Doesn't this change the current?

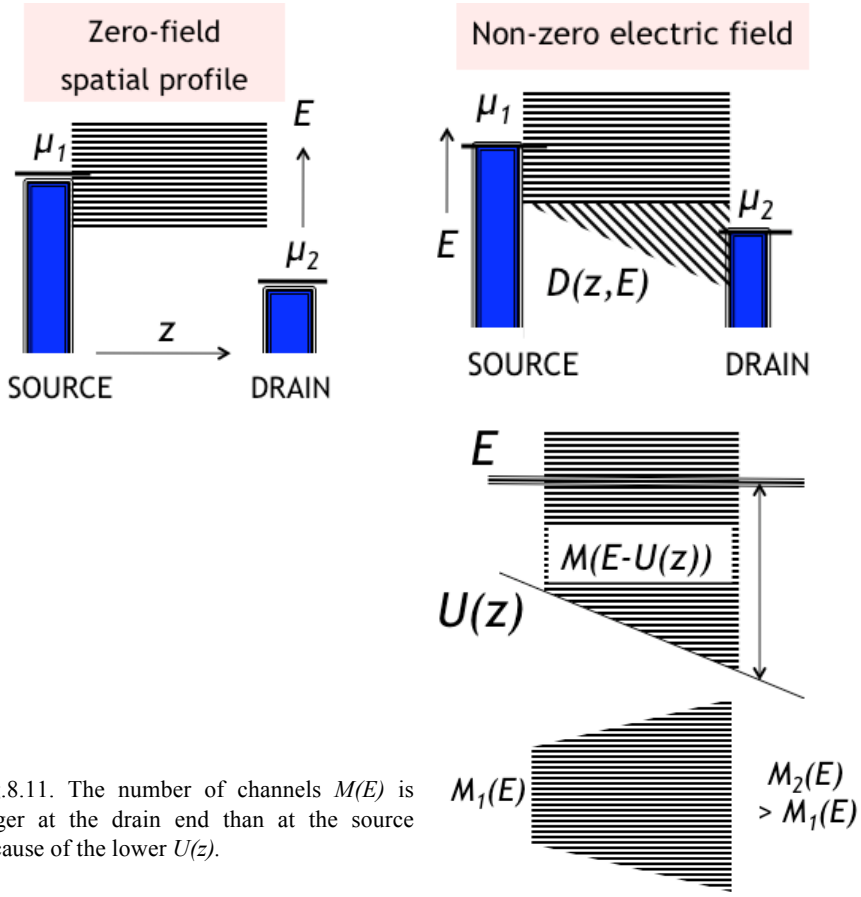


Fig.8.11. The number of channels $M(E)$ is larger at the drain end than at the source because of the lower $U(z)$.

For an elastic resistor one could argue that the additional states with the slanted (rather than horizontal) shading are not really available for conduction since (in an elastic resistor) every energy represents an independent energy channel and can only conduct if it connects all the way from the source to the drain.

But even for an elastic resistor there should be an increase in current because at a given energy E, the number of modes at the drain end is larger than the number of modes at the source end. This is because the number of modes at an energy E depends on how far this energy is from the bottom of the band determined by $U(z)$ which is lower at the drain than at the source.

The structure almost looks as if it were "wider" at the drain than at the source. For a ballistic conductor this makes no difference since the conductance function cannot exceed the maximum set by the "narrowest" point. But for a conductor that is many mean free paths long, the broadening at the drain could increase the conductance relative to that of an un-broadened channel.

In general we could write

$$\frac{q^2}{h}\frac{M_1\lambda}{L+\lambda} \leq G(E) \leq \frac{q^2}{h}M_1 \tag{8.7}$$

This effect is not very important for near ballistic elastic channels, since the minimum and maximum values of the conductance function in Eq.(8.6) are then essentially equal. Indeed this increase in conductance could be ascribed to a field-dependent mean free path which can be ignored in the low bias limit as we have done so far.

How do we include it in a quantitative model? We could simply take our "drift-diffusion" equation from Lecture 6 and modify it to include a spatially varying conductivity:

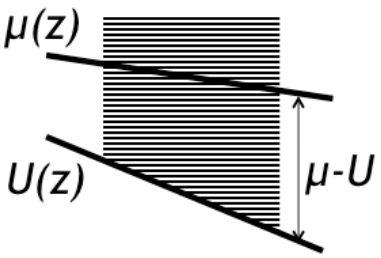
$$\frac{d}{dz}I = 0$$

$$\frac{I}{A} = -\frac{\sigma(z)}{q}\frac{d\mu}{dz} \tag{8.8}$$

What do we use for the conductivity, $\sigma(z)$? Our old expression

$$\sigma = \int_{-\infty}^{+\infty} dE\,\sigma(E)\left(-\frac{\partial f}{\partial E}\right)_{E=\mu_0} \quad \text{(same as Eq.(5.3))}$$

involved an energy average of $\sigma(E)$ over an energy window of a few *kT* around $E = \mu_0$.



The spatially varying *U(z)* shifts the available energy states in energy, so that one now has to look at the energy window around $E = \mu(z) - U(z)$ suggesting that we replace Eq.(5.3) with

$$\sigma = \int_{-\infty}^{+\infty} dE\,\sigma(E)\left(-\frac{\partial f}{\partial E}\right)_{E=\mu(z)-U(z)} \tag{8.9}$$

For low bias, we could replace $\mu(z)$ with μ_0 to obtain our earlier result in Eq,(7.12) from obtained by directly linearizing the BTE.

Note that to use Eqs.(8.7), (8.8) we have to determine $\mu(z) - U(z)$ from a self-consistent solution the Poisson equation (ε : Permittivity, n_0, *n:* electron density per unit volume at equilibrium and out of equilibrium)

$$\frac{d}{dz}\left(\varepsilon\frac{dU}{dz}\right) = q^2(n-n_0) \tag{8.10}$$

The electron density per unit length entering the Poisson equation is calculated by filling up the density of states (per unit length) shifted by the local potential *U(z),* according to the local electrochemical potential, so that we can write

$$n(z) \equiv \int_{-\infty}^{+\infty} dE \, \frac{D(E-U(z))}{L} \, \frac{1}{1+\exp\dfrac{E-\mu(z)}{kT}} \tag{8.11a}$$

$$n_0 = \int_{-\infty}^{+\infty} dE \, \frac{D(E)}{L} \, \frac{1}{1+\exp\dfrac{E-\mu_0}{kT}} \tag{8.11b}$$

Solving Eq.(8.11)) self-consistently with the Poisson equation (Eq.(8.10)) is indeed the standard approach to obtaining the correct $\mu(z)$, $U(z)$, which can then be used to find the current from Eq.(8.8). We could view this procedure as the extended channel version of the point channel model in Fig.8.7 as shown below.

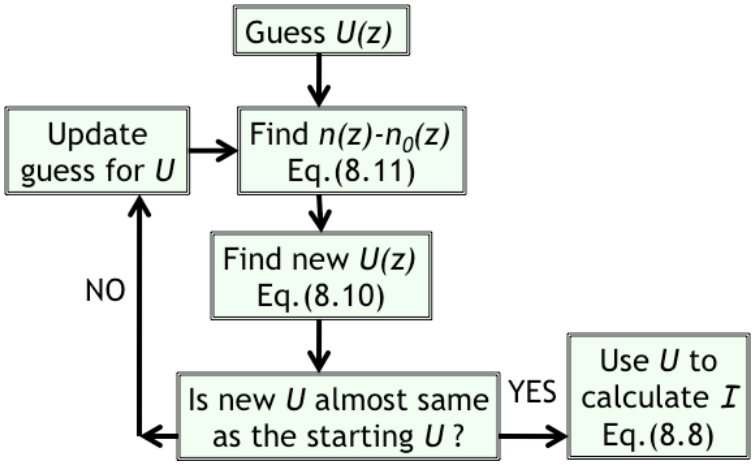


Fig.8.12. Extended channel version of the point channel model in Fig.8.7.

Note that this whole approach is based on the assumption of local electrochemical potentials $\mu^{\pm}(z)$ describing right and left-moving electrons whose average is the $\mu(z)$ appearing in Eq.(9.1). In general, electron distributions can deviate so badly from Fermi functions that an

electrochemical potential may not be adequate and one needs the full semiclassical formalism based on the Boltzmann Transport Equation (BTE) and much progress has been made in this direction. However, full-fledged BTE-based simulation is time-consuming and the drift-diffusion equation based on the concept of a local potential $\mu(z)$ continues to be the "bread and butter" of device modeling.

What our bottom-up approach adds is that Eq.(8.8) can be used even to model ballistic channels if the boundary conditions are modified appropriately (Eq.(6.4)) to include the interface resistance, a result that was obtained by carefully accounting for the distinction between $\mu^{+}(z)$ and $\mu^{-}(z)$ (Lecture 6).

Lecture 9

Smart Contacts

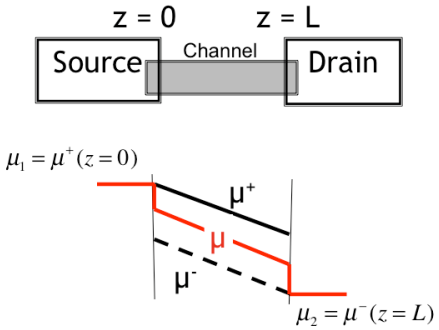
9.1. Why p-n Junctions are Different
9.2. Contacts are Fundamental

We are now ready to finish up with part one of these lectures, which I entitled "the new Ohm's law" referring to

$$R = \frac{\rho(L+\lambda)}{A} \qquad \text{(same as Eq.(4.2))}$$

which includes an extra contact resistance $\rho\lambda / A$ that depends solely on the properties of the channel and cannot be eliminated by better contacting procedures.

As we saw in Lecture 6, the key concept in identifying this interface resistance was the recognition that when a current flows, the electrochemical potentials μ^+ and μ^- for the drainbound and sourcebound states are different (Fig.6.3, also reproduced here for convenience).



From Eqs.(6.3b) and (6.6c) we could write (Note: $\mu_1 - \mu_2 = qV$)

$$\delta\mu \equiv \mu^+ - \mu^- = \frac{\mu_1 - \mu_2}{1 + L/\lambda} \qquad (9.1)$$

The contacts held at different potentials μ_1 and μ_2 drive the two groups of states (drainbound and sourcebound) out of equilibrium, while backscattering processes described by the mean free path λ try to restore equilibrium. Eq.(9.1) describes the result of these competing forces.

Normally we do not like to deal with multiple electrochemical potentials. The diffusion equation for example (see Eq.(6.1)),

$$\frac{I}{A} = -\frac{\sigma(z)}{q}\frac{d\mu}{dz} \tag{9.2}$$

works in terms of a single potential $\mu(z)$ and what we saw in Lecture 6 was how we could sweep the two potentials $\mu^+(z)$ and $\mu^-(z)$ under the proverbial rug, by defining $\mu(z)$ as the average of the two and including interface resistances into the boundary conditions by replacing Eq.(6.2) with Eq.(6.4).

The point I wish to make in this Lecture is that this separation of the electrochemical potentials for different groups of states is really far more ubiquitous and cannot always be swept under the rug. Indeed I would like to go further and argue that the most interesting devices of the future will be the ones where multiple electrochemical potentials will represent the essential physics and cannot be swept under the rug.

This is not really as exotic as it may sound. For example, all semiconductor device texts start with the p-n junction for which the need for two separate electrochemical potentials is well-recognized. Let me elaborate.

9.1. Why p-n Junctions are Different

Fig.9.1 shows a grayscale plot of the density of states $D(z,E)$. The white band indicates the bandgap with a non-zero DOS both above and below it on each side which are shifted in energy with respect to each other. A

positive voltage is applied to the right with respect to the left, so that μ_2 is lower than μ_1 as shown.

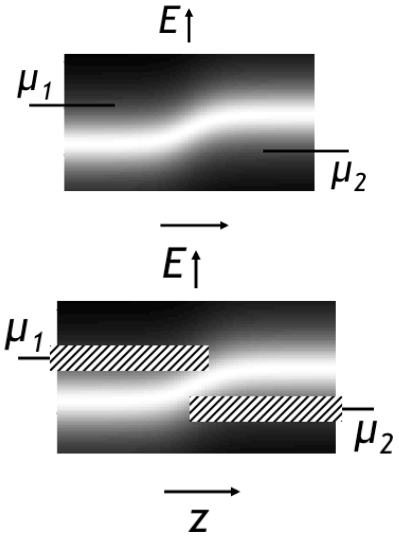


Fig.9.1. Simplified grayscale plot of the spatially varying density of states *D(z,E)* across a p-n junction.

If we look at a narrow range of energies around μ_1 (see shaded area on the left) it communicates primarily with contact 1. If we look at a narrow range of energies around μ_2 (see shaded area on the right) it communicates primarily with contact 2.

We could draw an *idealized* diagram with each of these two groups communicating just with one contact and cut off from the other as shown in Fig.9.2. In reality of course neither group is completely cutoff from either contact, and people who design real devices often go to great lengths to achieve better isolation, but let us not worry about such details.

Would the idealized device in Fig.9.2 allow any current to flow? None at all, if we it were an elastic resistor. There is no energy channel that will let an electron get all the way from left to right. The ones connected to

the left are disconnected from the right and those connected to the right are disconnected from the left.

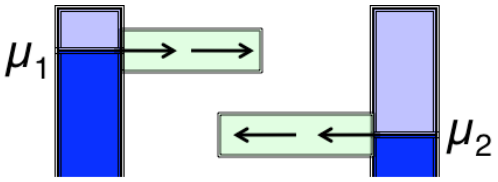


Fig.9.2. An idealized version of the p-n junction in Fig.9.1.

But current can and does flow because of inelastic processes that allow electrons to change energies along the channel. Electrons can then come in from the left, change energy and then exit to the right as sketched in Fig.9.3.

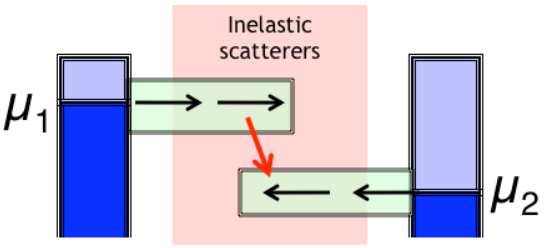


Fig.9.3. Current flow in the idealized device of Fig.9.2 is facilitated by distributed inelastic processes.

Indeed this is exactly how currents flow in p-n junctions, by transferring from the upper group of states down to the lower group by inelastic processes, which are generally referred to as recombination-generation (R-G) processes, since people like to think in terms of electrons in the upper group recombining with a "hole" in the lower group. But as we mentioned in Lecture 5, this is really an unnecessary complication and one could simply think purely in terms of electrons transferring inelastically from one group of states to another.

The point to note is that this class of devices cannot be described with one electrochemical potential and to capture the correct physics, it is

essential to treat the two groups of states separately, introducing ***two different electrochemical potentials***, labeled with the index *n*

$$I_n = -\frac{\sigma_n}{q}\frac{d\mu_n}{dz} \tag{9.3}$$

These currents are all coupled together by inelastic processes generally called "RG processes" in the context of p-n junctions

$$\frac{dI_n}{dz} = \sum_m [RG]_{m\to n} - [RG]_{n\to m} \tag{9.4}$$

that take electrons from one group of states *m* to the other *n*. This is indeed the way p-n junctions are modeled.

It is well-known that the current in a p-n junction is given by an expression of the form

$$I = I_0\,(e^{qV/\nu kT} - 1) \tag{9.5}$$

where the number ν as well as the coefficient I_0 are determined by the nature of the inelastic or RG processes. The conductivities σ_n of either of the two groups of states plays hardly any role in determining this current.

The physical reason for this is clear. The rate-determining step in current flow is the inelastic process transferring electrons from one group of states to the other. Transport within any of these groups only adds an unimportant resistance in series with the basic device.

Everything we have talked about in these lectures has been about the conductivities σ_n of the homogeneous p-type or n-type materials. And this is exactly the physics that is relevant to the operation of the most popular electronic device today, namely the Field Effect Transistor (FET) whose conductivity is controlled by a gate electrode through the electrostatic potential *U*.

But the p-n junction is a totally different device from the FET both in terms of its current-voltage characteristics and the physics that underlies it. It is the basic device structure used to construct solar cells and the principle it embodies is key to a broad class of energy conversion devices. So let me take a short detour to elaborate on this principle.

9.1.1. Current-Voltage Characteristics

Consider for example the device in Fig.9.4 assuming that the upper group of states (labeled A) is clustered around an energy ε_A while the lower group (labeled B) is clustered around ε_B .

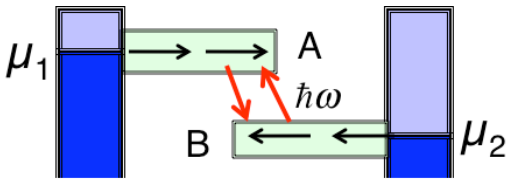


Fig.9.4. Same as Fig.9.3 with the two groups of states labeled A and B. Electronic transitions between A and B are facilitated by inelastic interactions.

The essential physics of such p-n junction like devices is contained not in Eq.(9.3), but in Eq.(9.4) which for two levels A and B can be written as

$$I \sim \overbrace{D_{B\leftarrow A}\, f_A(\varepsilon_A)\,(1-f_B(\varepsilon_B))}^{A\rightarrow B} - \overbrace{D_{A\leftarrow B}\, f_B(\varepsilon_B)\,(1-f_A(\varepsilon_A))}^{B\rightarrow A} \quad (9.6)$$

where the coefficients D_{BA} and D_{AB} denote the strength of the inelastic processes inducing the transitions from A to B and from B to A

respectively (note that the transition occurs from the second subscript to the first).

Interestingly these two rates D_{AB} and D_{BA} are generally NOT equal. D_{AB} involves absorbing an amount of energy

$$\hbar\omega = \varepsilon_A - \varepsilon_B$$

from the surroundings, while D_{BA} involves giving up the same amount of energy. A fundamental principle of equilibrium statistical mechanics (see Lecture 16) is that if the entity causing the inelastic scattering is at equilibrium with a temperature T_0, then it is always harder to absorb energy from it than it is give up energy to it and the ratio of the two processes is given by

$$\frac{D_{AB}}{D_{BA}} = \exp\left(-\frac{\hbar\omega}{kT_0}\right) \tag{9.7}$$

We can write the current from Eq.(9.4) in the form

$$I \sim D_{AB}\, f_B(\varepsilon_B)\,(1 - f_A(\varepsilon_A))\,\left(X - 1\right) \tag{9.8}$$

where

$$X \equiv \frac{D_{BA}}{D_{AB}} \frac{f_A(\varepsilon_A)}{1 - f_A(\varepsilon_A)} \frac{1 - f_B(\varepsilon_B)}{f_B(\varepsilon_B)} \tag{9.9}$$

Making use of Eq.(9.8), Eq.(9.9) and the following property of Fermi functions (Eq.(2.2))

$$\frac{1 - f_0(\varepsilon)}{f_0(\varepsilon)} = \exp\left(\frac{\varepsilon - \mu_0}{kT}\right) \tag{9.10}$$

we can rewrite Eq.(9.9) as

$$X = \exp\left(\frac{\hbar\omega}{kT_0} - \frac{\hbar\omega}{kT}\right) \exp\left(\frac{\mu_A - \mu_B}{kT}\right) \tag{9.11}$$

Since Level A is connected to contact 1 and Level B to contact 2, if the inelastic processes taking electrons from A to B are not too strong, level A is almost in equilibrium with contact 1 and level B with contact 2 , so that

$$\mu_A - \mu_B \;\cong\; \mu_1 - \mu_2 \;=\; qV$$

If $T_0 = T$, we can write the current from Eq.(9.8) as

$$I \;\sim\; (X-1) \;\sim\; e^{qV/kT} - 1$$

which is the standard *I-V* relation for p-n junctions stated earlier (see Eq.(9.5)) with $\nu = 1$. Other values of ν would be obtained if we consider more elaborate RG processes rather than the direct "band-to-band" processes considered here.

But the more important point I want to stress is that this device can be used for ***energy conversion***. If the scatterers are at a temperature different from that of the device ($T_0 \neq T$) then one can have a current flowing even without any applied voltage. This short circuit current is given by

$$I_{sc} \;\equiv\; I(V=0) \;\sim\; \exp\frac{\hbar\omega}{k}\left(\frac{1}{T_0} - \frac{1}{T}\right) - 1 \tag{9.12}$$

One could in principle use a device like this to convert a temperature difference ($T_0 \neq T$) into an electrical current. The short circuit current has the opposite sign for $T_0 > T$ and for $T > T_0$. Readers familiar with Feynman's ratchet and pawl lecture (Feynman 1963, cited in Lecture 17 of these notes) may notice the similarity. The ratchet reverses direction depending on whether its temperature is lower or higher than the ambient.

One could view more practical devices like solar cells as embodiments of the same principle, the light from the sun having a temperature $T_0 \sim$

6000^0C characteristic of the surface of the sun, much larger than the ambient temperature.

From Eq.(9.8) it is easy to see that under open circuit conditions ($I=0$), we must have $X=1$, so that from Eq.(9.11) we have

$$\frac{qV_{oc}}{\hbar\omega} = 1-\frac{T}{T_0}$$

The left hand side represents the energy extracted per photon under very low current (near open circuit) conditions, so that this could be called the Carnot efficiency of a solar cell viewed as a "heat engine". However, since $T_0 >> T$, this Carnot efficiency is very close to 100% and my colleague Ashraf often points out that other factors related to the small angular spectrum of solar energy are important in lowering the ideal efficiency to much lower values.

9.2. Contacts Are Fundamental

The point I want to make is how important the discriminating contacts are in the design of this class of devices which we could generally refer to as "solar cells" (Fig.9.5a). The external source raises electrons from the B states to the A states from where they exit through the left contact, while the empty state left behind in B is filled up by an electron that comes in through the right contact. Every electron raised from B to A thus causes an electron to flow in the external circuit.

But if the contacts are connected "normally" injecting and extracting equally from either group (Figure 9.5b) then we cannot expect any current to flow in the external circuit, from the sheer symmetry of the arrangement. After all, why should electrons flow from left to right any more that they would flow from right to left?

It is this asymmetric contacting that makes p-n junctions fundamentally different from the Field Effect Transistor (FET) that we started our

lectures with, both in terms of the current-voltage characteristics and the physics underlying it. It is of course well recognized that the physics of p-n junctions demands two different electrochemical potentials. What is not as well recognized is the generic nature of this phenomenon. Let me explain.

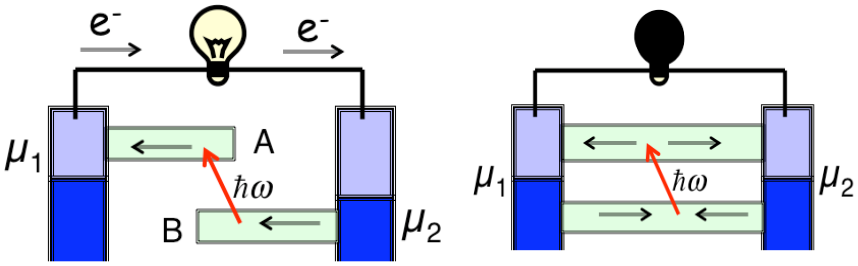


Fig.9.5. (a) Asymmetric contacts are central to the operation of the "solar cell". (b) If contacted symmetrically no electrical output is obtained.

For most of these lectures we have discussed how the contacts in an ordinary device drive drainbound and sourcebound states out of equilibrium faster than backscattering processes can restore equilibrium. In p-n junctions we just saw how the contacts drive the two bands out of equilibrium, faster than R-G processes can restore equilibrium. In Lecture 14 we will talk about spin valve devices where magnetic contacts drive upspin and downspin states out of equilibrium faster than spin-flip processes can restore equilibrium.

In every case there are groups of states A, B etc that are driven out of equilibrium by smart contacts that can discriminate between them.

More and more of such examples can be expected in the coming years, as we learn to control current flow not just with gate electrodes that control the electrostatic potential, but with subtle contacting schemes that engineer the electrochemical potential(s). Many believe that nature does

just that in designing many biological 'devices', but that is a different story. In the context of man-made devices there are many possibilities. Perhaps we will figure out how to contact s-orbitals differently from p-orbitals, or one valley differently from another valley, leading to fundamentally different devices.

But this requires a basic change in approach. Traditionally the work of device design has been divided neatly between two groups of specialists: physicists and material scientists who innovate new materials using atomistic theory and device engineers who worry about contacts and related issues using macroscopic theory. Future "solar cells" that seek to function effectively at the microscopic level may well require an approach that integrates materials and contacts at the atomistic level. Perhaps then we will be able to create devices that rival the marvels of nature like photosynthesis.