

## 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).

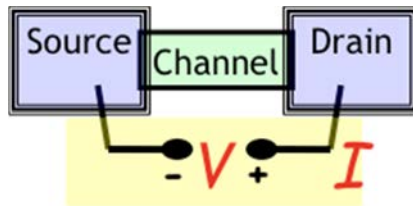


Fig.1.1a.

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  $R = \text{Voltage } (V) / \text{Current } (I)$  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.

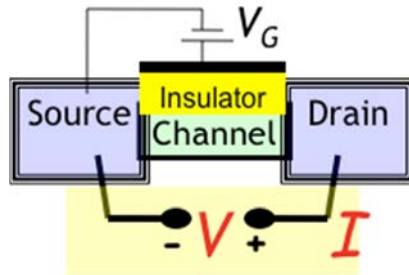


Fig.1.1b.

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.

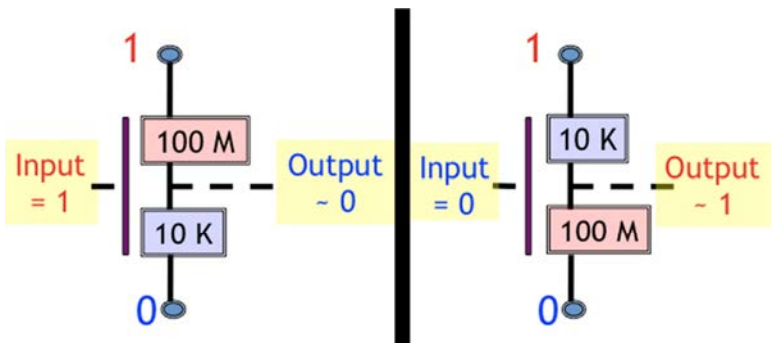


Fig.1.2.

A complementary pair of FET's form an inverter switch.

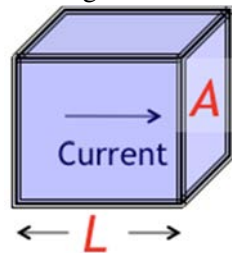
Both the increasing number and the speed of transistors are consequences of their ever-shrinking size and it is this continuing miniaturization that has driven the industry from the first four-function calculators of the 1970's to the modern laptops. For example, if each transistor takes up a space of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

$$3\ \text{cm} / 10\ \mu\text{m} = 3000$$

That is where things stood back in the ancient 1990's. But now that a transistor takes up an area of  $\sim 1\ \mu\text{m}^2$ , we can fit 900 million (nearly a billion) of them into the same  $3\text{cm} \times 3\text{cm}$  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



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

$\rho$  being a geometry-independent property of the material that the channel is made of.

The reciprocal of the resistance is the conductance

$$\frac{I}{V} = \frac{\sigma A}{L} \quad (1.1b)$$

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,  $\lambda$  that an electron travels before getting scattered is typically less than a micrometer (also called a micron =  $10^{-3}$  mm, denoted  $\mu\text{m}$ ) in common semiconductors, but it varies widely with temperature and from one material to another.

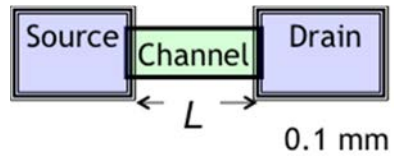
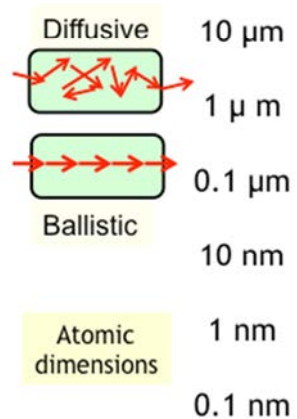
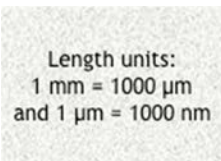


Fig.1.3.

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  $\sim 50$  nm, which amounts to a few hundred atoms!



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

$$R_B = \frac{h}{\underbrace{q^2}} \frac{1}{M} \quad (1.2)$$

$\sim 25 \text{ K}\Omega$

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 ( $\lambda$ : mean free path)

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

$R_B$

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

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

with a new expression

$$\rho = \frac{h}{q^2} \frac{A}{M \lambda} \quad (1.5)$$

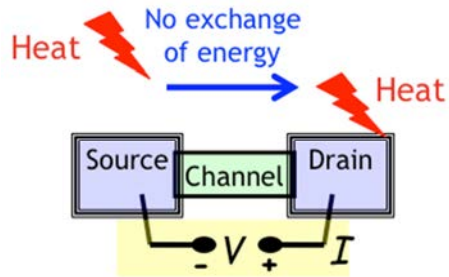
that provides a different view of resistivity in terms of the number of modes per unit area and the mean free path.

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

$$\text{Classical Dynamics} + \text{⚡} = \text{BTE}$$

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 Schrödinger equation with heat generating processes. This is exactly what is achieved in the non-equilibrium Green function (NEGF) method

$$\text{Quantum Dynamics} + \text{⚡} = \text{NEGF}$$

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.

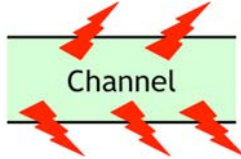


Fig.1.5. Resistance in long conductors primarily arise from distributed heat generating processes along the channel. Prior to 1990, papers dealing with basic transport theory seldom considered the actual physical contacts.

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.

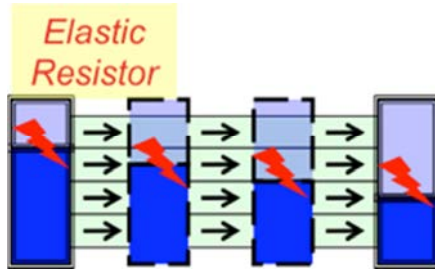


Fig.1.6. Long resistors can be approximately viewed as a series of elastic resistors, as discussed in Section 3.3.

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.

**This page intentionally left blank**

## Lecture 2

# Why Electrons Flow

2.1. *Two Key Concepts*

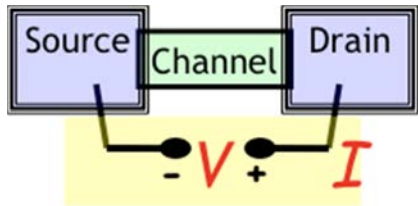
2.2. *Fermi Function*

2.3. *Non-equilibrium: Two Fermi Functions*

2.4. *Linear Response*

2.5. *Difference in “Agenda” Drives the 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.

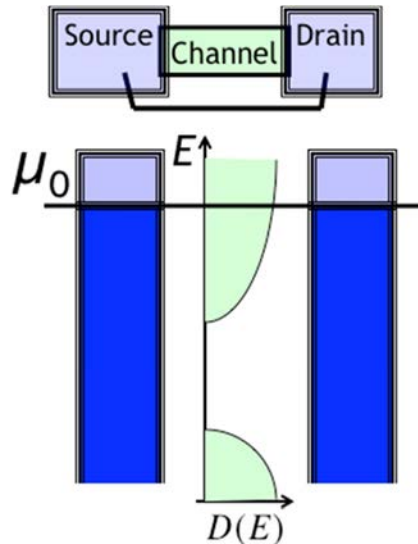


Fig.2.1.

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  $\mu_0$  separating the filled from the empty states.

The second key input is the *location of the electrochemical potential*,  $\mu_0$  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*”.

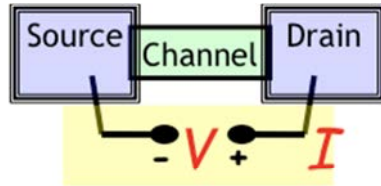
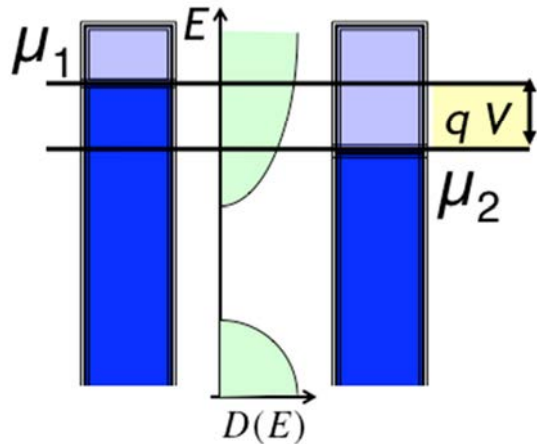


Fig.2.2.

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:

$$\mu_1 - \mu_2 = qV \quad (2.1)$$

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  $\mu_1$  and  $\mu_2$  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  $\mu$  and emptying all states with energies greater than  $\mu$ .

Consider the states with energy  $E$  that are less than  $\mu_1$  but greater than  $\mu_2$ . 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  $\mu_1$  and  $\mu_2$ . 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  $\mu_1$  and  $\mu_2$  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  $\mu_1$  and  $\mu_2$ , or more correctly outside  $\pm$  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  $\mu_0$  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 ( $\sim 80 \mu\text{eV/K}$ ) and  $T$  is the absolute temperature. Mathematically this transition is described by the Fermi function :

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

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).

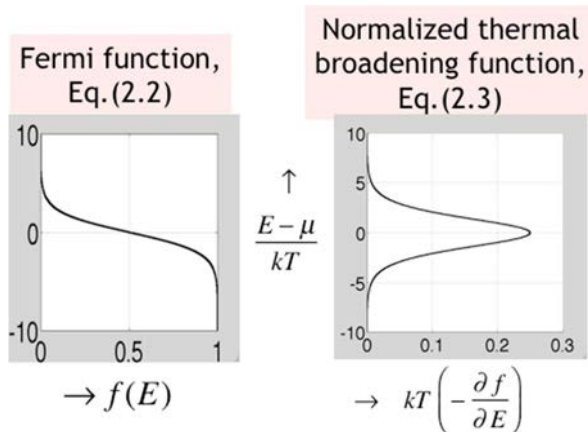


Fig.2.3. Fermi function and the normalized (dimensionless) thermal broadening function.

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 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  $\sim \pm 2kT$  around  $\mu_0$ .

### 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:

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

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

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

Note:

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

$$F_T(E, \mu) = F_T(E - \mu) = F_T(\mu - E) \quad (2.4a)$$

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

$$F_T = f(1 - f) \quad (2.4b)$$

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

$$\begin{aligned} \int_{-\infty}^{+\infty} dE F_T(E, \mu) &= kT \int_{-\infty}^{+\infty} dE \left( -\frac{\partial f}{\partial E} \right) \\ &= kT \left[ -f \right]_{-\infty}^{+\infty} = kT (1 - 0) = kT \end{aligned} \quad (2.4c)$$

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  $\sim 4kT$ .

### 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):

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

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

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  $\mu_1, \mu_2$  since  $f_1$  and  $f_2$  are

both zero. It also goes to zero when  $E$  lies way below  $\mu_1, \mu_2$  since  $f_1$  and  $f_2$  are both one. Current flow occurs only in the intermediate energy window, as we had argued earlier.

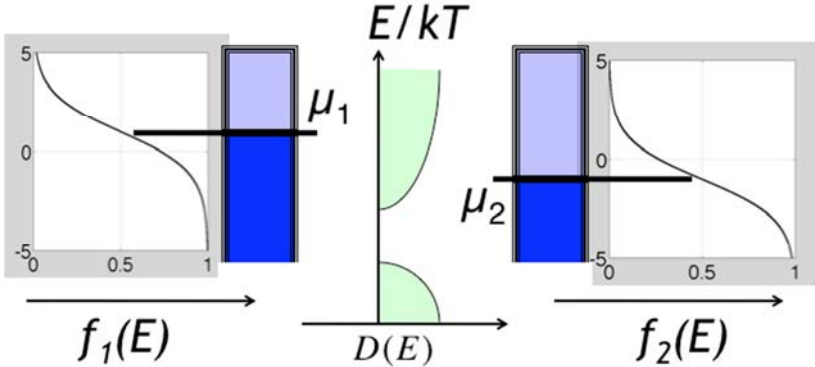


Fig.2.4.

Electrons in the contacts occupy the available states with a probability described by a Fermi function  $f(E)$  with the appropriate electrochemical potential  $\mu$ .

## 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

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

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:

$$F(E) \rightarrow F_T(E), \text{ as } qV / kT \rightarrow 0$$

so that from Eq.(2.6)

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

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

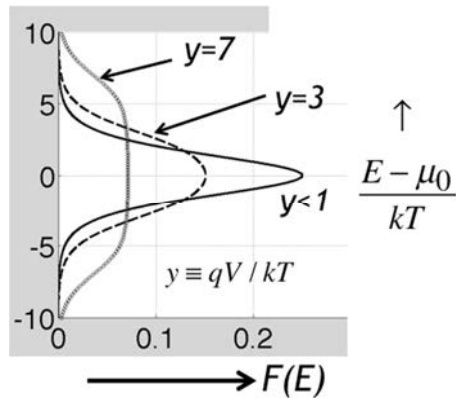


Fig.2.5.  $F(E)$  from Eq.(2.6) versus  $(E - \mu_0) / kT$  for different values of  $y = qV / kT$ .

The validity of Eq.(2.7) for  $qV \ll kT$  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

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

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

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

The two functions look similar over a limited range of energies

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

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 field  $F$  as the driving term and adding a frictional term to Newton’s law ( $\tau_m$  is the so-called “momentum relaxation time”)

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

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  $\sim 10^{20}$  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  $\sim kT$  around the electrochemical potential  $\mu_0$ , which can vary widely from one material to another. This is well-known to experts and comes mathematically from the dependence of the conductivity

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

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 stop 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.

**This page intentionally left blank**



## Lecture 3

# The Elastic Resistor

### 3.1. How an Elastic Resistor Dissipates Heat

### 3.2. Conductance of an Elastic Resistor

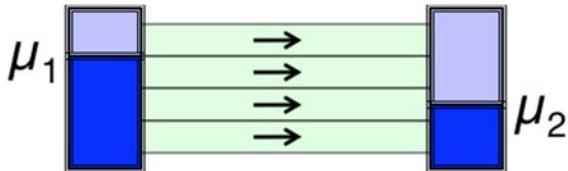
### 3.3. Why an Elastic Resistor is Relevant

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

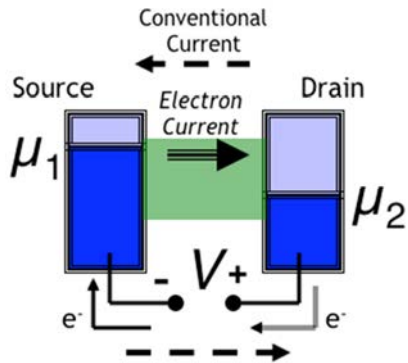
$$\frac{I}{V} = \int_{-\infty}^{+\infty} dE \left( -\frac{\partial f_0}{\partial E} \right) G(E) \quad (3.1)$$

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  $\sim \pm 2kT$  around the electrochemical potential  $\mu_0$  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  $D(E)$ .

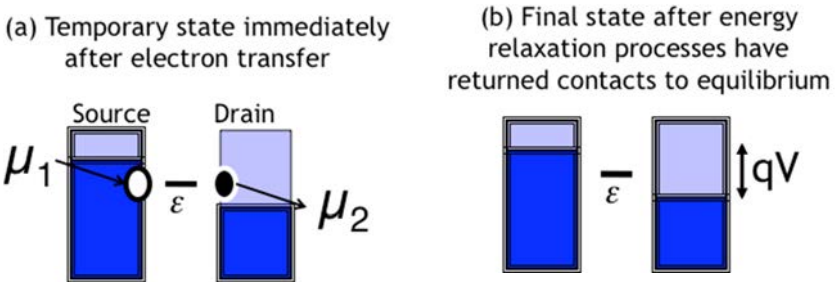
The concept of an elastic resistor is not only useful in understanding nanoscale devices, but it also helps understand transport properties like the conductivity of large resistors by viewing them as multiple elastic resistors in series, as explained in **Section 3.3**. This is what makes the bottom-up 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  $\varepsilon$ . Every time an electron crosses over through the channel, it appears as a "hot electron" on the drain side with an energy  $\varepsilon$  in excess of the local electrochemical potential  $\mu_2$  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  $\varepsilon$  that is much less than the local electrochemical potential  $\mu_1$ , 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,

*an energy  $(\mu_1 - \epsilon)$  is dissipated in the source*

*an energy  $(\epsilon - \mu_2)$  is dissipated in the drain*

The total energy dissipated is

$$\mu_1 - \mu_2 = qV$$

which is supplied by the external battery that maintains the potential difference  $\mu_1 - \mu_2$ . The overall flow of electrons and heat is summarized in Fig.3.3 below.

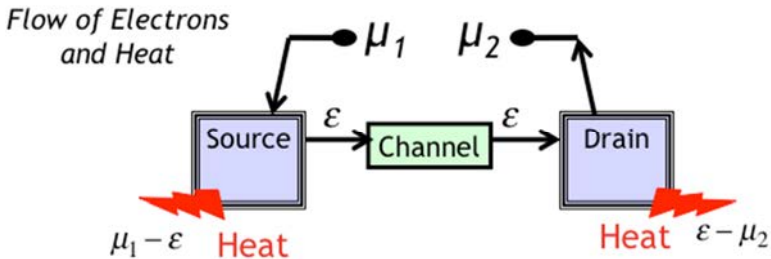


Fig.3.3. Flow of electrons and heat in a one-level elastic resistor having one level with  $E = \epsilon$ .

If  $N$  electrons cross over in a time  $t$

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

since

$$\text{Current} = q * N / t$$

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

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  $\mathcal{E}$  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

$$I_{one\ level} = \frac{q}{t} (f_1(\mathcal{E}) - f_2(\mathcal{E})) \quad (3.2)$$

*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:

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

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

where

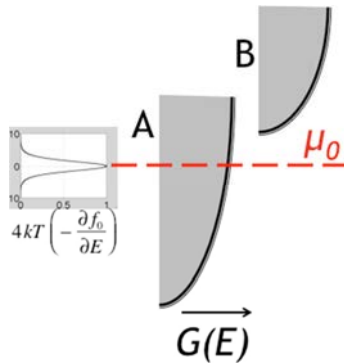
If the applied voltage  $\mu_1 - \mu_2 = qV$  is much less than  $kT$ , we can use Eq.(2.7) to write

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

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

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

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  $\mu_0$  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

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

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?

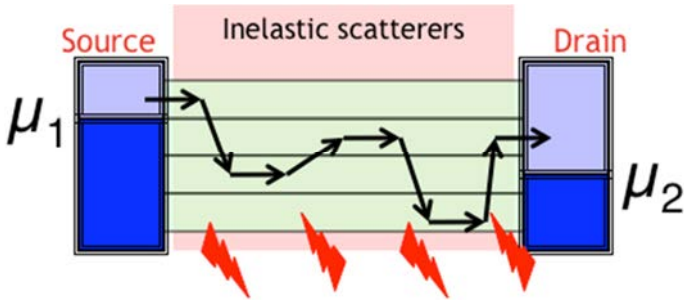


Fig.3.4  
Real conductors have inelastic scatterers distributed throughout the channel.

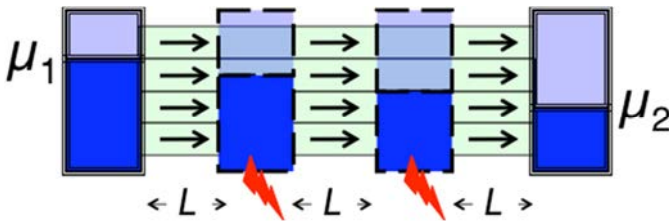


Fig.3.5  
A hypothetical series of elastic resistors as an approximation to a real resistor with distributed inelastic scattering as shown in Fig.3.4.

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

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

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  $\lambda$  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 ( $L \gg \lambda$ ).

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

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

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