

A Primer on Semiconductor Fundamentals

Mark Lundstrom
Purdue University

The two lectures that follow quickly summarize some important semiconductor fundamentals. For those acquainted with semiconductors, they may be useful as a brief refresher. For those just getting started with semiconductors, my hope is that these two lectures provide just enough understanding to allow you to get started in understanding semiconductor devices. A much deeper understanding of semiconductor fundamentals will be necessary to get very far in this field. The two references below are good places to start.

*Mark Lundstrom
Purdue University
December, 2015*

For an introduction to semiconductor fundamentals and to semiconductor devices, see:

[1] Robert F. Pierret *Semiconductor Device Fundamentals*, 2nd Ed., Addison-Wesley Publishing Co, 1996..

For a more advanced treatment of semiconductor fundamentals:

[2] Robert F. Pierret *Advanced Semiconductor Fundamentals*, 2nd Ed., Vol. VI, Modular Series on Solid-State Devices, Prentice Hall, Upper Saddle River, N.J., USA, 2003.

Chapter 1

Equilibrium

- 1.1 Introduction
- 1.2 From energy levels to energy bands
- 1.3 Density-of-states and Fermi function
- 1.4 Carrier densities and Fermi levels
- 1.5 Doping and carrier densities
- 1.6 Energy band diagrams
- 1.7 Bandstructure
- 1.8 Quantum confinement
- 1.8 Summary
- 1.9 References

1.1 Introduction

A physical understanding of how semiconductor devices work can be conveyed without going too deeply into semiconductor physics, but a basic understanding of some key concepts at the level of an introductory course in solid state physics or semiconductor devices is necessary. This lecture and the next will summarize a few essential concepts. Results will be stated, not derived. These two lectures will be a review for those who have already had a basic course in semiconductors and should provide just enough background for others who may be encountering semiconductor devices for the first time. The references steer the reader to more extensive discussions of semiconductor physics.

1.2 From energy levels to energy bands

Silicon has an atomic number of 14; it has 14 electrons whose negative charges are balanced by the positive charges of 14 protons in the nucleus. We learn in Freshman chemistry that these fourteen electrons occupy various energy levels associated with molecular orbitals. The lowest energy level is the $n = 1$ s orbital, which can accommodate two electrons (one with spin up and the other with spin down). The next energy level is the $n = 2$ s orbital, then the three $n = 2$ p orbitals p_x , p_y , and p_z . Each of the three p orbitals can accommodate two electrons so the $n = 2$ p orbitals can accommodate six electrons. Next are the $n = 3$ s and p orbitals.

The 14 electrons occupy orbitals with the lowest energy first, then the next lowest, etc. The orbital configuration for Si is $1s^2 2s^2 2p^6 3s^2 3p^2$. The $n = 1$ and $n = 2$ levels are completely filled. The $n = 3$ s and p orbitals can accommodate 8 electrons, but only 4 of the orbitals need to be filled to accommodate all 14 electrons, so in the 3s and 3p orbitals, we have 8 “states” that can accommodate electrons, but only 4 are filled. All 14 electrons are accommodated in a set of molecular orbitals with discrete energies. We are most interested in the highest occupied orbitals, which contain the “valence” electrons, because they are responsible for chemical bonding. In addition to the highest occupied orbitals, we are also interested in the

lowest unoccupied orbitals, because these are the energy levels whose electron populations we can modify to produce electronic devices. The lower energy electrons, closer to the atom's “core” will not be of interest to us.

The semiconductor used for most semiconductor devices such as transistors is silicon, and most silicon transistors are made with high quality, single crystal silicon in which each silicon atom covalently bonds with four nearest neighbors in the so-called *diamond lattice* structure. The density of Si atoms is about about $N_A = 5 \times 10^{22} \text{ cm}^{-3}$, and the nearest neighbor spacing is 0.235 nm. Since there were eight valence electron states with 4 of them occupied for each Si atom, we expect to find $8N_A$ valence states in the crystal with $4N_A$ of them occupied. That is what happens, but the quantum mechanical interaction of the electrons shifts the energies and broadens them. Half of the states decrease in energy and become the “bonding states” responsible for the covalent bonds in the Si lattice, and the other half of the states (the anti-bonding states) increase in energy.

The $4N_A$ filled electron energy levels become $4N_A$ occupied states in a range of energies called the *valence band* (Fig. 1.1). The $4N_A$ unoccupied states become a band of unoccupied states known as the *conduction band* (Fig. 1.1). As shown in Fig. 1.1, the top of the valence band is separated from the bottom of the conduction band by an “energy gap” (or *band gap*), which is about 1.12 eV for Si. At $T = 0 \text{ K}$, all of the $4N_A$ states in the valence band are occupied and all of the $4N_A$ states in the conduction band are empty.

At a finite temperature, the Si lattice is vibrating and has an average thermal energy of $3k_B T/2$, where k_B is Boltzman's constant. At room temperature, $T = 300 \text{ K}$, $k_B T = 0.026 \text{ eV}$, which is about 40 times smaller than the bandgap of Si. On average, there is not enough thermal energy available to break the covalent bonds and promote electrons from the valence band to the conduction band, but there is a small probability of thermally exciting an electron from the valence band to the conduction band. As shown in Fig. 1.1, this means that there will be a small probability of empty states (called *holes*) in the valence and a small probability of filled states near the bottom of the conduction band. The density of electrons in the conduction band is equal to the density of holes in the valence band and is known as the *intrinsic carrier concentration*, n_i . For silicon at room temperature, $n_i \approx 10^{10} \text{ cm}^{-3}$ [1, 2].

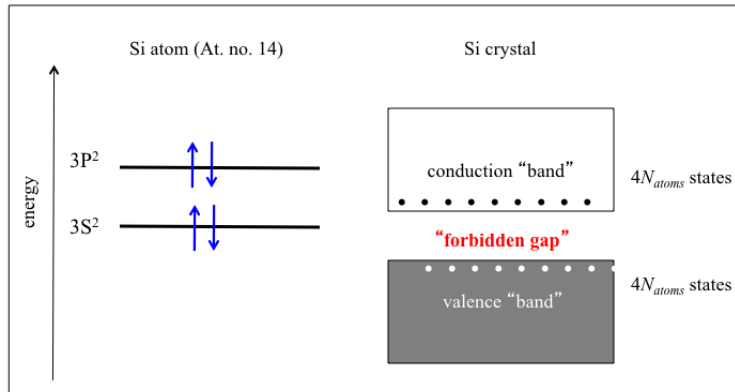


Figure 1.1: Illustration of how the energy levels of the Si valence electrons become the valence and conduction bands in a Si crystal.

A Si MOSFET (one of the most important semiconductor devices) consists of a metal gate electrode and source and drain contacts, an SiO_2 gate insulator, and the Si substrate. Figure 1.2 illustrates the differences between the energy bands of these materials. In an insulator, the valence band is filled with electrons, but the bandgap is very large, so there is virtually no probability of promoting an electron from the valence band to the conduction band, so no holes exist in the valence band and no electrons in the conduction band. A metal is different in that the electron states are filled up to the middle of a band. Metals conduct electricity very well, and insulators have negligible conductivity. In a semiconductor, we can change the number of holes in the valence band and electrons in the conduction band either by varying the temperature or by introducing a small concentration of *dopants* in the Si lattice. The conductivity of a semiconductor can be made to vary over many orders of magnitude – from almost metallic to nearly insulating, which is what

makes them so useful.

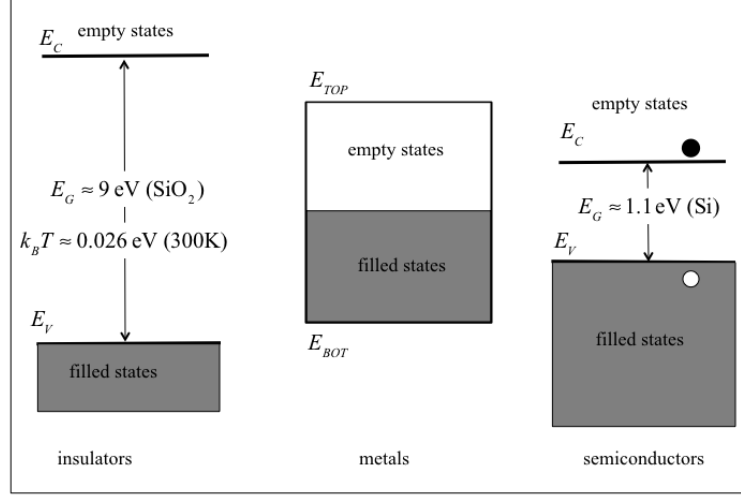


Figure 1.2: Sketches of the energy bands in an insulator, semiconductor, and metal.

1.3 Density-of-states and Fermi function

We have seen that in Si (our prototypical semiconductor), the valence band contains $4N_A$ states per unit volume and the conduction band contains the same number. The states are distributed in energy from the bottom of the valence band to the top, and from the bottom of the conduction band to the top. In an energy range, dE , the number of states per unit volume is $D(E)dE$, where $D(E)$ is the *density-of-states*, the number of states per unit volume per unit energy. In general, the density-of-states can be a complicated function of energy. We know that it must go to zero at the bottom of the band, where a *forbidden gap* begins and at the top of the band where another forbidden gap begins. We also know the total number of states in the conduction and valence bands of Si must satisfy

$$4N_A = \int_{E_1}^{E_2} D(E) dE, \quad (1.1)$$

where E_1 is the energy at the bottom of the band and E_2 the energy at the top of the band.

The density-of-states is, in general a complicated function of energy that can be computed numerically. As indicated in Fig. 1.1, however, most of the valence band is filled, except for a few empty states near the top of the valence band, and most of the conduction band is empty, except for a few occupied states near the bottom of the conduction band. we are, therefore, most interested in the density-of-states near the top of the valence band and near the bottom of the conduction band. Simple calculations give a density-of-states near the band edges that is adequate for many semiconductor problems. For a simple three-dimensional semiconductor, we find [1 – 3].

$$D_c^{3D} = \frac{m_n^* \sqrt{2m_n^* (E - E_c)}}{\pi^2 \hbar^3} \Theta(E - E_c), \quad (1.2)$$

where m_n^* is a material-dependent property known as the *effective mass* of electrons in the conduction band, and $\Theta(E - E_c)$ is the step function, which is zero for $E < E_c$ and one for $E > E_c$. Similarly, the density-of-states near the top of the valence band is given by

$$D_v^{3D} = \frac{m_p^* \sqrt{2m_p^* (E_v - E)}}{\pi^2 \hbar^3} \Theta(E_v - E), \quad (1.3)$$

where m_p^* is a material-dependent property known as the effective mass of holes in the valence band. Equations (1.2) and (1.3) give the densities-of-states in the energy ranges of interest. Different materials will have different effective masses and, therefore, different densities-of-states [2, 3]. A three-dimensional volume of a material, a two-dimensional sheet of the same material, and a one-dimensional wire of the same material, will all have different densities-of-states. The density-of-states depends, therefore, on the material and on the dimensionality of the material.

Knowing the density-of-states is not enough – we must also know the probability that a state at energy, E , is occupied. Intuitively, we expect that the lowest energy states will have a high probability of being occupied and the highest energy states will have a low probability of being occupied. We fill up the states with two electrons per state according to the *Pauli exclusion principle* so that the energy is minimized. At $T = 0$ K, the highest filled state is at an energy, E_F , the *Fermi energy*. At finite temperatures, the probability that a state of occupied is given by the well-known, *Fermi function*,

$$f_0(E) = \frac{1}{e^{(E-E_F)/k_B T} + 1}, \quad (1.4)$$

where E_F is the *electrochemical potential*. (The subscript, “0”, reminds us that we are assuming thermal equilibrium conditions.) Strictly speaking, the electrochemical potential at $T = 0$ K is the Fermi energy, but it is common in electrical engineering to call the electrochemical potential at finite temperature the temperature-dependent Fermi energy, so for our purposes, we will denote the electrochemical potential by the symbol, E_F .

Figure 1.3 is a plot of the Fermi function. It shows that states far below E_F have a high probability of being occupied ($f_0 \rightarrow 1$ for $E \ll E_F$) and states far above E_F have a small probability of being occupied, ($f_0 \rightarrow 0$ for $E \gg E_F$). States at the Fermi energy have a probability of 0.5 of being occupied. The transition from high probability to low probability occurs in an energy range of a few $k_B T$ around E_F .

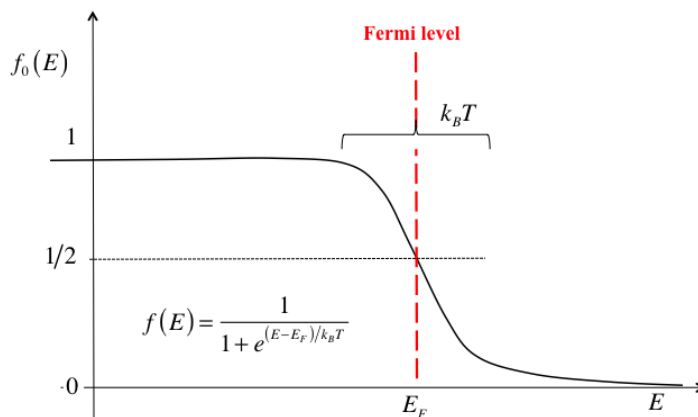


Figure 1.3: Plot of the Fermi function vs. energy.

1.4 Carrier density and Fermi level

There is a one to one relation between the Fermi level and the number of electrons (or electron density per unit volume) in a material. This is expected because to accommodate an increasing number of electrons, more and more energy levels must be filled up according to the Paul exclusion principle. The Fermi level must increase. To compute the electron density for a given Fermi energy, we multiply the number of states in an energy range, dE , which is $D(E)dE$, by the probability that the states at this energy are occupied, $f_0(E)$, and integrate from the bottom of the band to the top. The result is

$$n_0 = \int_{E_c}^{\infty} D_{3D}^c(E) f_0(E) dE. \quad (1.5)$$

Note that instead of integrating to the top of the band, we have set the upper limit to infinity. We can do this because under most conditions, the Fermi function insures that states far above the bottom of the conduction band have essentially zero probability of being occupied.

Using eqns. (1.2) and (1.4), we express the integral in eqn. (1.5) as

$$\begin{aligned} n_0 &= \int_{E_c}^{\infty} \frac{m_n^* \sqrt{2m_n^* (E - E_c)}}{\pi^2 \hbar^3} \frac{1}{1 + e^{(E - E_F)/k_B T}} dE \\ &= \frac{m_n^* \sqrt{2m_n^*}}{\pi^2 \hbar^3} \int_{E_c}^{\infty} \frac{\sqrt{(E - E_c)}}{1 + e^{(E - E_F)/k_B T}} dE \end{aligned} \quad (1.6)$$

Making a change of variables,

$$\begin{aligned} \eta &= (E - E_c)/k_B T \\ \eta_F &= (E_F - E_c)/k_B T, \end{aligned} \quad (1.7)$$

we find

$$\begin{aligned} n_0 &= \frac{m_n^* \sqrt{2m_n^*}}{\pi^2 \hbar^3} (k_B T)^{3/2} \int_0^{\infty} \frac{\eta^{1/2}}{1 + e^{\eta - \eta_F}} d\eta \\ &= N_C \frac{2}{\sqrt{\pi}} \int_0^{\infty} \frac{\eta^{1/2}}{1 + e^{\eta - \eta_F}} d\eta \end{aligned} \quad (1.8)$$

where

$$N_C \equiv \frac{1}{4} \left(\frac{2m_n^* k_B T}{\pi \hbar^2} \right)^{3/2} \quad (1.9)$$

is known as the *effective density-of-states*.

In general, the integral in eqn. (1.8) cannot be done analytically, but integrals of this type occur so often in semiconductor work that they have been given a name – *Fermi-Dirac integrals* [4]. In this case, the integral in eqn. (1.8) is a Fermi-Dirac integral of order one-half,

$$\mathcal{F}_{1/2}(\eta_F) \equiv \frac{2}{\sqrt{\pi}} \int_0^{\infty} \frac{\eta^{1/2}}{1 + e^{\eta - \eta_F}} d\eta, \quad (1.10)$$

so we can relate the equilibrium density of electrons in the conduction band to the location of the Fermi level by

$$n_0 = N_C \mathcal{F}_{1/2}(\eta_F). \quad (1.11)$$

Fermi-Dirac integrals

The Fermi-Dirac integral of order j is defined as

$$\mathcal{F}_j(\eta_F) \equiv \frac{1}{\Gamma(j+1)} \int_0^{\infty} \frac{\eta^j d\eta}{1 + e^{\eta - \eta_F}}, \quad (1.12)$$

where the Γ -function is defined for integer arguments of zero or greater as

$$\Gamma(n) = (n-1)!. \quad (1.13)$$

We also have the following useful relations,

$$\begin{aligned} \Gamma(1/2) &= \sqrt{\pi} \\ \Gamma(p+1) &= p\Gamma(p) \end{aligned} \quad (1.14)$$

For non-degenerate semiconductors, the Fermi level is several $k_B T$ below the band edge, so $\eta_F = (E_F - E_c)/k_B T \ll 0$. Under these conditions, Fermi-Dirac integrals of any order reduce to exponentials:

$$\mathcal{F}_j(\eta_F) \rightarrow e^{\eta_F} \quad \eta_F \ll 0. \quad (1.15)$$

Another useful property involves taking the derivative of a Fermi-Dirac integral,

$$\frac{d\mathcal{F}_j(\eta_F)}{d\eta_F} = \mathcal{F}_{j-1}(\eta_F). \quad (1.16)$$

These few definitions and rules are all we need for most semiconductor problems. One warning – don’t confuse the “script” Fermi-Dirac integral as defined in eqn. (1.12) with the “Roman” Fermi-Dirac integral, $F_j(\eta_F)$, which does not include the Γ -function normalization. For a good introduction to Fermi-Dirac Integrals – including approximations and scripts to compute them, see [4].

Equation (1.11) relates the density of electrons in the conduction band to the location of the Fermi level. In a non-degenerate n-type semiconductor, $E_F \ll E_c$, so the states in the conduction band lie far above the Fermi level in energy. In this case, the Fermi function, eqn. (1.4) can be approximated by an exponential,

$$f_0(E) = \frac{1}{e^{(E-E_F)/k_B T} + 1} \approx e^{(E_F-E)/k_B T}. \quad (1.17)$$

In this case, the states in the conduction band are nearly empty and the Pauli exclusion principle plays no role; Fermi-Dirac statistics as described by the Fermi function can be replaced by Maxwell-Boltzmann statistics. For non degenerate conditions, $\eta_F \ll 0$, and the Fermi-Dirac integral reduces to an exponential, so we find from eqns. (1.11) and (1.15)

$$\boxed{n_0 = N_C e^{(E_F - E_c)/k_B T}}, \quad (1.18)$$

which describes a non-degenerate semiconductor for which Maxwell-Boltzmann carrier statistics can be used instead of Fermi-Dirac statistics. Similar arguments allow us to relate the density of empty states (holes) near the top of the valence band to the Fermi level under non-degenerate conditions as

$$\boxed{p_0 = N_V e^{(E_v - E_F)/k_B T}}. \quad (1.19)$$

The product of the equilibrium electron and hole densities under non-degenerate conditions can be found by multiplying eqn. (1.18) by eqn. (1.19) to find

$$\boxed{n_0 p_0 = n_i^2}, \quad (1.20)$$

where n_i^2 is a material and temperature dependent parameter that is independent of the specific location of the Fermi level (for a non-degenerate semiconductor). From eqns. (1.18) and (1.19), we find

$$n_0 p_0 = n_i^2 = N_C N_V e^{(E_v - E_c)/k_B T} = N_C N_V e^{-E_G/k_B T}.$$

or

$$\boxed{n_i = \sqrt{N_C N_V} e^{-E_G/2k_B T}}. \quad (1.21)$$

According to eqn (1.21), semiconductors with a large bandgap have a small density of intrinsic electron-hole pairs. Note also that n_i increases exponentially with temperature. For Si at room temperature, $n_i \approx 10^{10} \text{ cm}^{-3}$ [2].

1.5 Doping and carrier densities

What makes semiconductors useful is the fact that in a semiconductor it is readily possible to change the densities of holes in the valence band and electrons in the conduction band. Stated another way, one can place the Fermi level from near the top of the valence band to near the bottom of the conduction band or anywhere

in between. One way to do this is by *doping* the semiconductor. (Another way is by using a gate to change the electrostatic potential within the semiconductor – so-called gating.) Figure 1.4 illustrates how to dope a semiconductor by substituting a small concentration of dopant atoms for Si atoms. The two-dimensional cartoon is intended to illustrate the three-dimensional bonding in which each Si atom forms chemical bonds with four nearest neighbors. If a Si atom (in column IV of the periodic table with four valence electrons) is replaced by an element such as phosphorus or arsenic from column V with five valence electrons, then the dopant forms covalent bonds with the four silicon neighbors, but the fifth valence electron is left over and is weakly bound. The small amount of thermal energy available at room temperature can break this weak bond and promote the electron to the Si conduction band leaving behind a positively charged dopant atom because it has lost one electron.

Figure 1.4 also shows what happens when a dopant from column III of the periodic table (e.g. boron) is substituted for a Si atom. In this case, the dopant has three valence electrons, so it can form covalent bonds with three of the four neighbors. It takes only a little thermal energy to move an electron from a nearby Si:Si bond and place it on the dopant site and fill the missing covalent bond. The dopant is now negatively charged because it has an extra electron, but we have one missing Si:Si bond, so we have created a hole in the valence band. By introducing small quantities of column V or column III impurities in the Si lattice, we can control the quantity of electrons in the conduction band and holes in the valence band.

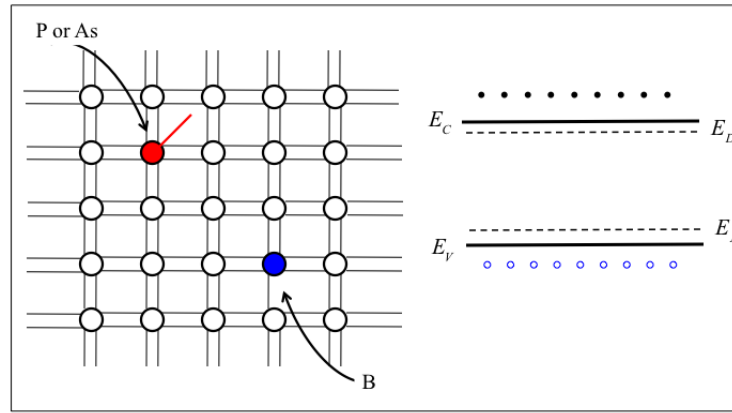


Figure 1.4: Illustration of semiconductor doping showing a Si lattice with each Si atom surrounded by four nearest neighbors and a column V, n-type dopant and a column III, p-type dopant. The energy band diagram on the right shows that dopants introduce states in the forbidden gap of Si. $E_c - E_D$ is the energy it takes to ionize a donor and place a free electron in the conduction band. $E_A - E_v$ is the energy it takes to place an electron on the acceptor, which ionizes it and places a mobile hole in the valence band.

The concentration of electrons and holes in a semiconductor is determined by the density of dopants. Assume that we have a concentration of column V *donors* that is equal to $N_D \text{ cm}^{-3}$ and a concentration of column III *acceptors* that is equal to $N_A \text{ cm}^{-3}$. Also assume that the thermal energy is sufficient to ionize these dopants, so that the concentration of ionized dopants is $N_D^+ = N_D$ and $N_A^- = N_A$. For an intrinsic semiconductor, the concentration of electrons in the conduction band and holes in the valence band is $n_0 = p_0 = n_i$. These concentrations can be changed by doping the semiconductor.

The net charge in the semiconductor is

$$\rho = q(p_0 - n_0 + N_D - N_A) \quad \text{C/cm}^3. \quad (1.22)$$

Because the electrons and hole are mobile, (the dopant atoms are fixed in the lattice and cannot move), they will move about in order to cancel the net charge and make the semiconductor neutral,

$$\rho = q(p_0 - n_0 + N_D - N_A) = 0. \quad (1.23)$$

Equation (1.23) for the charge density in C/cm^2 is one equation for the two unknowns, n_0 and p_0 , but eqn.

(1.20) provides the second equation. The result is that eqn. (1.23) can be expressed as an equation for n_0 ,

$$\frac{n_i^2}{n_0} - n_0 + N_D - N_A = 0; \quad (1.24)$$

a similar equation can be written for p_0 . By solving the resulting quadratic equation for n_0 or p_0 and then using eqn. (1.20) for the other concentration, we can determine both n_0 and p_0 . If the location of the Fermi energy is needed, it can be obtained from eqn. (1.18) or (1.19).

Consider an *extrinsic semiconductor* at a temperature high enough to ionize the dopants and for which $N_D - N_A \gg n_i$. In this case, we find

$$\begin{aligned} n_0 &\approx N_D - N_A \\ p_0 &= n_i^2/n_0 \end{aligned} \quad (1.25)$$

In the case of an extrinsic p-type semiconductor ($N_A - N_D \gg n_i$), we find analogous expressions.

1.6 Energy band diagrams

An energy band diagram is a plot of the top of the valence band and the bottom of the conduction band vs. position. Energy band diagrams play an important role in understanding the operation of semiconductor devices [1,5]. When drawing energy band diagrams, only the bottom of the conduction band and the top of the valence band are plotted, because we understand that the filled and empty states of interest are very close to these band edges. Consider a uniformly doped, p-type semiconductor. In equilibrium the Fermi level is constant, and the separation between the Fermi level and the valence band edge is also constant, as given by eqn. (1.19). The energy band diagram for this uniform semiconductor in equilibrium is a set of horizontal lines, one for E_c , one for E_v , and another for E_F . The lowest possible energy of an electron in the conduction band is E_c , the potential energy of the electron. If an electron has an energy above the bottom of the conduction band, then it has a kinetic energy of $E - E_c$.

Now let us assume that there is a position-dependent electrostatic potential, $\psi(x)$, within the semiconductor. (There are different ways to produce a position-dependent electrostatic potential, such as non-uniform doping or by the application of a gate voltage.) A positive electrostatic potential lowers the energy of an electron according to

$$E_c(x) = E_{c0} - q\psi(x), \quad (1.26)$$

where E_{c0} is the energy of the conduction band edge in the absence of an electrostatic potential. Accordingly, we expect the bands to bend in the presence of an electrostatic potential. A positive electrostatic potential bends both the conduction and valence bands down. If the semiconductor is still in equilibrium, then the Fermi level remains constant – even in the presence of a non-uniform electrostatic potential. An example is shown in Fig. 1.5 (left). From the energy band diagram, we determine the electrostatic potential by flipping either the conduction or valence band upside down as shown in Fig. 1.5 (right). It is also easy to show that the slope of $E_c(x)$ is proportional to the electric field.

Energy band diagrams are extremely useful in understanding the operation of semiconductor devices. Semiconductor device researchers and engineers make extensive use of energy band diagrams.

1.7 Bandstructure

The semiconductor's *bandstructure* relates the electron energy to its momentum. Recall that for a classical particle,

$$E(p) = \frac{p^2}{2m}, \quad (1.27)$$

where p is the momentum of the particle, and m is its mass. For electrons in a crystal, energy is related to *crystal momentum* by the effective mass. A simple example is shown in Fig. 1.6 on the left. Electrons in the conduction band have a minimum energy of E_c , and the energy increases with momentum according to

$$E(p) = E_c + \frac{p^2}{2m_n^*} = E_c + \frac{\hbar^2 k^2}{2m_n^*}. \quad (1.28)$$

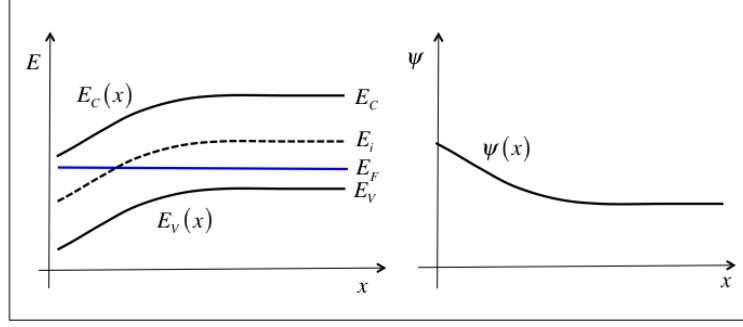


Figure 1.5: Semiconductor energy band diagram, a plot of the band edges vs. position. The fact that the Fermi level is flat shows that the semiconductor is in equilibrium. Left: An energy band diagram showing “band banding,” which indicates that a non-uniform electrostatic potential, $\psi(x)$, is present in the semiconductor. Right: The electrostatic potential, $\psi(x)$, as determined by flipping the conduction or valence band upside down. The electric field vs. position is proportional to the slope of E_C or E_V .

Electrons in the valence band are described by

$$E(p) = E_v - \frac{p^2}{2m_p^*} = E_v - \frac{\hbar^2 k^2}{2m_p^*}. \quad (1.29)$$

We say that these electrons have a negative effective mass, which is equivalent to holes with a positive effective mass. In these equations, $p = \hbar k$ is the crystal momentum, which will be discussed next.

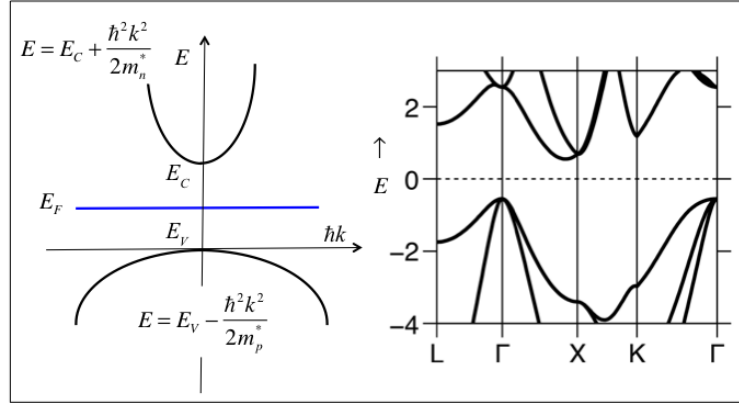


Figure 1.6: Plot of electron energy vs. crystal momentum. On the left is shown a simple, parabolic $E(k)$. On the right is shown an actual band structure for silicon. While the band structure is complex, the parabolic band assumption describes electrons near the bottom of the conduction band and near the top of the valence band, which are the two regions of interest. (The full bandstructure was computed by Dr. Jesse Maassen, Purdue University using the program, VASP.)

In Fig. 1.6, we labeled the momentum by $p = \hbar k$, where k is the electron’s wavevector. The electron is a quantum mechanical wave described by

$$\psi_n(\vec{r}) = u_{n\vec{k}}(\vec{r}) e^{i\vec{k} \cdot \vec{r}}, \quad (1.30)$$

where n is the quantum number, and the first term on the right arises from the crystal potential and is periodic, $u(\vec{r}) = u(\vec{r} + \vec{a})$, where $\vec{a}$ is the lattice vector. The quantity, $\hbar \vec{k}$ has the units of momentum. It is not the actual momentum of the electron, but it satisfies conservation laws that look like momentum

conservation. This quantity is known as the *crystal momentum*. A plot of energy vs. momentum for electrons in a crystal is really a plot of energy vs. $\vec{k}$ or crystal momentum $\hbar\vec{k}$. The $E(\vec{k})$ characteristic is known as the electronic band structure, or dispersion.

The band structure of a semiconductor is considerably more complicated than the simple parabolic band structure shown in Fig. 1.6 on the left. The full bandstructure is shown on the right in Fig. 1.6. Because the crystal is periodic in real space, the bandstructure is periodic in k -space. The volume of k -space that is repeated is known as the *Brillouin zone*. The figure on the right in Fig. 1.6 is a plot of $E(\vec{k})$ along some high symmetry lines in the Brillouin zone, and $E = 0$ is the location of the Fermi level for intrinsic Si. The top of the valence band is the first band below $E = 0$, which is seen to occur at the Γ point in the Brillouin zone of Si (we also see that there are two degenerate valence bands at the Γ point, the *light and heavy hole bands*). The bottom of the conduction band occurs along the X directions, near the edges of the Brillouin zone. There are eight of these locations in the Si Brillouin zone, and electrons have an equal probability to be in any of the eight locations. We say that the *valley degeneracy* is eight, $g_V = 8$, for the conduction band of Si.

Bandstructures like the one shown on the right of Fig. 1.6 are computed by solving the Schrödinger wave equation with the crystal potential of the material while applying periodic boundary conditions. Although the band structure is complex, the electrons and holes of interest reside in states near the bottom of the conduction band or top of the valence band. The parabolic assumption shown on the left of Fig. 1.6 is often adequate for electrons near these band minima or maxima.

Finally, we should mention how the energy band diagram, a plot of the band edges vs. position, and the band structure, a plot of electron energy vs. crystal momentum, are related. It is usually a good assumption that the potential vs. position changes slowly on the scale of the lattice potential. In this typical case, at any location in the device, $E(\vec{k})$ is the same as in the bulk, except that the conduction and valence band edges have been moved up or down in energy according to the value of the electrostatic potential at that location.

1.8 Quantum Confinement

Quantum mechanics tells us that an electron behaves both as a particle and as a wave and that the wave aspects become important where the potential energy changes spatially on the scale of the electron's wavelength (the so-called de Broglie wavelength, λ_B). We can estimate the average electron wavelength from

$$p = \hbar k = \hbar \frac{2\pi}{\lambda_B}, \quad (1.31)$$

where p is the crystal momentum, k the electron's wave vector, and λ_B the electron's wavelength. The energy of the electron is $E = p^2/2m^*$, and the thermal equilibrium average electron energy is $3k_B T/2$. Using these relations, we obtain a rough estimate of the thermal average de Broglie wavelength as

$$\langle \lambda_B \rangle \approx \frac{h}{\sqrt{3m^* k_B T}} \approx 6 \text{ nm}, \quad (1.32)$$

where we have assumed for a rough estimate that $m^* = m_0$. Electrostatic potential wells to confine electrons to dimensions less than 10 nm are readily produced with a gate voltage, and semiconductor layers less than 10 nm thick are also readily achieved. The behavior of electrons confined in these *quantum wells* is different from the behavior of electrons in the bulk, and it is important to understand the differences.

Figure 1.7 sketches two quantum wells; the one on the left is a rectangular quantum well with infinitely high barriers on the sides, and the one on the right is a triangular quantum well. The direction of confinement is the z -direction, but we assume that electrons are free to move in the x - y plane. Just as the Coulomb potential of the nucleus of a hydrogen atom confines the electron to the vicinity of the nucleus, which leads to the occurrence of discrete energy levels of the hydrogen atom, we find that the energies of electrons in these quantum wells consist of discrete *subbands* associated with confinement in the z -direction.

The time independent Schrödinger equation for electrons is

$$\left[-\frac{\hbar^2}{2m^*} \nabla^2 - E_c(x, y, z) \right] \psi(x, y, z) = E \psi(x, y, z). \quad (1.33)$$

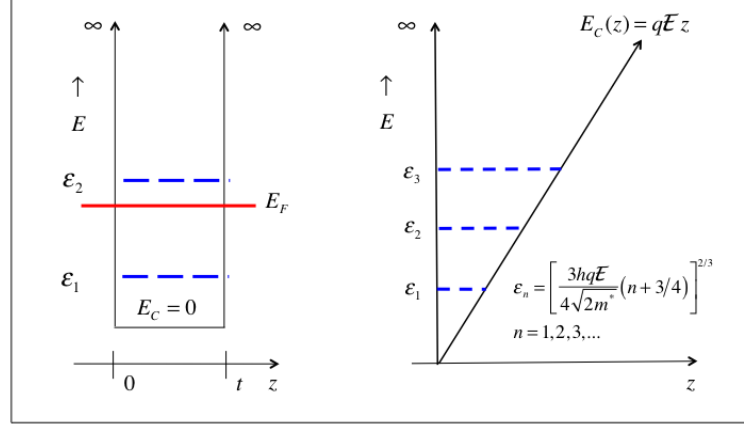


Figure 1.7: Illustration of simple quantum wells. The direction of confinement is the z -direction and electrons are free to move in the x - y plane. Left: Rectangular quantum well with infinitely high barriers. Right: Triangular quantum well with infinitely high barriers.

If E_c is a constant, then the solutions are plane waves,

$$\psi(x, y, z) = \frac{1}{\sqrt{\Omega}} e^{i\vec{k} \cdot \vec{r}}, \quad (1.34)$$

where Ω is an arbitrary normalization volume chosen so the the integral of $\psi\psi^*$ over the volume is one. The wvector, $\vec{k}$, is obtained from

$$\frac{\hbar^2 k^2}{2m^*} = (E - E_c). \quad (1.35)$$

The solution to the wave equation for a quantum well is the product of a plane wave in the x - y plane times a function in the z -direction that depends on the shape of the quantum well in the z -direction,

$$\psi(x, y, z) = \frac{1}{\sqrt{A}} e^{i(\vec{k}_{\parallel} \cdot \vec{\rho})} \times \phi(z), \quad (1.36)$$

where A is an area in x - y plane used to normalize the wavefunction in the x - y plane. To find $\phi(z)$, we solve

$$\left[-\frac{\hbar^2}{2m^*} \frac{d^2}{dz^2} - E_c(z) \right] \phi(z) = E\phi(z). \quad (1.37)$$

Consider the rectangular quantum well on the left of Fig. 1.7 and take $E_c = 0$. The solutions to eqn. (1.37) are $\sin(k_z z)$ and $\cos k_z z$, where

$$\frac{\hbar^2 k_z^2}{2m^*} = \epsilon = E - \frac{\hbar^2 k_{\parallel}^2}{2m^*}. \quad (1.38)$$

The boundary conditions are $\phi(0) = \phi(t) = 0$ because the assumed infinitely high barriers force the wave function to zero at the boundaries. Only $\sin(k_z z)$ satisfies the boundary condition at $z = 0$, and to satisfy the boundary condition at $z = t$, k_z must take on discrete values of

$$k_z t = n\pi, \quad (1.39)$$

where $n = 1, 2, \dots$. The result is that the energy in eqn. (1.38) becomes quantized; only the energies

$$\epsilon_n = \frac{\hbar^2 n^2 \pi^2}{2m^* t^2}, \quad (1.40)$$

are allowed. The total energy is

$$E = \frac{\hbar^2 (k_{||}^2 + k_n^2)}{2m^*} = \epsilon_n + \frac{\hbar^2 k_{||}^2}{2m^*}. \quad (1.41)$$

Quantum confinement produces a set of subbands in the conduction band (and a corresponding set in the valence band). For $n = 1$, the lowest energy is ϵ_1 , but there is an additional kinetic energy of $\hbar^2 k_{||}^2 / 2m^*$ associated with the electron's velocity in the x - y plane. Quantum confinement effectively raises the bottom of the conduction band. The number of subbands that are occupied depends on the location of the Fermi level. The subband energies are determined by the shape of the potential well and by the height of the barriers. For the triangular quantum well shown on the right of Fig. 1.7, we expect subbands, but the values of ϵ_n are different and the wavefunctions are Airy functions rather than sine functions. In general, however, light effective masses and thin quantum wells give high subband energies, as illustrated by Eq. (1.40) for the rectangular quantum well with infinite barriers.

In addition to changing the energies of electrons in the conduction and valence bands, quantum confinement also changes the spatial distribution of electrons. For the rectangular quantum well, $n(z) \propto \sin^2(k_n z)$. The contributions from all of the occupied subbands should be added to get the total electron density. The electrons in a quantum well are free to move in the x - y plane, but can move very little in the z -direction. They are *quasi-two-dimensional* electrons. The two-dimensional nature of the electrons changes the density of states. Instead of eqn. (1.2) for 3D (unconfined) electrons, we have for each subband

$$D_c^{2D} = g_V \frac{m_n^*}{\pi \hbar^2} \Theta(E - \epsilon_n). \quad (1.42)$$

Instead of eqn. (1.11) for the carrier density, we have

$$n_S^n = N_{2D} \mathcal{F}_0(\eta_F^n) m^{-2}, \quad (1.43)$$

for the sheet carrier density, where

$$N_{2D} \equiv g_V \frac{m^* k_B T}{\pi \hbar^2}, \quad (1.44)$$

is the effective density-of-states in 2D, and

$$\eta_F^n = (E_F - \epsilon_n) / k_B T, \quad (1.45)$$

$$\mathcal{F}_0(\eta_F) \equiv \frac{1}{\Gamma(1)} \int_0^\infty \frac{\eta^0 d\eta}{1 + e^{\eta - \eta_F}} = \ln(1 + e^{\eta_F}). \quad (1.46)$$

To get the total electron density, we add the contributions of all of the occupied subbands.

Some interesting effects occur when electrons in the conduction band of Si are confined in a quantum well. Figure 1.8 shows the constant energy surfaces for electrons in the conduction band of Si. The lowest energies occur at six different locations in the Brillouin zone along the three coordinate axes (the valley degeneracy is $g_V = 6$). The constant energy surfaces are ellipsoids of revolution described by

$$E = \frac{\hbar^2 k_x^2}{2m_{xx}^*} + \frac{\hbar^2 k_y^2}{2m_{yy}^*} + \frac{\hbar^2 k_z^2}{2m_{zz}^*}. \quad (1.47)$$

There are two different effective masses, a heavy *longitudinal effective mass*, m_l^* and a light, *transverse effective mass*, m_t^* . For Si, $m_l^* = 0.90m_0$ and $m_t^* = 0.19m_0$. For example, for the ellipsoids oriented along the x -axis, $m_{xx}^* = m_l^*$ and $m_{yy}^* = m_{zz}^* = m_t^*$.

According to eqn. (1.40), the subband energies are determined by the effective mass, but which effective mass should we use? The answer is to use the effective mass in the direction of confinement, the z -direction in this case. From Fig. 1.8, we see that two of the six ellipsoids have the heavy, longitudinal effective mass in the z -direction and four of the six have the light, transverse effective mass in the z -direction. The result is two different series of subbands – an *unprimed ladder* of subbands with the energies determined by m_l^* and a degenerate of $g_V = 2$, and a *primed ladder* of subbands with the energies determined by m_t^* and a valley degeneracy of 4. The lowest subband is the $n = 1$ unprimed subband. In the x - y plane, electrons in these two degenerate subbands respond with the light, transverse effective mass.

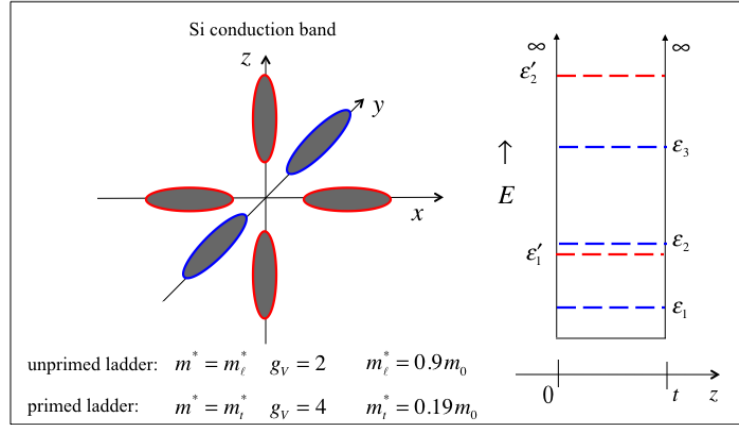


Figure 1.8: Left: A sketch of the constant energy surfaces for electrons in silicon. Right: the corresponding unprimed and primed "ladders" of subbands.

1.9 Summary

Our goal in this lecture has been to remind you of a few basic concepts in elementary semiconductor physics. The key points are:

1. The discrete energy levels for isolated atoms become bands of energies in a solid. Electrons in the valence and conduction bands of a semiconductor are "delocalized" and free to move within the crystal.
2. The states in a band are described by a density-of-states; the number of states per unity volume in a small energy range is $D(E) dE$.
3. In general, the density-of-states is a complicated function of energy that must be determined numerically, but the regions of interest are energies near the top of the valence band and near the bottom of the conduction band. For these energies, simple expressions for $D(E)$ can be used.
4. In equilibrium, the probability that a state is occupied is given by the Fermi function, The parameter in the Fermi function is the Fermi level, or electrochemical potential.
5. The higher the Fermi level, the higher the probability that states in the conduction band are occupied with mobile electrons and the lower the Fermi level, the higher the probability that states in the valence band contain empty states or mobile holes.
6. There is a direct relation between the location of the Fermi energy and the concentration of electrons in the conduction band and holes in the valence band.
7. The concentration of electrons and holes in an intrinsic semiconductor is n_i , a material parameter. The electron and hole concentrations can be changed by doping the semiconductor.
8. The concentration of electrons and holes can be altered by doping, but the product, $n_0 p_0 = n_i^2$, does not change in equilibrium (for a non-degenerate semiconductor).
9. An energy band diagram is a plot of the conduction and valence band edges vs. position.
10. The semiconductor bandstructure relates the electron energy in a crystal to its wave vector, $\vec{k}$ or crystal momentum, $\hbar\vec{k}$. In the simplest case of parabolic energy bands, energy and crystal momentum are related by the effective mass.
11. Quantum confinement introduces subbands in the conduction and valence bands and produces quasi-two-dimensional electrons with a dispersion, $E(\vec{k})$, and a density-of-states that differs from those of unconfined, three-dimensional electrons in the bulk.

1.10 References

For an introduction to semiconductor theory as used for analyzing electronic devices, see:

- [1] Robert F. Pierret *Semiconductor Device Fundamentals*, 2nd Ed., , Addison-Wesley Publishing Co, 1996.
- [2] Robert F. Pierret *Advanced Semiconductor Fundamentals*, 2nd Ed., Vol. VI, Modular Series on Solid-State Devices, Prentice Hall, Upper Saddle River, N.J., USA, 2003.

Lecture 4 of the following online course discusses the density-of-states.

- [3] Mark Lundstrom, “ECE 656: Electronic Transport in Semiconductors,” Purdue University, Fall 2013, [//https://www.nanohub.org/groups/ece656_f13](https://www.nanohub.org/groups/ece656_f13).

For a quick summary of the essentials of Fermi-Dirac integrals, see:

- [4] R. Kim and M.S. Lundstrom, “Notes on Fermi-Dirac Integrals,” 3rd Ed., <https://www.nanohub.org/resources/5475>.

Pierret discusses energy band diagrams. Their importance is emphasized by Herbert Kroemer in the lecture he gave after accepting his Nobel prize.

- [5] Herbert Kroemer, “Quasi-electric fields and band offsets: Teaching electrons new tricks,” Nobel Lecture, Dec. 8, 2000.

Chapter 2

Transport

2.1	Introduction
2.2	Thermal velocities
2.3	Near-equilibrium transport
2.4	High-field transport
2.5	Non-local transport
2.6	Ballistic and quasi-ballistic transport
2.7	Quantum transport
2.8	The semiconductor equations
2.9	Summary
2.10	References

2.1 Introduction

The operation of a semiconductor device like a transistor is determined by: 1) electrostatics, which determines how the energy bands are controlled by the terminal voltages, and 2) transport, which determines how quickly electrons and holes can flow through the device. Understanding electrostatics is non-trivial, but straight-forward. The basic considerations have not changed for several decades. Carrier transport is more challenging.

The traditional approach to carrier transport assumes that carriers are nearly in equilibrium with the semiconductor lattice and that they frequently scatter off of defects, lattice vibrations, etc. Semiconductor device theory usually begins with the *drift-diffusion* equation, which describes carrier transport as the sum of a component corresponding to carriers drifting in an electric field and a component due to carriers diffusing down a concentration gradient. The key parameter in this equation is the *mobility*, a material-dependent parameter. Extensions to high electric fields can be done, as long as the electric fields vary slowly with position. Under these conditions, the drift-diffusion equation can describe high-field velocity saturation by using a carrier mobility that depends on the local electric field.

As device dimensions have shrunk over the past several decades, the approximations that made the drift-diffusion equation applicable to devices began to lose validity for many devices. In short devices, *non-local transport* occurs, by non-local we mean that transport does not depend simply on the local value of the electric field and concentration gradient. The velocity does not saturate in high electric field regions if they are short enough; instead, the velocity overshoots the bulk saturation velocity. Effects like this are difficult to describe simply, and sophisticated simulation techniques have been developed to treat them. A simple, but very physical approach, known as the *Landauer approach* is often used to treat the ballistic and quasi-ballistic transport that occurs in modern devices.

The approaches discussed above are semiclassical, by which we mean that the charge carriers are treated as semi-classical particles with the quantum mechanics embedded into the effective masses that describes their dynamics in the crystal. As devices continue to scale it is becoming increasingly important to treat

charge carriers as quantum mechanical particles, so that effects such as tunneling and quantum mechanical reflections can be included.

In this lecture, we will briefly review the drift-diffusion approach to carrier transport, because it is the way most electrical engineers are introduced to semiconductor devices. We will briefly introduce the Landauer approach to transport, which is a useful and very physical way to describe small devices. Finally, non-local, ballistic, and quantum transport will be very briefly described.

2.2 Thermal velocities

Consider a two-dimensional sheet of carriers (this might represent the channel of a MOSFET, in which carriers flow in a thin layer near the surface). As show in Fig. 2.1, electrons scatter randomly from lattice vibration (phonons), impurities, etc. Since equilibrium is assumed, the average velocity in any direction is zero, $\langle v_x \rangle = \langle v_y \rangle = 0$. The electrons are in thermodynamic equilibrium with the vibrating lattice. Electrons and phonons exchange energy through electron-phonon collisions, and thermodynamic equilibrium is achieved when the average energies of electrons and the lattice are equal. In 2D, the average kinetic energy per electron is $\langle KE \rangle = k_B T$, because statistical mechanics tells us that there is $k_B T/2$ energy per degree of freedom, and we are assuming two-dimensional electrons. (We are also assuming a dilute, non-degenerate gas of electrons that satisfies Maxwell-Boltzmann statistics.) Equating the thermal energy to the electron's average kinetic energy, we find

$$\frac{1}{2} m_n^* \langle v^2 \rangle = k_B T, \quad (2.1)$$

from which we can find the *rms thermal velocity* as

$$\sqrt{\langle v^2 \rangle} = v_{\text{rms}} = \sqrt{\frac{2k_B T}{m_n^*}}. \quad (2.2)$$

A typical value for the rms thermal velocity in a semiconductor is $v_{\text{rms}} \approx 10^7$ cm/s, which is a relatively high value. The point is that at thermal equilibrium, electrons are zipping about at high velocity in random thermal motion.

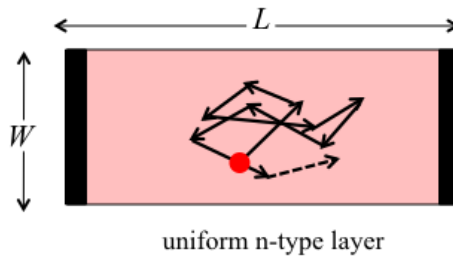


Figure 2.1: Sketch of electrons in a two-dimensional n-type semiconductor in equilibrium. Electrons exchange thermal energy with the lattice until the temperature of the electrons equals the temperature of the lattice. In thermal equilibrium electrons are in random thermal motion, moving at high velocity with an average velocity that is zero in any direction.

Average thermal velocity can be defined in several different ways. We have seen one way, the rms thermal velocity, but this thermal velocity is not the most useful one for analyzing nanoscale devices. Figure 2.2 illustrates how we define the *unidirectional thermal velocity*. Consider again a non-degenerate semiconductor.

Since the Fermi level is located well below the bottom of the conduction band, $E \gg E_F$ for all energies of interest, and the Fermi function simplifies to

$$f_0 = e^{(E_F - E)/k_B T}. \quad (2.3)$$

If we assume parabolic energy bands, then $E = E_c + m_n^* v^2/2$, and we can write

$$f_0 = e^{(E_F - E_c)/k_B T} e^{-m_n^* v^2/2k_B T} = e^{(E_F - E_c)/k_B T} e^{-m_n^* (v_x^2 + v_y^2)/2k_B T}, \quad (2.4)$$

which is just the expected Maxwellian distribution of velocities for a dilute gas. Figure 2.2 is a plot of this Maxwellian velocity distribution along the x-axis. Because the distribution is symmetrical, $\langle v_x \rangle = 0$, but the average velocity of electrons with a positive v_x is finite. The value of this velocity, $\langle v_x^+ \rangle$, is known as the *unidirectional thermal velocity*, v_T , and is given by

$$\langle v_x^+ \rangle = v_T = \sqrt{\frac{2k_B T}{\pi m_n^*}}. \quad (2.5)$$

The unidirectional thermal velocity has a numerical value of $\approx 10^7$ cm/s for electrons in non-degenerate Si. For a degenerate semiconductor, an analogous quantity, $\tilde{v}_T \geq v_T$, can be defined. Under non-degenerate conditions, v_T does not depend on the location of the Fermi level, but for a degenerate semiconductor, we must first determine the location of E_F to determine $\tilde{v}_T$.

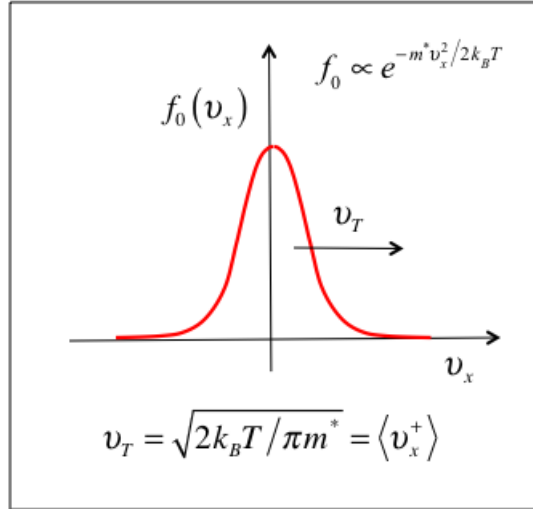


Figure 2.2: Sketch of an equilibrium Maxwellian distribution of electron velocities in a non-degenerate semiconductor illustrating what is meant by the unidirectional thermal velocity.

2.3 Near-equilibrium transport

Figure 2.3 illustrates what we mean by the term, *diffusive transport*. In this case, a small voltage has been applied across the semiconductor sketched in Fig. 2.1. Electrons are still in random thermal motion, but now there is a slightly higher chance for electrons that enter from the left contact to exit on the right contact, as opposed to returning to the left contact. The average distance between scattering events is known as the *mean-free-path*, and we assume that the resistor is much longer than a mean-free-path. We wish to compute the current through this resistor and will begin with the definition of current,

$$I = Q/t_t, \quad (2.6)$$

where Q is the total charge in the device and t_t is the time it takes to move this charge out of the device, the so-called *transit time*. We can write the charge in the device in terms of n_s , the sheet carrier density per m^2 and Q_n , the charge per m^2 as

$$Q = -qn_sWL = Q_nWL. \quad (2.7)$$

The transit time can be written in terms of the length of the resistor and the average *drift velocity*, v_d , as

$$t_t = \frac{L}{v_d}. \quad (2.8)$$

Finally, by using eqns. (2.7) and (2.8) in (2.6), we find the current as

$$I = WQ_nv_d, \quad (2.9)$$

which is intuitively sensible. The current is proportional to the width of the resistor, because increasing the width is like putting more resistors in parallel, to the amount of charge in the device (per m^2) and to how fast that charge is moving, v_d .

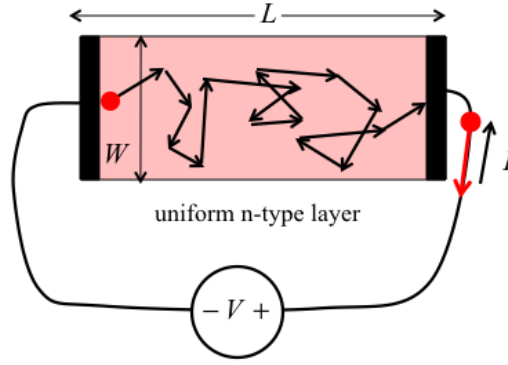


Figure 2.3: Illustration of diffusive transport in a semiconductor with a small voltage applied.

We seek to relate the current in the resistor to the applied voltage. The approach we will follow is the classical approach, which applies when the resistor is many mean-free-paths long [1-3]. We should not expect it to apply to modern transistors for which the channel lengths are very short.

Since the resistor is assumed to be uniform, we can relate the current to the electric field in the device, $\mathcal{E} = V/L$. Classical mechanics tells us that the momentum of a particle increases when an external force is applied according to

$$\frac{dp}{dt} = F_e = -q\mathcal{E}. \quad (2.10)$$

If the net momentum is destroyed by scattering in a time, τ , then the increase in momentum over this time is

$$\Delta p = -q\mathcal{E}\tau = m_n^* \Delta v. \quad (2.11)$$

Solving for the increase in velocity, we find

$$\Delta v = -\frac{q\tau}{m_n^*} \mathcal{E}. \quad (2.12)$$

Averaging over all scattering times and assuming that the average velocity at the start of the acceleration in the electric field is zero, we find

$$v_d = -\frac{q \langle \tau \rangle}{m_n^*} \mathcal{E} = -\mu_n \mathcal{E}, \quad (2.13)$$

where v_d is the drift velocity and μ_n is the electron mobility in $\text{m}^2/\text{V} \cdot \text{s}$.

This simple derivation indicates that we expect the average velocity of electrons in eqn. (2.9) to be proportional to the electric field with the constant of proportionality being the *mobility*.

$$\mu_n = \frac{q \langle \tau \rangle}{m_n^*} \quad \text{m}^2/\text{V} \cdot \text{s}. \quad (2.14)$$

Equations (2.13) and (2.14) are known as the *Drude model* for electrical conduction, and its use in eqn. (2.13) gives a good description of carrier transport as long as the electric field is not too large (and as long as the electron density, n_s is uniform).

Figure 2.4 illustrates a different case, for which the electric field is zero, but n_s varies with position. In this case, we inject a flux of electrons on the left side and collect them on the right. Electrons diffuse from the left to the right, and the current density is given by

$$J_n = I/W = (-q) \left[-D_n \frac{dn_s}{dx} \right] = qD_n \frac{dn_s}{dx}, \quad (2.15)$$

where D_n is the *diffusion coefficient* in m^2/s .

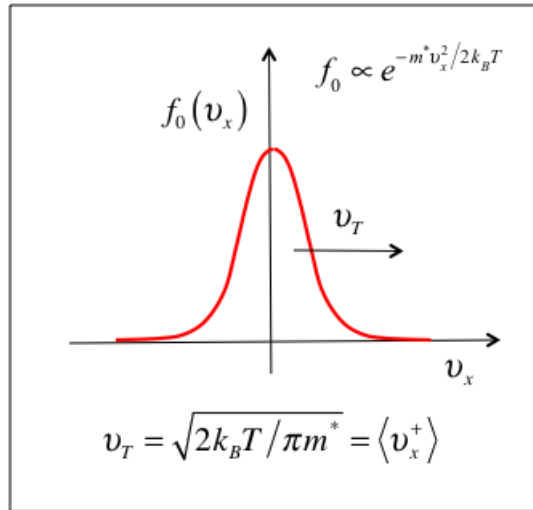


Figure 2.4: Illustration of electron diffusion in a semiconductor that is many mean-free-paths long. Electrons are injected at the left and collected at the right. The result is a diffusive flux of electrons from the left to the right.

Diffusive transport occurs in the absence of an electric field or in the presence of a small electric field. Since these are independent processes near equilibrium, we can add them to find the drift-diffusion equation,

$$\boxed{J_n = I/W = n_s q \mu_n \mathcal{E} + q D_n \frac{dn_s}{dx} \quad \text{A/cm}.$$

$$D_n / \mu_n = k_B T / q$$
(2.16)

The relation between D_n and μ_n in eqn. (2.16) is known as the *Einstein relation*. The drift-diffusion equation is the cornerstone of semiconductor traditional device theory, but it must be used with great caution when applying it to nanoscale devices whose dimensions are typically only a few mean-free-paths long.

We conclude this section by writing the transport equation in a different way. Recall from Lecture 1 that we can relate the equilibrium carrier density to the Fermi level. In this case we find,

$$n_{S0} = N_{2D} e^{(E_F - E_c)/k_B T}, \quad (2.17)$$

where N_{2D} is the 2D effective density of states and the superscript, “0”, reminds us that we are in equilibrium. If we are out of equilibrium, but still near equilibrium, then we can write a similar expression

$$n_s = N_{2D} e^{(F_n - E_c)/k_B T}, \quad (2.18)$$

where F_n is the electrochemical potential for electrons. In equilibrium, thermodynamics tells us that the Fermi level (the electrochemical potential in equilibrium) is constant, but out of equilibrium, the electrochemical potential may vary with position. If we use eqn. (2.18) in the drift-diffusion equation, (2.16), we find a simple expression for the current density

$$J_n = n_s \mu_n \frac{dF_n}{dx}. \quad (2.19)$$

Equation (2.19) is an important result. It states that gradients in the electrochemical potential drive current flow. It looks as though we have derived this result from the drift-diffusion equation, but it is really much more general and can be derived from non-equilibrium thermodynamics [4]. In fact, the drift-diffusion equation, (2.16), should be derived from (2.19). The approach used here simply shows that eqn. (2.16) is consistent with a more fundamental description of transport.

Relation between MFP, diffusion coefficient, and mobility

Near equilibrium there is a simple relation between the average mean-free-path, the diffusion coefficient, and the mobility. Since the mobility is frequently known, these relations can be used to estimate the mean-free-path from the measured mobility. The relation between mean-free-path and diffusion coefficient is the most direct. For a non-degenerate semiconductor, one can show, that [6]

$$D_n = \frac{v_T \lambda_0}{2}, \quad (2.20)$$

where λ_0 is a specially defined, average *mean-free-path for backscattering* [6].

Since the mobility and diffusion coefficient are related by the Einstein relation, we can write

$$\mu_n = \frac{D_n}{k_B T/q} = \frac{v_T \lambda_0}{2 (k_B T/q)}. \quad (2.21)$$

Equation (2.21) gives the mobility in terms of the mean-free-path. In practice, it is usually the mobility that is known, and the mean-free-path is deduced by solving eqn. (2.21). Reference [6] discussed the relation between mobility and mean-free-path when the semiconductor is not non-degenerate.

Exercise: Estimating the mean-free-path from the mobility

To determine whether ballistic transport is possible, we should compare the mean-free-path between scattering events to the channel length (or to the critical part of the channel that controls the current). Assume a near-equilibrium mobility of $250 \text{ cm}^2/\text{V} \cdot \text{s}$ (typical for a Si MOSFET) and estimate the mean-free-path.

Solving eqn. (2.21) for the mean-free-path, we find:

$$\lambda_0 = \frac{2 (k_B T/q) \mu_n}{v_T}. \quad (2.22)$$

The unidirectional thermal velocity is defined in eqn. (2.5). For electrons in a Si MOSFETs under the on state, we should use $m_n^* = m_t^* = 0.19m_0$. By inserting numbers in eqn. (2.22), we find a mean-free-path for backscattering of about 13 nm - close to the length of the channel in current state-of-the-art MOSFETs. We conclude that operation near the ballistic limit is a real possibility for nanoscale transistors.

2.4 High-field transport

Figure 2.5 is a sketch of the average electron velocity vs. electric field for electrons in bulk silicon. By “bulk”, we mean many mean-free-paths long. For small fields (less than about 10^4 V/cm, the velocity is proportional to the electric field, as expected from eqn. (2.13). As the electric field increases well above 10^3 V/cm, the drift velocity begins to increase sub linearly, and when the electric field reaches about 10^4 V/cm, the drift velocity saturates at $v_{\text{sat}} \approx 10^7$ cm/s in Si. Electron transport under these conditions is known as *high-field* or *hot carrier* transport [7].

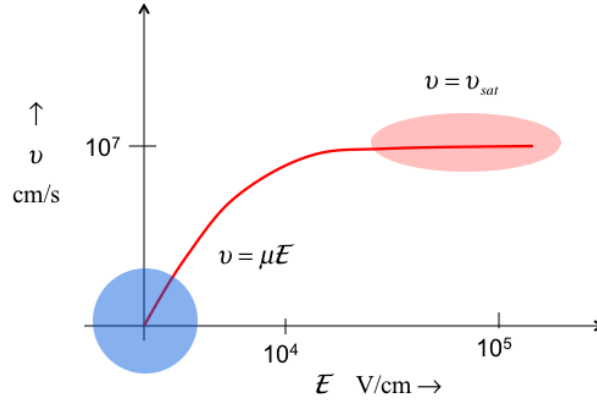


Figure 2.5: Typical velocity vs. electric field characteristic for electrons in bulk silicon at room temperature.

It can be difficult to compute the velocity vs. electric field characteristic from very low to high electric fields, but the general features are readily understood. For low fields, the drift velocity is proportional to the electric field, as discussed in Sec. 2.3. As the electric field increases, the average electron energy increases above the equilibrium value of $k_B T_L$, where T_L is the temperature of the lattice. As the energy increases, the scattering rate generally increases, because the density-of-states increases with energy, so there are more final states for the electrons to scatter to. The increased scattering means that the average time between scattering events, $\langle \tau \rangle$, decreases. According to eqn. (2.14), this means that the mobility decreases for large electric fields. The decreasing mobility with increasing electric field produces a sublinear increase in velocity. At very high fields, the decrease in mobility becomes linear with electric field in bulk silicon, and the velocity saturates. In other semiconductors, the velocity vs. electric field characteristics can be more complicated, but the high field velocity is always considerably less than what would be estimated from the near-equilibrium mobility.

To determine whether modern transistors operate in the low or high electric field regions, we can do a simple estimate. State of the art MOSFETs have channel lengths of about 20 nm and operate at a drain to source voltage of about 1 V. The electric field is nonlinear within the channel, but we can assume that it is linear to estimate the average electric field in the channel. Accordingly, we find,

$$\mathcal{E} \approx \frac{1}{20} \frac{\text{V}}{\text{nm}} = 500,000 \text{ V/cm} . \quad (2.23)$$

The electric fields in modern transistors are very high, so it would seem reasonable to assume that the velocity saturates in the channel of modern MOSFETs when operated at high V_{DS} (but see the discussion in Sec. 2.5).

One common approach to the modeling of electron transport under high electric fields is to use eqn. (2.16) but replace the near-equilibrium mobility with an electric field dependent mobility,

$$J = n_s q \mu_n(\mathcal{E}) \mathcal{E} + k_B T \mu_n(\mathcal{E}) \frac{dn_s}{dx} \quad \text{A/cm} . \quad (2.24)$$

By specifying the field-dependent mobility such that the measured velocity vs. field characteristic is reproduced, this approach can treat high high electric fields as well as low ones. But this approach assumes that the carrier velocity at a given location depends only on the electric field at that location, which can be justified when the electric field varies slowly in space. This assumption breaks down in modern transistors, which have high electric fields that vary rapidly with position.

2.5 Non-local transport

Because the electric field in the channel is so high, we might expect the velocity to saturate in the channel of a Si MOSFET. Figure 2.6 shows the results of a numerical simulation that accurately captures the physics of high-field transport with electric fields that vary rapidly with position. This simulation is of electron transport in a 30 nm channel length MOSFET [8]. Each dot in the image at the left corresponds to an electron, which is tracked through the device as it is accelerated by the electric field and as its energy and momentum are affected by random scattering processes. Surprisingly, the average velocity vs. position plotted on the right shows no sign of velocity saturation at all. The average velocity is well over $v_{sat} \approx 10^7$ cm/s over the entire channel. The reason is that in short devices, there is not enough time for scattering to reduce the velocity to v_{sat} – electrons scatter only a few times before they leave the device. In bulk Si with a high electric field, electrons scatter often, and their velocity saturates.

Understanding how this *velocity overshoot* or *non-local transport* affects the IV characteristics of sub-micron channel length MOSFETs was the subject of a great deal of research in the 1980's and 90's. The physics are readily understood [7], but sophisticated simulations are required to calculate a specific velocity vs. electric field characteristic. The average velocity at a specific position is not simply related to the electric field at that location; a drift-diffusion equation with a local field-dependent mobility cannot describe velocity overshoot. Simulations that capture these effects are indispensable in designing transistors.

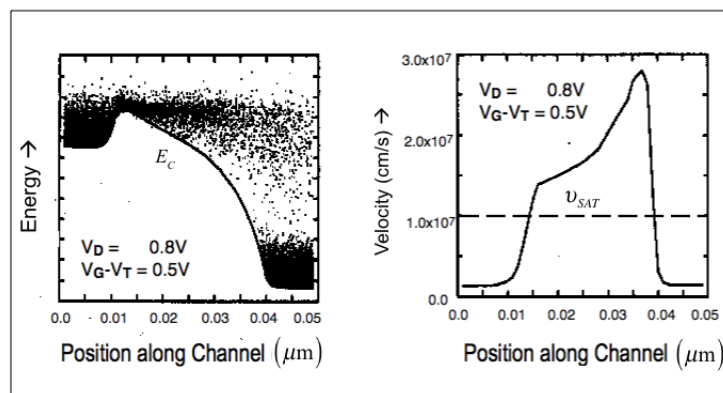


Figure 2.6: Results of a Monte Carlo simulation of electron transport in a 30 nm channel length Si MOSFET under high gate and drain bias. Left: A “snapshot” in time showing the location and energy of electrons in the device. Right: Average electron velocity vs. position. (From: D. Frank, S. Laux, and M. Fischetti, Int. Electron Dev. Mtg., Dec., 1992.)

2.6 Ballistic and quasi-ballistic transport

We have discussed near-equilibrium transport in which the average electron velocity is proportional to the electric field, high-field transport in a long semiconductor, for which the velocity saturates, and non-local transport in short, high-field regions, where the velocity can be much higher than the bulk saturation velocity. In each of these cases, the resulting average velocity is an interplay between the effect of the electric field and the scattering processes, but channel lengths have become so short that we should also consider the

possibility of *ballistic transport*, in which electrons travel from the source to the drain without scattering at all.

To determine whether ballistic transport is possible, we should compare the *mean-free-path* (the average distance between scattering events) to the device length (or to the critical part of the device that controls the current). As discussed earlier, the near-equilibrium mean-free-path for backscattering is close to the length of the channel in current state-of-the-art MOSFETs. The real situation is, however, more complicated; near the drain, electrons gain a lot of kinetic energy and scatter more, so their mean-free-path is reduced. It turns out that well-designed MOSFETs have a short bottleneck region that controls the current, and the length of this bottleneck is only a small fraction of the channel length. As a result, drain currents near the ballistic limit are expected in nanoscale transistors.

During the 1980's and 1990's, our scientific understanding of ballistic electron transport in nano structures advanced considerably. For large structures, we write the conductance of a two-dimensional resistor (like the channel of a MOSFET) as

$$G = \sigma_s W/L = n_s q \mu_n W/L, \quad (2.25)$$

where σ_s is the sheet conductance, W is the width of the conductor, and L is its length. As the length approaches zero, the conductance should become infinite, but it is found to approach the finite value

$$G = \frac{2q^2}{h} M \mathcal{T}, \quad (2.26)$$

where q is the magnitude of the charge on an electron, h is Planck's constant, M is the number of conducting "channels", and $\mathcal{T}$ is the transmission, a number between zero and one. (If you think of the conductor as a highway carrying electrons, the number of lanes is analogous to the number of channels, and the probability of getting from the beginning to the end of your trip is the transmission.)

Figure 2.7 shows the result of a remarkable experiment. The conductance of a 2D conductor was measured as the width of the conductor was varied (electrically by a reverse biased Schottky barrier). The measured conductance was observed to increase in steps of $2q^2/h$ (the *quantum of conductance*) as the width of the conductor increased and the number of channels, M , increased in steps. These measurements were done at a temperature of 4 K to minimize scattering and assure ballistic transport ($\mathcal{T} = 1$). Device dimensions have scaled to such short dimensions that ballistic transport and quantized conduction now need to be considered in modern transistors operating at room temperature. The first two volumes in this lecture notes series discuss the theory [5] and application [6] of the Landauer approach, which we will use in these lectures to describe transport in nanoscale channels.

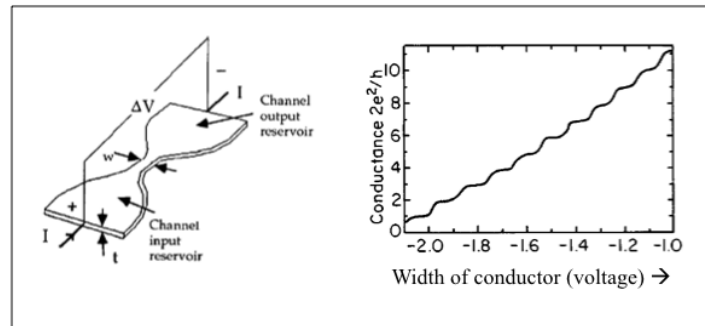


Figure 2.7: Experiments of van Wees, et al. experimentally demonstrating that conductance is quantized. Left: sketch of the device structure. Right: measured conductance. (Data from: B. J. van Wees, et al., *Phys. Rev. Lett.* **60**, 848851, 1988. Figures from D. F. Holcomb, "Quantum electrical transport in samples of limited dimensions", *Am. J. Phys.*, **67**, pp. 278-297, 1999. Reprinted with permission from *Am. J. Phys.* Copyright 1999, American Association of Physics Teachers.)

2.7 Quantum transport

The transport effects discussed so far are semiclassical – they consider electrons to be particles with the quantum mechanics being embedded in the band structure or effective mass, but as devices continue to shrink, it is becoming important to consider explicitly the quantum mechanical nature of electrons. We expect that quantum mechanical effects will become important when the potential energy changes rapidly on the scale of the electron's de Broglie wavelength. As discussed in Sec. 1.8, this happens under high gate voltage in the direction normal to the channel and for all gate voltages in structures that have a thin Si channel. The resulting quantum confinement alters the energy levels and changes the density-of-states.

Electron transport can also be affected by quantum mechanics. For example, in very short channel devices, the potential along the direction of current flow may vary rapidly, and *quantum transport* becomes important. A simple estimate of the de Broglie wavelength of thermal equilibrium electrons in Si (recall eqn. (1.34) gives about 10 nm, which is not much less than the present day channel lengths of transistors. Powerful techniques to treat the quantum mechanical transport of electrons in transistors have been developed [9]. Figure 2.8 is an example. As MOSFET channel lengths shrink below 10 nm, it is becoming increasingly necessary to describe electron transport quantum mechanically, but for channel lengths above about 10 nm, the semiclassical picture generally works well.

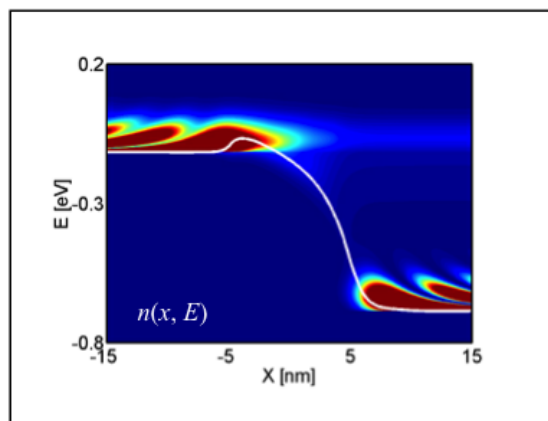


Figure 2.8: A quantum mechanical simulation of electron transport in a 10 nm channel length Si MOSFET. The white line is the electron potential energy (the bottom of the conduction band) under high gate and drain bias. The plot illustrates the energy resolved electron density. Quantum mechanical reflection from the potential energy barrier is observed as is quantum mechanical tunneling of electrons under the energy barrier. This figure is the result of a simulation by the nanoMOS program [9].

2.8 The semiconductor equations

We conclude by very briefly summarizing the process for analyzing or simulating a semiconductor device. As shown in Fig. 2.9, there are two parts to the procedure. To solve for the current, we solve a transport problem that describes how the charge flows in the device. In the simplest case, this might involve a drift-diffusion equation like eqn. (2.16) for the mobile charge density. But the drift-diffusion equation depends on the electrostatic potential, which depends on the fixed (e.g. dopant) charge and the mobile charge itself, which we are trying to compute. If we knew the mobile charge throughout the device, then we would solve the Poisson equation for the electric field (or electrostatic potential) and then use that electric field in the transport equation. Because these two processes are coupled, an iterative solution is needed.

Figure 2.9 summarizes the procedure. If we begin with the transport problem, then we need to guess the electrostatic potential throughout the device. Knowing the electrostatic potential, we can solve the transport problem to find the spatial distribution of mobile charge in the device, $n(\vec{r})$ and $p(\vec{r})$. Knowing the spatial

distribution of charge, we can then solve the Poisson equation for the electrostatic potential throughout the device, $V(\vec{r})$. It is not likely that the computed $V(\vec{r})$ will agree with the value guessed to start the process, so the transport problem has to be solved again for updated values of $n(\vec{r})$ and $p(\vec{r})$, which can then be used in the Poisson equation to update $V(\vec{r})$ again. The process continues until $n(\vec{r})$, $p(\vec{r})$ and $V(\vec{r})$ converge to within some tolerance.

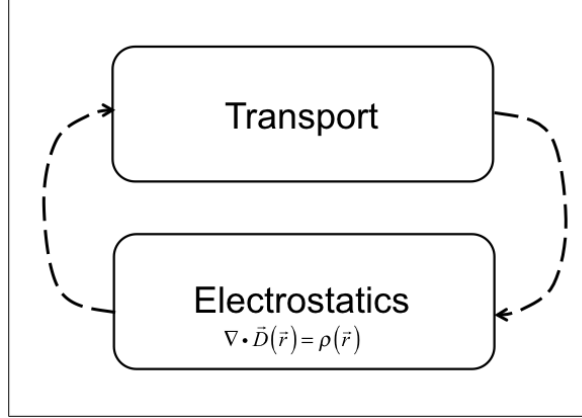


Figure 2.9: A simple representation of the process used to analyze or simulate a semiconductor device. One part is the solution of a transport problem to determine how mobile charge flows in the device. The second part is the solution of the Poisson equation to determine the electrostatic potential that results from a distribution of mobile charge. The two parts are coupled, so an iterative process is necessary.

For the electrostatics part of the process, the Poisson equation is almost universally used. For some very small structures, many body corrections are sometimes used, but the Poisson equation is generally adequate for devices. For the transport part of the process, there are several options with drift-diffusion based approaches being the simplest (and still most widely-used) and dissipative quantum transport approaches being the most sophisticated (and seeing increased usage). To summarize the self-consistent process, we will adopt the drift-diffusion approach.

For the transport part of the problem, we need two bookkeeping equations, one for electrons and one for holes. For electrons, the bookkeeping equation is the continuity equation for electrons, which is written as

$$\frac{\partial n(\vec{r})}{\partial t} = -\nabla \cdot \left(\frac{\vec{J}_n(\vec{r})}{-q} \right) + G_n(\vec{r}) - R_n(\vec{r}). \quad (2.27)$$

In words, this equation is interpreted as follows. The time rate of change of the electron concentration at the location, $\vec{r}$, is the sum of three components. The first is minus the divergence of the electron flux. (The divergence represents the net outflow of a quantity, so minus the divergence represents the net flow of electrons in to the position, $\vec{r}$.) The term, $G_n(\vec{r})$, describes the generation of electrons by, for example, optical excitation or impact ionization. Finally, the terms, $R_n(\vec{r})$, describes the recombination of electrons with holes through radiative, defect-assisted, or other processes. Equation (2.27) is simply a bookkeeping equation that tells us how to account for the increase or decrease of the electron density. A similar equation for holes can be written as

$$\frac{\partial p(\vec{r})}{\partial t} = -\nabla \cdot \left(\frac{\vec{J}_p(\vec{r})}{q} \right) + G_p(\vec{r}) - R_p(\vec{r}). \quad (2.28)$$

To actually solve these equations, constitutive relations are needed. For example, for the electron current densities, we could use eqn. (2.16) and a similar drift-diffusion equation for holes. Expressions for the generation and recombination rates in terms of the electron and hole densities and the electrostatic potential would also need to be specified.

For the second part of the process, we solve the Poisson equation:

$$\nabla \cdot \vec{D} = -\nabla \cdot (\epsilon \nabla V(\vec{r})) = \rho(\vec{r}), \quad (2.29)$$

where the space charge density is

$$\rho(\vec{r}) = q [p(\vec{r}) - n(\vec{r}) + N_D^+(\vec{r}) - N_A^-(\vec{r})] , \quad (2.30)$$

where N_D^+ and N_A^- are the ionized dopant densities.

The combined set of equations, the two continuity equations, the Poisson equations, and the constitutive equations for the electron and hole currents and the space charge density,

$$\boxed{\begin{aligned} \frac{\partial n(\vec{r})}{\partial t} &= -\nabla \cdot \left(\frac{\vec{J}_n(\vec{r})}{-q} \right) + G_n(\vec{r}) - R_n(\vec{r}) \\ \frac{\partial p(\vec{r})}{\partial t} &= -\nabla \cdot \left(\frac{\vec{J}_p(\vec{r})}{q} \right) + G_p(\vec{r}) - R_p(\vec{r}) \\ \nabla \cdot \vec{D} &= -\nabla \cdot (\epsilon \nabla V(\vec{r})) = \rho(\vec{r}) \\ J_n &= nq\mu_n \mathcal{E} + qD_n \frac{dn}{dx} \quad \text{A/cm}^2 \\ J_p &= pq\mu_n \mathcal{E} - qD_p \frac{dp}{dx} \quad \text{A/cm}^2 \\ \rho(\vec{r}) &= q [p(\vec{r}) - n(\vec{r}) + N_D^+(\vec{r}) - N_A^-(\vec{r})] \quad \text{C/cm}^3 \end{aligned}} , \quad (2.31)$$

are often referred to as “the semiconductor equations.” When supplemented by appropriate constitutive relations for the generation and recombination processes, these equations are three coupled, nonlinear partial differential equations in three unknowns, $n(\vec{r})$, $p(\vec{r})$ and $V(\vec{r})$. Computer programs that solve these equations are widely used. When we analyze a semiconductor device analytically, we generally simplify the equations, but are essentially following the same process.

Finally, we should note that the semiconductor equations are not a fundamental description of semiconductor devices in the same way that Maxwell’s Equations fundamentally describe electromagnetic fields. There are several simplifying assumptions in eqns. (2.31), especially in the transport equations. Two major themes of research in the past few decades have been on the development of efficient numerical algorithms to solve equations like (2.31) and improved descriptions of the underlying physics in these equations.

2.9 Summary

Our goal in this lecture has been to remind you of the traditional approach to carrier transport in semiconductor devices and to quickly review some topics in advanced transport theory. The key points to remember are:

1. The cornerstone of traditional semiconductor device theory is the drift–diffusion equation: $J = n_s q \mu_n \mathcal{E} + q D_n dn_s / dx$ A/cm.
2. At more fundamental level, current is related to the gradient of the electrochemical potential according to $J = n_s \mu_n dF_n / dx$.
3. The above two equations are valid near-equilibrium. For high electric fields, we can extend these equations with acceptable accuracy, as long as the electric field does not vary too rapidly with position.
4. When the electric field is low, the average electron velocity is proportional to the electric field, but for high electric fields, the velocity of electrons in bulk Si saturates at $\approx 10^7$ cm/s.
5. The electric field in a short channel MOSFET is very high, but the velocity does not saturate because the electrons leave the device before they have a chance to scatter very often. The velocity in the channel overshoots the saturated velocity.

6. For very short channels, ballistic transport is possible. The Landauer approach provide simple, physical way to describe ballistic transport. It can also describe diffusive transport in which scattering is strong. Because it works from the ballistic to diffusive limit, the Landauer approach provides a suitable framework for modeling transport in modern devices.
7. When the potential in the channel changes abruptly on the scale of the electron's wavelength, electrons cannot be treated a semi-classical particles. Quantum confinement changes the energy levels and density-of-states, and quantum transport along the channel describes quantum reflections and tunneling.
8. Analyzing a semiconductor device consists of two parts. First, solving a transport problem to describe the flow of mobile charge and second, solving the Poisson equation to compute the electrostatic potential that results from a given distribution of mobile charge.

2.10 References

For a description of the traditional approach to carrier transport, drift-diffusion equation, Drude equation for mobility, etc., see:

- [1] Robert F. Pierret *Semiconductor Device Fundamentals*, 2nd Ed., Addison-Wesley Publishing Co, 1996..
- [2] Robert F. Pierret *Advanced Semiconductor Fundamentals*, 2nd Ed., Vol. VI, Modular Series on Solid-State Devices, Prentice Hall, Upper Saddle River, N.J., USA, 2003.
- [3] N.W. Ashcroft and N.D. Mermin, *Solid-State Physics*, Saunders College, Philadelphia, PA, 1976.

Smith, Janek, and Adler provide a good derivation of $J_n = n\mu_n dn/dx$.

- [4] A.C. Smith, J. Janak, and R. Adler, *Electronic Conduction in Solids*, McGraw-Hill, New York, NY 1965.

The Landauer approach to carrier transport at the nanoscale is discussed in Vols. 1 and 2 of this series. In particular, see Chapter 3 of [6] for a derivation of eqn. (4.26).

- [5] Supriyo Datta, *Lessons from Nanoelectronics: A new approach to transport theory*, World Scientific Publishing Company, Singapore, 2011.
- [6] Mark Lundstrom, *Near-Equilibrium Transport: Fundamentals and Applications*, World Scientific Publishing Company, Singapore, 2012.

For a discussion of high-field and non-local transport, see:

- [7] Mark Lundstrom, *Fundamentals of Carrier Transport*, 2nd Ed., Cambridge Univ. Press, Cambridge, U.K., 2000.

The following two references discuss physically detailed MOSFET device simulation - the first semiclassical and the second quantum mechanical.

- [8] D. Frank, S. Laux, and M. Fischetti, "Monte Carlo simulation of a 30 nm dual-gate MOSFET: How short can Si go?," Intern. Electron Dev. Mtg., Dec., 1992.
- [9] Z. Ren, R. Venugopal, S. Goasguen, S. Datta, and M.S. Lundstrom "nanoMOS 2.5: A Two -Dimensional Simulator for Quantum Transport in Double-Gate MOSFETs, *IEEE Trans. Electron. Dev.*, **50**, pp. 1914-1925, 2003.