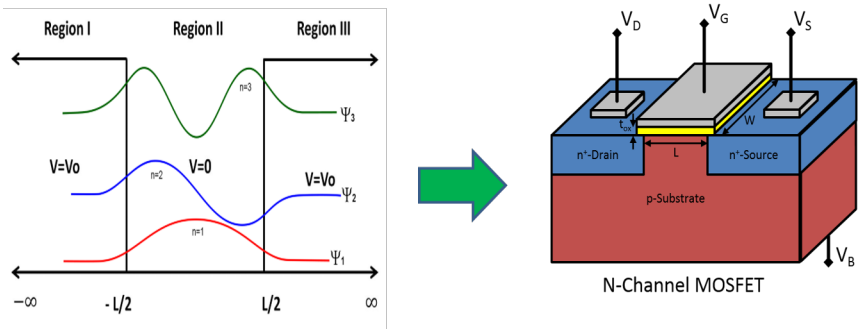
SEMICONDUCTOR AND DEVICE PHYSICS: A CONCISE INTRODUCTION

Neil Goldsman
and
Chris Darmody

April 29, 2020

Under Contract for Publication by Springer, Fall 2019



Preface

This text is meant for students starting to learn about semiconductor devices and physics, as well as those who are interested in a review. It is meant to be a concise description of what the authors think are the key areas of this subject. The goal is to acquaint readers with the information to give them a sufficient understanding of semiconductor devices and physics so it can either serve as a starting point from which much more studying can be done in the area, or it can serve as a stand alone final course that complements other areas of study and practice.

The authors are grateful to Dr. Zeynep Dilli and Dr. Anshu Sarge for their help at the beginning of this project; Dr. Danilo Romero for his insightful discussions on teaching semiconductor physics; Dr. Rob Valente for editing; Casey Goldvale for providing figures; Dylan Goldvale for comments; Yumeng Cui and Usama Khalid for being excellent graduate teaching assistants; and to UMD ENEE 313 students Hannah Watsky and Mauricio Perez-Oviedo for proofreading the text.

Contents

List of Tables

List of Figures

Chapter 1

Crystal Structure

1.1 Introduction

Crystal is a periodic array of atoms arranged in a lattice.
A crystal is a periodic array of atoms. There are many crystals in nature. Most of the solid elements in the periodic table are crystals. Silicon is an example of a crystal that is very important in electronics because it is also a semiconductor. Other crystals include metals like aluminum and salts such as sodium chloride. As a first step to defining crystal structure, we need to talk about lattices.

1.2 Lattice

Lattice is a periodic array of points in space.
The lattice is a mathematical concept. It is not the crystal. The atoms are arranged on a lattice to define the crystal. However, the lattice itself is just a regular periodic array of points in space arranged according to the following vector relationship:

$$\vec{R} = \mu\vec{a} + \nu\vec{b} + \omega\vec{c} \tag{1.1}$$

where, $\vec{a}, \vec{b}, \vec{c}$ are the fundamental lattice translation vectors and μ, ν, ω are integers.

$\vec{R}$ is a vector that represents or lands on the set of points that defines the lattice. Different values of the integers correspond to different points on the lattice. The lattice looks the same as when viewed from any point $\vec{R}$
in space.

Lattices can be in one dimension, two dimensions or three dimensions. Of course, a physical crystal typically will be described using a 3-D lattice. However, for illustration purposes, it is useful to show 2-D lattices. Figure 1.1 shows an example of a rectangular 2-Dimensional lattice. The figure also shows the lattice translation vectors $\vec{a}$ and $\vec{b}$.

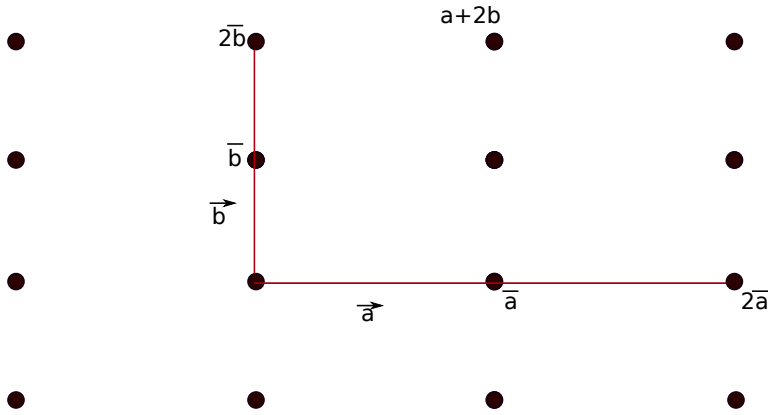


Figure 1.1: A rectangular lattice. $\vec{a}, \vec{b}$ are the fundamental translation vectors for the set of points which define the lattice.

Another 2-D example is shown in Figure 1.2 where a 2-Dimensional diamond-like lattice is illustrated.

Lattice Primitive Unit Cell is geometric structure that when translated by the primitive vectors will fill up the entire space of the lattice. The Lattice Primitive unit cell will contain only one lattice point. An example of a Lattice Primitive unit cell is shown in Figure 1.2 denoted with the primitive vectors $\vec{a}$ and $\vec{b}$.

Lattice Conventional Unit Cell is typically not a lattice primitive unit cell but a convenient unit cell that describes the lattice. An example of a lattice conventional (often called convenient) unit cell is shown in Figure 1.2 denoted with the unit cell vectors $\vec{a'}$ and $\vec{b'}$. Note that the conventional lattice unit cell contains more than one point (in this case two), which has the effect of increasing the number of atoms in the basis. The atomic basis is discussed later in Section 1.3.

Example 1.1:
Primitive Unit Cell Descriptions in 2-D

Given the lattice structure shown in Figure 1.2, assign primitive lattice vectors in four different ways. Remember, the choice of primitive vectors is not unique.

Any set of two (linear independent) vectors which contain a single lattice site can define the primitive cell of a crystal. In this example, we can choose any of the lattice vector pairs shown in Figure 1.3 labeled $1-4$ as well as all combinations of $\pm\vec{a}_x$ and $\pm\vec{b}_x$. The pair of vectors labeled 5 and 6 do not define a primitive cell. Vectors 5 are not primitive because they enclose two full lattice sites and vectors 6 are not valid because they are not linearly independent and thus do not define a 2-D expanse.

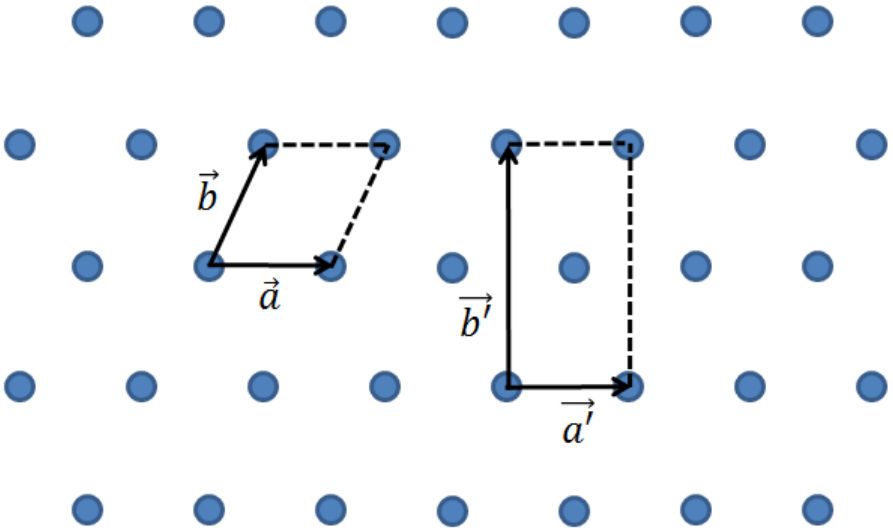


Figure 1.2: A 2-dimensional diamond-like lattice. $\vec{a}, \vec{b}$ are the fundamental translation vectors for the set of points which define the lattice. These are better known as the primitive lattice vectors, and they enclose one full lattice point in the primitive cell. A conventional or convenient unit cell description of the crystal is shown using vectors $\vec{a'}, \vec{b'}$. In this description of the crystal, two lattice points are contained within the unit cell.

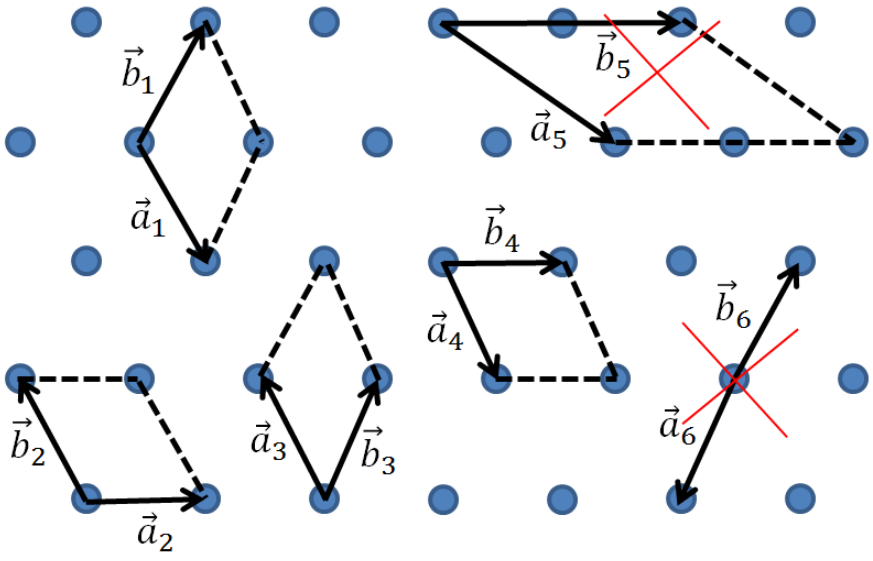


Figure 1.3: Primitive vectors for a 2-D Lattice

1.2.1 Common 3D Lattices and Lattice Unit Cells

For real crystals, we have 3-Dimensional lattices. There are a number of important 3-D lattices. While an entire class can be devoted to the geometrical and symmetry properties of these 3-D lattices, we will confine ourselves to mentioning four important ones that are especially relevant to most crystalline solids, especially semiconductors. These are the:

- **Simple Cubic Lattice**, which has lattice points on the corners of a cube, and the 3-D lattice is formed by translating the cube through space, using the orthogonal vectors $\vec{a}, \vec{b}$ and $\vec{c}$, where $|\vec{a}| = |\vec{b}| = |\vec{c}|$ There is one complete lattice point in the cube unit cell. (This is a primitive lattice unit cell.) Figure 1.4 shows the a unit cell from of the simple cubic lattice.

- **Body Centered Lattice**, which has lattice points on the corners of a cube, and one in the center, and the 3-D lattice is formed by translating the cube through space, using the orthogonal vectors $\vec{a}, \vec{b}$ and $\vec{c}$. There are two complete lattice points in the body centered cubic unit cell. (This is not a primitive unit cell for the body centered lattice.) Figure 1.5 shows the a unit cell from of the body centered cubic lattice.

- **Face Centered Lattice**, which has lattice points on the corners of a cube, and one each of the six faces, and the 3-D lattice is formed by translating the cube through space, using the orthogonal vectors $\vec{a}, \vec{b}$ and $\vec{c}$. There are four complete lattice points in the face centered cubic unit cell. (This is not a primitive unit cell for the face centered lattice.) Figure 1.6 shows the unit cell of the face centered cubic lattice.

1.3 Crystal

Now that we have talked a little about lattices, let's move on to crystals. A crystal is the actual physical entity which consists of a periodic arrangement of atoms on a lattice. The crystals that we talk about in this class are largely crystalline solids and therefore will typically contain close to Avogadro's number of atoms. Recall, Avogadro's number is 1.602×10^{23} number of atoms/molecules in one mole of a substance. The separation of atoms in the crystal is typically about 2 angstroms. We typically describe the crystal as the lattice plus the atoms arranged on each lattice point. Each lattice point will have the same arrangement of atoms associated with it. The set of atoms associated with each lattice point is called the **Atomic Basis**. The atomic basis is given by these atoms and their coordinates with respect to the lattice point.

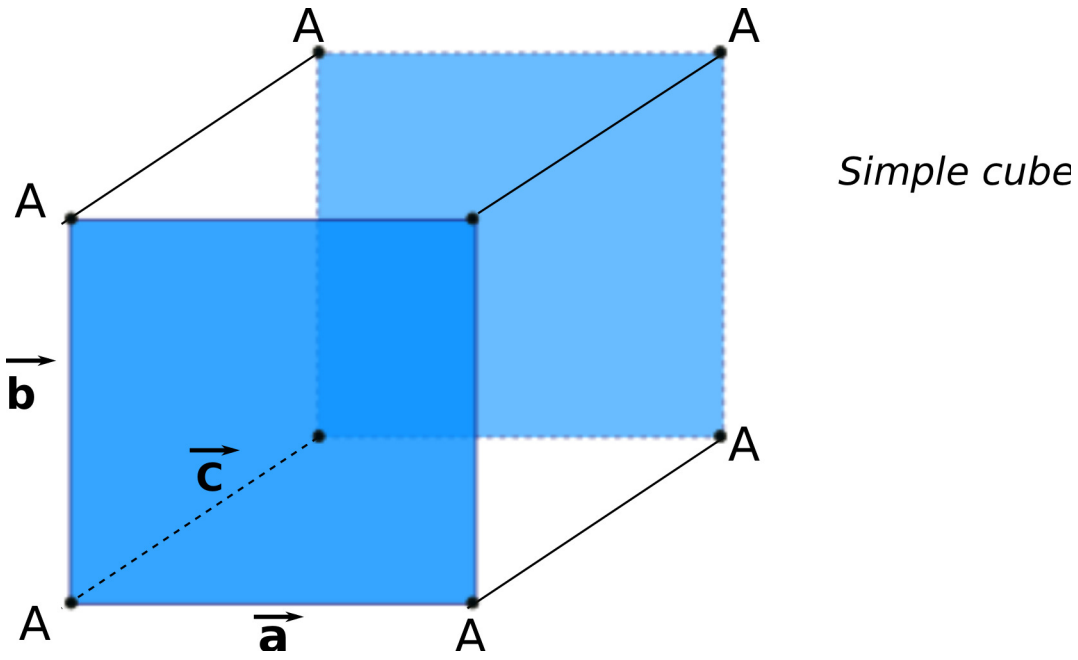


Figure 1.4: Primitive cell of SC. A cube with atoms at its each corner.

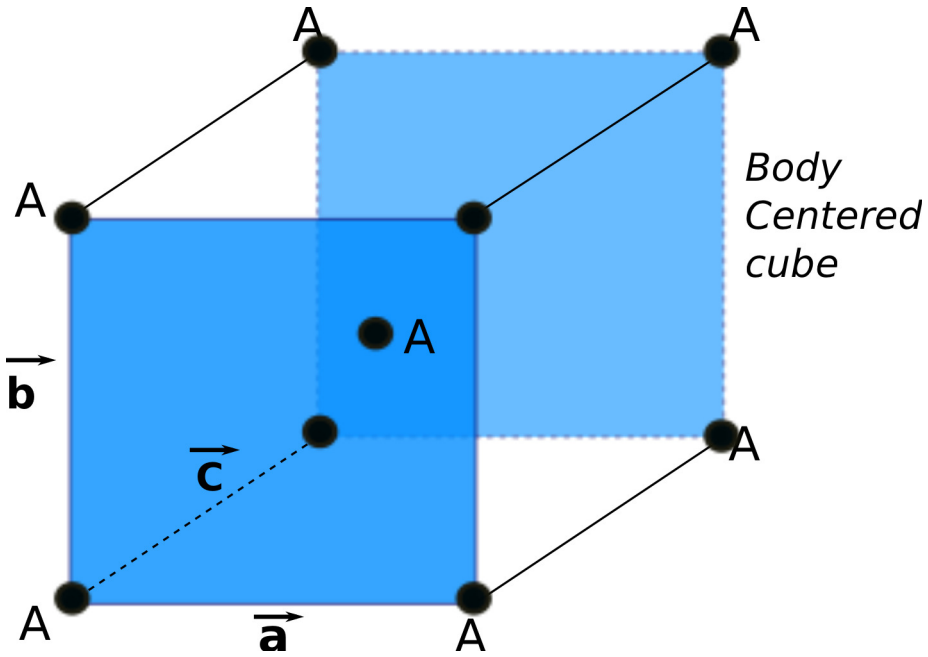


Figure 1.5: A cube with 1/8 lattice point at its each corner and one in the center.

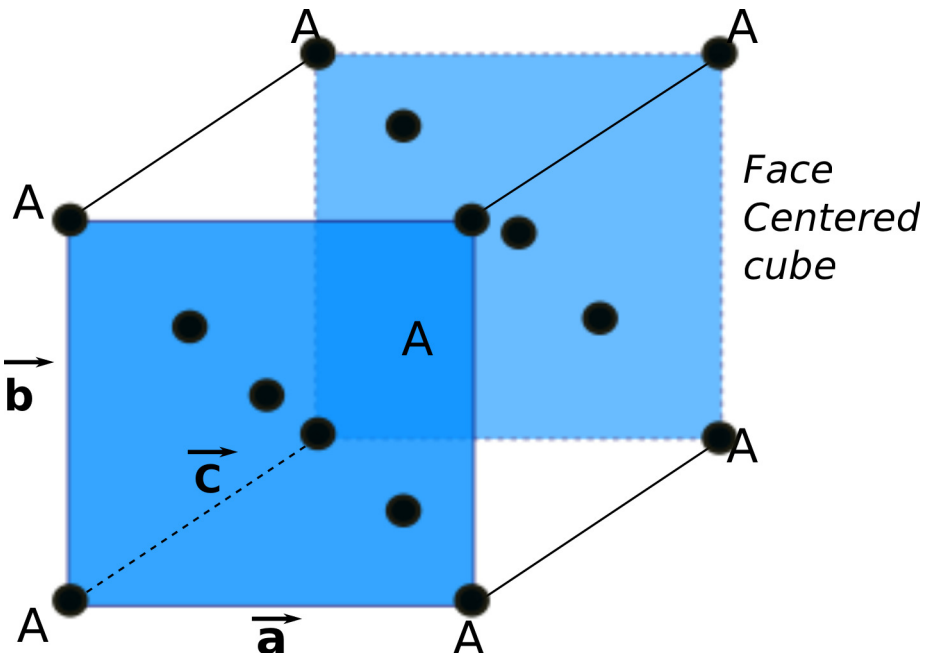


Figure 1.6: A cube with lattice points at its each corner and one at the center of each of the six faces.

Atomic Basis or often just called the **Basis** is arrangement of a group of atoms in the crystal with respect to a lattice point.

$$\textbf{Crystal} = \textbf{Lattice} + \textbf{Atomic Basis}$$

Crystal Unit Cell: Fundamental building block which when translated by integer multiples of lattice vectors produces the complete crystal. There are two basic types of unit cells, a crystal primitive unit cell and a crystal convenient unit cell.

Note: Lattice Unit Cell versus Crystal Unit Cell: Earlier in the chapter we talked about Lattice Unit Cells, and in this section we are talking about Crystal Unit Cells. The main difference is that Lattice Unit Cells are just points in space, while Crystal Unit Cells have atoms associated with them. In other words, Crystal Unit Cells are real physical objects, while Lattice Unit cells are mathematical constructions. It is also important to understand that the Lattice Primitive Cell will only have one point. However, the Crystal Primitive Cell can have more than one atom as explained below.

Crystal Primitive Unit Cell: The smallest unit cell you can have and still have the material. For a monatomic crystal, the primitive unit cell usually contains only one atom. However, if the crystal contains more than one type of atom, the primitive unit cell will contain more than one atom. For diatomic crystals containing

for example, atom type A and atom type B, the primitive unit cell will typically contain two atoms, one of type A and one of type B. However, unless the lattice is rectangular, the primitive unit cell will typically have coordinate vectors that are not orthogonal, and thus are difficult to work with.

Crystal Convenient Unit Cell: When scientists and engineers work with crystals, they usually do not utilize the primitive unit cell to describe the crystal, but use alternative or non-primitive rectangular unit cells that is much more convenient to work with. The coordinate vectors for these conventional or convenient unit cells will typically be at right angles.

Example 1.2:
2D Monatomic Crystal on Rectangular Lattice

Probably the simplest 2-D crystal is the crystalline solid that contains only one type of atom arranged on a rectangular lattice. The structure is shown in Figure 1.7. Determine the primitive lattice vectors and atomic basis.

For this case there is only one atom in the atomic basis and the crystal is described as follows:

Lattice $\mu\vec{a} + \nu\vec{b}$
Basis: Atom A at (0,0)

Example 1.3:
2D Monatomic Crystal on Diamond Lattice

In Figure 1.8, we show 2-D crystal is the crystalline solid that contains only one type of atom arranged on a diamond lattice. The diamond type structure is the primitive lattice and the primitive unit cell. Give a primitive and convenient cell description of the crystal by specifying the lattice vectors and atomic basis.

For this case there is only one atom in the atomic basis and the crystal using the primitive unit cell and a one atom basis is described as follows:

Lattice $\mu\vec{a} + \nu\vec{b}$
Basis: Atom A at (0,0)

This monatomic crystal can also be described using a convenient rectangular lattice and a convenient unit cell that contains two identical atoms. For this case, the same crystal can be described by the following rectangular lattice and the two atom basis.

Lattice $\mu\vec{c} + \nu\vec{d}$
Basis: Atom A at (0,0); and atom A at $(\frac{1}{2}c, \frac{1}{2}d)$

Figure 1.8 shows both the primitive diamond and the convenient rectangular structures.

Example 1.4:
Diatomic 2D Crystal on a rectangular lattice

An example of a 2-D crystal is shown in Figure 1.9, where there is a 2-D rectangular lattice and an atomic basis consisting of two different types of atoms, A and B. Determine the primitive lattice vectors and atomic basis.

We can describe the crystal in the following way:

Lattice $\mu\vec{a} + \nu\vec{b}$
Basis: Atom A at (0,0); Atom B at $(\frac{1}{2}a, \frac{1}{2}b)$

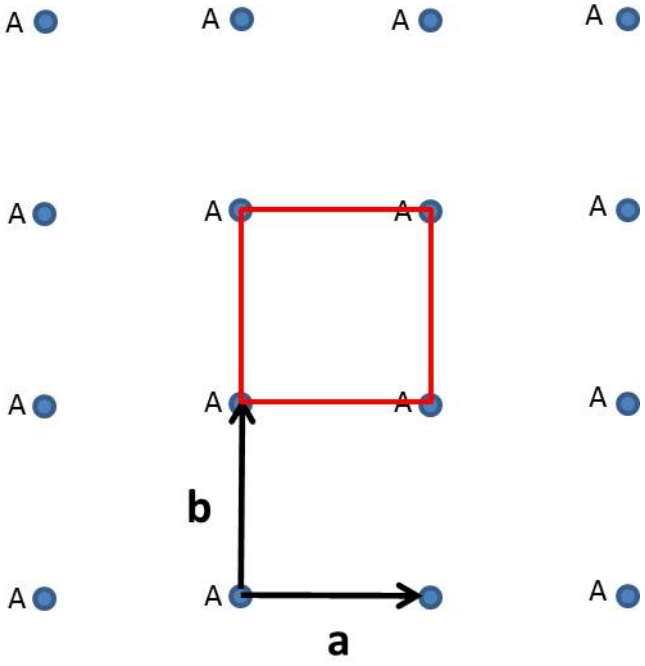


Figure 1.7: A 2-D Crystal on a Rectangular lattice with a one atom (A) basis. The lattice translation vectors and the unit cell are shown.

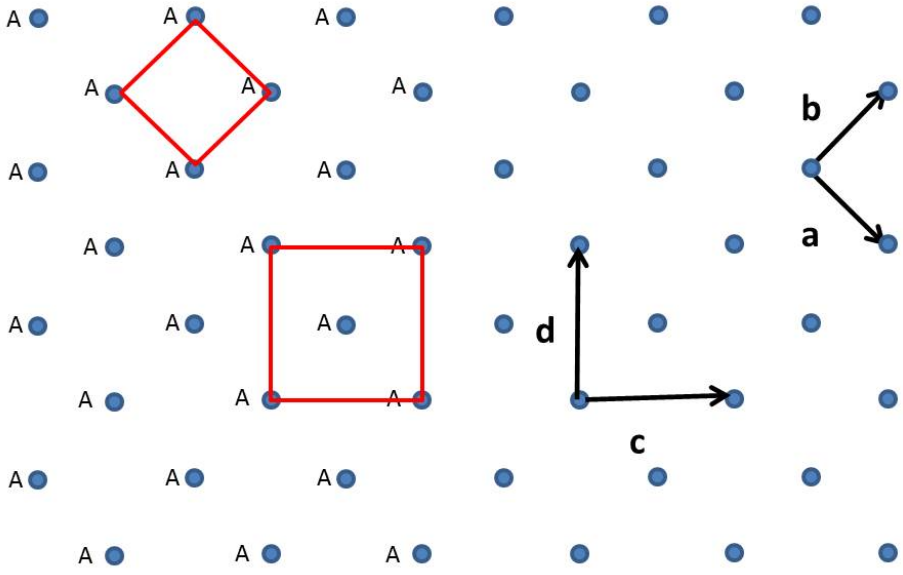


Figure 1.8: A 2-D Monatomic Crystal on a diamond-like lattice, with a one atom (A) basis. The lattice translation vectors and the primitive unit cell are shown. Also shown is the corresponding convenient rectangular lattice with a conventional rectangular unit cell and a basis consisting of 2 (A) type atoms.

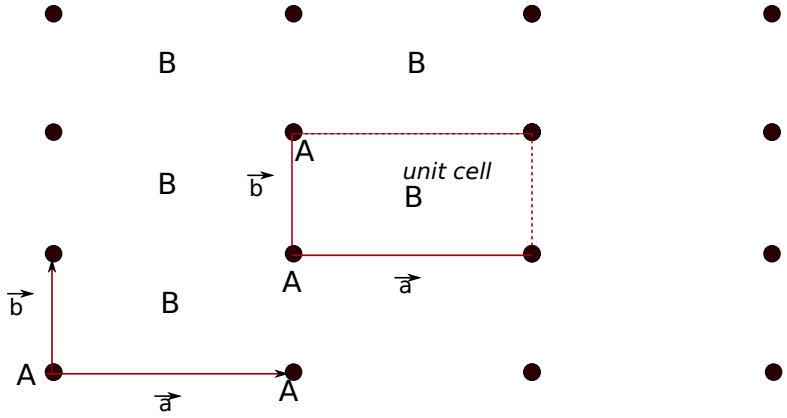


Figure 1.9: A 2-D Crystal on a Rectangular lattice with a two atom (A & B) basis

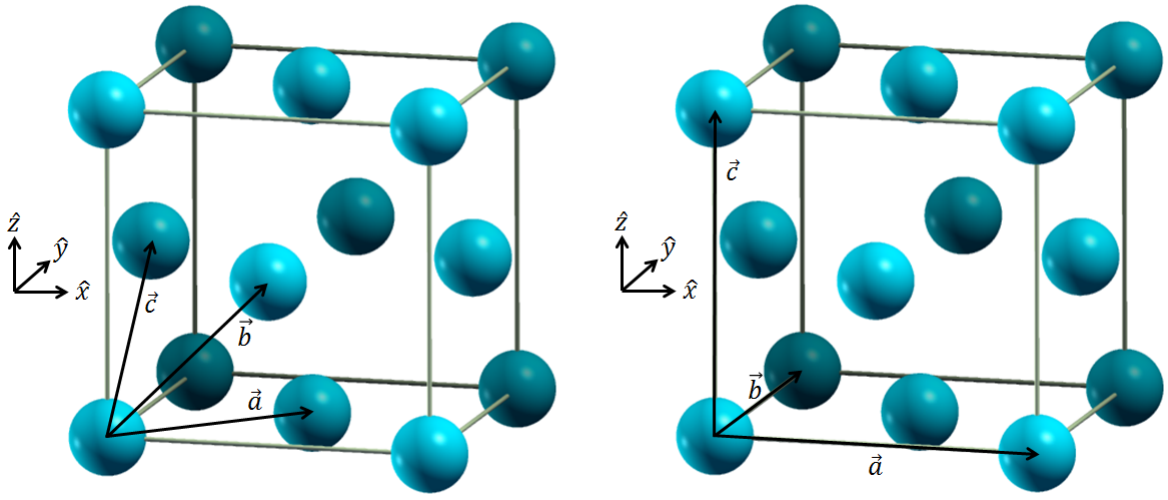


Figure 1.10: An FCC Lattice

Example 1.5:
Primitive and Convenient Unit Cell Descriptions in 3-D

Given the lattice structure shown in Figure 1.10, and assuming the side-lengths of the box are equal to **a**, where $\mathbf{a} = \|\tilde{\mathbf{a}}\|$:
Give a description of the lattice (lattice vectors + atomic sites in basis) for a primitive and convenient unit cell.

Using the FCC description of the crystal we obtain the primitive cell:

Primitive Lattice Vectors:
$\vec{a} = \frac{a}{2}\hat{x} + \frac{a}{2}\hat{y}$, $\vec{b} = \frac{a}{2}\hat{x} + \frac{a}{2}\hat{z}$, $\vec{c} = \frac{a}{2}\hat{y} + \frac{a}{2}\hat{z}$

Atomic Basis:
One atom at $(0, 0, 0)$

We only need to include one atom in the basis and translate it with various linear combinations of the primitive lattice vectors to fill out the entire crystal. The vectors are shown on the left of Figure 1.10.

For the convenient cell, we can use a SC lattice and increase the number of atoms in the basis:

Lattice Vectors:
$\vec{a} = a\hat{x}$, $\vec{b} = a\hat{y}$, $\vec{c} = a\hat{z}$

Atomic Basis:
Four atoms $(0, 0, 0)$, $(\frac{a}{2}, \frac{a}{2}, 0)$, $(\frac{a}{2}, 0, \frac{a}{2})$, $(0, \frac{a}{2}, \frac{a}{2})$

We need to include more atoms in the convenient cell basis because if we only translated the single atom from the primitive basis along linear combinations of these new lattice vectors, we would be missing atoms on the faces (and interior) of the unit cell. The vectors are shown on the right of Figure 1.10.

1.4 Crystal Planes and Directions

In a crystal atoms are arranged on the lattice and as a results we have layers or planes of atoms. Planes form along different directions. We have a method of referring to these planes and directions in a lattice and in a crystal structure.

Crystallographic Directions are written in brackets [**uvw**], where u, v and w represent integers. The crystallographic direction is generally given by the coefficients of the lattice translation vectors. Examples of crystallographic directions for a 2-D crystal are shown in Figure 1.11. with directions in 2-D given by values of [uv].

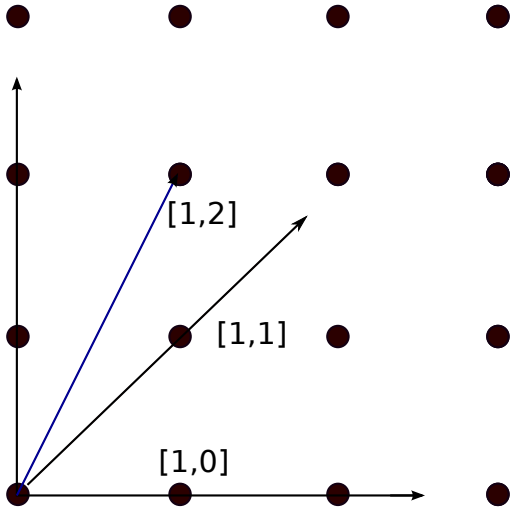


Figure 1.11: Figure showing examples of Crystallographic Directions and their labeling

Crystallographic Planes are written parenthesis (hkl), where h, k and l represent integers that are called the Miller Indices for that crystal plane. The following methodology is used to label crystallographic planes, and thus get the values of the integer Miller indices **h, k** and **l**.

1. Draw an orthogonal coordinate system somewhere on the crystal.
2. Determine the x,y,z coordinates where the plane intersects the axes. These will be integer values and will represent a number of lattice points.
3. Take the reciprocals of each of these points of intersection and then determine the smallest set of integers that give the same ratios between the 1/x, 1/y and 1/z and these will be your h,k and l indices and thus the **(hkl)** crystal plan. It is probably easiest to understand by example. Let's start with a plane in 2D which is actually a line.

Example 1.6:
Find the label for Plane I in Figure 1.12.

1. Intercepts: $4\vec{a}$, $\infty\vec{b}$
2. Taken plane is the reciprocals: $\frac{1}{4}, \frac{1}{\infty}$
3. Reduce to smallest set of integer that has same ratio $\frac{1}{4}, 0 \rightarrow (1, 0)$ Plane

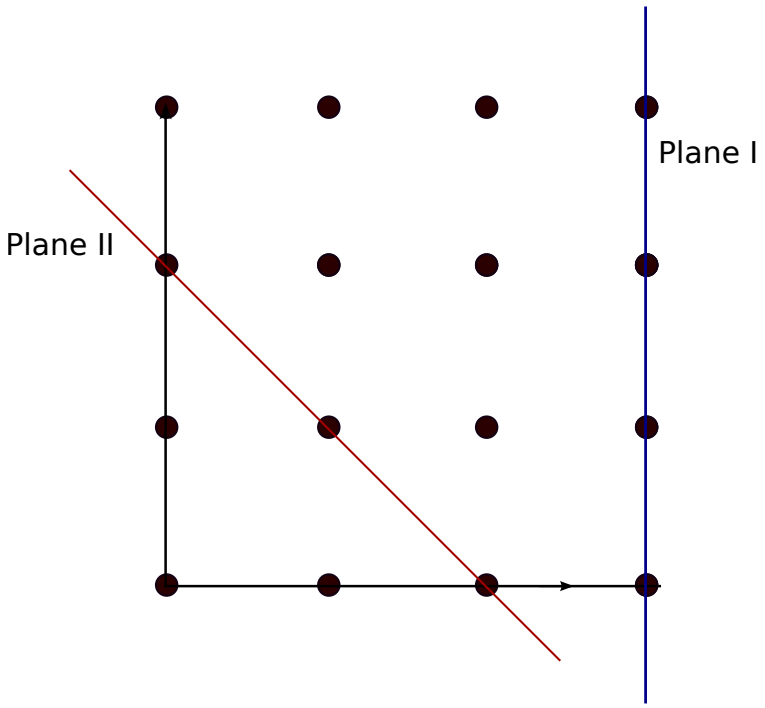


Figure 1.12: Figure showing examples of planes (1,0) and (1,1).

Example 1.7:
Find the label for Plane II in Figure 1.12.

1. Intercepts: (x=2) and (y=2)
2. Reciprocal $\frac{1}{2}, \frac{1}{2}$
3. Plane (1,1) (Figure 1.12 Plane II)

Example 1.8:
Find the label for a plane with indices (3,1).

1. Intercepts at (x=1, y=3)
2. Reciprocals: $\frac{1}{1}, \frac{1}{3}$
3. Plane (3,1) (See Figure 1.13)

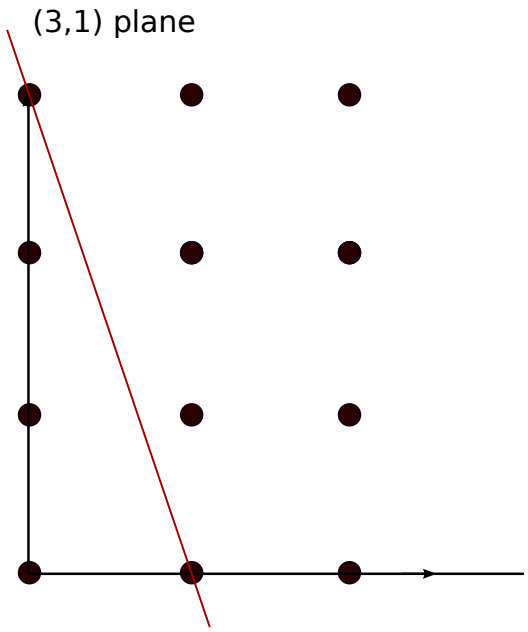


Figure 1.13: Figure showing plane (3,1).

Example 1.9:
Find the label for a 3D Crystal Plane with indices (2,1,4).

1. Intercepts at (x=2, y=4, z=1)
2. Reciprocals: $\frac{1}{2}, \frac{1}{4}, \frac{1}{1}$
3. Plane (2,1,4) (See Figure 1.14)

1.5 Semiconductor Crystals

1.5.1 Elemental Semiconductors

Silicon is the most important semiconductor. It has a diamond crystal structure which has 2 atoms in its Primitive Unit Cell. Germanium is also a semiconductor and has the same crystal structure as silicon. And of course, Carbon also forms in the diamond structure. These atoms make up the first three the elements of the 14^{th} column of periodic table and are arranged from top to bottom as C, Si, then Ge which indicate increasing atomic size. This size increase can be seen in the size of their corresponding crystal lattice constants: $\mathbf{a_C} = 0.3567\text{nm}, \mathbf{a_{Si}} = 0.5431\text{nm}, \mathbf{a_{Ge}} = 0.5658\text{nm}$.

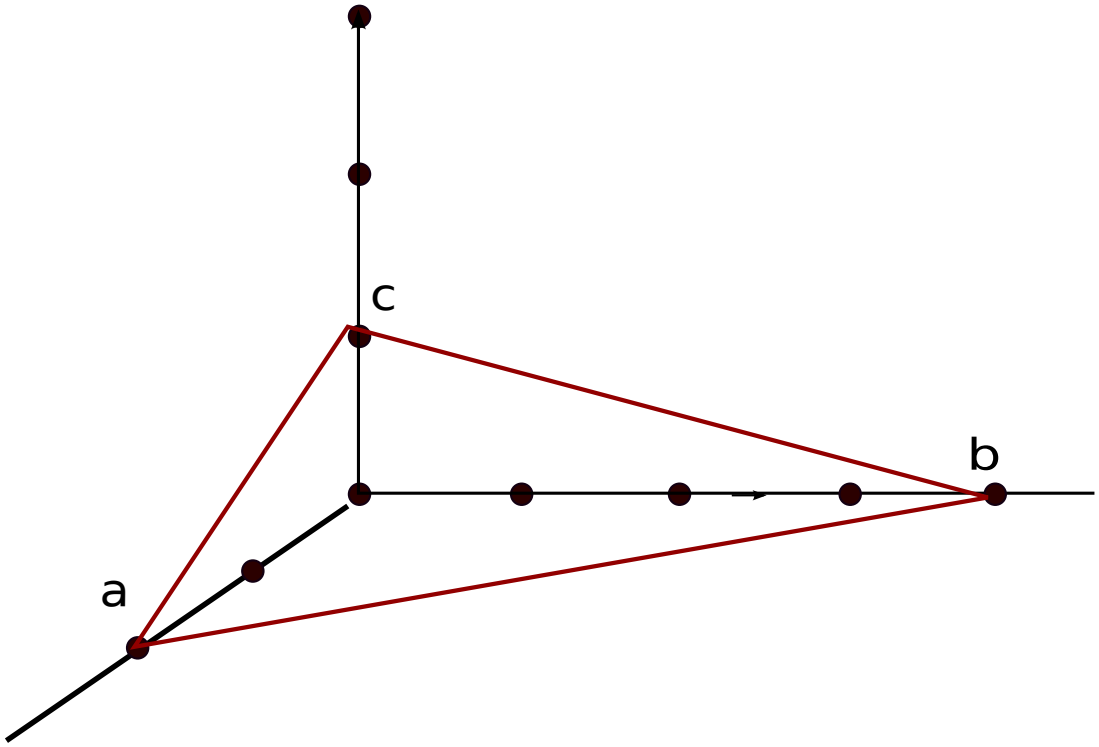


Figure 1.14: 3D Crystal plane (2,1,4).

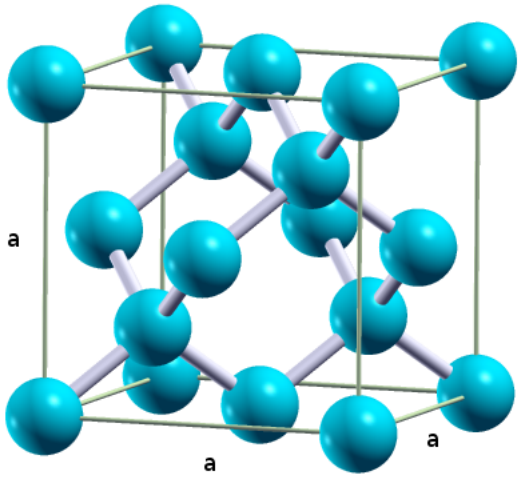


Figure 1.15: A cubic unit cell of the diamond lattice. It can be thought of as two interpenetrating FCC unit cells, with one translated one quarter way up along the diagonal faces. Lattice constants: $\mathbf{a_C} = 0.3567\text{nm}, \mathbf{a_{Si}} = 0.5431\text{nm}, \mathbf{a_{Ge}} = 0.5658\text{nm}$.

Diamond Crystal Structure is very important in electronics because the semiconductors silicon and germanium are arranged on a diamond lattice. The diamond lattice is more complicated than the aforementioned structures. It can be described as two interlaced Face Centered Cubic lattices, with one lattice being translated by $1/4\ \vec{a}, 1/4\ \vec{b}$ and $1/4\ \vec{c}$ along the Face Centered diagonal. The diamond lattice cubic nonprimitive unit cell is shown in Figure 1.15. This convenient unit cell contains an 8 atom basis. For a more detailed discussion of various lattice types, see a text on Solid State Physics such as Kittel [1].

Carbon, Silicon and Germanium are all in the 14^{th} group in the periodic table, so they all have 4 valence electrons. When they form crystals, the 4 valence electrons hybridize into the four *sp3* orbitals, which give rise to bonding and the diamond crystal structure. Each silicon atom will be bonded to four other silicon atoms to form the diamond lattice, forming covalent bonds and satisfying the octet rule for pairing of valence electrons.

1.5.2 Compound Semiconductors

In addition to elemental semiconductors, there are compound semiconductors that also form as diamond like structures. Gallium Arsenide (GaAs) and Gallium Nitride (GaN) differ from the elemental structures of Si by the following. They are composed of two interlaced FCC crystals, however, for GaAs one FCC crystal is composed of Ga and the other translated FCC structure is composed of As. A similar description can be given for GaN. This diamond like structure is called **Zincblende** These compounds are called three-five semiconductors because they come from the 13th and 15th columns of the periodic table.

1.6 Problems

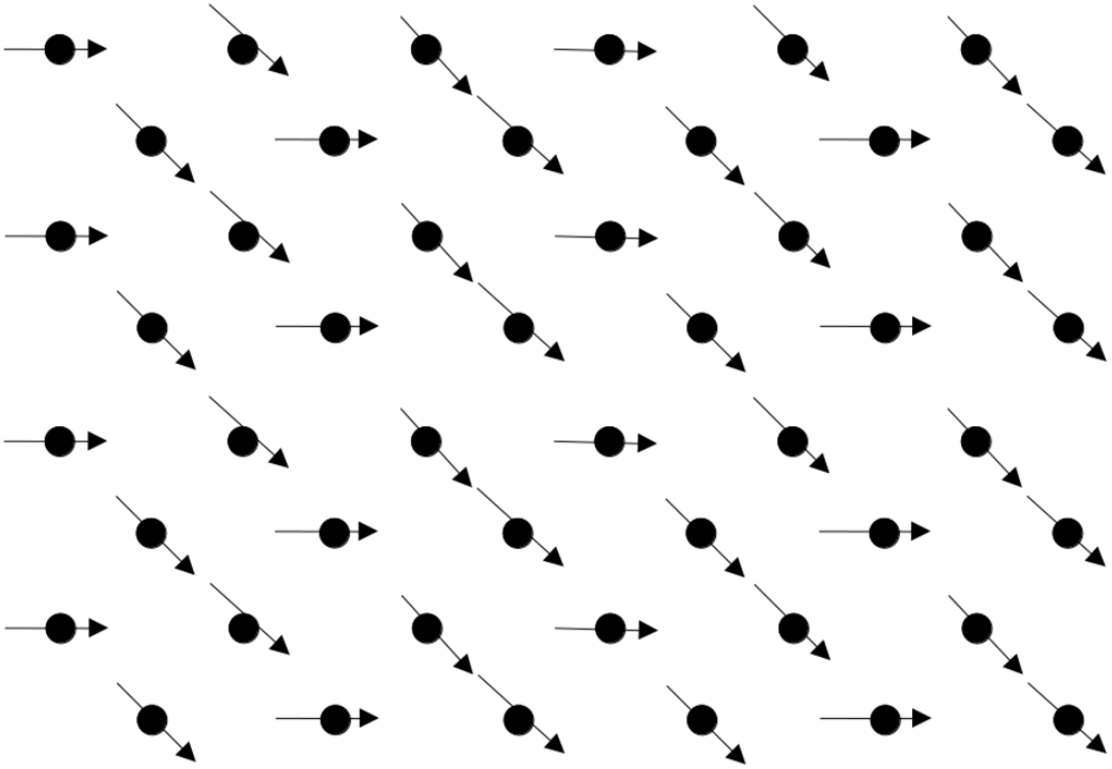
1.1 Primitive Unit Cell vs. Convenient Rectangular Unit Cell with Basis.
Recall, the primitive unit cell is the smallest unit cell we can have of a material, and still have that material. However, the shape of the primitive can be difficult to deal with. Often, we use a convenient larger rectangular unit cell with a different atomic basis since it is easier to work with than the primitive unit cell. Now, suppose we have a 2-Dimensional material composed of a single type of atom (Atom A). The material has a convenient rectangular unit cell with vectors: $\mathbf{a} = 1$nm, and $\mathbf{b} = 2$nm. Also, the atomic basis of this convenient rectangular unit cell is (A: 0,0) and (A: $\frac{1}{2}$,$\frac{1}{2}$).

(a) Sketch the lattice of this crystalline material and outline the convenient rectangular unit cell described above. Include the lattice vectors a and b

on the cell.

(b) Outline a primitive unit cell for this material on your sketch and label the primitive vectors.

1.2 Consider the structure illustrated below, which might represent the structure of a (two dimensional) molecular crystal with each arrow representing a heteropolar molecule, and the direction of the arrow designating the orientation of a molecule. The molecules are arranged on an equilateral triangular network.



(a) Define the structure in terms of a convenient rectangular cell and a suitable basis.

(b) Define the structure in terms of a primitive cell and a basis.

(c) Repeat (a) and (b) if the molecules are replaced by atoms (i.e. dots instead of arrows). (Note that this would also be an appropriate description!)

1.3 Miller Indices and Lattice Planes

(a) For the 2-Dimensional lattice, sketch the (1,0), (1,1) and (1,2) planes. (These are actually lines for 2-D lattices)

(b) For a 3-Dimensional cubic lattice, sketch the (1,0,0), (1,1,0) and (3,2,1) planes

1.4 A 2-Dimensional crystalline material with the chemical formula AB_2 has its atoms arranged on square lattice. The three atom atomic basis of this structure is (A: 0,0), (B: $\frac{1}{2}$,0), (B: 0,$\frac{1}{2}$). The length of the primitive vector **a** is 0.5nm. (Since the primitive unit cell is a square, both primitive vectors have the same length.) The radius of atom A = 0.1nm, and the radius of Atom B = 0.05nm.

(a) Sketch this crystal

(b) Sketch a primitive unit cell of this crystal and discuss it in one or two sentences.

(c) What is the area of the primitive unit cell filled by atoms, and what is the empty area? (We are dealing with areas since this is a 2-D system).

1.5 A 3-Dimensional crystalline material has a simple cubic primitive unit cell that has the chemical formula AB. The length of the primitive vector = **a** is 0.5nm. The atomic basis is (A: 0,0,0) and (B: $\frac{1}{2}$,$\frac{1}{2}$,$\frac{1}{2}$). The radius of atom A = 0.1nm, and the radius of Atom B = 0.05nm.

(a) Sketch this simple cubic primitive unit cell.

(b) What is the total volume of this primitive unit cell?

(c) What is volume of the cell that is occupied by atoms?

(d) If instead of having the atom B in the center, you had atom A, would this still be a primitive unit cell? Explain why or why not.

1.6 Sketch the unit cell for a:

(a) simple cubic lattice

(b) body centered cubic lattice

(c) face centered cubic lattice

1.7 Most semiconductors are arranged on a diamond lattice, shown in 1.15. How many atoms are in the convenient rectangular (diamond structure) unit cell of silicon? By giving the location of each of the atoms in this unit cell, justify your answer.

Chapter 2

Atoms, Electrons and the Beginning of Quantum Mechanics

2.1 Introduction

Quantum Mechanics is the new physics. To be a good engineer, you really need to have some fundamental understanding of basic quantum mechanics. Quantum mechanical properties become important when we start working with small objects, which typically have the dimensions of nanometers or smaller. We learned in basic chemistry that quantum mechanics is important in describing the energy levels of an atom. In addition, quantum mechanics plays a key role the operation of modern electronic devices. The operation of the fundamental electronic devices, such as transistors and diodes, is described by a combination classical mechanics and quantum mechanical principles. As I write this in the year 2014, the latest generation MOSFET transistor in production, which is the key transistor building block of most electronics, has a gate length of 22nm. Other devices are even smaller, and have critical dimensions of a few nanometers or less, and thus are strongly influenced by the principles of quantum mechanics. Electrons and holes, the basic charge carriers in electronics are quantum mechanical entities. While virtually all electronic devices are influenced by quantum mechanics, some devices operate virtually totally on quantum mechanical principles. These devices include Flash Memory Sticks, Solar Cells, solid state Lasers, LEDs and LED lighting.

From a historical perspective, quantum mechanics started to come into being in the early 20th century. Then, for the next one hundred years, most discovery in physical science has been guided by quantum mechanical issues. Quantum Mechanics arose to satisfy the need for a new science to describe the observations

that were being discovered about one hundred years ago. Two of these key areas of the discovery were the **Photoelectric Effect** and the **Light Emission Spectra of Atoms**. These discoveries opened up the fundamental quantum mechanical properties that small particles, like electrons, exhibit both wavelike and particle-like behavior, and that light, long thought of as a wave, can exhibit both particle and wave-like properties as well.

There are numerous excellent texts on Modern Physics and Introductory Quantum Mechanics. Students who want to delve deeper into these topics than is presented in this text might want to check out the following books [2, 3], for example.

2.2 Photoelectric Effect

In 1887 Heinrich Hertz discovered that if he focused ultraviolet light on a metal surface electric sparks could more easily be emitted. Further investigations showed that if you shined light of high enough frequency electric current could be emitted from the metal surface. Furthermore, as the intensity of the light increased, the electric current, or the number of electrons collected per unit time, would increase. However, the energy of the individual electrons would not change unless you changed the color or frequency of the incident light. If you shined light of higher frequency, then the energy of the emitted electrons would be greater. Ultimately, it was discovered that low frequency light (in the red part of the spectrum for example) would not give rise at all to any emission of this 'photocurrent' or photoelectrons, no matter what the intensity of the incident light was. This was puzzling to scientists at the time because light intensity, which is defined as the light power per unit area (Watts/m^2), had always been taken to be proportional to the square of the amplitude of the light. According to this classical view of the energy and light, it had nothing to do with the frequency of the light, and therefore the frequency should not have affected the energy of the emitted electrons.

Let's summarize the Photoelectric Effect experiment (Figure 2.1):

1. Shine light at a metal plate.
2. Detect electrons that are emitted if the frequency of light is high enough.
3. Measure energy of emitted electrons.
4. Measurement showed that the energy of the emitted electrons is proportional to the frequency of light, and the proportionality constant had a particular value and was given the name **Planck's Constant** and the symbol **h**

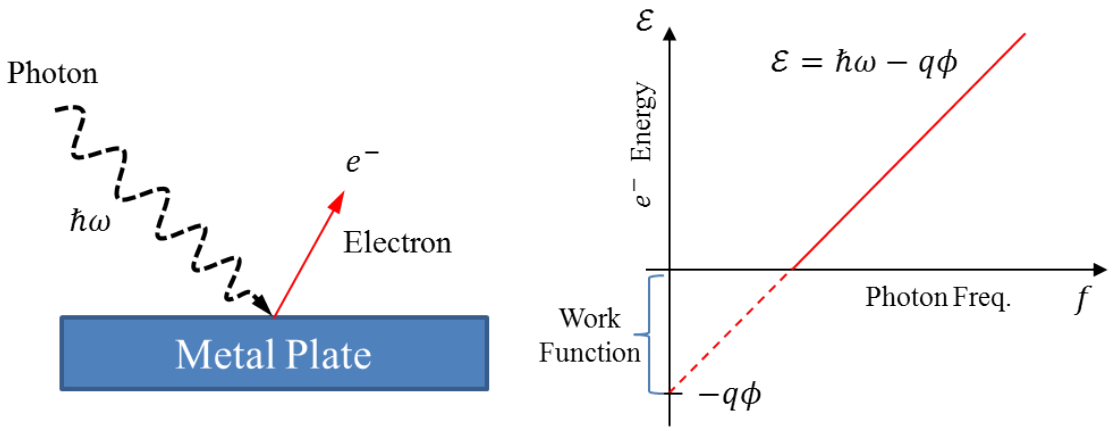


Figure 2.1: Left diagram: Electrons are emitted when light strikes the material. Right graph: Energy of the electrons as a function of frequency (energy) of light striking it.

5. Measurement also showed that the number of electrons per unit time emitted (electric current) for a fixed frequency (color) of light was proportional to the intensity of the incident light.

When the energy of the emitted electrons was plotted versus frequency of light a straight line was generated and the y-intercept was at a negative value. Mathematically this was given by:

$$\mathcal{E} = hf - \mathcal{E}_0 \tag{2.1}$$

or written in terms of angular frequency ω and the work function as defined below:

$$\mathcal{E} = \hbar\omega - q\phi \tag{2.2}$$

Where h = Planck's constant $= 6.63\times10^{-34} Jsec = 4.14\times10^{-15} eVsec$, $\hbar = \frac{h}{2\pi}$, f is the frequency of the light, and $\omega = 2\pi f$.

The y-intercept of the energy versus frequency plot, $\mathcal{E}_0 = q\phi$, is defined as the **Work function**, which is the minimum energy required to remove an electron from an atom.

Because there is a minimum photon energy required to remove an electron, that means there is also a corresponding minimum photon frequency and wavelength. The wavelength λ can be calculated from the frequency f using the relation $c = \lambda f$, where c is the speed of light.

Conclusions about Quantum Mechanics (QM) and the particle nature of light:

1. Light is composed of wave packets or particles of light called photons. A single electron is emitted when the material absorbs a single photon. Energy from the photon is transferred to the electron.

2. Light intensity is given by number of photons present. That is why the increase in light intensity gives rise to more electrons emitted per unit time.

3. Energy of photon ($\hbar\omega$) is proportional to light f (frequency), and is governed by the following fundamental relationship:

$$\boxed{\mathcal{E} = hf \qquad or \qquad \mathcal{E} = \hbar\omega} \tag{2.3}$$

Example 2.1:
UV light of frequency $f = 3\times10^{15}s^{-1}$ strikes a nickel plate with work function $\mathcal{E}_0 = 5eV$. What is the energy of the emitted electrons? What if the incoming light had $f = 1 \times 10^{15}Hz$ instead?

First calculate the energy of the incoming light:

$$\mathcal{E}_{photon} = hf = 4.14 \times 10^{-15}eVs \cdot 3 \times 10^{15}s^{-1} = 12.4eV$$

Then subtract the work function of the metal to determine extra energy given to the emitted electron:

$$\mathcal{E} = hf - \mathcal{E}_0 = 12.4eV - 5eV = \mathbf{7.4eV}$$

If the frequency is $f = 1 \times 10^{15}s^{-1}$:

$$\mathcal{E}_{photon} = hf = 4.14 \times 10^{-15}eVs \cdot 1 \times 10^{15}s^{-1} = 4.14eV$$

The photon does not have enough energy to free an electron from the metal.

Intensity of Light

As we discussed above, for electrons that are ejected from a metal surface after being hit by light, the energy of the emitted electron only depends on the frequency of the incident light, and not the intensity. However, if we increase the intensity of the light, and keep the frequency or wavelength fixed, then more electrons are emitted with the same energy. This leads to the quantum idea that light intensity is proportional to the number of photons flowing in the light beam. Quantitatively, the particle nature of light gives rise to the following definition of intensity I_L.

$$I_L = N_{ph}\hbar\omega \tag{2.4}$$

Where N_{ph} is the number of photons flowing across a unit area of surface per unit time or photon flux. So N_{ph} is typically in units of $1/m^2sec$ and if $\hbar\omega$ is in units of $Joules$, then the intensity is in units of $Joules/m^2sec$ or $Watts/m^2$.

It is also worth noting that from electromagnetics, the intensity of an electromagnetic wave is classically given by $\frac{1}{2}c\epsilon_o E^2$, where c is the speed of light, ϵ_o is the permittivity of free space, and E_o is the amplitude of the electric field of the electromagnetic wave. Thus, we can conclude that classically, the intensity of an electromagnetic wave is proportional to the square of the amplitude, and quantum mechanically, the intensity is proportional to the photon flux.

2.3 Atomic Spectra

When gases are electrically excited, they emit light at specific frequencies (called the emission spectra of the gas).

2.3.1 Hydrogen Spectrum

An electron in an excited hydrogen atom jumps from one orbit to a lower orbit releasing energy at specific wavelengths. When hydrogen gas was put in a tube and electric current ran through the tube, the H-spectra was observed to have the following discrete set light frequencies, shown below in Equations 2.5-2.7. The discrete frequencies of the light obeyed certain relationships which were governed by reciprocals of integers squared. The relationships between the sets of spectral lines are classified as a set of light emission series. These are called the Lyman, Balmer and Paschen series, where the set of observed series of light colors are named after scientists that observed them.

- Lyman Series (Ultraviolet):

$$f = cR\left[\frac{1}{1^2} - \frac{1}{n^2}\right], \qquad n = 2, 3, 4.... \tag{2.5}$$

- Balmer Series (Visible):

$$f = cR\left[\frac{1}{2^2} - \frac{1}{n^2}\right], \qquad n = 3, 4, 5.... \tag{2.6}$$

- Paschen Series (Infrared):

$$f = cR\left[\frac{1}{3^2} - \frac{1}{n^2}\right], \qquad n = 4, 5.... \tag{2.7}$$

Where f is the frequency of the emitted spectral line, R is the Rydberg constant ($R = 109,678/cm$) and c is the speed of light ($c = 3 \times 10^{10} cm/sec$).

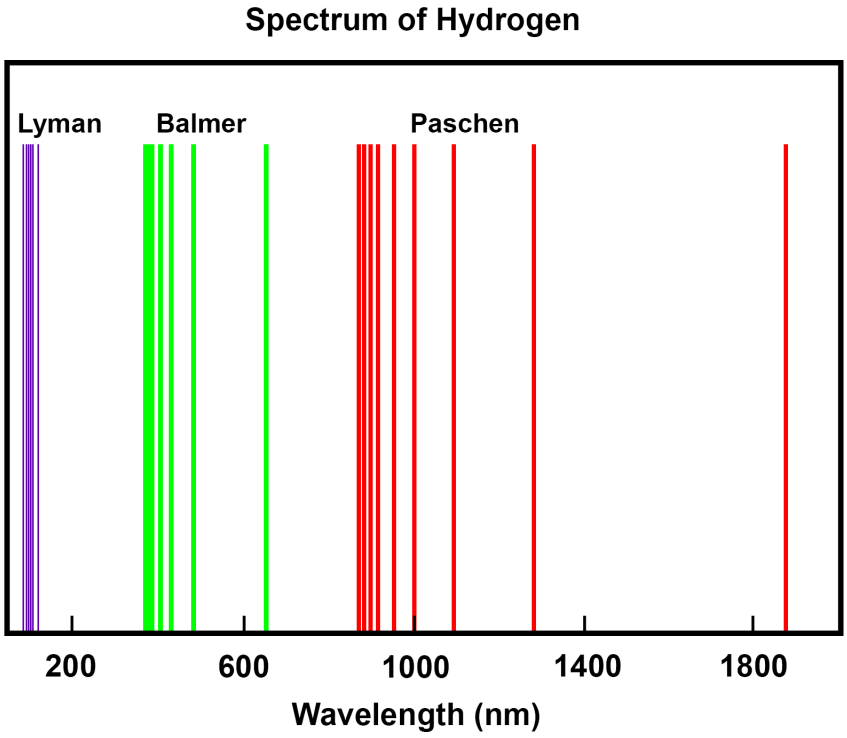


Figure 2.2: Hydrogen spectrum indicating the Lyman (UV), Balmer (Visible) and Paschen (IR) series.

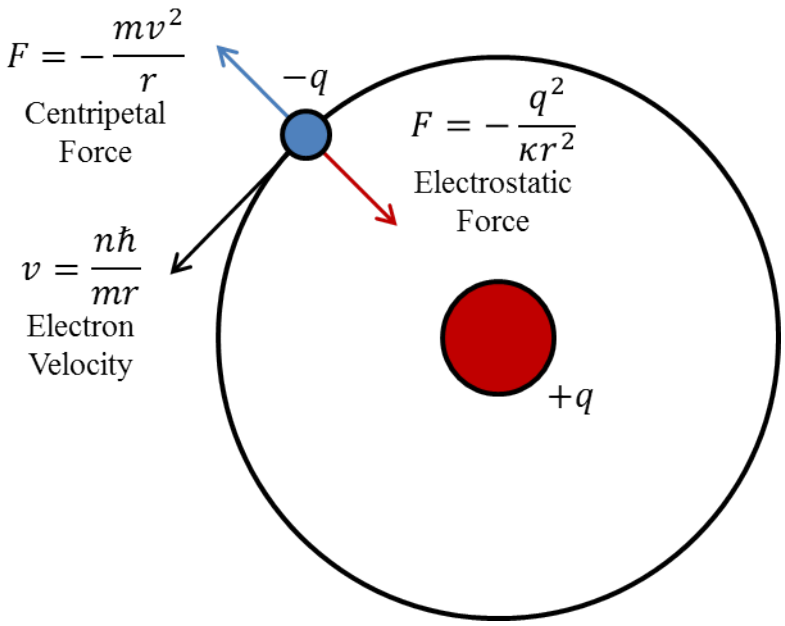


Figure 2.3: Centripetal force of orbiting electron provided by electrostatic attraction between positively charged nucleus and negatively charged orbiting electron. Note that this diagram is not at all to scale.

.

2.4 Bohr Model

The emission of light at discrete frequencies, that followed the inverse square integer relationship described above led the Danish scientist Niels Bohr in 1913 to construct a model of the hydrogen atom that explained the main spectral lines. This model became know as the 'Bohr Atom', and was the first quantized model of the atom and said that electrons could occupy only specific orbits of specific discrete values of angular momentum and energy. Also, when the electrons made transitions from a higher orbit of higher energy to a lower orbit of lower energy they emitted light, and the frequency of emitted light followed the hydrogen spectra described above. A key postulate made by Bohr was that the negative electron that orbited the positive nucleus could only have specific values of angular momentum, which were integer values of Planck's constant divided by 2π or $\hbar$. This idea of quantized angular momentum can be visualized somewhat as having an integer number of electron full de-Broglie wave-lengths fitting into an orbit.

$$mvr = n\hbar, \qquad n = 1, 2, 3, ... \tag{2.8}$$

Where mvr is the angular momentum of the electron, and m is the mass of the electron, v is the magnitude of the electron velocity and r is the radius of the electron orbit around the nucleus. Bohr said that the electrons can occupy specific

orbits only (discrete). Therefore, angular momentum is quantized.

Solving 2.8 for v gives velocity magnitude in terms of discrete values as related to angular momentum.

$$v = \frac{n\hbar}{mr}, \qquad n = 1, 2, 3..... \tag{2.9}$$

Bohr next equated centripetal force of the orbiting to the electrostatic force between the electron and the nucleus:

$$-\frac{q^2}{\kappa r^2} = -\frac{mv^2}{r} \tag{2.10}$$

where, $\kappa = 4\pi\epsilon_0$ and ϵ_0 is the permittivity of free space.

Substituting the expression for velocity magnitude from 2.9 into equation 2.10 and solving for r, gives us what is known as the **Bohr Radius when the value of integer n = 1.** As calculated, the Bohr radius should be the radius of the smallest electron orbit around the nucleus.

$$r_n = \frac{\kappa n^2 \hbar^2}{q^2 m} \tag{2.11}$$

$$\text{Bohr Radius} \equiv r_1 = 0.529\text{Å} \tag{2.12}$$

Bohr next introduced kinetic and potential energy into the model. We know that the total energy is the sum of a particle's kinetic plus its potential energy, or:

E_{Total}=Kinetic Energy (K.E.) + Potential Energy (P.E.), where

Kinetic Energy (K.E.)=$\frac{1}{2}mv^2$

Potential energy (P.E.)= $-\frac{q^2}{\kappa r_n}$

Now, using equation 2.9 for velocity, and equation 2.11 for r_n, we can express the kinetic energy as:

$$K.E. = \frac{mv^2}{2} = \frac{1}{2}\frac{n^2\hbar^2}{mr_n^2} = \frac{1}{2}\frac{q^4 m}{\kappa^2 n^2 \hbar^2} \tag{2.13}$$

Similarly for Potential energy, using equation 2.11 for the r_n, we can express the potential energy as:

$$P.E. = -\frac{q^4 m}{\kappa^2 n^2 \hbar^2} \tag{2.14}$$

Utilizing the fact that total energy is equal to kinetic plus potential, we add equations 2.13 and 2.14 to obtain:

$$\mathcal{E}_{Total} = K.E. + P.E. = -\frac{1}{2}\frac{q^4 m}{\kappa^2 n^2 \hbar^2} \tag{2.15}$$

The negative sign here means that this is a 'bound energy' (in other words, energy at infinity is the zero reference).

More important is that the energy difference between two values of 'n' (or 2 orbits) given by the following expression:

$$\boxed{\mathcal{E}_{n_2} - \mathcal{E}_{n_1} = \frac{q^4 m}{2\kappa^2 \hbar^2}\left[\frac{1}{n_1^2} - \frac{1}{n_2^2}\right]} \tag{2.16}$$

The above equation is very important because it is observable. This equation predicted and largely explained the hydrogen emission spectrum. It also helped to give more validation to the relationship of electron energy and photon frequency given in the photoelectric effect. From the photoelectric effect and the Bohr atom transition energy between orbits of equation 2.16, one can infer that the frequency of emitted light corresponds to the transition of an electron from orbit n_2 to orbit n_1 is:

$$\omega_{n_1,n_2} = \frac{1}{\hbar}\left[\mathcal{E}_{n_2} - \mathcal{E}_{n_1}\right] \tag{2.17}$$

where $\mathcal{E}_{n_2}$ and $\mathcal{E}_{n_1}$ are from equation 2.16, and where, $\omega = 2\pi f$. These emission frequencies can be translated into wavelengths using the relation $c = \lambda f$ to find the emitted light wavelengths. For the hydrogen atom, Figure 2.4 shows the emitted photons for various energy level transitions. In general, the color of a photon with any given wavelength can be found on the electromagnetic spectrum shown in Figure 2.5.

Example uses of these equations include constructing lasers, and for optical communications like the fiber network based Verizon internet network 'fios'. Also, astronomers use similar emission spectra to study the composition of different stars.

Example 2.2:
An electron in the 5^{th} excited state ($n = 6$) falls down to the 2^{nd} excited state ($n = 3$) in a hydrogen atom. What is frequency of the emitted photon?

Calculate the energy difference of the two states:

$$\begin{aligned}\mathcal{E}_{n_2} - \mathcal{E}_{n_1} &= \frac{q^4 m}{2\kappa^2 \hbar^2}\left[\frac{1}{n_1^2} - \frac{1}{n_2^2}\right] = \frac{(1.6\times10^{-19}C)^4 \cdot (9.1\times10^{-31})}{2\cdot(4\pi\epsilon_0)^2\cdot(1.054\times10^{-34}Js)^2}\left[\frac{1}{3^2} - \frac{1}{6^2}\right] \\ &= 1.13eV\end{aligned}$$

Convert photon energy to frequency:

$$f = \frac{\mathcal{E}}{h} = \frac{1.13eV}{4.14eVs} = \mathbf{2.73\times10^{14}s^{-1}}$$

Importance of Derivation and Equations 2.16 and 2.17: The picture of the atom as mostly empty space with a very solid positively charged center and

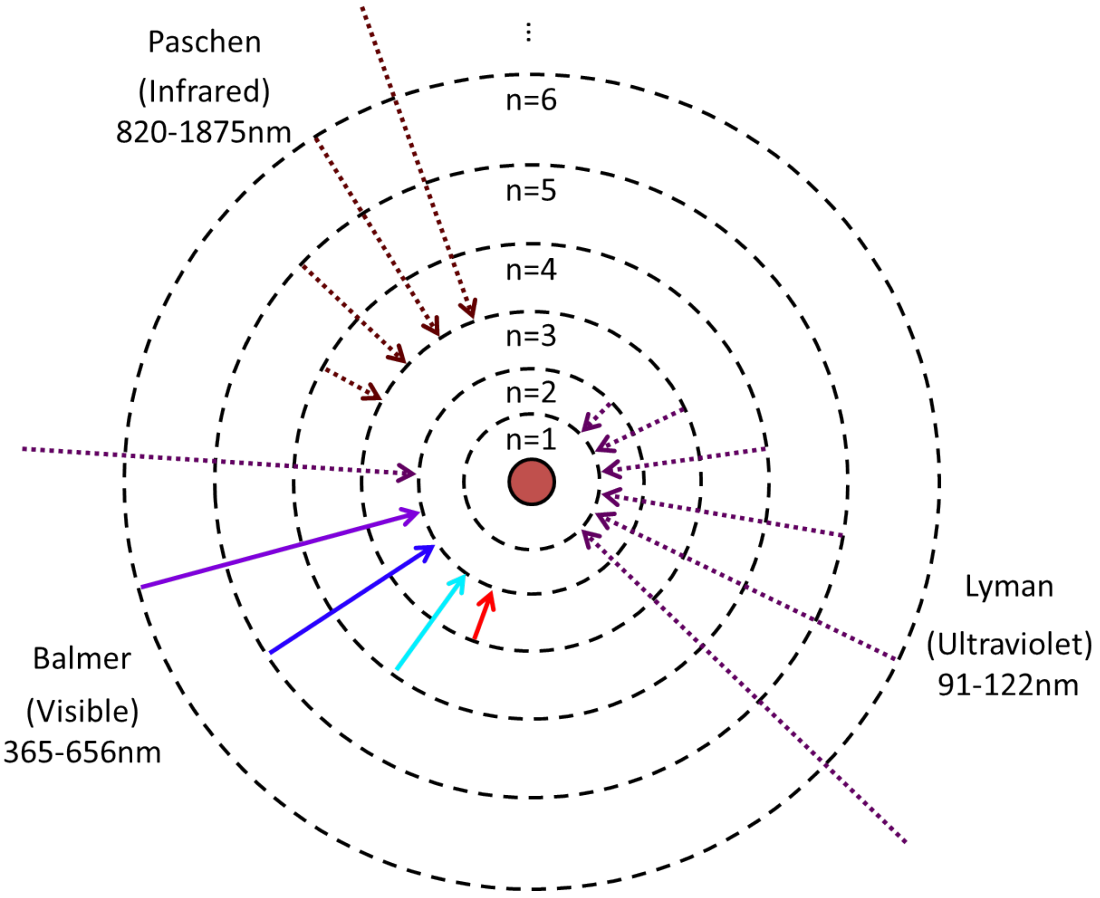


Figure 2.4: Electron transitions from higher energy orbitals to lower causes emission of photons with specific wavelengths. The dotted arrows indicate invisible colors and solid arrows show the visible colors.

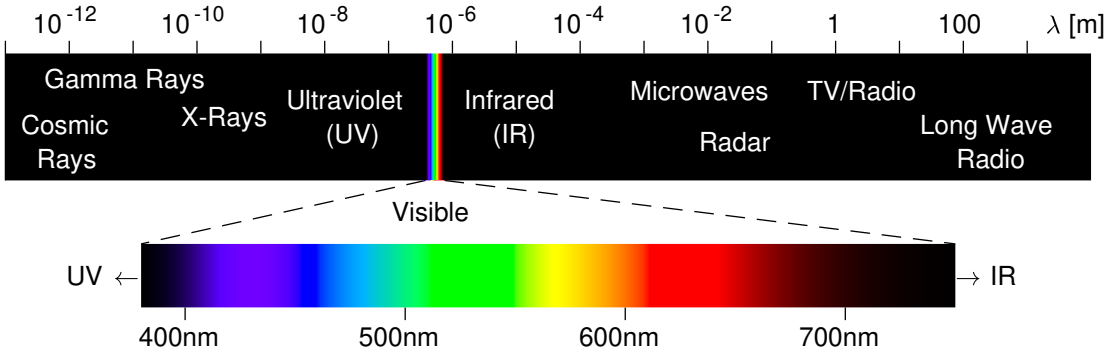


Figure 2.5: The electromagnetic spectrum.

negatively charged electrons some distance away was strengthened by Bohr's model. It gave rise to the 'planetary' picture of the atom, which consisted of a positively charged center or nucleus, and negatively charged electrons circling around at some distance away, and in very well defined orbits. Equation 2.16 gave the difference in the energies of respective orbits. While this view was later revised with a more complicated picture derived from more formal quantum mechanics, it provided a foundation for the picture and discrete energy levels of an atom that we often still utilize today.

2.5 De Broglie Wavelength

Soon after the photoelectric effect was explained by Albert Einstein, for which he received the Nobel Prize, and an atomic model was put forth by Niels Bohr, a physicist named Louis de Broglie suggested that if light could have particle-like characteristics, then objects that we typically think of as particles can have wave-like characteristics. De Broglie then presented a relationship between the wavelength λ of a particle and its momentum be given as:

$$\lambda = \frac{h}{p} \tag{2.18}$$

p= momentum of the particle, which in this text is typically an electron.

Hence, the higher the momentum of the electron, the shorter the wavelength, and greater the frequency of the matter-wave and the greater the energy of the particle. This very powerful postulate is at the heart of the "wave-particle duality" idea that we often hear discussed at parties, sports events and at the dinner table. The idea that an electron acts like a wave is the basis on which electron microscopes operate. The electrons are accelerated to very high momenta, and thereby have very small associated wavelengths and can therefore resolve features that can not be seen by light microscopes of longer wavelengths. The idea of the de Broglie wavelength was also utilized in the Bohr atom, which we will see below.

Relation to Bohr Atom

De Broglie also showed that this matter-wave and its relationship to a particle's momentum helped to explain Bohr's hypothesis that angular momentum was quantized by the relation $mvr = n\hbar$. De Broglie showed that if an integer number of wavelengths were made to fit into one of the atomic orbits, then angular momentum could be quantized using the same relationship that Bohr postulated. More specifically,

$$2\pi r = n\lambda = n\frac{h}{p} = n\frac{h}{mv} \tag{2.19}$$

Now, re-arranging gives

$$mvr = n\frac{h}{2\pi} = n\hbar \tag{2.20}$$

Thus, by allowing only integer numbers of wavelengths in an atomic orbit, De Broglie gets the same relationship as the one postulated by Bohr when he said only values of allowed angular momentum are given by an integer n times Planck's constant $\hbar$.

Example 2.3:
Let's calculate the first three Bohr radii and their associated de-Broglie wavelengths to get a feeling for the size scales that we are dealing with in the Bohr atom.
Remember that we can show the allowed orbital radii must obey $r_n = \frac{\kappa n^2 \hbar^2}{q^2 m}$ by balancing forces and using Bohr's quantized momentum postulate. Using the de-Broglie relation for momentum and wavelength $\lambda = \frac{h}{p}$ and the quantized angular momentum equation $mvr = n\hbar$, we can solve for the wavelength in terms of n.

$$p = \frac{h}{\lambda} = mv = \frac{n\hbar}{r_n}$$
$$\lambda = \frac{2\pi}{n} r_n = 2\pi \frac{\kappa n \hbar^2}{q^2 m}$$

Plugging in $n = 1, 2, 3$ for the first three orbitals we find:

$$r_1 = 0.529\mathring{A} \qquad \lambda_1 = 3.32\mathring{A}$$
$$r_2 = 2.12\mathring{A} \qquad \lambda_2 = 6.65\mathring{A}$$
$$r_3 = 4.76\mathring{A} \qquad \lambda_3 = 9.97\mathring{A}$$

The value we find here for r_1 comes out to the value defined in literature as the 'Bohr Radius' ($\mathbf{a_0}$). It is also interesting to note that the experimental bond length in a hydrogen molecule (H_2) is $0.74\mathring{A}$ and for the singly ionized molecule (H_2^+), the bond length is almost exactly twice the Bohr radius calculated by this simple formula.

Note: Intuitively, one might think that increasing energy should lead to a shorter wavelength but here we seem to have the opposite. For those curious as to why this is the case, higher energy orbitals with a larger radius move the electron farther from the atomic core which, in addition to increasing total energy, increases the potential energy and decreases the kinetic energy. This can be seen in Equations 2.13 and 2.14 where both the kinetic energy and potential energy are proportional to $1/n^2$ except the potential energy is negative so it is actually increasing with larger n as it becomes less negative. The decrease in kinetic energy corresponds to

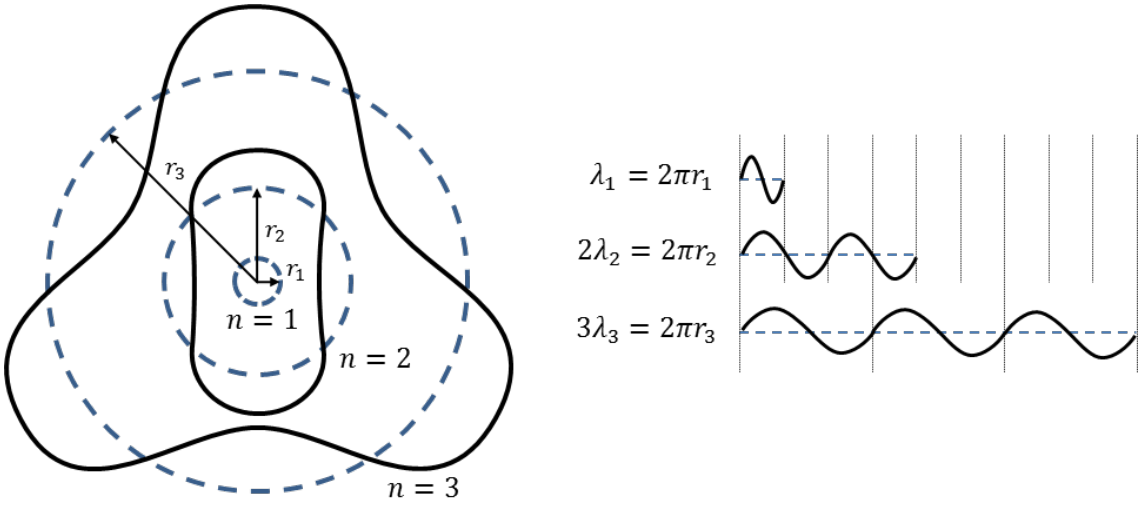


Figure 2.6: Bohr Model of the atom showing how an integer multiple of the corresponding wavelengths fits the around circumference for each energy level.

an increase in the wavelength because the two are inversely proportional which is shown below.

$$K.E. = \frac{mv^2}{2} = \frac{p^2}{2m} = \frac{h^2}{2m\lambda^2} \tag{2.21}$$

2.6 The Heisenberg Uncertainty Principle

With all these new ideas about particles and waves floating around in the 1920's, the physicist Werner Heisenberg came up with his famous Heisenberg Uncertainty Principle:

$$\boxed{(\Delta x \Delta p) \geq \frac{\hbar}{2}} \tag{2.22}$$

where,
Δx =uncertainty of particles position 'x'.
Δp =uncertainty of particles momentum 'p'.

An approximate derivation of the uncertainty principle is as follows. If an electron has wave characteristics, it will have at wavelength λ. So the electron is spread out over this wavelength, so the uncertainty in its position is $\Delta x = \frac{\lambda}{2}$.

Now, from the De Broglie wavelength we have $\lambda = \frac{h}{p}$. Solving this expression for p, and introducing the concept that since the wave is spread out over some length, then we can say that $\Delta p = \frac{h}{\lambda}$. Now, taking the product:

$$\Delta x \Delta p = (\frac{\lambda}{2})(\frac{h}{\lambda}) = \frac{h}{2} \tag{2.23}$$

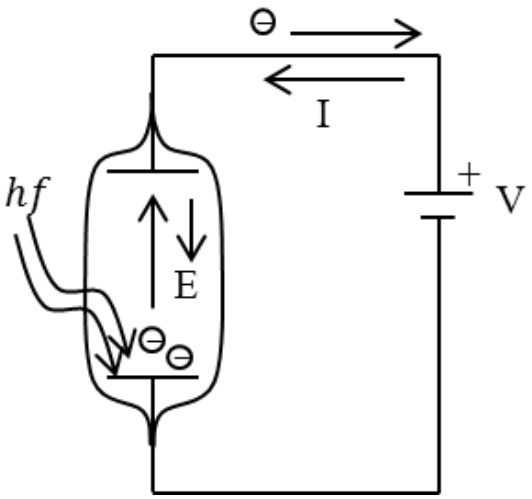
While equation (2.23) is not the exact uncertainty principle, it motivates the development of the uncertainty principle, and it can then be slightly modified to give the actual Heisenberg Uncertainty Principle of equation (2.22). The physical interpretation of the Heisenberg Uncertainty Principle is as follows. It says that since electrons have a wave like characteristics, you can not really localize it to a specific point in space, just like you can not localize a wave to a single point in space. You can only sort of say that there is a probability of finding it at a certain location. This becomes especially important when performing measurements or operations on electrons. It indicates that since the electron is spread out like a wave, if you try to perform an experiment on the electron, the experimenter will have to wind up interacting with the wavelike properties of the electron and therefore influence the experiment. This idea of knowing only the probability of somewhere finding an electron is at the heart of quantum mechanics, which we will find out in the next chapter.

2.7 Problems

2.1 Explain in your own words the difference between quantum and classical physics with respect to the physical phenomena they each tend to explain. Give three examples of phenomena explained by classical and quantum physics, respectively.

2.2 Aluminum has a work function of $4.06eV$. What is the minimum frequency of light that will liberate an electron from that metal surface? What classification of light is this (UV, IR or Visible)?

2.3 A mercury vapor lamp emits light at several wavelengths. Using a diffraction grading the $184nm$ line is separated out and is directed toward the surface of a one of the nickel electrodes that is in a tube that is transparent to this wavelength, as shown below. The intensity of the light incident onto the electrode is $0.1Watt/m^2$. The work function of nickel is $5.04eV$. (Nickel is often used in semiconductor fabrication because it does not easily oxidize.) The surface dimensions of each electrode are $5mm \times 5mm$. There is a battery attached to the electrodes as shown to establish an electric field in the tube that pulls the emitted electrons from one electrode to the other.



(a) What is the energy of the emitted electrons?

(b) What is the current in the wire?

2.4 In a Hydrogen atom, what is the energy and wavelength of the emitted photons (in Joules and eV) if electrons transition from:

(a) The 4^{th} energy level to the 1^{st} energy level?

(b) The 4^{th} energy level to the 2^{nd} energy level?

2.5 Electron microscopes work on the De Broglie wavelength of an electron. If the lenses of an electron microscope have a potential difference of $10,000V$, what is the minimum size of an object that can be resolved by the microscope.

2.6 Using the Bohr atom:

(a) What is the energy needed to ionize a hydrogen atom if the electron is in the lowest orbital (the ground state)?

(b) What is the energy needed to ionize hydrogen if the electron is in the first excited state (n=2)?

2.7 How was Bohr's atom different from classical physics? What were his postulates?

2.8 Calculate the frequency of light emitted when an electron transitions from the second lowest energy level to the lowest in a hydrogen atom. Calculate the energy required to remove an electron from a hydrogen atom when the electron is in the lowest energy level.

2.9 If you can measure the energy of an electron within an error range of 1.0eV, what is the uncertainty in the location of the electron?

2.10 In the Bohr model of the atom, the derivation is performed in a non-relativistic framework. Is this reasonable approximation? Justify your answer by calculating an estimation for the velocity of an electron in the lowest energy level for a hydrogen atom.

2.11 In a hydrogen atom, if you consider the electron to be a small particle, how far is it away from the nucleus under the following conditions:

(a) When the electron is in the first energy level (also called the ground state)?

(b) When the electron is in the second energy level?

2.12 The proton in a hydrogen atom has a radius of 8.5×10^{-16}m and a mass of 1.67×10^{-27}kg. An electron has a mass of 9.1×10^{-31}kg. Assume for these calculations that the electron can be treated as a spherical particle, not a wave.

(a) What is the ratio of the proton mass to the electron mass? (This is a famous number.)

(b) If we make the approximation that the density of the electron and proton are the same, what is the size of the radius of the electron?

(c) If the electron is in the ground state, what is the total volume of the hydrogen atom?

(d) What is the percentage of occupied space in the atom?

(e) From your answer to part (d), what can you conclude about "matter" with respect to occupied and empty space?

2.13 What is the De Broglie wavelength of the following:

(a) A free electron moving randomly at room temperature? ($27\text{C} = 300\text{K}$, $\frac{1}{2}\text{mv}^2 = \frac{3}{2}\text{k}_\text{B}\text{T}$. Recall that at $T = 300\text{K}, \text{k}_\text{B}\text{T} = 0.026\text{eV}$.)

(b) A baseball thrown by Stephen Stausburg at 100mph? (A baseball is 0.145kg.)

(c) Compare your answers for parts (a) and (b) in this problem and comment.

Chapter 3

Quantum Mechanics and The Schrodinger Wave Equation

3.1 Introduction

While Bohr's theory provided a model for the atom, it did not fully explain all the spectral lines of hydrogen. Also, his model was limited to the atomic structure, what about other small objects? For example, what physics was underlying the photoelectric effect? Why does red light pass right through certain materials, while blue light is absorbed? Why does an electron not radiate electromagnetic energy and not spiral into the nucleus like a classically accelerating charge? Other phenomena like the change in a material's heat capacity that occurred at very low temperatures could not be explained by classical thermodynamics. Newton's equations provided an entire framework for classical mechanics, Maxwell's equations provided a comprehensive theory for classical electromagnetism. A theory that provided a comprehensive framework to describe the physics of nanoscale particles, and this idea of wave particle duality, was needed by the early 20th century. This new theory came about in 1926, when the physicist Erwin Schrodinger published a paper using the concept of the particle's **Wave Function** and introduced the **Schrodinger Wave Equation**.

3.2 Key Concepts of Quantum Mechanics

The concepts and mathematics associated with Quantum Mechanics are a little different than classical physics and takes a little getting used to. Below we

summarize some of these key ideas.

The Wave-Function: The concept of the particle's "Wave-Function" is at the core of the quantum mechanics. To quantify the properties of a small particle, like an electron, you typically determine its wave-function. The concept of the wave-function is in line with the idea that a small physical object like an electron will have both particle and wave characteristics. The wave-function of different electrons in general will not be the same. The wave function of an electron is determined by the environment the electron is in. For example, the wave-function in the first orbit of the Bohr atom will be different than the wave function for the electron if it is in the second orbit of the Bohr atom. Also, the electrons that conduct electricity in a metal will have different wave-functions than the electrons in isolated atoms in a gas. From a semiconductor perspective, the wave-functions of electrons in a MOSFET will be different from the wave functions for electrons in a Bipolar Junction Transistor (BJT). The differences between electron's wave function in different environments is due to the differences in the potential energy of the system that the particle is in.

Wave-Function and the Schrodinger Wave Equation: One obtains the wave-function by solving the Schrodinger Wave Equation, which is typically shortened and referred to as the Schrodinger Equation. The Schrodinger equation is the governing equation in quantum mechanics and is analogous to $\vec{F} = m\vec{a}$ in classical mechanics and Maxwell's equations in electromagnetics. This will become much more evident in the sections below.

Quantum Mechanics and Probability: A very important aspect of Quantum Mechanics is that for most of the particle's physical attributes, you can not ascribe exact values, only probabilities that they have those values. For example, you cannot tell exactly where an electron is at a given instant in time, you can only tell the probability of a particle being in that region. Similarly, you cannot typically say exactly what the momentum of a particle is, you can only determine the probability of finding a particle around that momentum or you can provide the average or expected value of momentum. This concept of probability is in line with the concept that a small particle has wave-like characteristics when one recalls that a wave is spread over a large region of space and, in contrast to a particle, one does not typically say that a wave is located at a specific location.

Quantum Mechanics and Discrete Energy Levels: Another one of the characteristics of quantum mechanics is that under many conditions, a particle can only have specific values of energy. More specifically, when a particle is bound by some potential energy, it can only have specific energy values. In other words, it can only have specific values or 'quanta' of energy. For example, an electron in an atom can only take on specific energy values, that are determined by the pull of the nucleus. It cannot take on a continuum of values. Having only allowed values is another concept that is at the core of quantum mechanics.

Mathematics of Quantum Mechanics: Instead of largely based on vector calculus like classical mechanics and classical electromagnetics, quantum mechanics relies on the mathematics of partial differential equations, eigenvalue equations, differential operators, linear algebra and orthogonal functions.

3.3 Mathematical Description

$\Psi(x, y, z, t)$ is the wave-function and it is essentially a complete description of the physical nature of the small particle. In other words, once you know the wave-function of the particle, from it you can pretty much extract any information that you might need to know about the characteristics of the particle in a particular system. Calling it the wave-function comes from the idea that electrons can have both wave-like and particle-like characteristics. The wave-function depends on the independent variables in space and time x, y, z and t, and has been traditionally written as the Greek letter Psi (Ψ).

Probability of Finding the Particle

While Ψ is the fundamental small particle function, the observable physics comes from working with the properties of the wave-function, especially its square magnitude $|\Psi|^2$. It is probably best to just list these properties and then offer some explanation.

The square magnitude of the wave-function gives the probability density of finding a particle at the point (x, y, z, t)

$$\Psi^* \cdot \Psi = |\Psi|^2 = |\Psi(x, y, z, t)|^2 \tag{3.1}$$

where Ψ^* is the complex conjugate of Ψ.
It follows that the probability of finding an electron in the small volume element $dxdydz$ around the point (x, y, z) at time t is:

$$Probability = |\Psi(x, y, z, t)|^2 dxdydz \tag{3.2}$$

The integral over all space of the probability density is equal to one, which says that the particle must exist somewhere.

$$\int_{-\infty}^{\infty} |\Psi(x, y, z, t)|^2 dxdydz = 1 \tag{3.3}$$

	Classical Physics	Quantum Mech.
Position	x	x
Function	f(x)	f(x)
Momentum	p	$-j\hbar\frac{\partial}{\partial x}$
Kinetic Energy	$\frac{p^2}{2m}$	$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}$
Total Energy	$\mathcal{E}$	$j\hbar\frac{\partial}{\partial t}$

Table 3.1: Table comparing classical physics with quatum mechanics.

Example 3.1:
Say you have an electron in a state with normalized wave-function $\Psi(x) = \sqrt{\frac{2}{L}}\sin\left(\frac{2\pi}{L}x\right)$ defined on $(0 \leq x \leq L)$. What is the probability of finding the electron in the center half of the region?

To find the probability in this region, we simply integrate the probability density from $L/4$ to $3L/4$.

$$\int_{L/4}^{3L/4} |\Psi(x)|^2\, dx = \frac{2}{L}\int_{L/4}^{3L/4} \sin^2\left(\frac{2\pi}{L}x\right) dx = \frac{2}{L}\left[\frac{x}{2} - \frac{L\sin\left(\frac{4\pi}{L}x\right)}{8\pi}\right]_{L/4}^{3L/4} = \mathbf{\frac{L}{4}}$$

Now we must get an idea of the quantum world representation of physical quantities. To do this, we use *differential operators*.

Expectation Value of Position

As mentioned above, Quantum Mechanics deals with probabilities, and we usually cannot determine the precise position x or momentum p of a particle. So as a result, we usually calculate their average or "Expectation" values. The expectation value of position $\langle x\rangle$ is given by:

$$\langle x\rangle = \int_{-\infty}^{\infty} \Psi^*\, x\, \Psi dx \tag{3.4}$$

Similarly, the expectation value of any function of position $f(x)$ is given by:

$$\langle f(x)\rangle = \int_{-\infty}^{\infty} \Psi^*\, f(x)\, \Psi dx \tag{3.5}$$

Momentum Operator and Expectation Value:

Similarly, we usually cannot determine the exact momentum of a particle, so we calculate its average or expected value. Finding the average momentum p of the system is a little different from finding the expected value of position, because in quantum mechanics, momentum takes on the form of a differential operator, and not just an algebraic expression. At first this appears to be a little weird, but you will eventually get used to it.

$$\langle p \rangle = \int_{-\infty}^{\infty} \Psi^* \, \hat{p} \, \Psi dx \tag{3.6}$$

Where the momentum operator $\hat{p}$ is:

$$Momentum\ Operator = -j\hbar \frac{\partial}{\partial x} \tag{3.7}$$

$$\langle p \rangle = \int_{-\infty}^{\infty} \Psi^* (-j\hbar \frac{\partial}{\partial x}) \Psi dx \tag{3.8}$$

The operator then works on the next term in the expression. In other words, the operation performed to obtain the expected value of momentum is simply to take the derivative of the wave function with respect to position to the right of the momentum operator, and then just evaluate the integral.

There are other operators in quantum mechanics as well. Quantities that usually involve momentum in some way are typically expressed in quantum mechanics as operators. For example, kinetic energy in classical physics is $(\frac{p^2}{2m})$. Thus by analogy with the momentum operator, one can see that the in QM, kinetic energy take on the following operator form: $\hat{KE}_{op} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}$

In general for a differential operator $\hat{Q_{op}}$, we determine its expected value to be:

$$\langle \hat{Q_{op}} \rangle = \int \Psi^* \, \hat{Q_{op}} \, \Psi dx \tag{3.9}$$

Energy Operator for Time Dependent Case: We have not talked about the energy operator before, and we will see later that it typically does not play a major role in our QM analyses. The energy operator works on the variable of time. Since we typically work with time-independent phenomena to obtain key results, the energy operator does not often arise. However, energy is still extremely important at the heart of many QM calculations. In particular, the Energy Eigenvalue has great importance which will become more meaningful and evident in later sections. For now, just know that the energy operator in the time dependent Schrodinger equation is given by:

$$\hat{\mathcal{E}} = j\hbar \frac{\partial}{\partial t} \tag{3.10}$$

Remember, operators have a differential form, not an algebraic form.

Example 3.2:
For the following wave-functions, calculate the expected value of position and momentum for each:

$$\Psi_1(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{\pi}{L}x\right) \qquad (0 \leq x \leq L)$$
$$\Psi_2(x) = \exp\left(jkx\right) \qquad (-\infty \leq x \leq \infty)$$

For the first wavefunction we only need to integrate over the range 0 to L because this is a confined state:

$$\langle x \rangle = \int_0^L \Psi_1^* x \Psi_1 dx = \frac{2}{L}\int_0^L x \sin^2\left(\frac{\pi}{L}x\right) dx = \mathbf{\frac{L}{2}}$$
$$\langle p \rangle = \int_0^L \Psi_1^*(-j\hbar\frac{\partial}{\partial x})\Psi_1 dx = \frac{2}{L}\int_0^L -j\hbar \sin\left(\frac{\pi}{L}x\right)\frac{\partial}{\partial x}\sin\left(\frac{\pi}{L}x\right) dx$$
$$= \frac{2}{L}\int_0^L -j\hbar \sin\left(\frac{\pi}{L}x\right)\frac{\pi}{L}\cos\left(\frac{\pi}{L}x\right) dx = \mathbf{0}$$

As it turns out, this first wave-function is a type of confined or **bound** state. Due to symmetry, an electron in this state is expected to be in the center of the confining region - as one might expect. Also due to symmetry, the electron's expected momentum is zero. This can be thought of as the electron 'bouncing' back and forth inside the well i.e. it has equal and opposite components of $\pm$ momentum - the average of which is no net momentum.

For the second wave-function, we have a state which extends over all space. In this case, the state is said to be 'non-normalizable' because $\int_{-\infty}^{\infty} \Psi_2^* \Psi_2 dx = \infty$. Because the state is not normalized, the expectation value formula must be divided by the normalization integral which we will introduce here in the form of a limit.

$$\langle x \rangle = \lim_{L->\infty} \frac{\int_{-L}^{L} \Psi_2^* x \Psi_2 dx}{\int_{-L}^{L} \Psi_2^* \Psi_2 dx} = \lim_{L->\infty} \frac{1}{2L}\int_{-L}^{L} \exp(-jkx) x \exp(jkx) dx$$
$$= \lim_{L->\infty} \frac{1}{2L}\int_{-L}^{L} x dx = \mathbf{0}$$
$$\langle p \rangle = \lim_{L->\infty} \frac{\int_{-L}^{L} \Psi_2^*(-j\hbar\frac{\partial}{\partial x})\Psi_2 dx}{\int_{-L}^{L} \Psi_2^* \Psi_2 dx} = \lim_{L->\infty} \frac{-j\hbar}{2L}\int_{-L}^{L} \exp(-jkx)\frac{\partial}{\partial x}\exp(jkx) dx$$
$$= \lim_{L->\infty} \frac{-j\hbar}{2L}\int_{-L}^{L} \exp(-jkx) jk \exp(jkx) dx = \frac{1}{2L}\int_{-L}^{L} \hbar k dx = \frac{\hbar k 2L}{2L} = \hbar\mathbf{k}$$

This type of wave-function is known as a **plane wave** where we see that this state has a well defined momentum and it is traveling in the positive direction.

3.4 The Schrodinger Wave Equation

As was stated above, the Schrodinger Wave Equation (or Schrodinger equation for short), is at the heart of the formal mathematical theory of quantum mechanics. The solution of the Schrodinger equation for a specific physical environment, gives the wave function for the particle in that environment. And from the wave function, most of the particle's physical attributes can be extracted.

Just like in classical physics, where we do not derive Newton's laws which are largely determined from observation, we do not rigorously derive the Schrodinger equation. We do, however, come to the Schrodinger equation by using arguments of conservation of energy and quantum mechanical operators. We obtain Schrodinger wave equation (SWE) by using arguments of energy conservation:

$$Kinetic\ Energy\ +\ Potential\ Energy\ =\ \mathcal{E}_{Total} \tag{3.11}$$

Classically we write this as:

$$\frac{p^2}{2m} + V(x) = \mathcal{E}_{Total} \tag{3.12}$$

The next thing we do is put in the operators for the specific terms in the above statement of energy conservation, and also, we include the wave function Ψ, because it is the function that the operators act upon.

Schrodinger Wave Equation: This yields the following equation which is the Schrodinger Wave Equation:

$$\boxed{-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2} + V(x,t)\Psi(x,t) = j\hbar\frac{\partial\Psi(x,t)}{\partial t}} \tag{3.13}$$

The first term of equation 3.13 is the kinetic energy operator, operating on the wave-function. The next term reflects the potential energy and is algebraic since it is not a function of momentum, and the final term is the total energy operator, acting on the wave function.

Since we live in a 3-dimensional world, it makes sense to also write the Schrodinger equation in 3-D:

$$\boxed{-\frac{\hbar^2}{2m}\nabla^2\Psi(x,y,z,t) + V(x,y,z,t)\Psi(x,y,z,t) = j\hbar\frac{\partial\Psi(x,y,z,t)}{\partial t}} \tag{3.14}$$

Where the Laplacian operator ∇^2 is:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{3.15}$$

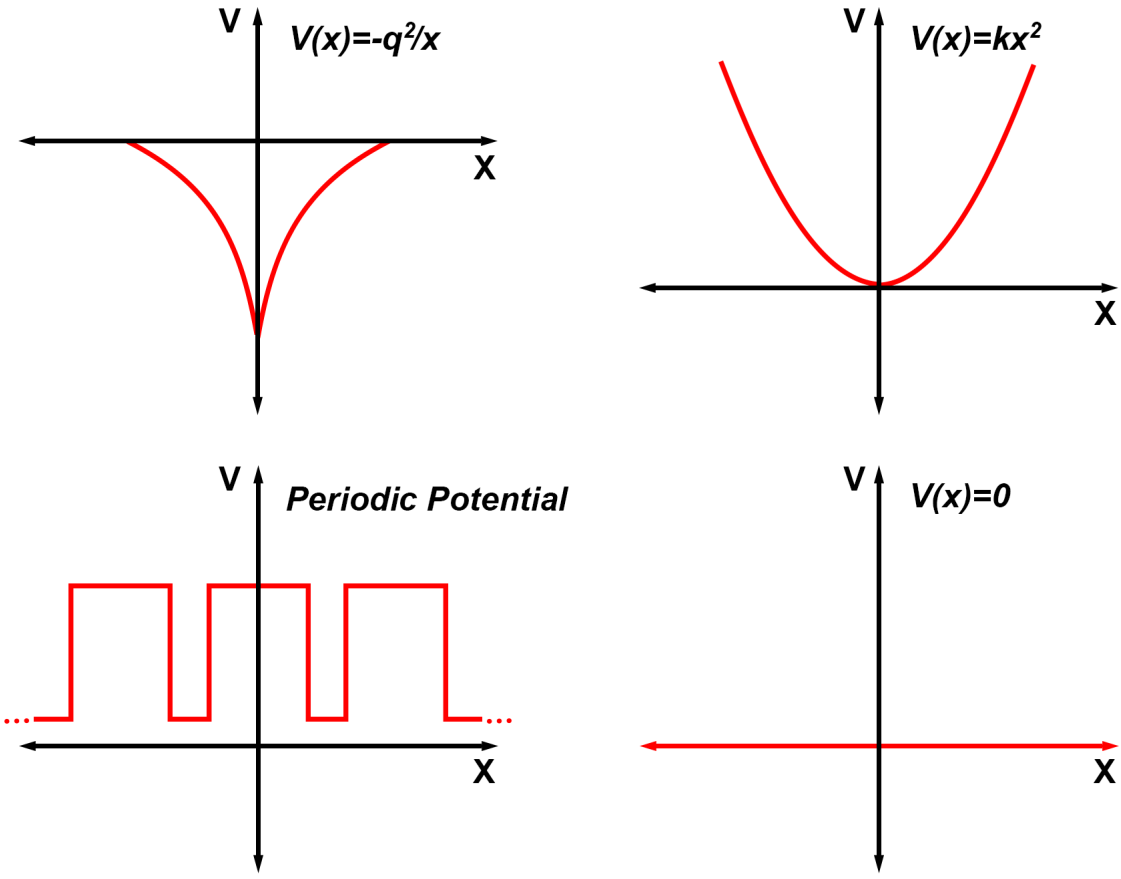


Figure 3.1: Various possibilities for potential in space.

Potential Energy $V(x, y, z)$: In one spatial dimension, we write potential energy as $V(x, t)$. The potential energy term in the Schrodinger equation is what accounts for the environment of the particle. The potential energy term is what makes the resulting wave function different for particles that have different potential energies. The potential can be a function of both space and time. However, we typically have situations where the potential term only depends on x, and not on t. We will see below that when potential energy is only a function of position, it will allow us to transform the Schrodinger equation into an eigenvalue equation that can provide discrete energy levels. So, it is the potential energy that makes each application of the Schrodinger equation distinct. Put in different physical situations (i.e. different expressions for V) in SWE and you'll get different expressions for Ψ. See the examples of various 1-D potentials in Figure 3.1.

3.4.1 Separation of Variables and the Time-Independent Schrodinger Equation

In this section we will derive perhaps the second most important equation in quantum mechanics which is the Time-Independent Schrodinger Equation. The time independent Schrodinger Equation (often referred to as just 'the Schrodinger Equation'), can be solved to provide the steady state wave-function and the allowed energy levels of the system. We will obtain the time-independent Schrodinger Equation by using the separation of variables method to transform the time-dependent Schrodinger equation, which is a single partial differential equation in space and time, into two separate equations: one equation in space and one equation in time. This can be achieved mathematically when the potential energy term only depends on space and not on time. In other words the potential energy term is $V(x)$.

With separation of variables, we re-write the solution of the Schrodinger equation as the product of two functions, one function of space and one function of time:

$$\Psi(x, t) = \psi(x)\phi(t) \tag{3.16}$$

Where $\psi(x)$ depends on space or x only, and $\phi(t)$ depends on time or t only. We can refer to $\psi(x)$ as the time-independent wave function or often just as the wave-function.

Now substitute the product in equation 3.16 into the Schrodinger equation 3.13:

$$-\frac{\hbar^2}{2m}\frac{\partial^2(\psi(x)\phi(t))}{\partial x^2} + V(x)(\psi(x)\phi(t)) = j\hbar\frac{\partial(\psi(x)\phi(t))}{\partial t} \tag{3.17}$$

Since the first term of equation 3.17 is independent of t, and the third term is

independent of x, we therefore obtain:

$$-\phi(t)\frac{\hbar^2}{2m}\frac{\partial^2\psi(x)}{\partial x^2} + V(x)(\psi(x)\phi(t)) = j\psi(x)\hbar\frac{\partial\phi(t)}{\partial t} \tag{3.18}$$

Multiplying eqn. 3.17 by $\frac{1}{\phi(t)\psi(x)}$, we get:

$$-\frac{1}{\psi(x)}\frac{\hbar^2}{2m}\frac{\partial^2\psi(x)}{\partial x^2} + V(x) = j\frac{1}{\phi(t)}\hbar\frac{\partial\phi(t)}{\partial t} \tag{3.19}$$

We have now obtained an important intermediate result which is that the left hand side (LHS) of equation 3.19 depends only on x and the right hand side (RHS) depends only on t. Separation has been achieved. (Note, that such separation of variables would not have been possible if the potential also depended on t as well as x.)

Now comes a part which might take a little thought. In equation 3.19 we have two independent variables, x and t, and they can thus take on any value. Also, in equation 3.19 the LHS and the RHS always have to be equal no matter what the values of x and t are assigned to be. The only way to ensure this is true is to have both sides of equation 3.19 equal to the same constant. So, LHS and RHS have to be equal to a constant, say $\mathbb{E}$, which we will call the **separation constant**. (Later we will find out that the separation constant is also the eigenvalue of the equation, which turns out to take on specific values which are the allowed energy levels).

Separating the LHS from the RHS and setting them both equal to the same separation constant $\mathbb{E}$ gives:

$$-\frac{1}{\psi(x)}\frac{\hbar^2}{2m}\frac{\partial^2\psi(x)}{\partial x^2} + V(x) = \mathbb{E} \tag{3.20}$$

$$j\frac{1}{\phi(t)}\hbar\frac{\partial\phi(t)}{\partial t} = \mathbb{E} \tag{3.21}$$

Now, re-arranging gives our final expressions which are the two separated Schrodinger equations, which are eigenvalue equations, one in space and one in time:

$$\boxed{-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = \mathbb{E}\psi(x)} \tag{3.22}$$

$$j\hbar\frac{d\phi(t)}{dt} = \mathbb{E}\phi(t) \tag{3.23}$$

(Note: After separation, $\frac{\partial}{\partial x}$ has been replaced with $\frac{d}{dx}$ since we are only working with one independent variable in each equation.)

While we did our analysis in 1-D for our space coordinate, it follows that if our potential varies in three dimensions in space, then the 3-D Time-Independent

Schrodinger equation is given by:

$$\boxed{-\frac{\hbar^2}{2m}\nabla^2\psi(x,y,z)+V(x,y,z)\psi(x,y,z)=\mathbb{E}\psi(x,y,z)} \tag{3.24}$$

Time Dependent Part: The time only part equation 3.23 is a simple first order ordinary differential equation. It can thus readily be solved to give:

$$\phi(t)=Ce^{-j\frac{\mathbb{E}t}{\hbar}} \tag{3.25}$$

The coefficient C and will be given by the boundary conditions for the specific application.

3.4.2 Physical Interpretation and Key Points of Time Independent Schrodinger Equation

Important physical outcomes:

1. $\mathbb{E} \equiv$ separation variable turns out to be the energy of that Quantum State: $\mathbb{E} \equiv \mathcal{E}$
2. $\mathcal{E}$ is the total energy of the state ($\mathcal{E} = K.E. + P.E.$).
3. Time independent Schrodinger Equation (3.22 or 3.24) is an eigenvalue equation.
4. The expression in square brackets is $\hat{\mathcal{H}}$: is a differential operator called the **Hamiltonian Operator**:

$$\hat{\mathcal{H}} \equiv \left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}+V(x)\right] \tag{3.26}$$

   where $\hat{\mathcal{H}}\psi(x)=\mathcal{E}\psi(x)$
5. $\psi(x)=$ eigenfunction, the wave function and also the eigenstate of the system.
6. $\mathcal{E}=$ eigenvalue, which is the energy of the state.
7. **Bound States:** We will see soon that if the particle is bound or held on to by the potential energy $V(x)$ (in other words if energy of the particle is less than the potential energy), then the particle is said to be bound, and the solution of the Schrodinger equation will be set of eigenfunctions $\psi(x)_i$ and each eigenfunction will correspond to an eigenvalue or eigenenergy $\mathcal{E}_i$. These are discrete energy values that correspond to the discrete **Quantum Energy Levels** of a system. For example, the discrete energy levels of and atom.

8. **Free States:** On the other hand, if the particle is not bound by the potential energy, then the particle is said to be free. In other words, if the total energy is greater than the potential energy for the entire system, the particle is considered to be free. Under these circumstances, we do not have discrete energy levels, but the system is typically characterized by a continuum of energy levels. An example of such a system is an electron microscope or a cathode ray tube.

3.4.3 Boundary Conditions for $\psi(x)$

To make sure that our solution to the Schrodinger equation corresponds to physically realistic situations, the wave function must satisfy the following set of boundary conditions:

1. At $x = -\infty$ and $x = +\infty$, $\psi(x) = 0$. This makes sense because the particle must exist somewhere in a finite region of space. Also, we must make sure that the wave function is finite, otherwise it would not be physical, so this also typically requires ψ to be zero at positive and negative infinity.

2. $\psi(x)$ must be continuous across all boundaries. This makes sense because as we look at the kinetic energy or first term of the Schrodinger equation, if the $\psi(x)$ were not continuous, then the derivative would not exist and then the wave-function could not be physical.

3. Similarly, the derivative $\frac{\partial \psi(x)}{\partial x}$ must be continuous to ensure that the kinetic energy term, which contains a second derivative, can be evaluated and thus give physical values for the wave function.

4. In regions where the potential is infinity, the wave function must be zero to keep the system finite and thus physically realistic.

3.4.4 System: Free Particle of the V(x) = 0 Potential

To begin feeling comfortable with the Schrodinger equation and the idea of the wave function, it helps to just start using it for some simple, but very informative cases. To start let's look at probably the simplest case for the Schrodinger equation, and that is for the case when the potential energy is zero or $V(x) = 0$. Under this condition, equation 3.22 becomes:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + 0 = \mathcal{E}\psi(x) \tag{3.27}$$

For convenience, we now multiply out by $\frac{-2m}{\hbar^2}$ and then make the change of variable $k^2 = \frac{2m\mathcal{E}}{\hbar^2}$ to obtain the form:

$$\frac{d^2\psi}{dx^2} = -k^2\psi(x) \tag{3.28}$$

This is an ordinary second order differential equation with constant coefficients. Since it is second order, there will be two solutions, and the solutions will be exponential functions. Also, since the LHS and RHS have different signs, the argument of the exponential is imaginary. Given these characteristics, we merely state that the general solution is given by the following:

$$\psi(x) = Ae^{jkx} + Be^{-jkx} \tag{3.29}$$

To verify that 3.29 is indeed the general solution you can substitute it back into to equations 3.27 and 3.28 and confirm that the RHS and LHS are identical algebraic expressions after performing the differentiations.

While 3.29 is the general solution, we still don't know the values of the unknown coefficients A and B, and nor the value of the argument parameter k, although we do know that it is proportional to the square root of the particle's energy. To get a little more physical insight into what the solution is saying, let's bring back the time variation with the help of equations 3.16 and 3.25 as follows:

$$\Psi(x,t) = \psi(x)\phi(t) = \left[Ae^{jkx} + Be^{-jkx}\right] Ce^{-j\frac{\mathcal{E}_t}{\hbar}} \tag{3.30}$$

or

$$\Psi(x,t) = Fe^{j(kx-\omega t)} + Ge^{-j(kx+\omega t)} \tag{3.31}$$

where we have renamed the unknown constants AC= F and BC=G, where of course F and G are also unknown constants. Also, we have made the substitution $\frac{\mathcal{E}}{\hbar} = \omega$.

Physical Interpretation: The general solution says that $\Psi(x,t)$ consists of two waves, one moving in the +x direction (with the F coefficient) and one traveling in the - x direction (with the G coefficient). The solution thus says that electron has a wave characteristics. The values of wave vector k and angular frequency ω, and the coefficients will be obtained from the boundary conditions.

Let's consider the example of an electron shot out of an electron filament with energy $\mathcal{E}_o$, and the electron is traveling in the positive x direction, and travels very far until it is collected by some electrode a very large distance L away. Where L is much much larger than the De Broglie wave length of the electron. Under these circumstances, the wave vector $k = \left[\frac{2m\mathcal{E}_o}{\hbar^2}\right]^{\frac{1}{2}}$ and the angular frequency is $\omega = \frac{\mathcal{E}_o}{\hbar}$. Also, since the electron is moving in positive x direction only, then $G = 0$, and the wave-function will be:

$$\Psi(x,t) = Fexp\left[j\left(\frac{\sqrt{2m\mathcal{E}_o}}{\hbar}x - \frac{\mathcal{E}_o}{\hbar}t\right)\right] \tag{3.32}$$

Finally, the coefficient F can be determined by requiring $\int_0^L \Psi^*\Psi(x)dx = 1$, where L is the long distance between the electrode from where the electron is emitted and the electrode where it is collected. Performing the integration gives $F^2L = 1$, so we wind up with the following form for the wave function for the free electron that is emitted with energy $\mathcal{E}_o$ and collected a distance L away.

$$\Psi(x,t) = \sqrt{\frac{1}{L}} exp\left[j\left(\frac{\sqrt{2m\mathcal{E}_o}}{\hbar}x - \frac{\mathcal{E}_o}{\hbar}t\right)\right] \tag{3.33}$$

Important: It is important to note that the length L used for the free particle here is usually taken as infinity in most quantum mechanics texts, and the wave function is left un-normalized. However, since we are engineers, we look for realistic situations. For us, L, which is the length of a cathode ray tube or an electron microscope, for example, may be a meter long, is much much larger than an electron wave length which is typically on the order of one nanometer. Thus, for all practical purposes our electrons in this example can be treated as free particles. You will find that this will not be true for the potential well cases in the following section where the wave length of the particle and the width of the well are of the same or similar orders of magnitude.

3.4.5 The Double Slit Experiment: Interference Pattern

The wave-function of an electron may seem like an abstract mathematical concept because it is not directly observable (though the magnitude squared is observable as a probability density), but the wave-like nature of a particle can be revealed experimentally. The famous experiment now known as the Double Slit Experiment was first performed by Thomas Young in 1803. It directly demonstrated the wave-like nature of photons by producing an interference pattern on a screen caused by a beam of light interacting with itself. The basic diagram of the experiment is shown in Figure 3.2. The peaks of the incoming wavefront shown as blue lines pass through the slit and diffract. Once the wave hits the double slit, each opening creates a separate set of wave-fronts which, when overlapping, interfere constructively to increase the wave intensity. In other locations, the peak of one wave interferes with the trough of another, causing destructive interference and as a result no wave. Constructive interference will occur at the screen when the path from slot **a** differs from path **b** by an integer multiple of the light wavelength. The experiment as performed using a beam of light may not be convincing as to the wave-like nature of a single particle because one might think that incident particle is merely interfering with all of the other particles simultaneously flowing through the opposite slit. Interestingly, the same pattern can be constructed by firing the particles one at a time though the double-slit and building up a distribution of where the final particles strike the screen. After many particles have been fired the

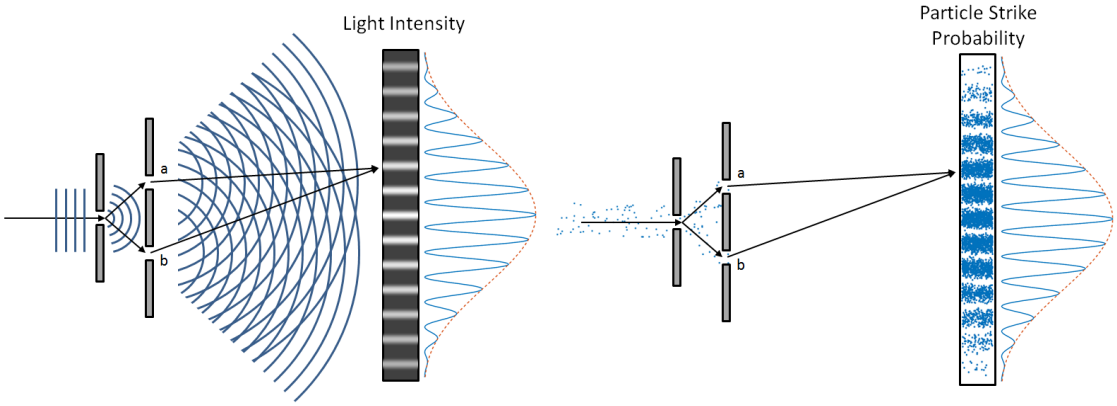


Figure 3.2: Double Slit Experiment: Left shows the intensity pattern seen when a beam of monochromatic light is shown on a double slit. Right shows the strike locations for single particles (such as electrons or photons) built up after firing the particles one at a time. The intensity/probability distribution (blue solid) follows the envelope of diffraction that would result from a single-slit (red dashed) but has the higher frequency oscillations caused by the interference.

probability distribution recovered matches the beam intensity pattern and it is clear that each particle must be able to interact and interfere with itself.

3.5 Solving the Time Independent Schrodinger Equation for Bound States

In this section we will give several important examples for solving the Schrodinger equation for some common potential energy terms. In these examples, the electrons are bound by the potential energy, and are thus not free to move anywhere. While the potentials for these examples may appear a little contrived, they do reflect real situations that actually do occur, especially in nanoscale electronics. We will see that for these bound electrons, they will only be allowed to have specific allowed energy levels.

3.5.1 System: The Particle in Infinite Potential Well

One of the most common, simple and also illustrative examples of solving the Schrodinger equation and a quantum system is the 1-Dimensional infinite potential energy well. In this situation, we have a region of zero potential energy that is bounded on either side by regions of infinitely repulsive potential energy. This is often referred to as the 1-D particle in a box problem where the particle is located

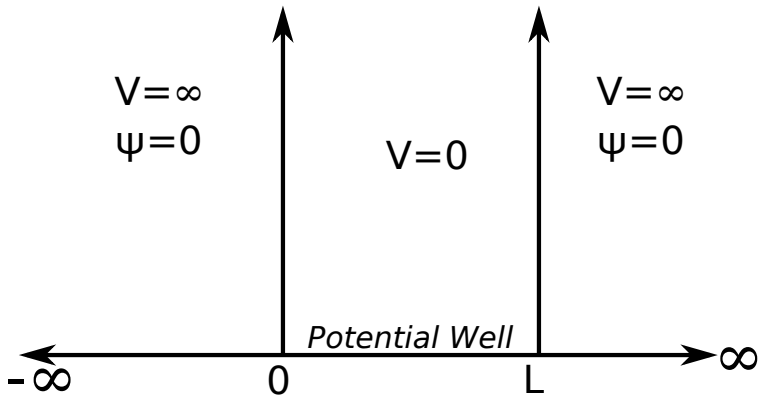


Figure 3.3: Infinite Potential Well.

in a box with walls that have infinitely high potential energy so the particle cannot escape the box. The potential is shown in Figure 3.3. While this actual potential does not really exist in nature, it is very useful for approximating many quantum systems. The general time independent Schrodinger equation for this potential energy is the following.

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = \mathcal{E}\psi(x) \tag{3.34}$$

Outside the Well

For, $x < 0$ and $x > L$, $V(x) = \infty$, then $\psi(x) = 0$ (particle does not exist in these regions). If it existed the energy of the particle would be infinite, which is not physical.

Inside the Well

Inside the well for x between 0 and L or $(0 < x < L)$, $V = 0$, so the Schrodinger equation in this region becomes

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + 0 = \mathcal{E}\psi(x) \tag{3.35}$$

This is a simple homogeneous second order differential eigenvalue equation which we solve with the following by first making the change of variables: $k^2 = \frac{2m\mathcal{E}}{\hbar^2}$ to obtain:

$$\frac{d^2\psi(x)}{dx^2} = -k^2\psi(x) \tag{3.36}$$

Since it is a second order equation we have two solutions and the general solution is the sum of the two individual solutions:

$$\psi(x) = A\sin(kx) + B\cos(kx) \tag{3.37}$$

You should verify that 3.37 is indeed a solution to differential equation by substituting it into the original the equation 3.35, performing the differentiation and ensuring that the two sides of the equation are indeed equal.

Obtaining values for unknown coefficients A, B, and the energy eigenvalues $\mathcal{E}_i$.

To obtain the values for the unknown coefficients and the energy eigenvalues, we apply boundary conditions (BCs) to equation 3.37:

1. BC1: at $x = 0$, $\psi(0) = 0$ so

$$\psi(0) = A\sin(0) + B\cos(0) = 0 \tag{3.38}$$

This means that the coefficient B must be zero, or $B = 0$ to satisfy equation 3.38

$$\psi(x) = A\sin(kx) \tag{3.39}$$

2. BC2: at $x = L,\ \psi(L) = 0$

$$\psi(L) = 0 = A\sin(kL) \tag{3.40}$$

This means that only certain values of k are allowed which satisfy equation 3.40, which are

$$k_n = \frac{n\pi}{L}, \qquad n = 1, 2, 3.... \tag{3.41}$$

So, $\psi_n(x) = A\sin(\frac{n\pi}{L}x)$. Note that we have put a subscript n with $\psi = \psi_n$ to reflect the numerous ψ's that satisfy equation.

Now, we need to find the coefficient A and use normalization to do so. To find A, recall $\int_{-\infty}^{\infty} \psi^*(x)\psi(x)dx = 1$. This yields the following value for A: $A = (\frac{2}{L})^{\frac{1}{2}}$

Exact Solution or Eigenfunctions

Finally, with the coefficient A ascertained using normalization, we have the set of wave-functions $\psi_n(x)$:

$$\psi_n(x) = \sqrt{\frac{2}{L}}\sin\left(\frac{n\pi}{L}x\right) \tag{3.42}$$

$\psi_n(x)$ are the eigenfunctions or particle wave-functions for the infinite 1-dimensional square well potential. Note, that all $\psi_n(x)$ have the same form, except for the integer n in the argument of the sine function. These reflect the different eigenfunctions that satisfy the Schrodinger equation for the given particle in a box potential. The first three wave functions for the values of $n = 1,\ 2\ 3$ are shown in Figure 3.4.

Allowed Energy Values $\mathcal{E}_n$

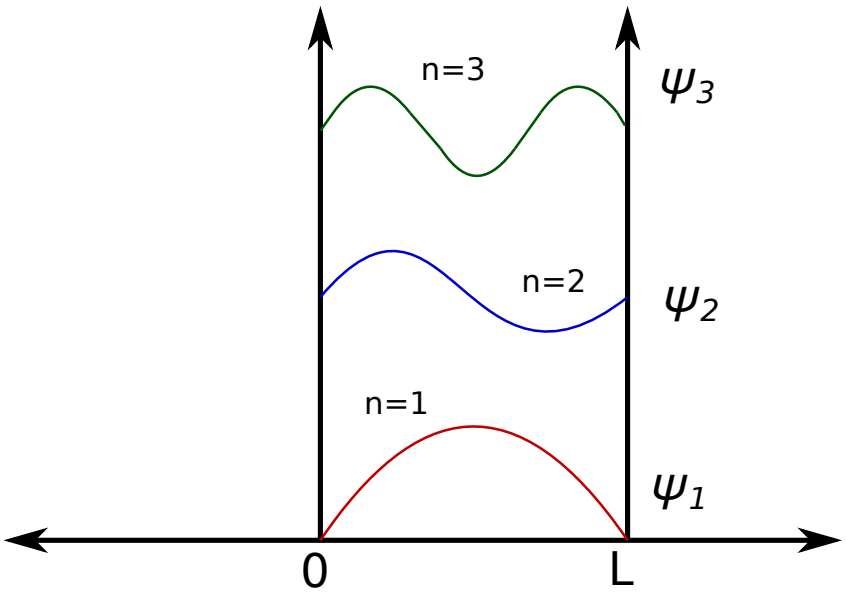


Figure 3.4: The first three wave functions for the values of $n = 1,\ 2\ 3$ for the Infinite 1-D Potential Well

Substituting back the energy as described by k, we have the following allowed value for energy, or the energy eigenvalues: $\mathcal{E}_n = \frac{\hbar^2 k_n^2}{2m}$ and using the allowed values of k_n gives:

$$\mathcal{E}_n = \frac{\hbar^2 \pi^2 n^2}{2mL^2} \tag{3.43}$$

$\mathcal{E}_n$ are the Eigenvalues or allowed energy values.

It is useful to observe that for large values of L, (size of room), the $\mathcal{E}_n$ for various n are close together making the energy function virtually continuous. but for small L (atom), the energy levels are separated far apart.

Additionally, the wavefunction solutions obtained here are all orthogonal (and orthonormal when normalized, as in Equation 3.42). Mathematically, this means:

$$\int_{-\infty}^{\infty} \psi_n(x)\psi_m(x)dx = \delta_{nm} \tag{3.44}$$

$$\delta_{nm} = \begin{cases} 0, & n \neq m \\ 1, & n = m \end{cases} \tag{3.45}$$

where δ_{nm} is the Kronecker delta function.

Example 3.3:
Quantum Well Lasers and Light Emitting Diodes

For a quantum well laser, or a quantum well Light Emitting Diode (LED), you generate very pure monochromatic light by constructing a nanoscale box of length $L = 1nm$ and inserting electrons inside. As electrons make transitions from the first excited energy level to the ground state, calculate the frequency of light you would expect to be emitted?

First, calculate the energy of the emitted light from making this transition:

$$\mathcal{E}_2 - \mathcal{E}_1 = \frac{\hbar^2\pi^2}{2mL^2}(2^2 - 1^2) = 1.13eV$$

Calculating the frequency of this light, recall $\omega = \mathcal{E}/\hbar$ and $f = \omega/(2\pi)$:

$$f = \frac{(\mathcal{E}_2 - \mathcal{E}_1)}{2\pi\hbar} = \frac{1.13eV}{2\pi \cdot 6.58 \times 10^{-16}eVs} = 2.73 \times 10^{14}Hz$$

This frequency corresponds to near infrared light.

3.5.2 System: Finite Potential Well

In Figure 3.5 we show a 1-Dimensional finite potential energy well. This is analogous to the previous example, except that the potential energy barriers outside the well are not infinity, but have a finite height which is equal to V_o. (Units here are typically either Joules or electron volts eV).

If the particle energy $\mathcal{E}$ is less than Vo, ($\mathcal{E} < Vo$), then quantum mechanics tells us that the particle can only have discrete values of energy $\mathcal{E}_i$and can be described by specific wave functions $\psi_i(x)$.

The other interesting thing that we will find by solving this problem is that quantum mechanics allows for the particle to enter the regions of potential energy Vo. This is totally different from classical physics which would never allow a particle that has lower energy than a blocking potential, to ever penetrate the blocking potential region. This is a fundamental QM result that we will see several times in this chapter. Furthermore, this phenomenon of barrier penetration is used for numerous electronic devices that you use daily, such as flash memory drives. It is also one of the operating mechanisms behind making electrical connections to MOSFET transistor devices. We will discuss this more later when we talk about "Tunneling".

To solve the finite well problem, we first divide our domain into three regions. We then solve the time independent Schrodinger equation in each region. We then

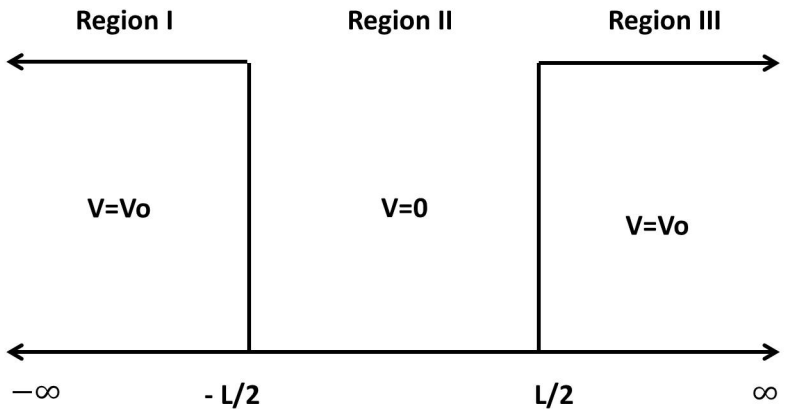


Figure 3.5: 1-Dimensional Finite Potential Energy Well

apply the boundary conditions on the wave function at (+) and (-) infinity, and at the well boundaries, $x = -\frac{L}{2}$ and at $x = \frac{L}{2}$.

We will solve the Finite Well problem for the case for which the energy of the particle is lower than the potential energy outside of the well ($\mathcal{E} < V_o$).

Region I: $(-\infty \leq x \leq -\frac{L}{2})$**,** $V = V_o$: We begin by substituting the potential V_o into the Schrodinger equation,

$$-\frac{\hbar^2}{2m}\frac{d^2\psi_I(x)}{dx^2} + V_o\psi_I(x) = \mathcal{E}\psi_I(x) \tag{3.46}$$

and then we re-arrange to get the following:

$$\frac{d^2\psi_I(x)}{dx^2} = \frac{2m(V_o - \mathcal{E})}{\hbar^2}\psi_I(x) \tag{3.47}$$

Making the change of variables for convenience we define K_I as the following expression. Also, since $V_o > \mathcal{E}$, K_I is a real number.

$$K_I = \sqrt{\frac{2m}{\hbar^2}(V_o - \mathcal{E})} \tag{3.48}$$

Since K_I is real, and the both sides of equation 3.47 have the same sign, and it is a second order differential equation, the general solution to 3.47 is the sum of two real exponentials.

$$\psi_I(x) = Ae^{K_I x} + Be^{-K_I x} \tag{3.49}$$

Where A and B are the unknown coefficients, which can be determined by applying boundary conditions that are specific for this problem.

Region II ($-\frac{L}{2} \leq x \leq \frac{L}{2}$), $V = 0$**:** In this region the potential is zero, which gives rise the following form of the Schrodinger equation.

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi_{II}(x) = \mathcal{E}\psi_{II}(x) \tag{3.50}$$

Since the equation is second order and the LHS and RHS are of opposite sign, the solutions are complex exponentials, which can also be expressed as sinusoidals: Thus,

$$\psi_{II}(x) = C\cos(K_{II}x) + D\sin(K_{II}x) \tag{3.51}$$

where the unknown coefficients C and D can be obtained by the boundary conditions, and K_{II} is:

$$K_{II} = \sqrt{\frac{2m}{\hbar^2}\mathcal{E}} \tag{3.52}$$

Region III: ($\frac{L}{2} \leq x \leq \infty$), $V = V_o$**:** This is the same as Region I, but with different coefficients:

$$\psi_{III}(x) = Fe^{K_{III}x} + Ge^{-K_{III}x} \tag{3.53}$$

$$K_{III} = \sqrt{\frac{2m}{\hbar^2}(V_o - \mathcal{E})} \tag{3.54}$$

Boundary Conditions (BCs): As we discussed in Section 3.4.3, the wave function must go to zero at infinity and it must be continuous across boundaries. Also, as required by the Schrodinger equation, the first derivative of the wave function must also be continuous across boundaries. These lead to following BCs:

1. ψ should be 0 at $\pm\infty$.
2. ψ is continuous at the intersection of Region I and Region II on the left, and Region II and Region III on the right,
3. $\frac{d\psi}{dx}$ is continuous at the intersection of Region I and Region II on the left, and Region II and Region III on the right,

Applying the BC at x=$-\infty$ that $\psi_I(-\infty) = 0$, we have $B = 0$. Similarly, using the BC at x=$+\infty$ that $\psi_{III}(+\infty) = 0$, we have $F = 0$. Rearranging the equations, we have

$$\psi_I(x) = Ae^{K_I x} \tag{3.55}$$

$$\psi_{II}(x) = C\cos(K_{II}x) + D\sin(K_{II}x) \tag{3.56}$$

$$\psi_{III}(x) = Ge^{-K_{III}x} \tag{3.57}$$

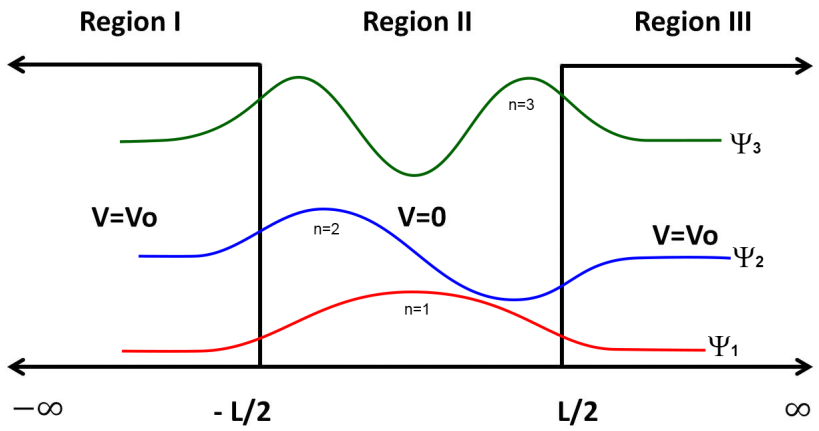


Figure 3.6: The first three wave functions (eigenfunctions) for the 1-Dimensional Finite Potential Energy Well. Notice that the wave function extends into the regions outside of the well, indicating that it is possible to find the particle outside of the well.

Using the BC's for the boundaries of $-\frac{L}{2}$ and $\frac{L}{2}$, we have

$$\psi_I(-\frac{L}{2}) = \psi_{II}(-\frac{L}{2}) \tag{3.58}$$

$$\psi_{II}(\frac{L}{2}) = \psi_{III}(\frac{L}{2}) \tag{3.59}$$

$$\left.\frac{d\psi_I(x)}{dx}\right|_{L=-\frac{L}{2}} = \left.\frac{d\psi_{II}(x)}{dx}\right|_{L=-\frac{L}{2}} \tag{3.60}$$

$$\left.\frac{d\psi_{II}(x)}{dx}\right|_{L=\frac{L}{2}} = \left.\frac{d\psi_{III}(x)}{dx}\right|_{L=\frac{L}{2}} \tag{3.61}$$

Also the overall wave function must be normalized to 1 since the particle exists somewhere, or $\int_{-\infty}^{\infty} \psi^*(x)\psi(x)dx = 1$. Applying the normalization requirement to the finite well system gives:

$$\int_{-\infty}^{-\frac{L}{2}} \psi_I^*(x)\psi_I(x)dx + \int_{-\frac{L}{2}}^{\frac{L}{2}} \psi_{II}^*(x)\psi_{II}(x)dx + \int_{\frac{L}{2}}^{\infty} \psi_{III}^*(x)\psi_{III}(x)dx = 1 \tag{3.62}$$

The wave function solutions 3.55 though 3.57 contain four unknown coefficients (A, C, D and G). They also have in them allowed energy values which are contained in K_I, K_{II} and K_{III}. To solve for the unknown coefficients, we would next apply five conditions given by equations 3.58 to 3.62 to obtain the four unknown coefficients and the allowed energy eigenvalues $\mathcal{E}_n$. This winds up being a lengthy exercise, the results of which are provided in Appendix A. We illustrate the results that you would obtain in Figure 3.6. The energy eigenvalues you would get are in a form that is similar to those of the infinite well as given by equation 3.43, but generally

have slightly lower values for a given solution integer n. Physically, this result makes sense because the system is very similar to the infinite well example, except that the wave function can penetrate or 'leak' into the high potential regions. This broadened wavefunction has the effect, in a sense, of increasing the effective "length" of the box, thus reducing the magnitudes of the allowed energy eigenvalues.

For an example finite well with $V_o = 5eV$ and $L = 2nm$, the first six allowed energies are compared to those in an infinite well of the same length below:

	1	2	3	4	5	6
Finite:	0.0795 eV	0.3175 eV	0.7128 eV	1.2626 eV	1.9625 eV	2.8038 eV
Infinite:	0.0940 eV	0.3761 eV	0.8462 eV	1.5043 eV	2.3505 eV	3.3848 eV

Important Observations and Results about the resulting wave function: It is important to notice that the wave function ψ is sinusoidal inside the well (RII region), and in Regions I and III ψ is given by decaying exponentials. This says that the particle can penetrate the wall of the potential energy well, even though the particle's energy is lower than the height of the potential energy barrier formed by the wall of the well. The particle can be found outside the box because $|\psi_I|^2$ and $|\psi_{III}|^2$ are not zero. This is totally in contrast with classical physics which says that if a particle has less energy than the potential barrier, it would never be able to penetrate it. However, experiments and basic electronic devices, like MOSFET transistors, that we use daily tell us that quantum mechanics gives us the correct picture.

3.6 Tunneling and Barrier Penetration

Tunneling is a purely quantum mechanical phenomenon that is very important in electronics. Tunneling is when a particle passes from one side of an energy barrier to the other side, for the case when the energy of the particle lower than the potential energy of the barrier. This is totally not allowed according to classical physics, but it happens all the time and is explained by quantum mechanics. It has many applications and is the mechanism behind various devices including the operation of flash memory sticks and solid state drives. Figure 3.7 illustrates the concept of tunneling. The top figure shows the particle wave function which depicts an electron wave function heading into the barrier. The wave then penetrates the barrier and comes out the other side. One thing to notice is that the incident wave in Region I has larger amplitude than the transmitted wave in Region III. This reflects the result that not all electrons wind up tunneling through the barrier, but some of the incident electrons are reflected back, so the probability of finding the particle in Region I is greater than finding it in Region III.

We will solve the tunneling barrier problem for the case for which the energy of the particle is lower than the potential energy of the barrier ($\mathcal{E} < V_B$). In other words, consider the case of an electron that is launched and travels in the positive x direction toward a barrier that has a blocking potential energy that is greater than the energy of which the particle is launched. To analyze this problem, we solve the Schrodinger equation in the three regions and apply boundary conditions. One thing to make sure you understand is that this is a free particle problem, not a bound particle situation.

Region I: $(x \leq 0)$**,** $V = 0$: We begin by substituting the potential $V = 0$ into the

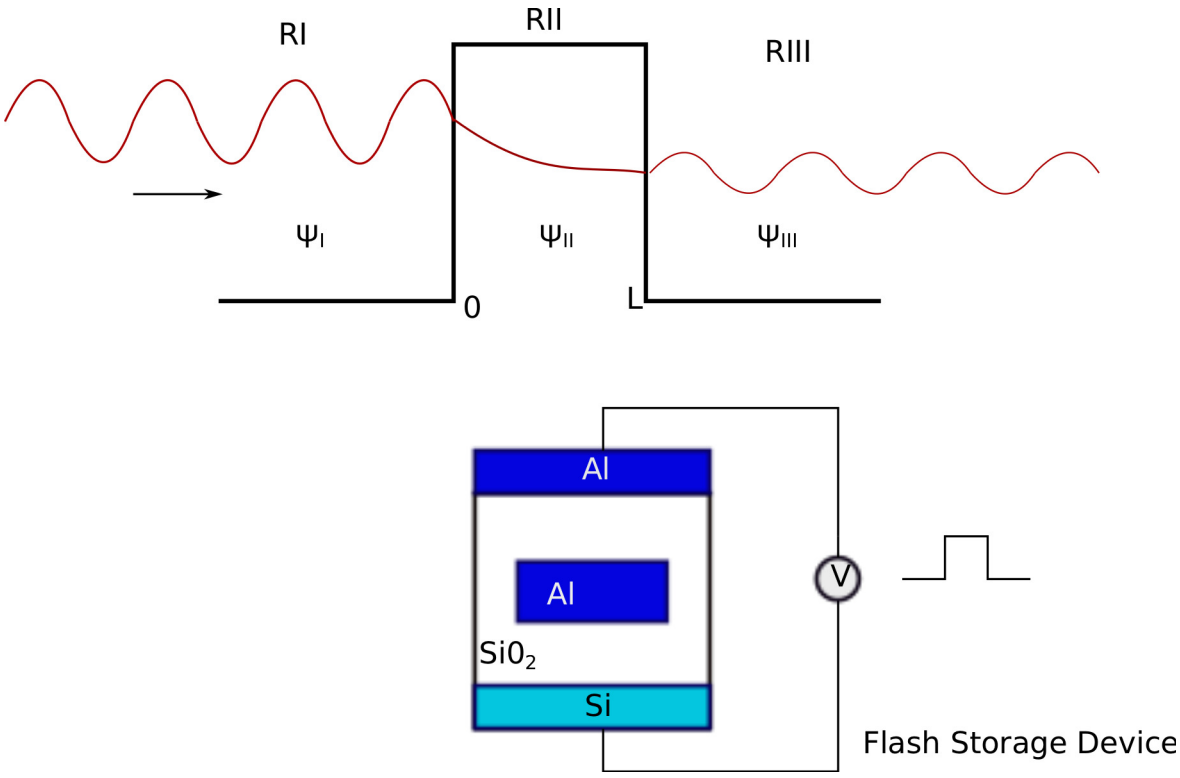


Figure 3.7: The top figure is an example wave function of a tunneling particle. The barrier is between 0 and L, and it has height of V_B units of energy (typically in Joules or eV). The bottom figure is an example of a Flash Memory Transistor that operates on the principle of tunneling.

Schrodinger equation, and rearranging to get the following:

$$\frac{d^2}{dx^2}\psi_I(x) = -\frac{2m}{\hbar^2}\mathcal{E}\psi_I(x) \tag{3.63}$$

Making the change of variables for convenience we define K_I as the following

$$K_I = \sqrt{\frac{2m\mathcal{E}}{\hbar^2}} \tag{3.64}$$

Since K_I is real, and the LHS and RHS equation 3.63 have the different sign, and it is a second order differential equation, the general solution to 3.47 is the sum of two complex exponentials and thus represents two waves: one moving in the $(+x)$ direction and the other moving in the $(-x)$ direction:

$$\psi_I(x) = Ae^{jK_Ix} + Be^{-jK_Ix} \tag{3.65}$$

Where A and B are the unknown coefficients, which can be determined by applying boundary conditions.

Region II: $(0 \leq x \leq L$**,** $V = V_B$**):** This is the region of the large potential energy barrier which has $V_B > \mathcal{E}$,

$$-\frac{\hbar^2}{2m}\frac{d^2\psi_{II}(x)}{dx^2} + V_B\psi_{II}(x) = \mathcal{E}\psi_{II}(x) \tag{3.66}$$

Making the change of variable

$$K_{II} = \sqrt{\frac{2m}{\hbar^2}(V_B - \mathcal{E})} \tag{3.67}$$

Rearranging which gives rise the following form of the Schrodinger equation.

$$\frac{d^2\psi_{II}(x)}{dx^2} = K_{II}^2\,\psi_{II}(x) \tag{3.68}$$

Since the equation is second order and the LHS and RHS are of the same sign, the solutions are real exponentials:

$$\psi_{II}(x) = Ce^{K_{II}x} + De^{-K_{II}x} \tag{3.69}$$

where the unknown coefficients C and D can be obtained by the boundary conditions.

Region III: $(L \leq x)$**,** $V = 0$**:** This is the same general wave-like as Region I, but with different coefficients:

$$\psi_{III}(x) = Fe^{jK_{III}x} + Ge^{-jK_{III}x} \tag{3.70}$$

$$K_{III} = \sqrt{\frac{2m\mathcal{E}}{\hbar^2}} \tag{3.71}$$

In Region III, the wave function represents a particle that has passed through the barrier and is traveling only in the $+x$ direction. Therefore, we can immediately set the coefficient $G = 0$, and we are left with the following wave function in Region III.

$$\psi_{III}(x) = Fe^{jK_{III}x} \tag{3.72}$$

Summarizing our results we get the following form of the wave functions for Regions I, II and III.

$$\psi_I(x) = Ae^{jK_Ix} + Be^{-jK_Ix} \quad (x \leq 0) \tag{3.73}$$

$$\psi_{II}(x) = Ce^{K_{II}x} + De^{-K_{II}x} \quad (0 \leq x \leq L) \tag{3.74}$$

$$\psi_{III}(x) = Fe^{jK_{III}x} \quad (x \geq L) \tag{3.75}$$

Physically 3.73 corresponds to the incident wave traveling toward the barrier (the first term) and then part of the wave that is reflected and thus traveling in the $-x$ direction away from the barrier (indicated by the second term). Equation 3.74 is an exponentially decaying part of the wave function the represents that the electron can enter and exist within the classically forbidden barrier. Equation 3.75 represents the fact that the particle can pass through the barrier and will continue to travel to the right (in the positive x direction). Note that we have not included the time dependence in our solution. However, the time dependence could easily be incorporated by multiplying all three of the above expressions by the time dependent part of the wave function $\phi(t) = Ce^{-j\frac{\mathbb{E}t}{\hbar}}$, as given by equation 3.25. Including the time dependence can be helpful in visualizing that we do indeed have traveling wave solutions in Regions I and III. You may want to do this on your own.

Finally, the actual energy of the particle will be set when the particle is first launched, and thus provides the values for K_I, K_{II} and K_{III}. The coefficients can be determined by the boundary conditions requiring the wave function and its derivative to be continuous across the region boundaries.

Using the BC's for the boundaries at $x = 0$ and $x = L$, we have

$$\psi_I(0) = \psi_{II}(0) \tag{3.76}$$

$$\psi_{II}(L) = \psi_{III}(L) \tag{3.77}$$

$$\left.\frac{d\psi_I(x)}{dx}\right|_{x=0} = \left.\frac{d\psi_{II}(x)}{dx}\right|_{x=0} \tag{3.78}$$

$$\left.\frac{d\psi_{II}(x)}{dx}\right|_{x=L} = \left.\frac{d\psi_{III}(x)}{dx}\right|_{x=L} \tag{3.79}$$

Now, explicitly substituting in the expressions for the wave functions (3.73), (3.74) and (3.75) and evaluating them at the boundaries ($x = 0$) and ($x = L$) as indicated above, gives the following four simultaneous equations:

$$A + B = C + D \tag{3.80}$$

$$Ce^{K_{II}L} + De^{-K_{II}L} = Fe^{jK_{III}L} \tag{3.81}$$

$$jK_IA - jK_IB = K_{II}C - K_{II}D \tag{3.82}$$

$$K_{II}Ce^{K_{II}L} - K_{II}De^{-K_{II}L} = jK_{III}Fe^{jK_{III}L} \tag{3.83}$$

Implementing the boundary conditions gives the four equations 3.80 through 3.83. However, we have five unknown coefficients: A, B, C, D and F. In our last example, which was the finite potential well, we implemented the normalization condition to give the fifth equation. We could do something similar here. However, for tunneling it is usually sufficient to not determine all the wavefunction coefficients, but to determine the all-important **Transmission** and **Reflection** coefficients **T** and **R**.

3.6.1 Tunneling Transmission and Reflection Coefficients

The transmission and reflection coefficients quantify the probability of particles being transmitted all the way through the barrier and being reflected back away from the barrier. The transmission probability is the square amplitude of the incident wave divided by the square amplitude of the fully transmitted wave, or ratios of the wavefunction coefficients as shown below. The actual determination of these coefficients, though a potentially useful exercise, is somewhat of an arduous endeavor so we will save that for when you are earning big bucks designing the next generation solid state computer memory drive. For those curious, the detailed expressions for all the transmission coefficients are given in Appendix A. Manipulating equations 3.80 through 3.83 we eliminate all but one unknown, and then by taking the ratio of $|F|^2/|A|^2$, we find the transmission coefficient.

Complete Transmission Coefficient:

$$T = \frac{|F|^2}{|A|^2} = \left(1 + \frac{V_B^2 \sinh^2(K_{II}L)}{4\mathcal{E}(V_B - \mathcal{E})}\right)^{-1} \tag{3.84}$$

Thick Barrier Approximation ($K_{II}L >> 1$): In the case where the barrier width L, is much wider than the De Broglie wavelength, the expression 3.84 can be approximated as the following product of a quadratic and a decaying exponential.

$$T = \frac{|F|^2}{|A|^2} \approx \frac{16\mathcal{E}}{V_B}\left(1 - \frac{\mathcal{E}}{V_B}\right) e^{-\frac{2L}{\hbar}\sqrt{2m(V_B - \mathcal{E})}} \tag{3.85}$$

Where we have directly substituted the expression (3.67) for the wave vector $K_{II} = \sqrt{\frac{2m}{\hbar^2}(V_B - \mathcal{E})}$ to more explicitly indicate the relationship between particle energy $\mathcal{E}$, barrier height V_B and barrier width L.

Thin Barrier Approximation ($K_{II}L << 1$): In the case where the barrier width L, is much thinner than the De Broglie wavelength, expression 3.84 can be approximated as the following simple decaying exponential, where again we substitute the explicit expression for K_{II}.

$$T = \frac{|F|^2}{|A|^2} \approx e^{-\frac{2L}{\hbar}\sqrt{2m(V_B-\mathcal{E})}} \tag{3.86}$$

Also, the reflection and transmission coefficients are related by:

$$R = 1 - T \tag{3.87}$$

Physical Interpretation: By observing the above expressions for the transmission coefficient, especially equation (3.86), for the thin barrier, we see that transmission will increase as the particle energy $\mathcal{E}$ increases toward the barrier height potential energy V_B. The transmission probability will also increase as the barrier gets wider. However, transmission will decrease as the barrier height increases. These results all make intuitive sense. The most direct application of the tunneling transmission coefficient is that it will quantify the number of particles that will actually pass through the barrier as compared to the number that are incident upon the barrier. For example, if $T = 0.05$, and a beam of $N = 20,000$ particles per second are launched or incident upon the barrier, then $T \times N = 0.05 \times 20,000 = 1,000$ particles per second on average will tunnel through and pass through to the other side. Also $(1-T) \times N = R \times N = 0.95 \times 20,000 = 19,000$ particles per second will be reflected back or bounce off the barrier and go back toward the original source.

Example 3.4:
A potential barrier of $V_B = 2eV$ is $1nm$ thick. Calculate the transmission probability for an electron incident with energy $1eV$.

$$K_{II} = \sqrt{2m(V_B - \mathcal{E})}/\hbar = \frac{\sqrt{2 \cdot 9.1 \times 10^{-31} kg \cdot (2eV - 1eV)}}{1.055 \times 10^{-34} Js} = 5.12nm^{-1}$$

$$T(1eV) = \left(1 + \frac{V_B^2 \sinh^2(K_{II}L)}{4\mathcal{E}(V_B - \mathcal{E})}\right)^{-1} = \left(1 + \frac{(2eV)^2 \sinh^2(5.12nm^{-1} \cdot 1nm)}{4 \cdot 1eV \cdot (2eV - 1eV)}\right)^{-1}$$
$$= \mathbf{1.42 \times 10^{-4}}$$

3.7 Problems

3.1 Using the method of separation of variables solve the Schrodinger Wave Equation for a free electron in one dimension. Recall that for the free electron V=0,

so that:

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2} = -\frac{\hbar}{j}\frac{\partial\Psi(x,t)}{\partial t}$$

(a) Assume the electron is moving in the positive x direction and show that the solution is that of a plane wave. Show also that for the solution to exist the following relation between E, our separation constant, and k must hold: $E = \frac{\hbar^2 k^2}{2m}$, where E is the total energy of the electron. (Note that k is the wavevector and $k = 2\pi/\lambda$ where λ the electron wavelength).

(b) Starting from the classical expression for kinetic energy, $E = \frac{1}{2}mv^2$, show that the momentum of the free electron can be written as $\hbar k$.

(c) We know the momentum operator is given by $\frac{\hbar}{j}\frac{\partial}{\partial x}$. Show that the average (or expected) momentum is given by $\hbar k$.

3.2 If an electron is in an infinite potential well of length 1.0 nanometers:

(a) Is the electron bound or unbound?

(b) Calculate the values of the first three energy levels that the electron can have in Joules and eV.

(c) Solve the Schrodinger equation to get the wave-functions for the first three energy levels.

(d) Plot the wave-functions for the first three levels.

(e) Plot the square magnitude of the wave-functions for the first three levels. Comment on the what each of these plots tell us with respect to finding the probability at a specific location in the well, for each energy level.

(f) Calculate the probability of finding the particle in the region $(0 < x < L/4)$ if the electron is in the nth energy level.

(g) Calculate the expectation value for position and momentum for a particle in the nth energy state of the well. Do your answers make physical sense? Compare this to what you would expect classically.

3.3 Using the quantum mechanical momentum operator $\hat{p} = -j\hbar\frac{\partial}{\partial x}$, derive the expression for the kinetic energy operator. (Hint: In classical physics, $K.E. = \frac{1}{2}mv^2$ and $p = mv$)

3.4 Lets say you are designing an optical communication device to be used in Fiber Optic Services (FIOS) and you want to use a higher frequency light for your optical communication.

(a) If you want to increase the frequency of the light that the electrons in the well could absorb, what could you do to the potential energy well. Explain qualitatively in a sentence or two.

(b) If your well is 1.0 nm in length, and can be approximated to have infinite potential on the sides, and you are using the first three energy levels for your optical system, what three wavelengths of light will you use for your optical communications?

3.5 A new kind of digital electronics is being developed. In this new technology, an electron is trapped in a 2-dimensional potential energy well and the different quantum states of the system are being used as different logic levels. Assume the well has infinitely high barriers on all four sides. Inside the well $V = 0$. Also, the well has length of $0.5nm$ and width of $1.0nm$.

(a) Solve the Schrodinger equation to get the wave functions of the well. (Hint, this is a 2-D problem, so use separation of variables: $\Psi(x, y) = X(x)Y(y)$, and then solve the two directions independently.)

(b) Derive and expression for the allowed energy levels.

(c) Calculate the energies of the first 3 levels in Joules and eV.

3.6 Considering a 1-D quantum well with finite barriers of height V on either side of the well:

(a) Write down the time-independent SWE for each of the three regions of the system.

(b) Solve the SWE for each of the three regions for the bound electrons, you do not have to obtain values for the unknown coefficients, except those that would keep the wave-function from going to infinity at $x = \pm\infty$.

(c) Sketch what you expect the wave-functions to look like. Remember the continuity requirements for the wave-function.

(d) Apply the boundary conditions and set up the explicit set of equations that you need to solve to determine the coefficients.

(e) What physically does it mean that the wave-function decays exponentially on either side of the quantum well. Compare this result to the infinite well case above and explain what the difference means physically.

3.7 A flash memory stick (also known as a travel-drive) contains billions of transistors for storing charge. Each charge-storing transistor gives rise to a 'bit', with a 1 or 0 depending on whether or not the charge is stored. The way charge is stored is by having electrons tunnel through an oxide barrier to charge a capacitor. When there is no voltage applied to the transistor, the barrier is too high for significant tunneling and thus charge storage does not take place. When a voltage is applied to the transistor, the tunneling barrier is lowered enough so that tunneling and charging and thus storage of 'bits' takes place.

(a) If the rectangular barrier is $1nm$ thick and $4.2eV$ high, write down the time-independent SWE for the three regions of interest. Take the potential energy on either side of the barrier to be $0eV$.

(b) Solve the SWE to obtain the wave-function in each of the three regions. Take the electron energy to be $1.1eV$, which is obviously lower than the barrier energy. You do not have to obtain values for the constant coefficients in front of the exponents.

(c) Now lower the barrier to $1.2eV$ by applying a bias of $3.0V$ to the device. How has the solution to the SWE changed? (The electron energy is still $1.1eV$). Describe quantitatively, by referring to your solutions above, the difference in the barrier penetration and thus charging between cases in parts (b) and (c).

3.8 Explain using concepts from quantum mechanics why glass is transparent.

3.9 Write down three examples of technology based on quantum mechanics that you use regularly. Explain qualitatively how their operations depend on quantum mechanical principles.

3.10 For the infinite well potential configuration shown in Figure 3.4:

(a) Confirm that the solutions given by Equation 3.42, do indeed work when substituted back into the Schrodinger Wave Equation.

(b) Show that the different solutions are orthonormal to each other, i.e. show that Equation 3.44 equals 0 for $n \neq m$ and 1 for $n = m$.

Chapter 4

Quantum Mechanics and The Hydrogen Atom

4.1 Introduction

An introduction to the key points and results of quantum mechanics would not be complete without a discussion on the hydrogen atom. The energy levels and the spectrum of the hydrogen atom are accurately predicted by quantum mechanics. While the Bohr atom predicted the gross properties of the hydrogen light emission spectrum, it was unable to predict the details known as the fine structure. Quantum mechanics overcomes this deficiency. While it gives the principal energy levels just as the Bohr atom, it is also able to predict the fine emission structure, which is a major success of quantum theory. The hydrogen atom consists of one large positively charged proton and a single electron some distance away. While it appears like a relatively simple system, it is actually fairly complicated. Thus, we will only outline the procedure analyzing the hydrogen and hydrogen-like atoms here. A full discussion is part of a course which has quantum mechanics as its main focus. Interested students are highly encouraged to take such an enriching course.

4.2 Schrodinger Equation for the Hydrogen Atom

The hydrogen atom consists of a negatively charged electron that is approximately separated from a positively charged nucleus by a distance r. The potential

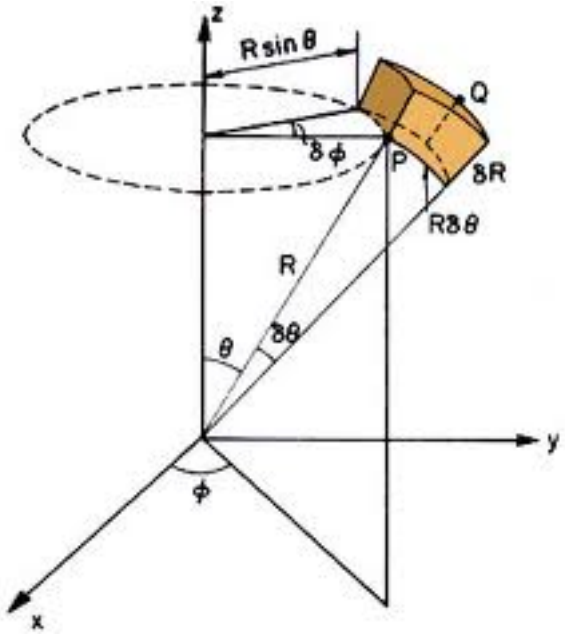


Figure 4.1: Spherical Coordinate System. Any point can be defined by the radial distance r; the polar angle θ from vertical axis, and the azimuthal angle ϕ, which is the angle between the x-axis and the projection of the radial vector onto the x-y plane. The volume element for spherical coordinates is shown in the figure.

energy is given by the Coulomb potential of electrostatics:

$$V(x,y,z) = -\frac{q^2}{4\pi\epsilon_0\sqrt{(x^2+y^2+z^2)}} \tag{4.1}$$

Potential V does not depend on the angle and is spherically symmetrical. Therefore it is best to not use rectangular coordinates, but write the potential using spherical coordinates:

$$V(r) = -\frac{q^2}{4\pi\epsilon_0 r} \tag{4.2}$$

where $r = \sqrt{(x^2+y^2+z^2)}$

Recall that the relationship between Cartesian and spherical coordinates is given by:

$x = r\,sin\theta\,cos\phi, \qquad y = r\,sin\theta\,sin\phi, \qquad z = r\,cos\theta$

And the volume element dV in spherical coordinates is:

$dV = r^2 sin\theta dr d\theta d\phi.$

Illustrations of the spherical coordinate system are given in figures 4.1 and 4.2.

Now, using the above expression for $V(r)$ we obtain the following form for the Schrodinger equation for the hydrogen atom, where the wave-function ψ is a

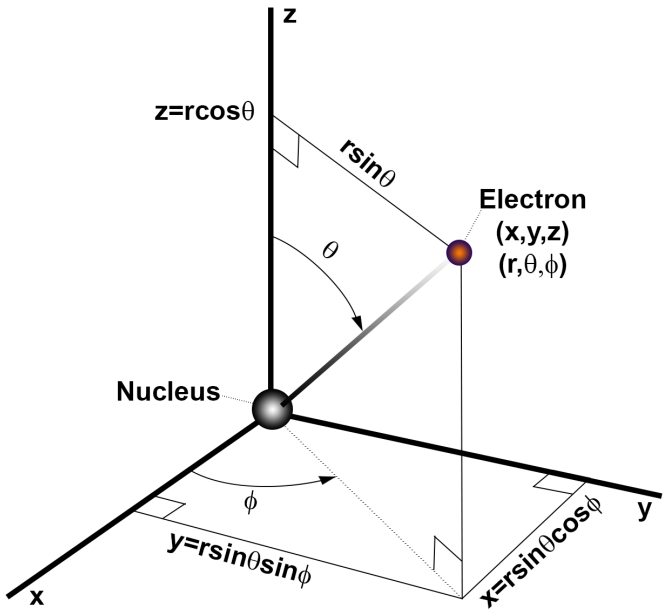


Figure 4.2: Spherical Coordinate System. Any point can be defined by r, the radial distance, the polar angle θ, and the azimuthal angle ϕ.

function of the spherical coordinates r, θ, ϕ:

$$-\frac{\hbar^2}{2M}\nabla^2\psi(r,\theta,\phi) - \frac{q^2}{4\pi\epsilon_0 r}\psi(r,\theta,\phi) = \mathcal{E}\psi(r,\theta,\phi) \tag{4.3}$$

Where M is the mass of the electron.

Here comes the complicated part. We now have to write the Laplacian operator ∇^2 in spherical coordinates to obtain the explicit form of the Schrodinger equation for the H atom. So, substituting ∇^2 using spherical coordinates gives:

$$\frac{-\hbar^2}{2M}\left[\frac{1}{r^2}\frac{\partial}{\partial r}(r^2\frac{\partial}{\partial r}) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\frac{\partial}{\partial\theta}) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right]\psi(r,\theta,\phi)$$
$$- \frac{q^2}{4\pi\epsilon_0 r}\psi(r,\theta,\phi) = \mathcal{E}\psi(r,\theta,\phi) \tag{4.4}$$

As said above, equation 4.4 is the Schrodinger equation for the hydrogen atom written in spherical coordinates, and the top part of the equation within the square brackets is the Laplacian operator ∇^2 also expressed in spherical coordinates.

4.3 Separation of Variables

Equation 4.4 looks pretty intimidating, however, there is something worth noting which is that the potential energy term only depends on the independent variable r. This property helps us to utilize the separation of variables technique where we can write the wave-function as:

$$\psi(r,\theta,\phi) = \mathbb{R}(r)\Theta(\theta)\Phi(\phi) \tag{4.5}$$

In equation 4.5 we write the wave, which depends on the three variables r, θ, ϕ as the product of three individual functions, with each depending only on a single independent variable. ($\mathbb{R}(r)$ depends on r, $\Theta(\theta)$ depends on polar angle $\theta, \Phi(\phi)$ depends only on azimuthal angle ϕ). Next we Substitute 4.5 into 4.4 and then do a fair amount of algebra and differentiation to transform the single Schrodinger equation 4.4 into three separate equations: one for $\Phi(\phi)$ only, one for $\Theta(\theta)$ only, and one for $\mathbb{R}(r)$ only. I will not include the details of this manipulation here, but it is not too bad and you can check it out in the references [2, 3]. The resulting three separate differential equations are:

$$\frac{d^2\Phi}{d\phi^2} = -m^2\Phi \tag{4.6}$$

$$-\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \frac{m^2}{\sin^2\theta}\Theta = l(l+1)\Theta \tag{4.7}$$

$$\frac{-\hbar^2}{2M}\left[\frac{1}{r^2}\frac{d}{dr}(r^2\frac{d}{dr}) - \frac{l(l+1)}{r^2}\right]\mathbb{R}(r) - \frac{q^2}{4\pi\epsilon_0 r}\mathbb{R}(r) = \mathcal{E}\,\mathbb{R}(r) \tag{4.8}$$

The separation of variables technique utilizes the separation constants of m and $l(l+1)$. Here, we used two separation constants because we have three equations. (Recall when we used separation of variables to separate space and time in Chapter 3, we used one separation constant because we had to separate the time dependent Schrodinger equation into two equations.) The exact values of these separation constants come from the boundary conditions and the requirements of continuity of the wave-function. This will become more clear later in this section especially when we describe the solutions of equation 4.7, which provides the polar angle θ dependence on the wave-function.

Eigenfunction and Eigenvalue Equations and their Solutions

Each one of the three separate equations has a similar special form. More specifically, each one of the three separated equations is an eigenvalue equation where a second order differential operator is acting on the function on the LHS, which is equal to a constant times the same function on the RHS. Equations of this form are called Sturm-Liouville equations after the two mathematicians who first studied them extensively in the mid 19th century. It has been found that equations in this form have a special type of solution. More specifically, they are satisfied by a set of solutions and each single solution of the set will give a specific Eigenvalue.

A simple example of a set like this are the solutions for the Schrodinger equation for the particle in an infinite potential well of length L. In section 3.5.1 we found that the set of solutions were $\psi(x) = (2/L)sin(n\pi x/L)$ (equation 3.42) and each solution gave rise to a specific eigenvalue $\mathcal{E}_n = \hbar^2\pi^2n^2/2mL^2$ (equation 3.43). The value of the integer n indicates the specific wave-function and its corresponding energy. Since equations 4.6 through 4.8 all have this general form, they will have a set of functions as their allowed solutions, and each one of the allowed functions of the set will correspond to a separate eigenvalue. Now let's look at the solution to each equation separately.

4.3.1 Solution of Φ Equation (Azimuthal Angle Dependence)

It is straightforward to integrate equation 4.6 and see that the solution is:

$$\Phi_m(\phi) = \frac{1}{\sqrt{2\pi}}e^{jm\phi} \quad m = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots \tag{4.9}$$

The normalization coefficient $\frac{1}{\sqrt{2\pi}}$ is obtained by the requirement that

$$\int_0^{2\pi} \Phi_m^*(\phi)\Phi_m(\phi)d\phi = 1 \tag{4.10}$$

Note the values of m are restricted to integers. This is because if we go around in a complete circle and come back to where we started, the wave-function must be the same. Otherwise it would not be physical.

The Eigenvalue of the azimuthal equation is m^2. More detailed study of quantum mechanics shows that $m\hbar$ is equal to the z component of the electron's angular momentum.

4.3.2 Solution of Θ Equation (Polar Angle Dependence)

The solution to equation 4.7 will give the θ or polar angle dependence on the wave-function. This may be easier said than done because equation 4.7 is very complicated. However, we are very lucky because equations of this type have been studied by mathematicians for about 200 years. This equation is called the Associated Legendre Equation and is named after the mathematician Adrien-Marie Legendre (1752-1833). Mathematicians have shown that solutions for this equation are given by polynomials. The only allowed solutions are when the equation 4.7 have specific integer values for the indices l and m, which are contained explicitly in the equation. Thus only very specific polynomials are the solutions and they correspond to different values of l and m. The system of polynomials are designated as $P_l^m(\cos\theta)$, and they are functions of $\cos\theta$ and $\sin\theta$ and are called the *Associated Legendre Polynomials*. Thus the solutions to equation 4.7 are

$$\Theta_{lm}(\theta) = N_{lm}P_l^m(cos\theta) \qquad\qquad l = 0, 1, 2, 3, ... \tag{4.11}$$

where N_{lm} is the normalization constant, that depends on the values of l and m. We won't worry about the exact values of N_{lm} in this class, but they satisfy the normalization condition:

$$\int_0^\pi \Theta_{lm}^2(\theta)\sin\theta d\theta = \int_0^\pi |N_{lm}P_l^m(cos\theta)|^2 \sin\theta d\theta = 1 \tag{4.12}$$

For those interested it turns out that the values of N_{lm} are the following [2]:

$$N_{lm} = \sqrt{\frac{(2l+1)(l-m)!}{2(l+m)!}} \tag{4.13}$$

Where ! is factorial. Also, one very important thing to know is that the values of m are restricted to be less than or equal to each specific value of l. In other words for a given value of l, we have the very important requirement that ($|m| \leq l$). It is also important to note that for the radial function $\Theta_{lm}(\theta)$ $m \geq 0$, which is true for the associated Legendre polynomials $P_l^m(\cos\theta)$ and the normalization coefficient N_{lm}. But m can take on negative values for the azimuthal angle component $\Phi_m(\phi)$ as described in the previous subsection.

The first six Associated Legendre Polynomials $P_l^m(cos\theta)$ are:

$s:$ $P_0^0 = 1$

$p:$ $P_1^0 = \cos\theta\,, \quad P_1^1 = -\sin\theta$

$d:$ $P_2^0 = \frac{1}{2}(3\cos^2\theta - 1) \quad P_2^1 = -3\sin\theta\,\cos\theta \quad P_2^2 = 3\sin^2\theta$

etc.

The set of Associated Legendre Polynomials will continue indefinitely. However, we are typically interested in the first few which correspond to the s, p, d and f atomic orbitals, which we will talk about more later in the chapter.

The Eigenvalue of the polar equation is $l(l+1)$. More detailed study of quantum mechanics shows that $\sqrt{l(l+1)}\hbar$ is equal to the electron's angular momentum.

Example:
Show that the polar normalization function N_{lm} is correct for normalizing $\Theta_{10}^2(\theta)$ and $\Theta_{11}^2(\theta)$.
When integrating the polar component of the wavefunction over $d\Theta$, we must include the θ dependence of the volume element $(\sin\theta)$ in the integrand.

$$\Theta_{lm}(\theta) = N_{lm}P_l^m(\cos\theta)$$

$$\int_0^\pi \Theta_{10}^2(\theta)\sin\theta d\theta = \int_0^\pi \left|\sqrt{\frac{(2+1)(1-0)!}{2(1+0)!}}\cos\theta\right|^2 \sin\theta d\theta =$$

$$= \int_0^\pi \frac{3}{2}\cos^2\theta\sin\theta d\theta = 1$$

$$\int_0^\pi \Theta_{11}^2(\theta)\sin\theta d\theta = \int_0^\pi \left|\sqrt{\frac{(2+1)(1-1)!}{2(1+1)!}}(-\sin\theta)\right|^2 \sin\theta d\theta =$$

$$= \int_0^\pi \frac{3}{4}\sin^3\theta\; d\theta = 1$$

Remember, when evaluating factorials $0! = 1$, also all instances of m in the polar equations are actually $|m|$.

4.3.3 Solution of the $\mathbb{R}$ Equation (Radial or Distance Dependence from the Nucleus)

The solution to equation 4.8 tells us how the wave-function varies as you move away from the center or nucleus of the atom regardless of the angle. It depends on absolute distance or r only. It generally describes the geometry of the overall electron orbit for each energy level n. The radial equation is also very complicated,

but is has been studied for quite a while and the solution has been determined using series solution methods. Like the equations for the ϕ and θ dependence, the radial equation has a set of solutions that will continue indefinitely. This set of functions are called the *Associated Laguerre Polynomials.* The specific form of each function of this set is provided by the integers n and l, and is given by the following:

$$\mathbb{R}_{nl}(r) = A_{nl}\, e^{-\frac{r}{na_o}} \left[\frac{2r}{na_o}\right]^l L^{2l+1}_{n-l-1}\left(\frac{2r}{na_o}\right) \quad n = 1, 2, 3, \ldots, \text{ and } (0 \leq l < n) \tag{4.14}$$

$$A_{nl} = \sqrt{\frac{4(n-l-1)!}{a_o^3 n^4 (n+l)!}} \tag{4.15}$$

Where r is the distance from the origin, A_{nl} are the normalization coefficients, $a_o = 0.53$ angstrom which is the Bohr radius that we found in Chapter 2, n is any integer greater than zero, and $L^{2l+1}_{n-l-1}(2r/na_o)$ are the Associated Laguerre polynomials and are given by

$$L^{2l+1}_{n-l-1}\left(\frac{2r}{na_o}\right) = \sum_{k=0}^{n-l-1} \frac{(-1)^k\,[(n+l)!]^2}{(n-l-1-k)!(2l+1+k)!k!}\left[\frac{2r}{na_o}\right]^k \tag{4.16}$$

Also, it is very important to note that the values of l are limited to be less than n or $(l < n)$. Furthermore just like for Φ and Θ, the function $\mathbb{R}(r)$ must be normalized:

$$\int_0^\infty \mathbb{R}^2_{nl}(r) r^2 dr = 1 \tag{4.17}$$

Note that the additional r^2 term in the integral arises due to the r dependence of the spherical coordinates volume elements. Graphs of the square magnitude of the radial distribution $|\mathbb{R}_{nl}|^2$ are shown in Figure 4.3 for the first three values of the principle quantum number n. These graphs give the probability of finding the electron at a distance r in angstroms from the nucleus. The values n =1, 2, 3 give the first three energy levels of the hydrogen atom, where n=1 is the ground state energy. We see at the ground state, the electron is most likely to be at about $0.5\mathring{A}$ from the nucleus which is about the same as the Bohr radius.

While the general form of the radial part of the wave-function $\mathbb{R}(r)$ is complicated (equation 4.14), and you don't need to know the details for this class, *it is IMPORTANT to know that the function is given by a polynomial multiplied by a decaying exponential function.*

$$\mathbb{R}_{nl}(r) = e^{\frac{-Zr}{na_o}} f_{nl}(r) \qquad n = 1, 2, 3..... \tag{4.18}$$

Where Z= nuclear charge (for hydrogen Z=1), and $a_o = \frac{4\pi\epsilon\hbar^2}{mq^2}$ = Bohr radius.

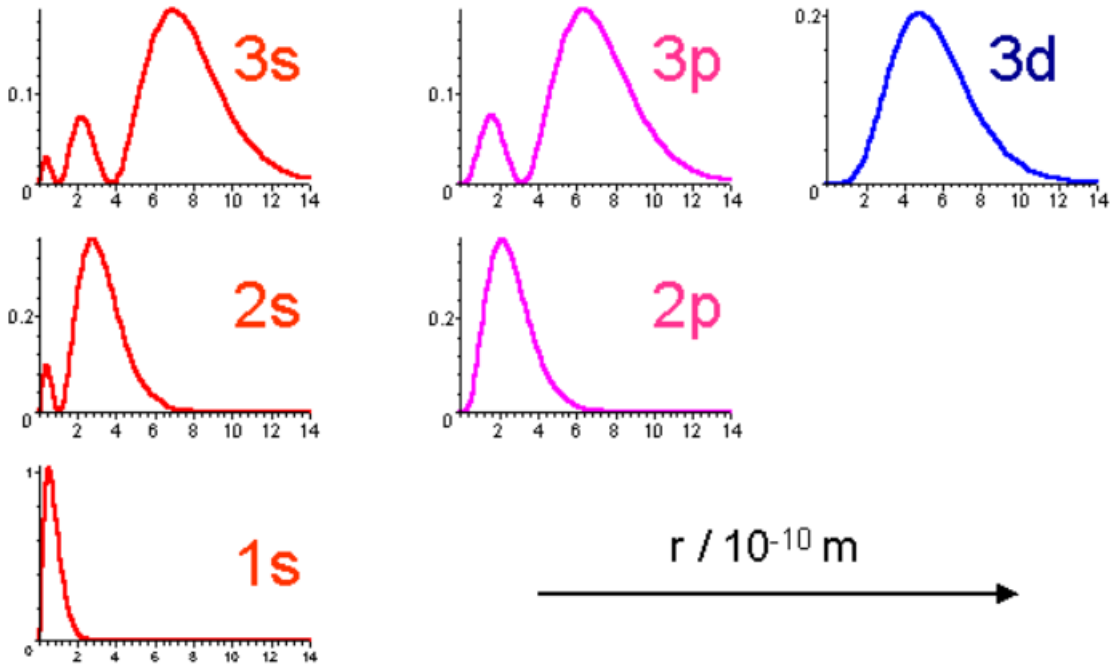


Figure 4.3: The square magnitude of the radial component of the hydrogen wave function $\mathbb{R}_{nl}(r)$ for the first three energy levels. Graphs show the plot of $r^2 R_{nl}^2$.

4.4 Energy Levels and the Total Wave-function

Energy Levels

The allowed energy levels of the hydrogen atom are given by the eigenvalues of the general Schrodinger equation for the hydrogen atom or equation 4.4. These are the same as the eigenvalues of the radial equation or equation 4.8 and are given by the following:

$$\boxed{\mathcal{E}_n = -\frac{MZ^2q^4}{(4\pi\epsilon_0)^2 2\hbar^2 n^2}} \tag{4.19}$$

Where M is the mass of the electron, and Z is the nuclear charge. For example, for hydrogen $Z = 1$, and for a helium ion He^+ $Z = 2$.

Note that the allowed energy values are negative and increase (get closer to zero) as the quantum number n increases. The negative value of energy just means that the electron is bounded by the positively charged nucleus. The larger the value of n, the higher the energy level and the less bound the electron. In the limit of infinite n, the energy becomes zero, which means that the electron is no longer bound to the nucleus. Also note that the values for the allowed energies are the same that Bohr obtained for his 'Bohr Atom'.

It is worth noting that the energy levels given by equation 4.19 only depend on the value of the quantum number n from the Radial equation and not on the values of l. This is actually a simplification. The levels do indeed depend on l as well as the spin of the electron. These are due to quantum relativistic effects and the interactions of polar angle wave function and the electron spin. However, this is probably something we should not spend too much time on here. Just know that these interactions do give rise to the various further splitting of energy levels which depend on l and the spin. This effect results in an energy that depends on the orbit the electron is in (n), as well as the suborbital (s, p, d, f), as defined by l.

Total Wave-function $\psi_{nlm}(r, \theta, \phi)$

Now that we have explained how we obtain the individual parts of the wave-function that we introduced for our separation of variables methodology, we can construct the total wave-function $\psi_{lmn}(r, \theta, \phi)$ by multiplying together the individual parts:

$$\psi_{nlm}(r, \theta, \phi) = \mathbb{R}_{nl}(r)\Theta_{lm}(\theta)\Phi_m(\phi) \tag{4.20}$$

We now start constructing the various hydrogen wave-functions by choosing values for n, l, m and multiplying functions together. The first three wave-functions, constructed for $nlm = 100$, $nlm = 210$ and $nlm = 21 \pm 1$, are given below. A large set of these hydrogen wave-functions can easily be found on the Internet and in many introductory quantum mechanics books [2, 3].

$$\psi_{100}(r, \theta, \phi) = \frac{1}{\sqrt{\pi}} \left(\frac{Z}{a_o}\right)^{3/2} e^{-Zr/a_o} \tag{4.21}$$

$$\psi_{200}(r, \theta, \phi) = \frac{1}{4\sqrt{2\pi}} \left(\frac{Z}{a_o}\right)^{3/2} \left(2 - \frac{Zr}{a_o}\right) e^{-Zr/2a_o} \tag{4.22}$$

$$\psi_{210}(r, \theta, \phi) = \frac{1}{4\sqrt{2\pi}} \left(\frac{Z}{a_o}\right)^{3/2} \frac{Zr}{a_o} e^{-Zr/2a_o} cos\theta \tag{4.23}$$

$$\psi_{21\pm1}(r, \theta, \phi) = \frac{1}{8\sqrt{\pi}} \left(\frac{Z}{a_o}\right)^{3/2} \frac{Zr}{a_o} e^{-Zr/2a_o} sin\theta e^{\pm j\phi} \tag{4.24}$$

Three-dimensional images illustrating the square magnitude of hydrogen wave-functions ψ_{nl} for n, l = 1,0; 2,1; 3,2; 4,3; 5,4 and the allowed values of the quantum number m for each case are shown in Figure 4.4.

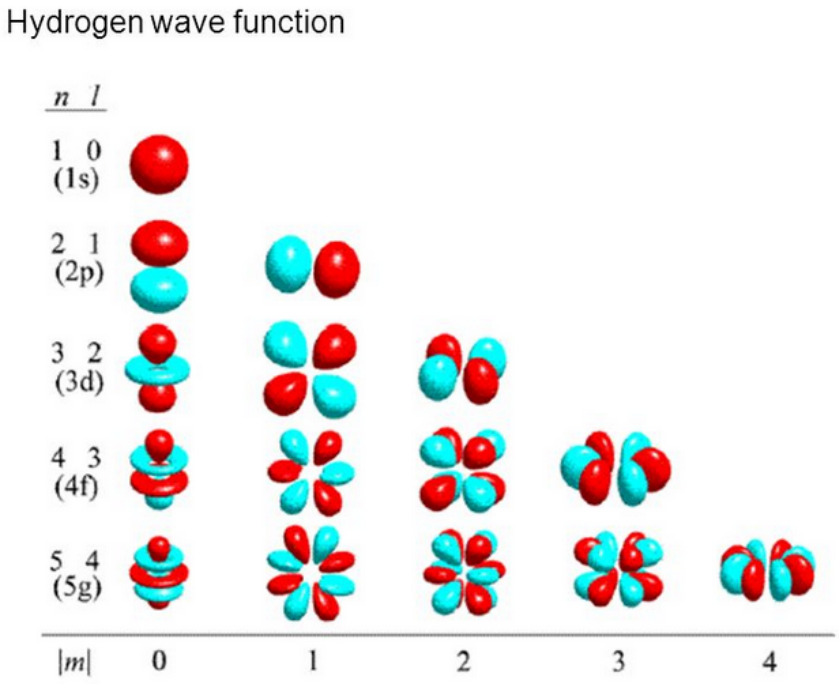


Figure 4.4: Three-dimensional images illustrating the square magnitude of hydrogen wave-functions $|\psi_{nl}|^2$ for n, l = 1,0; 2,1; 3,2; 4,3; 5,4 and the allowed values of the quantum number m for each case. The corresponding s, p and d indices are also shown. Recall the s,p,d indices correspond to specific values of the l and m quantum numbers.

4.5 The Indices: n, l and m, the Electron Spin s and the Pauli Exclusion Principle

These indices are also called the quantum numbers for the hydrogen atom. Each distinct combination of n, l and m corresponds to another wave-function and thus another quantum state of the hydrogen atom. (This does not include spin which we will talk about later.) The wave-functions for hydrogen and the energy values are also applicable to all the elements with minor modification to account for the different atoms having many electrons. Thus, the wave-functions above also correspond to all the known atoms and thus give structure to each of the atoms found in the periodic table.

These indices n, l and m are described as follows:

- n is principal quantum number. It describes the **orbit and energy level of the electron**, and comes from the possible solutions of the $\mathbb{R}_{nl}$ equation:

  $n = 1, 2, 3, 4.....$

- l is the polar angle quantum number. It **describes the suborbital** and comes from the $\Theta_{lm}(\theta)$ equation. The value of l corresponds to the orbitals: s, p, d, f,...as follows: $l = 0$ gives an s suborbital, $l = 1$ gives a p suborbital, $l = 2$ gives a d suborbital, $l = 3$ gives an f suborbital, etc. Values of l are integers, and l is restricted for a given value of n, and they give the allowed suborbits for a given orbit. For a given value of n, values of l can be:

  $l = 0, 1, 2, ..., n - 1$

  It is also important to note that the angular momentum L of the system is given by the expression: $L = \sqrt{l(l+1)}\hbar$

- m indicates the azimuthal dependence of the wave-function. It comes from $\phi_m(\phi)$ and $(-l \leq m \leq l)$. It gives the number of suborbitals there are for a given value of l. For each value of l, there will be $2l + 1$ values of m.

  $m = -l, -1 + 1, ... - 1, 0, 1, ..., l - 1, l$

  The eigenvalue of the azimuthal equation is actually m^2, and the z-component of the angular momentum $L_z = m\hbar$.

- Example: As an example consider the case where $n = 3$, $l = 1$, $m = 0$. This will correspond to a p suborbital in the third energy level or third row of the periodic table. When $m = 0$, this is traditionally taken to mean the p_z suborbital.

- **Electron Spin (s)**: We have not really talked about electron spin yet. Spin is a characteristic that largely determines the effect of the magnetic field on an

electron. All electrons have spin. There are two possible spin states that an electron will occupy: "**spin up**" and "**spin down**". The quantum numbers for spin are:

$s = \pm\frac{1}{2}$

Spin is regularly used in modern technology. One especially interesting and famous application is in Magnetic Resonance Imaging (MRI) which uses electron spin to produce tissue images inside the human body.

The Pauli Exclusion Principle says that no more than one electron can occupy the same quantum state. For the hydrogen atom, the specific combination of n, l, m, s determines the state the electron is in and will be unique for each state. (Often you will hear people say that two electrons can occupy a state. When people say this they are neglecting the effect of spin.) Particles that obey this restriction with respect to only one being able to occupy a given state are called "Fermions". The Pauli Exclusion Principle is extremely important to the world as we know it. If electrons did not obey the Pauli Exclusion Principle, they would all drop to the lowest energy level and there would not be interactions between atoms as we know it. Thus, life and the world as we know it would not exist!

Filling Up Atomic Orbitals

Atoms across the periodic table fill up their orbitals according to the number of electrons they have. Recall the following:

l=0, s orbital, 2 states

l=1, p orbital, 6 states

l=2, d orbital, 10 states

l=3, f orbital, 14 states

Atoms will continue to fill up suborbitals as the atomic number increases. It is probably easiest to show this by example. Filling up of the orbitals for Chlorine (atomic number $Z = 17$), for example, is as follows:

$1s^2 2s^2 2p^6 3s^2 3p^5$ (Total number of electrons=17)

The full table of quantum number combinations up to $n = 3$ is shown in table 4.1.

Table 4.1: **Allowed quantum number combinations up to $n = 3$.**

n	l	m	Spin	Total Suborbital States	Notation	Elements	Total States in Orbital
1	0	0	$\pm\frac{\hbar}{2}$	2	$1s^1$, $1s^2$	H, He	2
2	0	0	$\pm\frac{\hbar}{2}$	2	$2s^1$, $2s^2$	Li, Be	8
	1	-1	$\pm\frac{\hbar}{2}$	6	$2p^1 \ldots 2p^6$	B, C, N, O, F, Ne	
		0	$\pm\frac{\hbar}{2}$				
		1	$\pm\frac{\hbar}{2}$				
3	0	0	$\pm\frac{\hbar}{2}$	2	$3s^1$, $3s^2$	Na, Mg	18
	1	-1	$\pm\frac{\hbar}{2}$	6	$3p^1 \ldots 3p^6$	Al, Si, P, S, Cl, Ar	
		0	$\pm\frac{\hbar}{2}$				
		1	$\pm\frac{\hbar}{2}$				
	2	-2	$\pm\frac{\hbar}{2}$	10	$3d^1 \ldots 3d^{10}$	Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn*	
		-1	$\pm\frac{\hbar}{2}$				
		0	$\pm\frac{\hbar}{2}$				
		1	$\pm\frac{\hbar}{2}$				
		2	$\pm\frac{\hbar}{2}$				

*The 4s suborbital starts to fill before the 3d suborbital due to it's lower energy so the $4s^1$ and $4s^2$ electrons are needed to create all of these elements. Elements 19 and 20 (K and Ca) are not listed because they only have 4s electrons and no 3d electrons.

4.6 Problems

4.1 Write down the time independent Schrodinger equation for the hydrogen atom. Include the potential energy and the complete ∇^2 operator in spherical coordinates.

4.2 What is the probability density for the wave function in the azimuthal (ϕ) direction? Remember, the probability density is obtained from the magnitude of the wave function squared.

4.3 On a polar plot using the polar angle wave-function $\Theta(\theta)$, graph the probability density functions of the S_0, P_1^0 and P_1^1 orbitals.

4.4 Why are atoms typically stable when they have eight electrons in their outer orbits?

4.5 Briefly describe how the Schrodinger Equation is solved for the hydrogen atom. (Hint: describe how you would use separation of variables.)

4.6 Calculate the frequency of light emitted when an electron drops from the third energy level of hydrogen to the second.

4.7 Atomic Structure of Silicon (Atomic Number is 14),

(a) Write down the values of the quantum numbers: n, l, m, and spin for each electron in a silicon atom.

(b) Draw a diagram of how the electrons in a silicon atom are arranged into orbits (n) and suborbitals (s,p,...).

(c) How many bonds do you think a silicon atom can make and why?

4.8 The 1s orbital of a hydrogen atom is given by $\mathbb{R}_{10}(r) = Ae^{-r/a_0}$ where a_0 is the Bohr radius and A is a normalization constant.

(a) Find the value of the normalization constant A using equation 4.17.

(b) Calculate the expected radial distance one would measure the electron from the atomic core in this state. Leave your answer in terms of the Bohr radius.

(c) Calculate the most likely radial distance (where the probability density is maximum) and compare the value to the expectation. (Hint: As in part (a), you must multiply the radial probability density by the radial Jacobian for spherical coordinates before finding the maximum, i.e. $r^2 \left|\mathbb{R}_{10}(r)\right|^2$).

4.9 Hydrogen is excited by an electrical current and spectral measurements indicate that it has been excited to the second energy level ($n = 2$).

(a) What are the possible values of angular momentum that the hydrogen's electron can have?

(b) If the electron is in a p suborbital, and its z-component of angular momentum is equal to zero, what is the probability that the electron will be in a region that is between 0 and 30^o from the polar or z axis of the hydrogen atom?

4.10 This entire chapter has focused on the hydrogen atom. However, we are often most interested in other atoms, especially the semiconductors like silicon. Explain qualitatively why it is useful to learn all these things about hydrogen when we ultimately are more interested in other atoms in the periodic table.

4.11 If the Pauli Exclusion Principle did not exist, explain what you think atomic structure would look like. How would this affect chemical bonding and the world as we know it? Answer in about two or three sentences.

Chapter 5

Electrons, Holes and the Quantum Mechanics of Crystalline Solids

5.1 Introduction

In this chapter we will focus on electrical conduction in semiconductors, and we will answer many questions about the electrical characteristics of electronic materials. Why are some materials metals, others insulators, and even others semiconductors? Why do we make electronics out of mainly semiconductor crystalline solids? How does quantum mechanics determine the particles that act as charge carriers in semiconductors? How do we incorporate the periodic potential energy of the crystalline structure into the Schrodinger equation? What is an 'n-type' material and what is a 'p-type' material? In this chapter we will learn how quantum mechanics helps to answer many of these questions. We will learn about what is a conduction electron and what is a hole. We will learn about the valence band and the conduction band in crystalline solid materials. We will learn about the concept of effective mass. We will learn about the unique properties of semiconductors that allow for a process of doping from which we make PN junctions and virtually all electronics.

5.2 Electron Wave Function in a Crystal

Just like electrons around a single atom, and electrons in an infinite and finite potential well, electrons in a crystalline solid are described by ascertaining their wave functions by solving the Schrodinger equation.

5.2.1 Schrodinger Equation for Electrons in a Crystal

Periodic Potential Energy: Electrons in crystals are bound to the periodic array of atoms that make up the solid. In metals the electrons are very loosely bound,

in insulators the electrons are tightly bound, and in semiconductors they are bound somewhere in between. The thing that the electrons all have in common in different crystal solids is that the potential energy that the electrons feel is periodic with the periodicity of the crystal. Therefore, when we move from one unit cell to the same relative location on a different unit cell, since the environment is the same, the potential energy will be the same. Furthermore, since the crystal structure repeats itself over virtually an infinite number (on the order of Avogadro's Number) of unit cells, the electrical potential is periodic. We use the fact that we are dealing with a periodic potential to solve the Schrodinger equation to obtain the electron wave functions in all crystalline solids. In this class we will pay particular attention to semiconductor crystalline solids.

The time independent Schrodinger equation in a crystal solid is given by:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = \mathcal{E}\psi(x) \tag{5.1}$$

where $\psi(x)$ is the electron wave function, m is the mass of the electron, $\mathcal{E}$ is the electron energy, which is an allowed energy level, and $V(x)$ is the potential energy that the electron feels. The unique aspect of Schrodinger equation 5.1 is that the potential is periodic which means that it has the following property:

$$V(x) = V(x + na), \qquad n = 1, 2, 3, 4..... \tag{5.2}$$

In equation 5.2 a is a primitive lattice vector. In other words, the crystal is periodic in the length $'a'$. Note that our example here is for one dimension, but the same applies in three dimensions. In Figure 5.1 we illustrate the periodic potential for a 1-D array of atoms. Note that each atom has a form of $\frac{-1}{x}$ for its potential energy.

To get a more explicit idea about the Schrodinger equation for a periodic potential, we bring in Fourier series. Recall that any periodic function can be written as a Fourier series, which means it can be written as a sum of sinusoidal functions. Thus, the periodic potential can be written as the following complex series:

$$V(x) = \sum_{n=-\infty}^{\infty} V_n e^{j2\pi nx/a} \tag{5.3}$$

Where a is the lattice constant, and V_n are the coefficients for each term in the series. Substituting the Fourier series, which is a by definition a periodic function, we obtain the following, more explicit form for the Schrodnger equation for a crystalline solid like that of a semiconductor.

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + \left[\sum_{n=-\infty}^{\infty} V_n e^{j2\pi nx/a}\right]\psi(x) = \mathcal{E}\psi(x) \tag{5.4}$$

5.2.2 The Wave-Function and Bloch's Theorem

In general, it would be absurdly difficult to write down and solve the Schrodinger equation for electrons in a crystal because there are so many atoms and so many electrons. In his Ph.D thesis, in 1928, the Physicist Felix Bloch came up with a way to overcome this problem and greatly simplify solving the Schrodinger equation for so many atoms at once. Felix Bloch showed that the solution to the Schrodinger equation for an electron in a periodic potential like the crystal lattice is given by a plane wave function times another function that has the periodicity of the crystal:

$$\psi(x) = u(x)e^{jkx} : \tag{5.5}$$

where e^{jkx} is a plane wave, and $u(x)$ is periodic so that:

$$u(x) = u(x + na) \tag{5.6}$$

For a 1-Dimensional lattice, n is an integer and a is the distance between atoms.

5.3 Band Structure for the Material and Its Importance

If the Schrodinger equation is solved for a periodic potential, we will get the wave-function and a relation between $\mathcal{E}$ and k, that is in many ways similar to that of a free particle. Recall, for a free particle: $\mathcal{E} = \frac{\hbar^2 k^2}{2m}$ (parabola). For crystals, you get bands that are sort of parabolas, but have some specific differences that we will talk a little about below.

The relationship between energy and the electron wave-vector is called the **Band structure** of the material. An example picture of the band structure is given in the bottom part of Figure 5.1. In the figure the horizontal axis is k or the electron wave-vector. The vertical axis is the energy. Since we are dealing with quantum mechanics and electrons that are not totally free, bands contain many discrete states. Two electrons can exist in a given state of wave-vector k and energy $\mathcal{E}$, one with spin up and the other with spin down. (The spin up state and the spin down state are actually two different states with the same value of wave vector k.) The figure shows that there are gaps in the allowed energy that electrons can have between the specific bands. These spaces are called **Energy Bandgaps** and are very important for determining the electrical properties of the material. **There are no quantum states in the bandgap, they are forbidden zones, so electrons cannot exist in the energy bandgaps**. The figure also shows that k is not continuous either. This is illustrated by the little vertical lines in energy bands 3 and 4. In fact, all the bands have such restrictions on k, but they are only illustrated here in the top two bands. There are many states in a band. In general, for each band there are

the same number of states in the band as there are atoms in the entire crystal. And if we include spin, there are two times the number of states in each band as there are atoms in the crystal. Thus, if the crystal contains N atoms, then there will be in general 2N quantum states in each energy band where electrons can exist. Since N is the number of atoms in the crystal, it is very large, approximately the value of Avogadro's Number (N $\approx 10^{23}/cm^3$). Each state will have a unique value of the pair $(k, \mathcal{E})$. Let us clarify that we will have approximately 10^{23} States/cm^3 for a 3-dimensional crystal. In this class, we a simplifying and describing things in one dimension because the concept is the same. For a 1-D crystal, we would have $(N = 10^{23})^{1/3}$ or about 10^7 states/cm.

Again, since electrons are confined k will have discrete values. The allowed values of k are

$$k = \frac{2\pi n}{L} = \frac{2\pi n}{Na} \quad n = \pm 1, \pm 2, \pm 3... \pm \frac{N}{2} \tag{5.7}$$

N=number of atoms in crystal ($\approx$ Avogadro's number)
L=length of the crystal=Na
a= lattice constant or the spacing between primitive unit cells in the crystal.

Material's Band Structure Provides Most of its Electrical Characteristics:

The electrical characteristics of a particular crystalline solid can be extracted from its band structure. That is why the key objective for solving the Schrodinger equation for a particular crystal is to ultimately calculate its band structure. More specifically the following properties which we will talk more about in the next several pages, are contained in the band structure. The band structure describes:

1. Whether the material is a **conductor, insulator or a semiconductor** by the number of electrons in each band and the size of the distances between the bands or the bandgap.
2. The **instantaneous velocity** of the electrons in the material.
3. The **effective mass** of the electrons in the material.
4. The **intrinsic number of electrons and holes** there are in a semiconductor.
5. The **mobility of the electrons and holes** in semiconductors and ultimately the **electrical conductivity** of a specific semiconductor.

Effective Mass

The effective mass is a very important concept in semiconductors, especially with respect to electron transport and mobility. When an external electric field is applied to a semiconductor, the electrons feel the applied electric field and they also

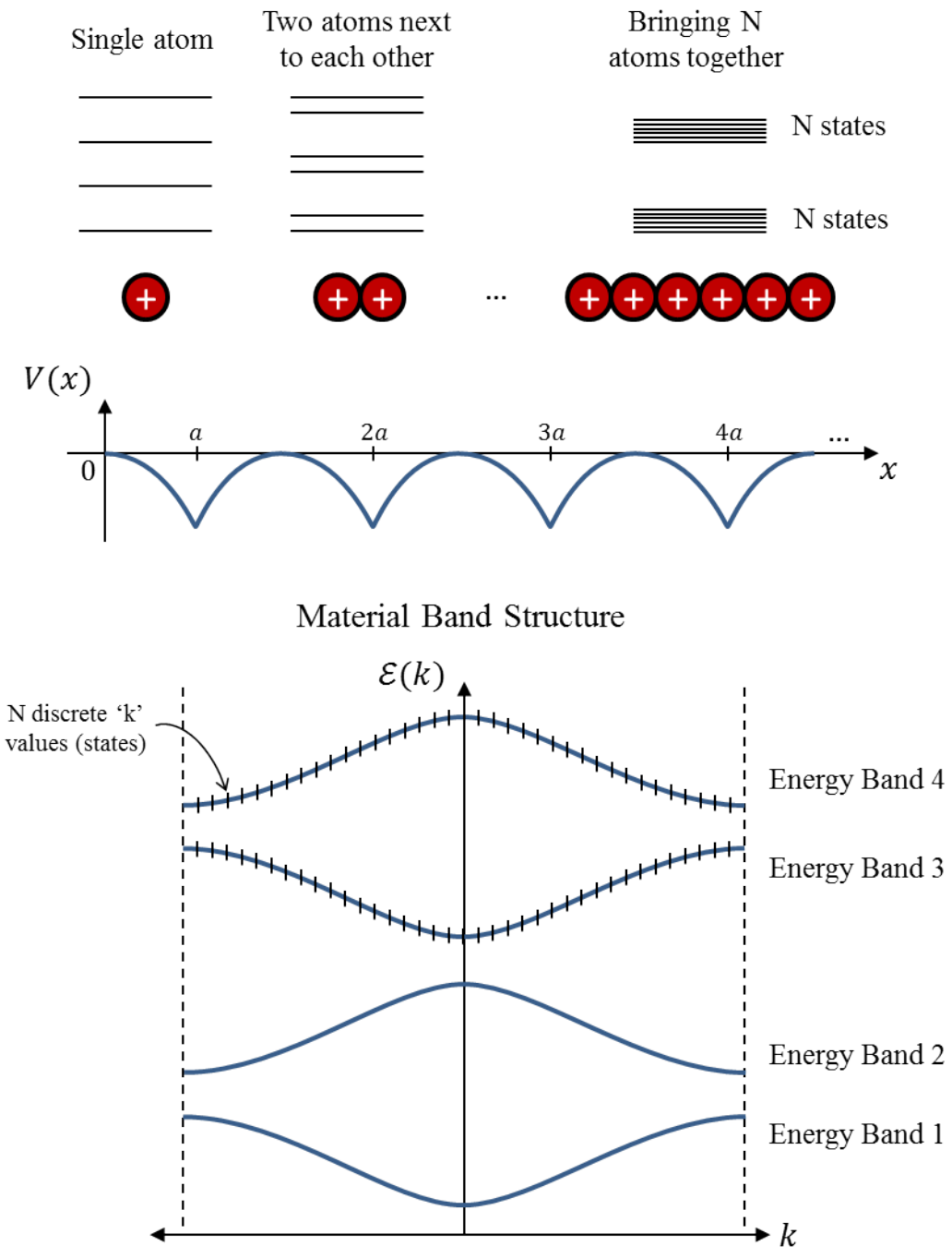


Figure 5.1: Figure showing how periodic potential varies and how orbital splitting takes place.

feel the periodic potential of the lattice. The electrons therefore accelerate due to combined effects of the applied field and the periodic internal potential in a complicated way. To take into account both these influences simultaneously, scientists found that we could use an "Effective Mass (m^*)" instead of the regular electron mass, and this effective mass could account for both, the internal microscopic periodic potential, as well as the macroscopic applied electric field. The effective mass is obtained directly from the second derivative of the band structure as follows:

$$\frac{1}{\hbar^2}\frac{d^2\mathcal{E}}{dk^2} = \frac{1}{m^*} \tag{5.8}$$

or

$$\frac{\hbar^2}{\frac{d^2\mathcal{E}}{dk^2}} = m^* \tag{5.9}$$

After solving the Schrodinger equation, and verifying with experiments, we find the $\mathcal{E}$ vs. k for the semiconductor, we can then compute $\frac{1}{\hbar^2}\frac{d^2\mathcal{E}}{dk^2}$ from which we obtain m^* for an electron in a particular band.

One important result of equation 5.9 is that if the band has high curvature, then the effective mass will be relatively low, while if the energy has low curvature, the effective mass will be relatively high.

Instantaneous Electron Velocity v_g:

The instantaneous velocity of an electron in a crystal at any instant in time and point in space is also obtained from the bandstructure. The symbol we give for instantaneous electron velocity is v_g, where the subscript g comes from the idea that we also call the instantaneous velocity the "group" velocity. We get this because if we describe the electron as a wave packet, the instantaneous velocity of the wave packet is called the group velocity. We obtain the instantaneous velocity for an electron in a particular state with wave vector k, from the derivative or slope of the band structure at that value of k:

$$v_{gi} = \frac{1}{\hbar}\left.\frac{d\mathcal{E}}{dk}\right|_i \tag{5.10}$$

where v_{gi} is the instantaneous velocity of the electron in the k_i quantum state. Note that the slope of the $\mathcal{E}\, vs. k$ at a particular value of k can be positive or negative. So the electron can be moving in either the positive (positive slope) or negative (negative derivative) direction instantaneously in the crystal.

Simplest Example: The Free Electron

In Chapter 3 we studied the free electron, which is the case where the potential equals 0 everywhere. We found in Chapter 3 that the relationship between $\mathcal{E}$ and k is given by $\mathcal{E} = \frac{\hbar^2 k^2}{2m}$.

Applying equation 5.10, we find that the instantaneous velocity for a free electron with wave vector k is:

$$\frac{1}{\hbar}\frac{d\mathcal{E}}{dk} = \frac{\hbar k}{m} = v_g \tag{5.11}$$

Now, if we apply expression 5.8 to the free electron we calculate the expected result for the mass of the electron.

$$\frac{1}{\hbar^2}\frac{d^2}{dk^2}(\frac{\hbar^2 k^2}{2m}) = \frac{2}{\hbar^2}(\frac{\hbar^2}{2m}) = \frac{1}{m} \tag{5.12}$$

k states in an Energy Band

Earlier in this section we said that the number of states allowed in an energy band is 2N, where N: number of atoms in crystal and the factor 2 comes from spin. We show this in the derivation below.

First of all, let's draw two identical crystals in 1-D, so that the end of one crystal is touching the beginning of the next crystal. Each crystal is of length L.

According to Bloch's theorem discussed above, the electron wave function in a particular state k at point x in the first crystal is given by $\psi_k(x) = U_k(x)e^{jkx}$, where $U_k(x)$ is a periodic function with the same periodicity of the periodic potential energy of the crystal. The subscript k in $U_k(x)$ says it is a function of k as well as x.

Now the wave function for the electron in the adjacent identical crystal is given by $\psi_k(x+L) = U_k(x+L)e^{jk(x+L)}$, since it is a distance L away.

Since the crystals are identical, $\psi(x) = \psi_k(x+L)$ and
$\Psi_k(x) = U_k(x)e^{jkx}$ and
$\Psi_k(x+L) = U_k(x+L)e^{jk(x+L)}$
where $L = Na$, (N=number of atoms in crystal).
Due to Bloch and periodicity: $U(x) = U(x+na) = U(x+Na)$.

So, $\psi(x+L) = U_k(x+Na)e^{jk(x+L)} = U_k(x)e^{jk(x+L)}$
Setting $\psi(x) = \psi(x+L)$
$U_k(x)e^{jkx} = U_k(x)e^{jk(x+L)}$
Canceling $U_k(x)$ from both sides gives:
$1 = e^{jkL}$
This allows k to take on many values, but restricts them such that kL must be values of $2\pi n = kL$, where n is an integer. Thus k is restricted to the following discrete values: $k_n = \frac{2\pi n}{L}$ n=$\pm$1, $\pm$2 ....$\pm\frac{N}{2}$
So, there are many k values and hence, lots of states. For a 1-D crystal $N \approx 10^7$/cm. For a 3D crystal $N \approx 10^{23}/cm^3$ or about Avogadro's Number. (One might ask the question, why we don't allow states with $n > |N/2|$. The answer is that these states

are physically the same as those that have values of $n < |N/2|$. This result becomes clear in follow up courses in solid state physics. For now just trust me.)

Distance Between States in k Space

The distance between adjacent quantum states or k states can be easily calculated as: $\frac{2\pi(n+1)}{L} - \frac{2\pi n}{L} = \frac{2\pi}{L} = \frac{2\pi}{Na}$.
Since N is such a large number, there are many states in the band and the states are very closely spaced. Since the states are so closely spaced, we typically think of the each band as a continuous function of $\mathcal{E}(k)$. The maximum and minimum values of k for a particular band are as follows: $k_{max} = \frac{2\pi}{L}\frac{N}{2}$ or $k_{max} = \frac{\pi}{a}$, and $k_{min} = -\frac{\pi}{a}$.

5.4 Material Classification: Conductor, Insulator or Semiconductor

Crystalline solids can be classified by their electrical conductivity. This is obviously important for their applications in electronics. As discussed before, the electrical conductive properties of the material are typically given by its energy band structure. These properties are typically determined by the electron occupancy of the bands and the size of the energy bandgaps between bands. Remember, no electrons exist at energies in the bandgaps because there are no quantum states in these bandgap regions.

Rules of Conduction: Occupancy of a Particular Band

First, let's summarize the conductivity of any single energy band as follows:

1. Full bands cannot conduct electricity. Full bands cannot conduct because electrons will have no place to go in k-space, where ever they would try to go is already occupied by another electron. And the Pauli Exclusion principle says that only one electron can occupy any given state.

2. Empty bands cannot conduct, they contain no electrons.

3. Only partially full bands can conduct. These bands have both occupied and empty states, so electrons can move from k-state to a different k-state, which can then give rise to current in the band.

Filling Up of Bands from Low Energy to High Energy

As we mentioned previously, in general each band has 2N states that can be occupied by electrons, where N is the number of atoms in the crystal. (Note, there will be exceptions to this, but we will not consider these kinds of details in this class.) In a crystal, electrons in the lowest orbits (those with quantum number n=1) will

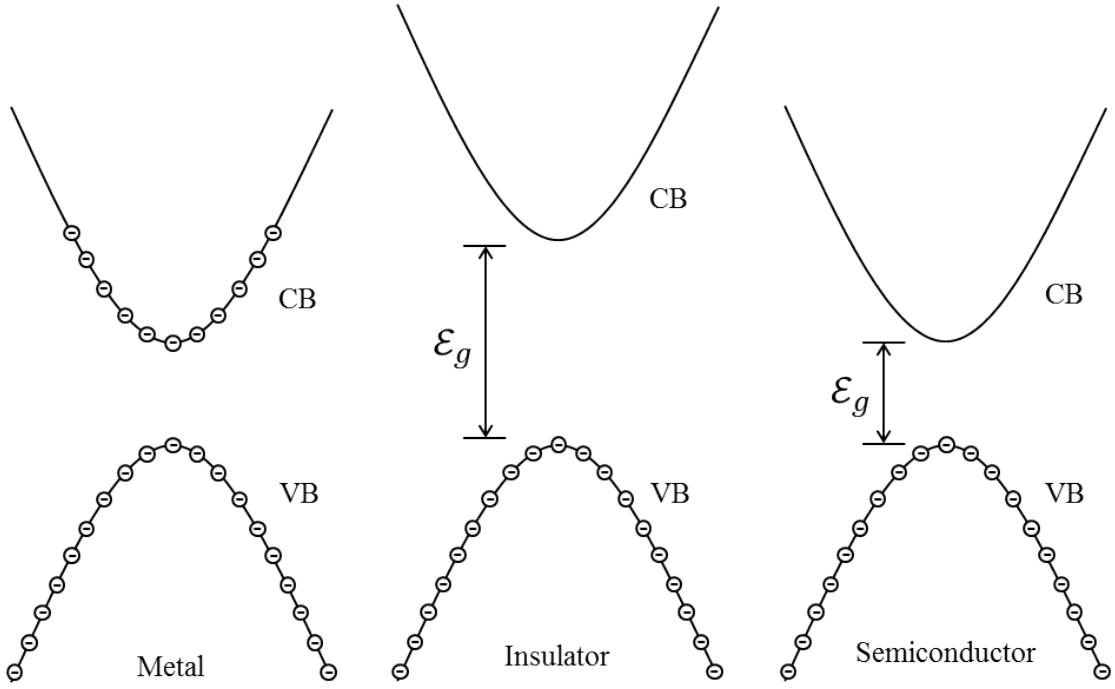


Figure 5.2: E-k diagram for conductor, insulator and semiconductor. VB, CB, Eg stand for valence band, conduction band, Energy Bandgap, respectively. The small circles represent electrons. The diagram is for absolute zero (T =0K). At higher temperatures the semiconductor would have some electrons in its conduction band, but nowhere near as many as the conductor has.

Material	T=0K		T>0K		Bandgap $\mathcal{E}_g$
	CB	VB	CB	VB	
Metal	$\frac{1}{2}$ Full	Full	$\frac{1}{2}$ Full	Full	Zero
Semiconductor	Empty	Full	Slightly Full	Slightly Empty	Small
Insulator	Empty	Full	Empty	Full	Large

Table 5.1: Table comparing metal, insulator and semiconductor.

fill the lowest energy bands, once those are filled the next higher bands will become populated with electrons from the higher orbits, and eventually the outer or valence electrons will fill the upper energy bands.

Valence Band: The highest band in energy that is totally full with all states occupied by electrons at absolute zero temperature ($T = 0K$) is called the *Valence Band (VB).*

Conduction Band: The next higher band above the valence band is called the *Conduction Band (CB).*

The Bandgap and Conductor, Insulator or Semiconductor

The occupancy of the conduction band at ($T = 0K$), as well as the size of the bandgap will determine whether or not the material is a conductor, insulator or semiconductor.

Conductors:

Materials that are good conductors have their conduction bands highly occupied at $T = 0K$. By highly, we mean that about 50% of conduction band k states are filled. These materials are metals and are very good conductors such as copper, silver or gold. Approximately each copper atom will contribute an electron to the conduction band giving approximately Avogadro's number of electrons per cm^3 mobile electrons which can contribute to electrical conduction and therefore very high conductivity. This is illustrated on the left diagram of Figure 5.2.

Insulators:

Insulators do not conduct electricity at low nor at high temperatures. Insulators have an empty conduction and a full valence band at low and high temperatures. Insulators also have a very large bandgap so that electrons cannot gain sufficient

thermal energy to jump from the VB to the CB at any temperature of practical importance. Thus the VB is always full so it cannot conduct electricity, and the CB is always empty and cannot conduct electricity either. This is illustrated on the middle diagram of Figure 5.2.

(Note that in addition to crystal insulators, there are also amorphous insulators like glass ($Si0_2$). Glass does not form a regular crystal structure so Bloch's theorem does not apply to it, so it does not have a well defined band structure. Also the bonding is in glass is covalent so electrons are held tightly to their atoms. As a result, amorphous materials are typically insulators.)

Semiconductors:

At absolute zero ($T = 0K$), the valence band of a semiconductor is full and the conduction band is empty. Thus, at absolute zero semiconductors do not conduct electricity. However, the bandgap between the valence and conduction bands is relatively small. In silicon for example, the bandgap is 1.1eV and in germanium it is only 0.66eV. (In contrast an insulator will typically have a bandgap in the 5eV to 10eV range.) As a result if the temperature is above absolute zero, a few electrons in the VB will acquire sufficient thermal energy to jump to the CB, leaving the VB with some empty states, called holes, and the CB populated with some electrons. This allows both bands to conduct electricity. However, since the actual number of electrons that have been excited to the CB is small compared to the number of states available states in the CB, or the number of atoms in the crystal, the intrinsic semiconductor winds up being a very poor conductor of electricity. (We will find later that we can use manufacturing to increase the natural or intrinsic low conductivity of semiconductors). At room temperature $300K$, the concentration of electrons in the CB of silicon is approximately $10^{10}/cm^3$. This concentration is called the intrinsic electron concentration (n_i), and it will be equal to the intrinsic concentration of holes (p_i) in the valence band. The higher the temperature, the higher the intrinsic concentration since more electrons can obtain the thermal energy to jump from the valence to the conduction band. Also, it follows that the larger the energy bandgap between the VB and CB, the smaller the intrinsic electron and hole concentrations. In general, this intrinsic population is actually very small when compared to the concentration of electrons in the CB of copper which is on the order of $10^{23}/cm^3$. It also largely explains why the conductivity of copper is approximately 10^{13} greater than the conductivity of intrinsic silicon. This is illustrated on the right diagram of Figure 5.2. Note that the elemental semiconductors are from column 14 (column 4 if you neglect transition elements) of the peridic table 5.3. They have four valence electrons and a bandgap.

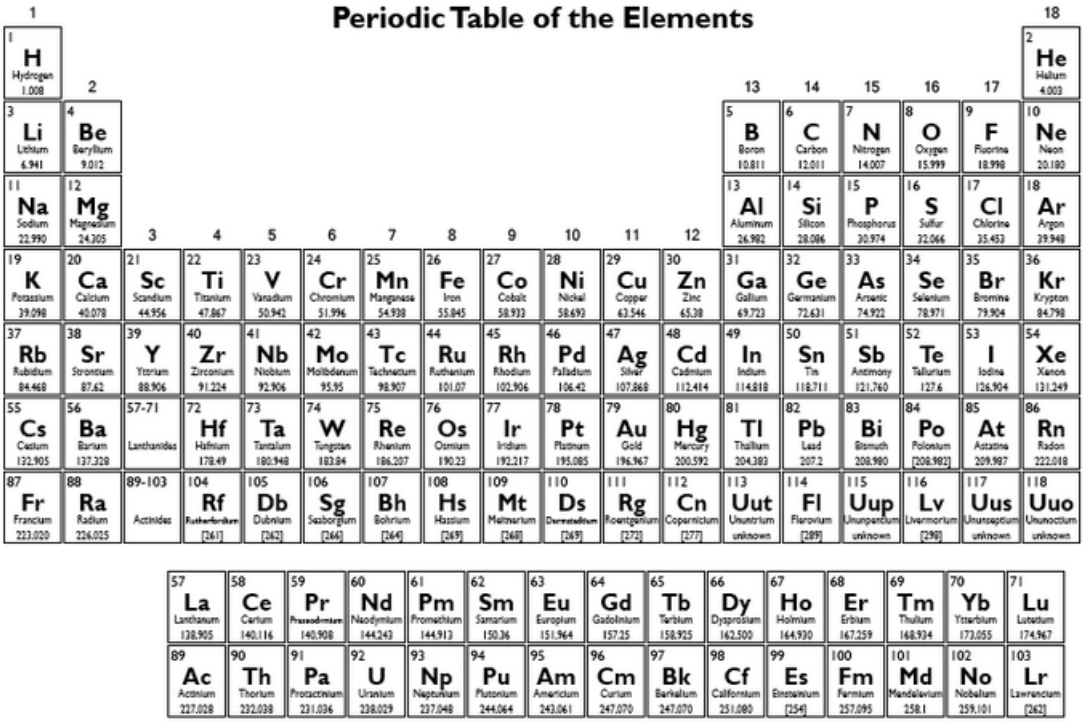


Figure 5.3: Periodic Table

5.5 Electron and Hole Transport in Semiconductor Energy Bands

5.5.1 Electron Current in the Conduction Band

Recall that current density is typically given by

$$\vec{J_n} = -qn < \vec{v} > \tag{5.13}$$

Where $\vec{J_n}$ is the electron current in a particular energy band, n= electron concentration, or electrons per unit volume, which we will use the units of $1/cm^3$, $< v >$ = average electron velocity and q is the magnitude of the electron charge (1.6×10^{-19} Coulombs.). Current density $\vec{J}$ is current per area, and in this class we will use the units of $Amps/cm^2$.

It is also possible to express the current density in a more detailed way. Instead of using the concentration times the average velocity, we can sum the velocity vectors of all the electrons per unit volume in the conduction band. In other words, let's write current density in an energy band as follows:

$$\vec{J_n} = -q \sum_{i}^{N_{ncb}} \vec{v}_{gi} \tag{5.14}$$

$\vec{J_n}$ is current density in conduction band, $\vec{v}_{gi}$ is the instantaneous velocity of the $i'th$ electron in the band. The sum is over all states in the band that are filled with electrons, and N_{ncb} is the number of electrons in the conduction band.

Now if we recall that $\vec{v}_{gi} = \frac{1}{\hbar}\frac{d\vec{\mathcal{E}}}{dk}$, then we can write the current in a particular band as follows:

$$\vec{J_n} = \frac{-q}{\hbar} \sum_{i}^{N_{ncb}} \left.\frac{d\mathcal{E}}{d\vec{k}}\right|_i \tag{5.15}$$

5.5.2 Hole Current in the Valence Band

Remember that when the temperature is greater than absolute zero, some electrons in the valence band will acquire sufficient energy and jump into the conduction band leaving empty states or holes in the valence band. Now that the valence band is not totally fully it can carry a current. Just like we did for electrons in the conduction band, let's sum over the velocities of all the electrons in the valence band to get the current density in that band.

$$\vec{J_p} = \frac{-q}{\hbar} \sum_{i}^{N_{nvb}} \left.\frac{d\mathcal{E}}{d\vec{k}}\right|_i \tag{5.16}$$

Where $\vec{J_p}$ is the current density in the valence band, and N_{nvb} is the number of electrons in the valence band.

Now remember, the number of empty states or holes in the valence band is very small compared to the total number of states and the total number of electrons in the band. In fact the valence band is still over 99.999% full of electrons. Therefore, it is more convenient to write the above sum as a summation over all states and then subtract off the states that are empty. So equation 5.16 can be written as follows:

$$\vec{J_p} = \frac{-q}{\hbar} \sum_{All} \left.\frac{d\mathcal{E}}{d\vec{k}}\right|_i - \left[\frac{-q}{\hbar} \sum_{Empty} \left.\frac{d\mathcal{E}}{d\vec{k}}\right|_j\right] \tag{5.17}$$

Now recall the a full band cannot conduct electricity, so the first summation in equation 5.17 is zero:

$$\vec{J_p} = 0 - \left[\frac{-q}{\hbar} \sum_{Empty} \left.\frac{d\mathcal{E}}{d\vec{k}}\right|_j\right] \tag{5.18}$$

Now distributing the negative sign and not writing the zero gives the following expression for current in the valence band:

$$\vec{J}_p = +\frac{q}{\hbar} \sum_{j}^{N_{pvb}} \frac{d\mathcal{E}}{d\vec{k}} \Bigg|_j \tag{5.19}$$

Where N_{pvb} is the number of empty states in the valence band. These empty states are called **holes**.

Important Result: Holes Behave Like Positively Charged Electrons.

It is important to examine equations 5.15 and 5.19 and to notice that they look identical except for the sign in front of q. This indicates that mathematically, the current in the valence band can be given by adding the velocities of the holes as opposed to summing over all the VB electrons, except that instead of the charge being negative, the charge is positive. This leads to the important result that holes in the valence band act as positively charged electrons when it comes to the flow of electrical current. In other words, those few empty states or holes in the valence band can be treated the same as positively charged electrons. So, even though holes are just empty states, they act like positively charged electrons.

Electron and Hole Summary: In semiconductors we have two types of charge carriers:

1. Negatively charged electrons in the conduction band.
2. Positively charged holes in the valence band.
3. Electrons and holes will typically have different instantaneous velocities and effective masses, which are determined by the slopes and curvatures of the conduction and valence bands, respectively.
4. It follows that semiconductors have current due to electrons in the conduction band and holes in the valence band. And the total current is the vector sum of both.

5.6 Doping, Intrinsic and Extrinsic Semiconductors

Intrinsic Semiconductor:

An intrinsic semiconductor is just a pure crystal of the material. It does not have any impurities called dopants. Intrinsic semiconductors are very poor

conductors of electricity because they have very few electrons in the conduction band and very few holes in the valence band. The intrinsic electron concentration in a semiconductor depends on temperature because they come from the electrons in the valence band that can get enough thermal energy to jump to the conduction band. The symbol for intrinsic electron concentration is n_i, and it is equal to the intrinsic hole concentration p_i in the valence band, because every electron that jumps to the conduction band from the valence band leaves behind a hole in the valence band. As an example, the intrinsic carrier concentration in silicon is approximately $10^{10}/cm^3$, while the concentration of conduction electrons in the conductor copper is approximately $10^{23}/cm^3$. It is therefore not surprising that the conductivity of intrinsic silicon is about 6×10^{11} times less than that of copper.

Extrinsic Semiconductors and Doping:

We change the number of mobile electrons and mobile holes by doping with impurities that give rise to electrons in the conduction band or holes in the valence band. Impurities that give rise to electrons in the conduction band are called **Donors**, and impurities that give rise to holes in the valence band are called **Acceptors**. These donors and acceptors are impurities that substitute for the intrinsic semiconductor atoms, typically silicon, in the crystal lattice. The process of substituting donors or acceptors into the pure semiconductor crystal lattice is called **Doping**.

N-Type Semiconductor: Let's use silicon in this discussion as the intrinsic semiconductor. As we know, silicon is a column four element and thus has four valence electrons, and the intrinsic electron and hole concentrations in silicon is approximately $10^{10}/cm^3$. To generate N-type silicon, we dope silicon with donor atoms. These atoms will have five valence electrons. Four of the donor's electrons will be tightly bound to the silicon atoms, the fifth will be very loosely bound. Since this fifth electron is so loosely bound, it will move away from its original donor atom and enter the conduction band of the silicon so that it now becomes a mobile electron. Donors are put in the semiconductor at concentrations $10^{14} < N_D < 10^{20}/cm^3$, which thereby greatly increases the mobile electron concentration and hence the conductivity. Typically, in silicon Phosphorous is used as a donor.

P-Type Semiconductor: To generate mobile holes, we dope silicon with a material that has three valence electrons. This give rise to unsatisfied bonds because silicon wants four pairs. An electron from the valence band will then leave the band and be used to satisfy the fourth pair. Since the electron has left the valence band, the band now has some empty states or holes. Acceptors are put in the semiconductor at concentrations $10^{14} < N_A < 10^{20}/cm^3$, which thereby greatly increases the mobile hole concentration and hence the conductivity. This gives rise to a P-type semiconductor. Typically, in silicon Boron is used as an acceptor.

Energy Levels of Dopants On an energy scale, the donor levels are found

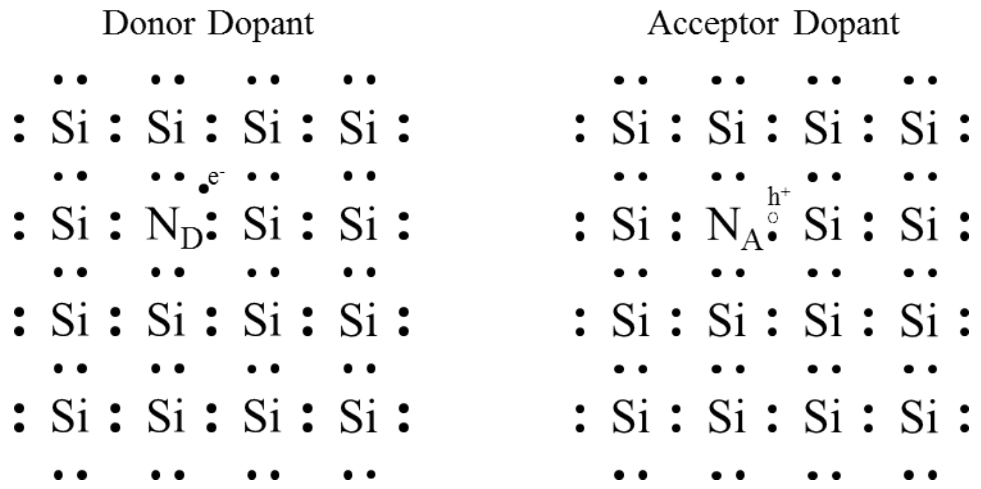


Figure 5.4: Dopant atoms incorporated into a Si lattice. Each Si atom comes with 4 valence electrons with which they uses to covalently bond to 4 neighboring atoms with 2 electrons per bond. Donors (left) contribute an extra mobile electron (e^-) in addition to making bonds with 4 neighbors. Acceptors (right) contribute 3 electrons for bonding with neighbor atoms which leaves a mobile hole (h^+).

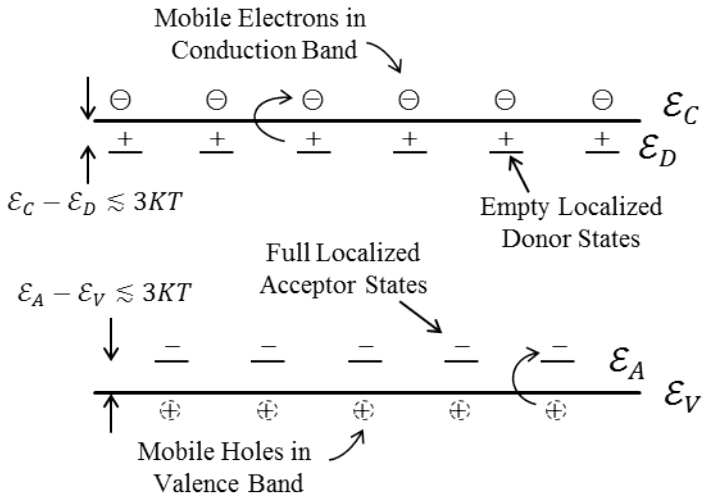


Figure 5.5: Donor level $\mathcal{E}_D$ and acceptor level $\mathcal{E}_A$ and their relative positions within the bandgap. Each dopant atom adds a localized state which becomes ionized by thermal energy.

just below the conduction band minimum, so they can easily gain sufficient thermal energy to jump to the conduction band. This gives rise to mobile electrons, and positively charged immobile donor ions that are fixed in the semiconductor crystal lattice.

The acceptor levels are found just above the valence band maximum, so electrons from the valence band can easily gain enough energy to attach to the acceptor

atoms. This gives rise to negatively charged immobile acceptor ions in lattice sites, and mobile positively charged holes in the valence band.

Charge Neutrality: It is important to note that uniformly doped semiconductors are not charged; overall they are charge neutral. By uniformly doped we mean that dopant atoms are distributed uniformly throughout the semiconductor, their density does not change with position. The reason for being charge neutral is that there will be the same concentration of ionized donors and acceptors as there will be electrons and holes. This leads to the following equation describing charge neutrality:

$$p - n + N_D^+ - N_A^- = 0 \tag{5.20}$$

5.6.1 Equilibrium Concentration of Electrons and Holes in Uniformly Doped Semiconductors

At equilibrium the following very important formula applies:

$$\boxed{n_i^2 = np} \tag{5.21}$$

where n_i is the intrinsic electron and hole concentrations. Now, once we dope, the electron and hole concentrations will be different. So, after doping we have the following:

- n is the mobile electron concentration (which is the concentration of electrons in the conduction band);
- p is the mobile hole concentration (which is the concentration of holes in the valence band);

Before going any further, let's remind ourselves what we mean by equilibrium. Generally, a semiconductor material is at equilibrium when it is just sitting there at a uniform temperature that is the same as its environment. Also, there should be no current flowing through it and there should not be any external light hitting it. Under these conditions, the material is typically at equilibrium. As we mentioned before, equation 5.21 only applies at equilibrium, and it comes from thermodynamics, and is analogous to thermal equilibrium in chemistry when the ratio of the concentration of the products to the reactants at chemical equilibrium is a constant for a given temperature.

Now let's apply charge neutrality and thermal equilibrium to obtain an expression for the mobile electron and hole concentrations in a doped semiconductor. Solving equations 5.20 and 5.21 simultaneously gives the following expression:

$$\boxed{n = \frac{N_D - N_A}{2} + \sqrt{\left(\frac{N_D - N_A}{2}\right)^2 + n_i^2}} \tag{5.22}$$

Once the electron concentration is calculated from equation 5.22, one can substitute it back into equation 5.21 to obtain the hole concentration at equilibrium.

Now there are several things to note here:

- If $N_D > N_A$ then the mobile electron concentration will be greater than the mobile hole concentration and the electrons are said to be the **majority carriers**, and holes are called the **minority carriers**. Similarly, if $N_A > N_D$, then holes are said to be the **majority carriers** and the mobile electrons are the **minority carriers**.

- If there is only one type of dopant atom, say for example $N_D = 10^{16}/cm^3$ and $N_A = 0$, then it is an excellent approximation to say that the electron concentration is equal to the donor concentration or $n = N_D = 10^{16}/cm^3$. Once you obtain n, you can immediately calculate the mobile hole concentration from equation 5.21 as $p = n_i^2/N_D$. So for example, if we have silicon where $n_i = 10^{10}/cm^3$ then $p = 10^4/cm^3$. Similarly, if the material is doped with acceptors, and $N_D = 0$, then $p = N_A$ and $n = n_i^2/N_A$.

- When a material is doped with both donors and acceptors, the higher doping will determine if the material is N-type or P-type. However, the precise value of n and p need to be obtained using equations 5.22 and 5.21.

  For example, if $N_D = 10^{17}$ and $N_A = 10^{15}$, and $n_i = 10^{10}$, all in units of $/cm^3$. Then using equations 5.21 and 5.22 gives $n = 9.9 \times 10^{16}$ and $p = 1.01 \times 10^3/cm^3$ and material is N-type since $N_D > N_A$, which results in $n > p$.

5.6.2 Semiconductor Carrier (Fermi) Statistics

The distributions of electrons and holes in equilibrium in the various energy states in the valence and conduction bands are described in detail by statistics. However, it is a specific type of statistics that takes into account the Pauli Exclusion Principle that says only one electron can exist in any quantum state. These statistics are called **Fermi Statistics** and they are generally very important in semiconductors, especially for more detailed analysis of concepts than we will do in this course. Fermi statistics tell us what the probability is that a particular quantum energy state is occupied by a single electron, and this probability is between zero and one. One of the key concepts in Fermi statistics is the **Fermi Level**. The Fermi level is the energy where Fermi Statistics says that the probability of a state being occupied is 1/2. At energies above the Fermi level we find that the probability of states being occupied is less than 1/2; and at energies below the Fermi level the probability of states being occupied is greater than 1/2. Figure 5.7 shows the location of the Fermi level for intrinsic, N-type and P-type semiconductors, respectively.

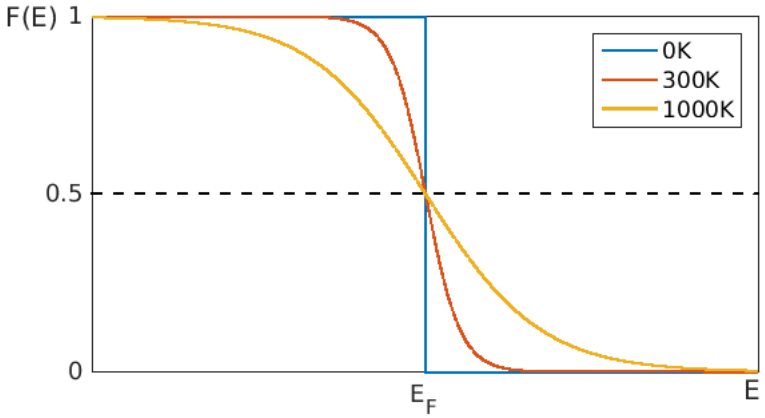


Figure 5.6: The Fermi function $F(\mathcal{E})$, depicting the greater than 50% chance of occupancy below the Fermi level $\mathcal{E}_F$ and the less than 50% chance above.

The form of the Fermi function is given in Figure 5.6. From this plot, we can see that as temperature increases, the Fermi function changes from a step function (at 0K) to a more gently changing slope, allowing more occupancy of the higher energy states (and consequently, less occupancy of the lower energy states) due to increased thermal energy.

Below are given some of the important expressions that relate mobile electron and hole concentrations to the conduction band, valence band and Fermi energies in equilibrium.

The probability that a quantum state is occupied by an electron is provided by the Fermi Probability Function which is given as the following expression:

$$F(\mathcal{E}) = \frac{1}{\exp\frac{\mathcal{E}-\mathcal{E}_F}{KT}+1} \tag{5.23}$$

Where $\mathcal{E}_F$ is the Fermi Level or Fermi Energy, which was described above, and K is the Boltzmann constant.

It follows that if there are N_q quantum states in a system at a particular energy $\mathcal{E}$, then the number of electrons n_e in those states is given by the number of quantum states N_q multiplied by the probability $F(\mathcal{E})$ that each state is occupied. This is expressed mathematically as:

$$n_e = N_q \times F(\mathcal{E}) = \frac{N_q}{\exp\frac{\mathcal{E}-\mathcal{E}_F}{KT}+1} \tag{5.24}$$

At this point the question often arises as to 'how do we know the Fermi Level', because without knowing it, we can not evaluate the above expression to calculate the number of electrons present. This is a good question and is typically answered by the fields of Statistical Mechanics and Chemical Thermodynamics. However, in semiconductors at the level of this text, we will treat the Fermi Level as a reference

point that helps describe the semiconductor system. The level will typically be given with respect to another known quantity, which we will see below.

Electrons Concentration in Conduction Band and Hole Concentration in Valence Band at Equilibrium

Fermi Statistics are often used to calculate the number of electrons in the conduction band or holes in the valence band. To calculate the number of electrons in the conduction band we apply the following logic.

$$n = \int_{\mathcal{E}_C}^{\infty} F(\mathcal{E}) DoS(\mathcal{E}) d\mathcal{E} \tag{5.25}$$

Where $DoS(\mathcal{E})$ is the density of states and $\mathcal{E}_C$ is the edge of the conduction band. The $DoS(\mathcal{E})$ is a function often seen while studying the electrical properties of crystalline solids. It gives the number of allowed quantum states that electrons can occupy per unit energy. Later courses and more detailed investigations in semiconductors often heavily involve the density of states. In this text it suffices to say qualitatively what it is, and how it is used with the Fermi function to calculate the number of electrons in the conduction band and the number of holes in the valence band. By multiplying the density of states in the conduction band by the probability of occupancy of each state, we integrate to obtain the total number of occupied states in the conduction band and thus the number of mobile electrons. Similarly, to calculate the number of mobile holes in the valence band, we multiply the valence band density of states by the probability the state is unoccupied $1 - F(\mathcal{E})$, then integrate. In the Appendix of this text, we describe in more detail the $DoS(\mathcal{E})$, and how these integrals are performed to obtain the expressions 5.27 and 5.28 below.

$$p = \int_{-\infty}^{\mathcal{E}_V} (1 - F(\mathcal{E})) DoS(\mathcal{E}) d\mathcal{E} \tag{5.26}$$

Relationship between Fermi Level, Electron and Hole Concentrations in Equilibrum and Effective Density of States

As it turns out, these integrals don't have a closed form expression but can be approximated by replacing the Fermi function with a Boltzmann function, assuming the Fermi level is more than a few KT below the conduction band edge (or above the valence band edge for holes). Making this approximation, we essentially create an effective density of states at the band edge denoted as N_C and N_V for the conduction and valence bands respectively.

$$n = N_C \exp[-(\mathcal{E}_C - \mathcal{E}_F)/KT] \tag{5.27}$$

$$p = N_V \exp[-(\mathcal{E}_F - \mathcal{E}_V)/KT] \tag{5.28}$$

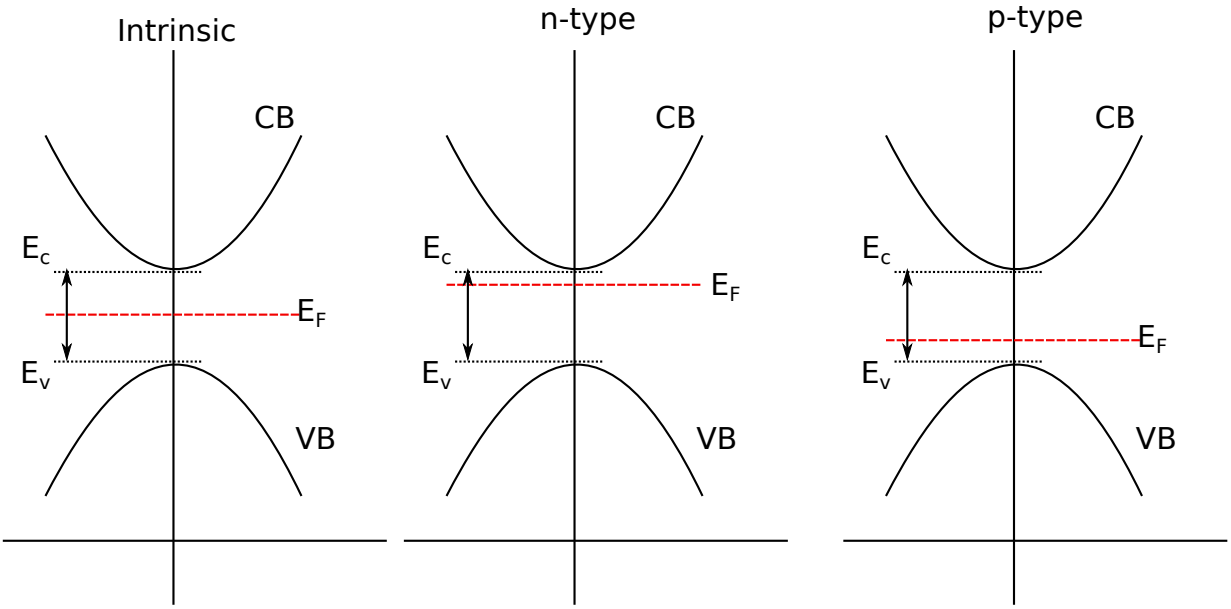


Figure 5.7: E-k diagram showing intrinsic (left), N-type (center) and P-type (right) materials as indicated by the position of the Fermi Level. The figure also indicates the conduction band, the valence band and the bandgap.

$$N_C = 2\left(\frac{2\pi m_n^* KT}{h^2}\right)^{3/2} \tag{5.29}$$

$$N_V = 2\left(\frac{2\pi m_p^* KT}{h^2}\right)^{3/2} \tag{5.30}$$

Where m_n^* and m_p^* are the density of states effective masses for electrons and holes respectively. By inserting the electron and hole concentrations obtained from equations 5.27 and 5.28 into the equation for intrinsic carrier concentration (Eq. 5.21), we obtain the formula for n_i in terms of the bandgap ($\mathcal{E}_g$) of the material and the temperature.

$$n_i = \sqrt{N_C N_V}\exp\left(\frac{-\mathcal{E}_g}{2KT}\right) \tag{5.31}$$

Intrinsic Fermi Level

Equations 5.27 and 5.28 give relationships between the Fermi level and the mobile electron and hole concentrations. By taking the product of n and p we also obtain a relationship between the intrinsic carrier concentration and the bandgap for a specific material at equilibrium. From this relationship, given by Equation 5.31, it becomes evident that n_i increases with increasing temperature. The intrinsic carrier concentrations increase as the bandgap decreases since since electrons will need less energy to be excited from the valence to the conduction band.

At this point it is also worth defining a specific value of the Fermi level when the carrier concentration is at its intrinsic value. We call this the **Intrinsic Fermi Level** $\mathcal{E}_{Fi}$ and the following expressions hold:

$$n_i = N_C \exp[-(\mathcal{E}_C - \mathcal{E}_{Fi})/KT] \tag{5.32}$$

$$p_i = N_V \exp[-(\mathcal{E}_{Fi} - \mathcal{E}_V)/KT] \tag{5.33}$$

And of course, $n_i = p_i$ since in intrinsic semiconductors at equilibrium, every electron excited into the conduction band will also leave behind or generate a hole in the valance band.

Fermi Level Location

In semiconductors, the Fermi level is found somewhere in the bandgap. In particular, in N-type semiconductors the Fermi level will be relatively close to the conduction band minimum $\mathcal{E}_C$. If a semiconductor is doped with acceptors so it is P-type, then the Fermi level will be located relatively close the valence band maximum $\mathcal{E}_V$. For intrinsic semiconductors, the Fermi level will be approximately at the center of the bandgap, and if the electron and hole effective masses are the same, $\mathcal{E}_{Fi}$ will be exactly at the bandgap center.

Fermi Level and the Chemical Potential

The Fermi level has its origins in chemistry and thermodynamics. It is formally the average energy required to add an electron to a material. In semiconductors in equilibrium the Fermi level is useful when different materials are joined. In this case electrons in the material with a higher Fermi level transfer to the electrons with the lower Fermi level to establish equilibrium between the two material. The difference in the two Fermi levels is called the work function or the electrochemical potential difference between the materials. We first were introduced to the work function when we studied the photoelectric effect for metals. When electrons transfer from one material to another to achieve equilibrium, the work function becomes equal to a built-in potential energy that will typically exist in a region where the the two materials come together. We will learn more about this built in potential in the later chapters of this book.

Example 5.1:
Find the intrinsic carrier concentration for silicon at room temperature $(300K)$.

Taking the effective masses to be $m_n^* = 1.08m_0$ and $m_p^* = 0.81m_0$, we find the effective density of states:

$$N_C = 2\left(\frac{2\pi m_n^* KT}{h^2}\right)^{3/2} = 2.8 \times 10^{19} cm^{-3}$$
$$N_V = 2\left(\frac{2\pi m_p^* KT}{h^2}\right)^{3/2} = 1.8 \times 10^{19} cm^{-3}$$

Values on the order of $10^{19} cm^{-3}$ are typical for most semiconductors. Substituting these into equation 5.31 using a bandgap of $1.1eV$:

$$n_i = \sqrt{N_C N_V} exp\left(\frac{-\mathcal{E}_g}{2KT}\right) = 1.3 \times 10^{10} cm^{-3}$$

The value reported in literature varies from around $1 - 1.5 \times 10^{10} cm^{-3}$.

Example 5.2:
A piece of silicon is doped with Donors $N_D = 1 \times 10^{18} cm^{-3}$ and Acceptors $N_A = 3 \times 10^{18} cm^{-3}$. Take $n_i = 10^{10} cm^{-3}$ and $N_V = 2 \times 10^{19} cm^{-3}$. Where is the Fermi level in this material?

First, because this material is doped with both types of dopant (also known as counter-doping), we must use equation 5.22 and solve for p using 5.21. Alternatively we can derive a similar formula to 5.22 by solving for p instead of n, since we know the material will be more p-type because $N_A > N_D >> n_i$.

$$p = \frac{N_A - N_D}{2} + \sqrt{(\frac{N_A - N_D}{2})^2 + n_i^2} = 2 \times 10^{18} cm^{-3}$$

This result makes sense because n_i is negligible compared to the doping concentration, so the equation simplifies to $p = N_A - N_D$. Continuing, we use equation 5.28 to solve for $\mathcal{E}_F - \mathcal{E}_V$:

$$p = N_V \exp[-(\mathcal{E}_F - \mathcal{E}_V)/KT]$$

Substituting the above values for p and N_V

$$2 \times 10^{18} cm^{-3} = \left(2 \times 10^{19} cm^{-3}\right) \exp(-(\mathcal{E}_F - \mathcal{E}_V)/KT)$$

Solving for $(\mathcal{E}_F - \mathcal{E}_V)$ gives

$$(\mathcal{E}_F - \mathcal{E}_V) = KT \ln(10) = 59.5 meV \approx 0.06 eV$$

In other words, the Fermi level is $0.06eV$ above the valence band edge. This is very close to the valence band as compared to in an intrinsic material where the Fermi level is around mid-gap or $0.55eV$ above the valence band edge for Si.

5.7 Problems

5.1 Questions about Band Structure of Solid Crystals

(a) Explain how energy bands arise in a crystal by bringing Avogadro's number of atoms together. How many quantum states are in each energy band?

(b) Sketch the band structure ($\mathcal{E}$ vs. k) for a metal, semiconductor and an insulator.

(c) What is the difference between an insulator, a metal and a semiconductor in terms of the bands? Explain this using bandgap and the filling of states by electrons.

(d) Why can't a full energy band conduct electricity.

(e) What is the difference between a conduction electron and a non-conduction electron?

(f) What is a hole?

(g) What is the general form of the wave-function for electrons in the periodic potential of a crystalline solid?

(h) What is the general form of the Schrodinger equation for a crystal? (Insert form of potential into SWE)

(i) How do you calculate the effective mass of an electron in a crystalline solid, and what does it account for?

5.2 Graph the intrinsic electron and hole concentrations of silicon as a function of temperature ranging from $-200°C$ to $+200°C$. Also, what is the carrier concentration at room temperature ($27°C$)? The bandgap of Silicon is 1.1eV, the effective masses for electrons and holes are $0.33m_0$ and $0.28m_0$, respectively, where m_0 is the mass of a free electron.

5.3 Calculate the number of atoms per cm^3 in Si.

5.4 Intrinsic Carrier Concentration and Fermi Level

(a) Graph the intrinsic electron concentration as a function of bandgap. For this calculation make the approximation that $m_n^* = m_p^* = m_0$ for all the semiconductors. Assume $\mathcal{E}_g$ ranges from $0.1eV$ to $6eV$. Comment on the values of n_i compared to the concentration of atoms in the crystals. (Assume the density of atoms is the same as what you calculated above for Si). Take the temperature to be $27°C$.

(b) Graph the Fermi Level with respect to the valence band edge $\mathcal{E}_V$ for the materials listed in the previous part of this problem.

(c) Graph the Fermi Level with respect to the conduction band edge $\mathcal{E}_C$ for the materials listed in the previous part of this problem.

(d) From your graphs in the previous to parts of this problem, what can we say about the Fermi level for an intrinsic semiconductor with respect to its location within the bandgap.

5.5 In your own words explain what intrinsic, N-type and P-Type semiconductors are.

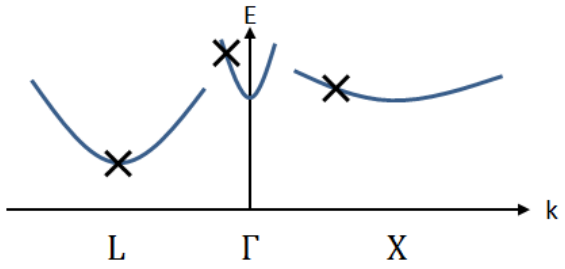
5.6 Why is Boron used as an acceptor in dopant and why is Phosphorous used as a donor dopant in Silicon.

5.7 In a silicon bar uniformly doped with 10^{16} phosphorus atoms per cm^3 and 5×10^{14} boron atoms per cm^3. Calculate the mobile electron and hole concentrations for this bar. (Note that phosphorus is a donor and boron is an acceptor for silicon.)

5.8 The silicon conduction band can be expressed as $\mathcal{E} = \frac{\hbar^2 k^2}{2m^*}$. Plot $\mathcal{E}$ vs k with the vertical axis $\mathcal{E}$. Plot for $\mathcal{E}$ less than $0.5eV$. Take $m^* = 0.26m_0$, where m_0 is the mass of a free electron. (Be careful with units.)

5.9 Using the energy vs. velocity relationship: $\nu = \frac{1}{\hbar}\frac{d\mathcal{E}}{dk}$, and the energy vs. k relationship in problem above, plot the velocity of a particle versus k for values of k between 0 and $2 \times 10^9/meter$.

5.10 The following band diagram illustrates the dispersion relation ($\mathcal{E}$ vs k) for conducting electrons.



(a) At which location marked × would an electron have the largest instantaneous velocity? Which would be the slowest?

(b) Order the valleys from lightest to heaviest effective mass.

5.11 What happens to the intrinsic carrier concentration as the temperature increases? If the intrinsic carrier concentration is larger than the dopant concentration, where does this put the Fermi level?

5.12 The energy bandgap of silicon is 1.1eV. For intrinsic silicon, graph the Fermi function $F(\mathcal{E})$ for an energy range starting from 1eV less than $\mathcal{E}_F$ to 1eV greater than $\mathcal{E}_F$ at T=300K. What is the maximum value of $F(\mathcal{E})$ and why does $F(\mathcal{E})$ become flat at its maximum value as $\mathcal{E}$ becomes several multiples of KT less than $\mathcal{E}_F$?

5.13 If $\mathcal{E}$ is several times KT greater than $\mathcal{E}_F$, show that $F(\mathcal{E})$ can be written as an exponential function.

5.14 N-Type Doping, P-Type Doping, and the Fermi Level

(a) If silicon is doped with phosphorus with concentration of $N_D = 1 \times 10^{17}/cm^3$, and all the phosphorus donors are ionized, what is the mobile electron mobile concentration? Calculate the position of the Fermi level with respect to the position of the conduction band minimum. Sketch the band diagram, the location of the Fermi Level and the location of the donor level in the diagram.

(b) If silicon is doped with boron with a concentration of $N_A = 1 \times 10^{17}/cm^3$, and all the acceptors are ionized, what is the mobile hole concentration? Calculate the position of the Fermi level with respect to the position of the valence band maximum. Sketch the band diagram, the location of the Fermi Level and the location of the acceptor level in the diagram.

(c) Comment on the location of the Fermi Level for N-type and P-type semiconductors with respect to conduction and valence band edges, $\mathcal{E}_C$ and $\mathcal{E}_V$, respectively.

Chapter 6

Semiconductor Currents: Drift and Diffusion

6.1 Introduction

Previously we learned about the quantum mechanics that explains the differences between semiconductors and metals and insulators; material band structures and how they give rise to electrons and holes in semiconductors, and other quantum related parameters like effective mass. We are now ready to use these concepts on a more classical level to describe electron and hole flow in semiconductors. We will learn about Drift and Diffusion, which are the two main mechanisms of current flow in semiconductors. We will also learn that a built-in electric field occurs when semiconductors are not uniformly doped. We will learn that the built-in electric arises in response to drift and diffusion currents tending to oppose each other. We will begin to find that non-uniform doping is at the heart of what makes solid state devices, such as transistors and diodes, work like they do.

6.2 Drift and Diffusion Currents

There are two mechanisms of current flow in semiconductors: **Drift** and **Diffusion**. Drift is the current that is driven by an electric field inside the semiconductor device or material. Diffusion is the current that is driven by a concentration gradient of electrons or holes inside a semiconductor.

6.2.1 Drift Current

Suppose we have a simple plain rectangular bar that is made up of silicon, and we apply a DC voltage between the two ends of the bar. Inside the bar an electric field will be establish because of the applied voltage. Also, since there are mobile electrons and holes, the electric field will give rise to an electric current. (Recall, that mobile electrons are the ones that are in the conduction band, and the mobile holes are those that are found in the valence band.) This current, which is in response to a macroscopic electric field, is called **Drift Current**. There will be a drift current for electrons and a separate drift current for holes, and they are given by the following expressions:

$$\vec{J}_{ndrift} = q\mu_n n\vec{E} \tag{6.1}$$

$$\vec{J}_{p_{drift}} = q\mu_p p\vec{E} \tag{6.2}$$

Where $\vec{J}_{ndrift}$ is the electron drift current density, typically in units of $amps/cm^2$; μ_n is the electron mobility, typically in units of $cm^2/Volt\,sec$; n is the mobile electron concentration, which is the same thing as the concentration of electrons in the conduction band and typically in units of electrons/cm^3; and $\vec{E}$ is the electric field, which is typically in units of $Volts/cm$.

$\vec{J}_{p_{drift}}$ is the hole drift current; μ_p is the hole mobility; p is the hole concentration, which is the same thing as the concentration of holes in the conduction band.

$q = 1.6 \times 10^{-19}\ Coulomb$ is the elementary charge which is a positive number.

Note that the sign on both the electron and hole drift currents are the same. This is because electrons and holes will move in opposite directions in response to the electric field, however, since the charge of an electron is negative, the negative charge and the negative direction multiply to give a positive drift current.

We will discuss mobility more, later in this chapter. For now, please remember that mobility is the proportionality factor that relates the electric field to the electron or hole average velocity that results in response to the field.

We can also express the drift currents using the electrostatic potential. Recall from your electromagnetism classes that the electric field is the negative gradient of the electrostatic potential ϕ, or:

$$\vec{E} = -\nabla\phi \tag{6.3}$$

Where the electrostatic potential ϕ is in units of Volts.

Substituting 6.3 for $\vec{E}$ in equations 6.1 and 6.2, we often find drift currents expressed as follows:

$$\vec{J}_{ndrift} = -q\mu_n n\nabla\phi \tag{6.4}$$

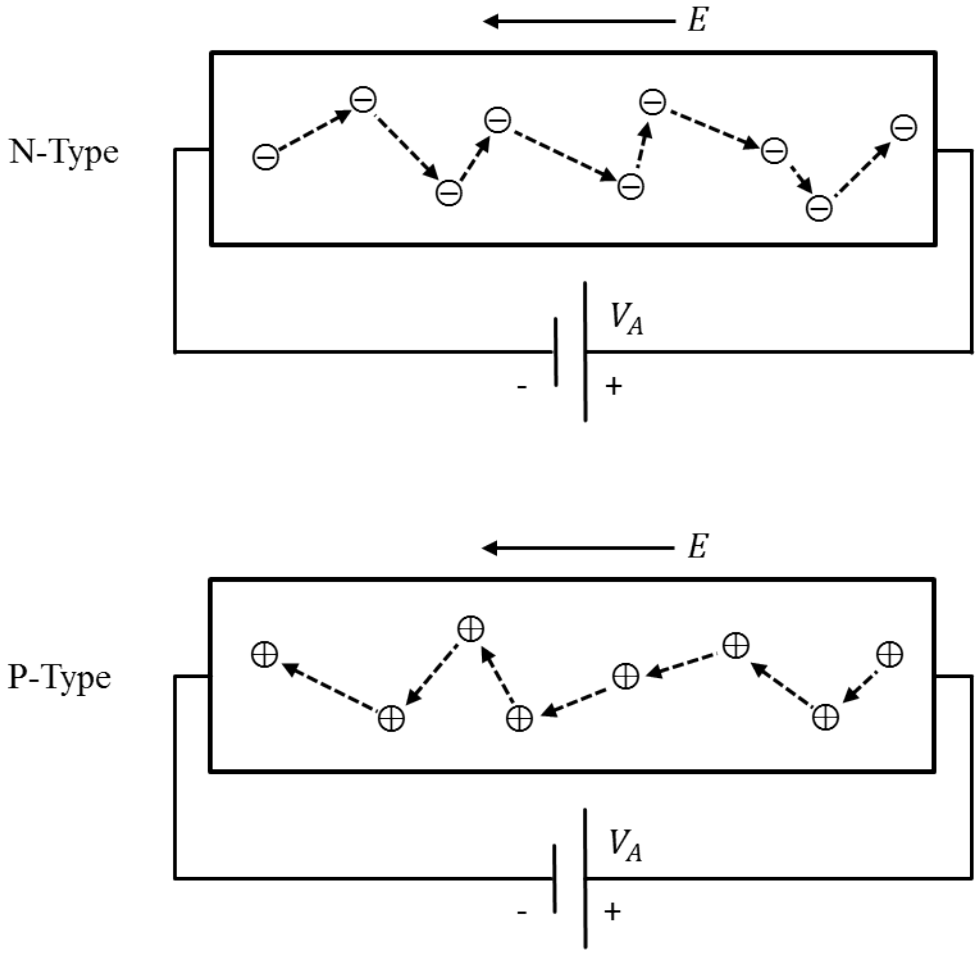


Figure 6.1: Electrons and holes drifting under the effect of an applied electric field. The mobility which relates the drift velocity to the electric field strength is inversely proportional to the amount of scattering the carrier does with the lattice.

$$\vec{J}_{p_{drift}} = -q\mu_p p \nabla\phi \tag{6.5}$$

Comment on Electric Field in Drift Current Expression

The electric field in equations 6.1 and 6.2 is a macroscopic field. It should not get confused with the field due to the period potential of the atoms in the lattice. The electric field in the current equations is typically due to an applied voltage. We will learn later that it will also arise when we have non uniform doping, such as in a PN junction. We call this field macroscopic because it does not vary in space nearly as much as the field due to the periodic potential of atoms. For drift currents in semiconductors we will always be talking about this macroscopic field.

Comment on Electron Drift and Scattering

It is important to point out that drift current is really due to the average motion of electrons and holes in response to a macroscopic electric field. This is illustrated in Figure 6.1. In general, electrons will flow toward the electric field, however, as they flow they are regularly being scattered through interactions with the crystal lattice, which randomize their momentum. So the overall motion will be something like this: The electron gets pulled toward the field, then it interacts with the lattice and gets scattered in another direction; it then gets pulled in the direction of the field again, then it gets scattered by the lattice, etc. This process continues until the electron, as shown in Figure 6.1, that is emitted at the negative contact on the semiconductor bar, finally makes its way to the other end where it is collected at the positive contact of the bar. A similar process occurs for hole transport, but in the other direction since holes are positively charged. The average distance and electron travels between scattering events is called the mean free path, and is about one nanometer. The average velocity of the electrons or holes, in the direction of the field after all the scattering, is called the drift velocity. We will talk about this more later in this chapter in the section on electron and hole mobility.

Example 6.1
A silicon bar is uniformly doped with $10^{17}/cm^3$ boron atoms. The bar is $100\mu m$ long. A positive voltage of $V_A = 10.0V$ is applied to the left end of the bar, and the right end of the bar is grounded. What is the current density in the bar?

Since the bar is uniformly doped with boron at $10^{17}/cm^3$, the hole concentration will be $p = 10^{17}/cm^3$, and the electron concentration $n = n_i^2/p = 10^3/cm^3$. Since n is fourteen orders of magnitude less than p, we will neglect it and use equation (6.4), which is one dimension is:

$$J_{p_{drift}} = -q\mu_p \frac{d\phi}{dx} = J.$$

Since the bar is uniformly doped, $\frac{d\phi}{dx} = \frac{V_A}{L}$. Substituting in values for p, L and V_A from above, taking $\mu_p = 450cm^2/V\,sec$ from the Appendix, and putting all lengths in centimeters gives:

$$J = -(1.6 \times 10^{-19}C)(450cm^2/V\,sec)(10^{17}/cm^3)\frac{10.0V}{1 \times 10^{-2}cm} = -7200A/cm^2$$

6.2.2 Diffusion Current

In semiconductors we also have diffusion current. This current arises due to random motion of electrons or holes in the presence of a concentration gradient. For example, if there is a larger concentration of electrons in one region in a semiconductor material than another region, then random thermal motion of the electrons will tend to have them move on average from the region of higher concentration to the region of lower concentration, giving rise to a diffusion current. Situations like this will also occur for holes as well. Note that semiconductors are different from metals. We generally do not have diffusion currents in conductors because the electron concentrations in conductors are uniform (do not vary with position) and thus do not have concentration gradients. Diffusion currents typically arise in semiconductors when we have non-uniform doping. The expressions for electron and hole diffusion currents are given by:

$$\vec{J}_{ndif} = qD_n\nabla n \tag{6.6}$$

$$\vec{J}_{p_{dif}} = -qD_p\nabla p \tag{6.7}$$

Where $\vec{J}_{ndif}$ is the electron diffusion current density; D_n is the diffusion coefficient for electrons; $\vec{J}_{p_{dif}}$ is the hole diffusion current and D_p is the diffusion coefficient for holes.

Note that the diffusion coefficient is the proportionality factor that relates the gradient in the charge carrier concentration to the current. We will discuss the diffusion constant later in this chapter.

6.2.3 Total Current

The total current density for electrons in the conduction band is given by the sum of the drift and diffusion electron currents. Similarly, the total hole current density in the valence band is given by the sum of the hole drift and diffusion current:

$$\vec{J}_{nT} = -q\mu_n n\nabla\phi + qD_n\nabla n \tag{6.8}$$

$$\vec{J}_{pT} = -q\mu_p p\nabla\phi - qD_p\nabla p \tag{6.9}$$

In our class, we will typically be doing our analyses in one dimension, where the gradient operators reduce to simple derivatives:

$$J_{nT} = -q\mu_n n\frac{d\phi}{dx} + qD_n\frac{dn}{dx} \tag{6.10}$$

$$J_{pT} = -q\mu_p p\frac{d\phi}{dx} - qD_p\frac{dp}{dx} \tag{6.11}$$

Now it is important to note that these expressions describe the current density, and they will typically be functions of position inside the semiconductor or the device. Similarly, the electrostatic potenital $\phi(x)$ and the carrier concentrations $n(x)$ and $p(x)$ are typically functions of position.

Total Current: At any given point x inside the semiconductor, we may have both electron current and hole current. The total current at any given point is the sum of the electron and hole currents at that point, or

$$J_T = J_{nT}(x) + J_{pT}(x) \tag{6.12}$$

6.3 Derivation of Mobility, Diffusion Coefficient and the Einstein Relation

6.3.1 Electron and Hole Mobility

As mentioned before, the electron mobility μ_n is the proportionality factor that relates the electric field to the average velocity of electrons due to drift in the electric field.

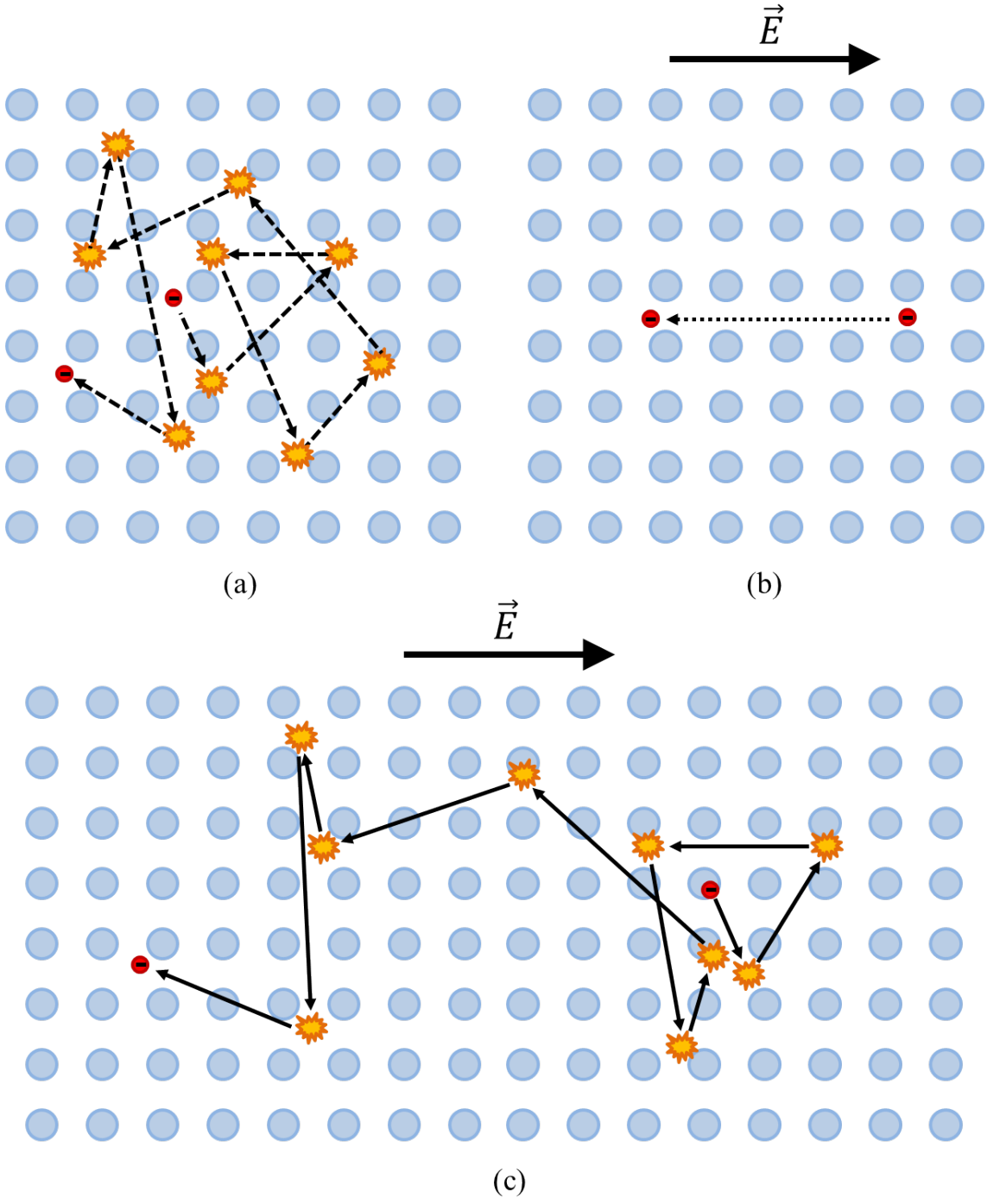


Figure 6.2: Diagram of drift motion inside of a semiconductor under an applied field. Explosion symbols represent scattering events which change the direction of motion. (a) Random thermal motion with average 0 displacement due to the symmetrical distribution of instantaneous velocities. (b) Drift motion of electron purely under the effect of the electric field. **Note: This is an idealized situation without scattering which under all practical instances won't happen.** (c) Total combined motion of electron drifting and scattering in a semiconductor.

Consider again the ordinary semiconductor bar that we described earlier. The bar has a voltage applied between the two ends so there will be a current due to drift in the bar. The drift current is actually an average flow of the electrons in response to the applied field. However, the electrons will also have an instantaneous random or thermal velocity v_t that is typically significantly larger than the average velocity due to drift. We talked about this velocity in the previous chapter where we calculated it from the slope of the $\mathcal{E}$ *vs.* k band diagram (Here we have renamed it and are calling it v_t as opposed to v_g, but it's the same thing, which is the instantaneous velocity). As electrons are pulled by the macroscopic electric field, their motion is quasi random due to electrons scattering with the lattice. Actually they scatter with thermal vibrations of the lattice called phonons. However, electrons will flow in response to the field with a net or average velocity $< v >$. The proportionality factor relating this average velocity to the electric field is call the **electron mobility**. We derive the expression for the electron mobility as follows:

The total average force on electrons in the semiconductor bar with the electric field inside is given by Newton's law:

$$m_n^* \vec{a} = \vec{F}_{field} + \vec{F}_{scattering} \tag{6.13}$$

Where m_n^* is the electron effective mass; $\vec{a}$ is the acceleration of the electrons; $\vec{F}_{Efield} = -q\vec{E}$ is the force on the electrons due to the macroscopic electric field.

$\vec{F}_{scattering} = \frac{-m_n^* < \vec{v} >}{\tau}$ is the average force on the electrons due to scattering with the crystal lattice as the electrons move about in the semiconductor and are pulled by the field. The scattering tends to oppose the force from the field. It is actually largely the process of electron scattering by the crystal lattice that gives rise to electrical resistance. The parameter τ = average time between collisions or scattering events and $\frac{1}{\tau}$ is the average scattering rate.

Overall, the electrons do not accelerate because whenever they gain velocity from the electric field, they then will lose that extra velocity as a result of scattering. Thus the total average force on the mobile drifting electrons is zero, or $m_n^* \vec{a} = 0$. So we have

$$0 = -qE - \frac{m^* < \vec{v} >}{\tau} \tag{6.14}$$

This can be solved for the average velocity:

$$< \vec{v} >= -\frac{q\tau}{m_n^*} \vec{E} \tag{6.15}$$

or

$$< \vec{v} >= -\mu_n \vec{E} \tag{6.16}$$

Where we have defined **electron mobility** as:

$$\boxed{\mu_n = \frac{q\tau}{m_n^*}} \tag{6.17}$$

Going through a similar analysis gives an analogous expression for hole mobility:

$$\boxed{\mu_p = \frac{q\tau}{m_p^*}} \tag{6.18}$$

Drift Velocity: By convention, we call the average velocity that arises in response to the electric field the **drift velocity** and designate it as : $< \vec{v} >= \vec{v}_{drift}$.

The motion of an electron drifting in an electric field is illustrated in Figure 6.2. In the illustration, (a) shows the random thermal motion of the electron scattering randomly with no preferred direction. On top of this motion, the electron also experiences a Coulombic force in the direction opposed to the electric field due to its negative charge. This force adds a strong bias to the direction the electron is moving resulting in an average drift velocity in the direction of the field with proportionality constant equal to the mobility μ_n.

Resistance and Scattering mechanisms: There are various scattering mechanisms which influence the motion of the electron traveling in the crystal. We have already discussed this scattering earlier in this chapter. These scattering events include collisions with defects and atomic vibrations called phonons which locally change the perfectly periodic lattice potential of the atoms. Such scattering is actually the main cause of resistance in a crystal. The more scattering there is, the higher the resistance. Also, affecting the resistance is the doping concentration. For example, the higher the donor concentration in an N-type silicon bar, the lower the resistance of the bar.

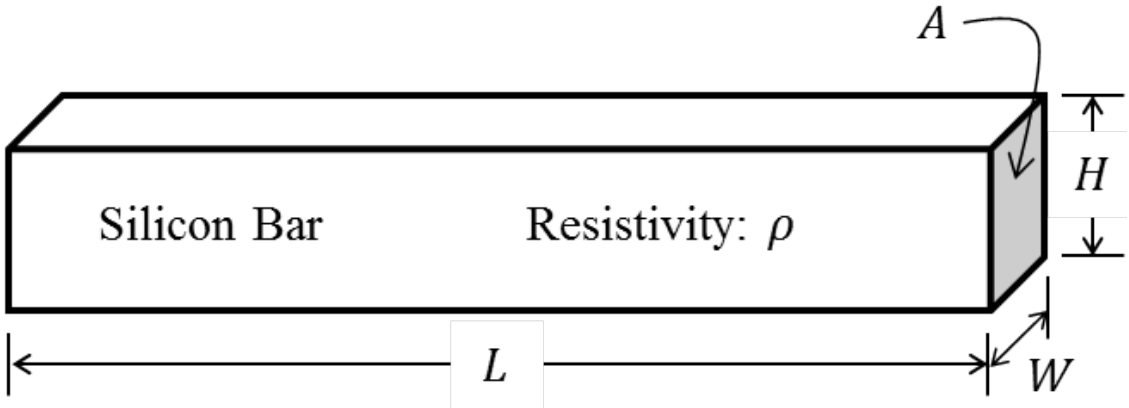


Figure 6.3: Dimensions of a resistive silicon bar.

The **resistivity** ρ of a bar of semiconductor is related to its conductivity σ which is in turn related to the mobility through the average scattering rate.

$$\rho = \frac{1}{\sigma} = \frac{1}{q(\mu_n n + \mu_p p)} \tag{6.19}$$

If the scattering rate for electrons and holes are both equal to τ, then resistivity can be written as:

$$\rho = \frac{1}{q^2\tau(\frac{n}{m_n^*} + \frac{p}{m_p^*})} \tag{6.20}$$

To calculate the full **resistance** of a bar of cross sectional area A and length L:

$$R = \frac{\rho L}{A} = \frac{L}{q(\mu_n n + \mu_p p)A} \tag{6.21}$$

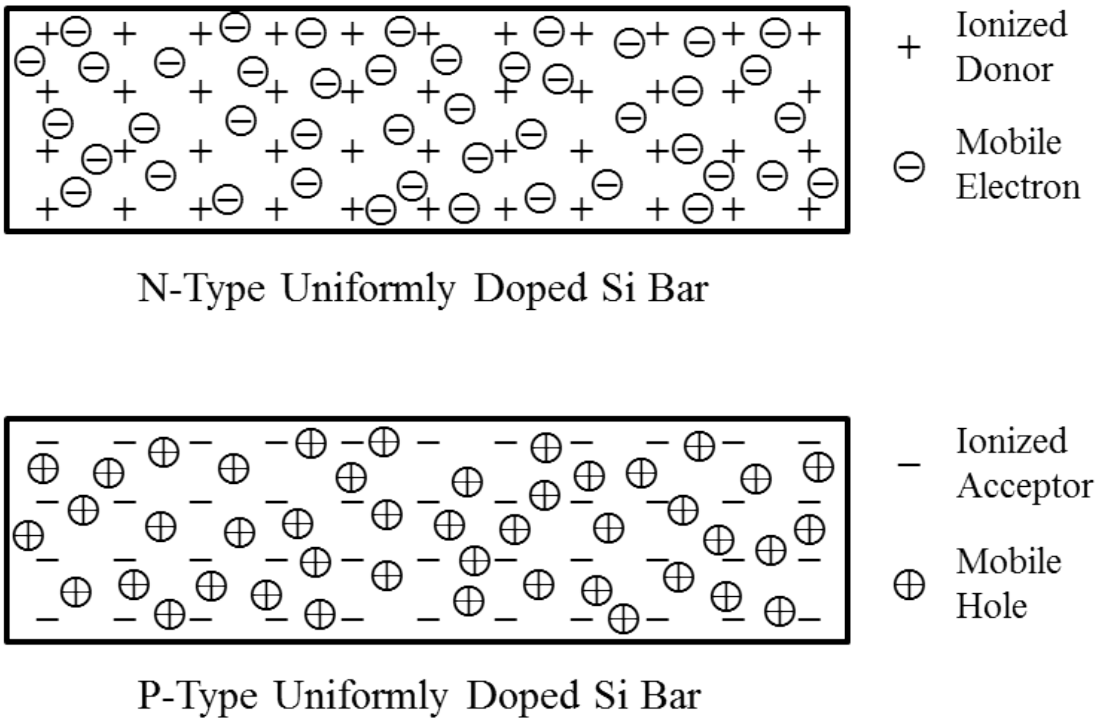


Figure 6.4: Charges in uniformly doped slabs of silicon. The ionized dopants are fixed in the crystal lattice but the mobile charges are free to move around. There are an equal number of ions and carriers so each bar is charge neutral as a whole.

6.3.2 Derivation of the Diffusion Current and Coefficient

In this section we derive the diffusion coefficients for electrons and holes: D_n and D_p. The derivation is aimed at addressing the following scenario: Let's determine the net flow of carriers across a plane, located at point $x = 0$, due to the random motion of carrier from the left and the right of the plane.

To do this we first realize that electrons are moving randomly with thermal velocity v_t. Due to this random motion we now define a flux of electrons across a plane at point $x = 0$ due to electrons on either side of this plane starting from locations $x = -l$ and $x = +l$, respectively. Here l is the 'mean free path', which is the short distance the electron travels between scatterings.

F(0) = net flux at x=0, which is composed of the flux from the left starting from the point of the last scattering $x = -l$ minus the flux from the right, again since the point of last scattering $x = +l$.

$$F(0) = \frac{1}{2}v_t n(-l) - \frac{1}{2}v_t n(l) \tag{6.22}$$

Where the fraction 1/2 comes from the idea that at any given point, approximately 1/2 of the randomly moving particles are moving toward the plane at $x = 0$ and the other half are moving away from it. Now, since the mean free path l is a small distance, we expand using Taylors series:

$$F(0) = \frac{v_t}{2}\left[n(0) - \frac{dn}{dx}l\right] - \frac{v_t}{2}\left[n(0) + \frac{dn}{dx}l\right] \tag{6.23}$$

This reduces to:

$$F(0) = -lv_t\frac{dn}{dx} \tag{6.24}$$

We now transform the net flux to the electron current density by multiplying it by the charge on an electron $-q$:

$$J_{n_{diff}} = -qF = qlv_t\frac{dn}{dx} \tag{6.25}$$

Now we define the quantity $lv_t = D_n$, we obtain the familiar expression for electron diffusion current density.

$$\boxed{J_{n_{diff}} = qD_n\frac{dn}{dx}} \tag{6.26}$$

where, D_n is the diffusion coefficient for electrons.

From Figure 6.5, it is clear that a **concentration gradient** will to produce a net diffusion current. Recall that the instantaneous velocity of the particles is defined by the $\mathcal{E}$ versus k relationship. Because this relationship is symmetrical for $\pm k$, it ensures that for each state traveling in the positive direction, there is always another state traveling in the negative direction with the same speed. This fixes the fraction of particles moving left or right at a given point to $\frac{1}{2}$, so the only way to produce a net flow of particles is to have a concentration gradient present.

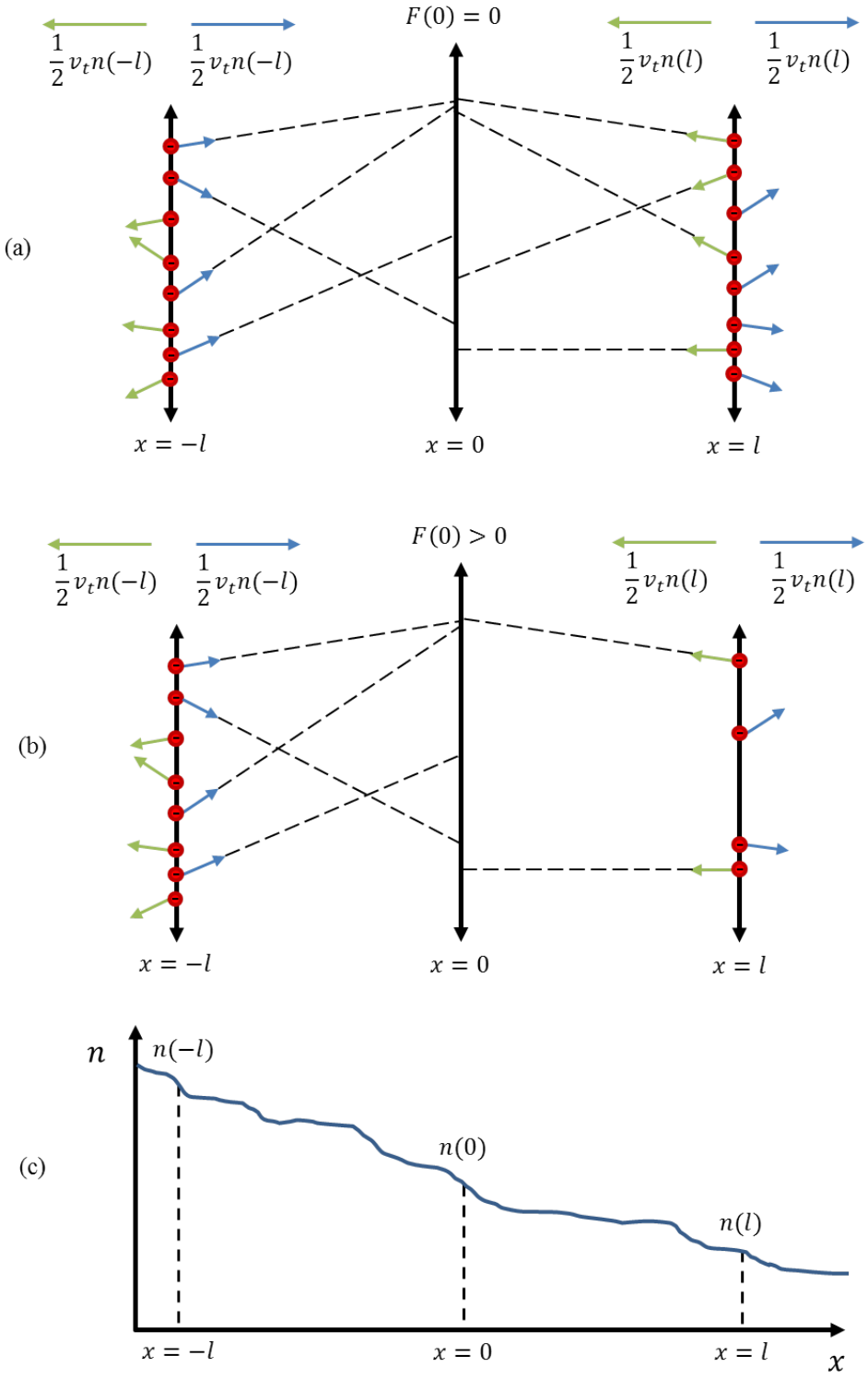


Figure 6.5: Electrons at any point move with random directions due to thermal motion so on average $\frac{1}{2}$ are moving to the left, and $\frac{1}{2}$ are moving to the right. (a) Total flux (and current) is zero because there are the same number of electrons moving left and right across $x = 0$. (b) More electrons are moving right across the $x = 0$ line so there is a net flux (and thus a diffusion current). (c) Actual electron concentration is a quantity which varies continuously with position.

6.3.3 The Einstein Relation between Mobility and Diffusivity

The Einstein relationship is important in semiconductors. It relates carrier mobility μ to carrier diffusivity D and is given by:

$$\boxed{D = \mu \frac{KT}{q}} \tag{6.27}$$

Where K is the Boltzmann constant and T is degrees Kelvin.

Now, we derive the Einstein relation. We first write the mean free path as the product of the thermal velocity and the time between collisions:

$$l = V_t \tau \tag{6.28}$$

We substitute this for l in the expression for D under equation 6.25. This gives

$$D = v_t^2 \tau \tag{6.29}$$

Now we relate kinetic energy to thermal energy in each dimension: $\frac{1}{2} m^* v_t^2 = \frac{KT}{2}$ or

$$v_t^2 = \frac{KT}{m^*} \tag{6.30}$$

Now, we substitute into the expression for D above to obtain the **Einstein Relation**:

$$\boxed{D = \frac{q\tau}{m^*} \frac{KT}{q} = \mu \frac{KT}{q}} \tag{6.31}$$

The diffusivity D is also related to the diffusion length L by the relation:

$$L = \sqrt{D\tau} \tag{6.32}$$

Example 6.2:
Find the diffusion current in a piece of silicon if the mobile electron concentration in the sample is found to have the following spatial variation:
$n(x) = n_0 \exp(\frac{-x}{L})$, where $n_0 = 10^{16}/cm^3$ and $L = 10^{-3}cm$, and x is in units of centimeters.

Using equation (6.6) for electron diffusion current density and substituting in for $n(x)$ we have:

$$J_n = qD_n \frac{dn}{dx} = -\frac{qD_n}{L} n_0 \exp\left(\frac{-x}{L}\right)$$

Next, we use the Einstein Relation $\frac{KT}{q}\mu_n$ for D_n. Where $KT/q = 0.026V$ at room temperature and $\mu = 1400cm^2/Vsec$ from the Appendix.
Substituting these values into the electron diffusion current expression gives:

$$J_n(x) = -\frac{(1.6 \times 10^{-19}C)(0.026V)(1400cm^2/Vs)}{10^{-3}cm}(10^{16}/cm^3)\exp\left(\frac{-x}{10^{-3}cm}\right)$$

Evaluating the numbers gives the following expression for electron diffusion current density as a function of position:

$$J_n(x) = -41.6\exp\left(\frac{-x}{10^{-3}cm}\right) A/cm^2$$

Note that the negative value of the current indicates that the current is flowing to the left, while the electrons are diffusing to the right.

6.4 Problems

6.1 A semiconductor has a mobile electron concentration of $1 \times 10^{16}/cm^3$ and the electron mobility is $\mu_n = 800cm^2/Vsec$. If an electric field of $100V/cm$ is applied to the material, calculate the electron drift current density J_{ndrift} in the semiconductor. If the semiconductor has a cross-section of $100\mu m$ by $100\mu m$, what is the current?

6.2 A semiconductor has an internal mobile electron concentration gradient of $1 \times 10^{20}/cm^4$. The electron diffusivity is $D_n = 20cm^2/sec$. Calculate the electron diffusion current density J_{ndif} in the semiconductor. If the semiconductor has a cross-section of $100\mu m$ by $100\mu m$, what is the current?

6.3 Describe what is meant by average drift velocity for electrons in the presence

of an electric field as opposed to the instantaneous velocity $\frac{1}{\hbar}\frac{d\mathcal{E}}{dk}$ from the band structure. What is the average force on electrons that are drifting in an electric field in a uniform semiconductor. Explain your answer using words like scattering and mobility.

6.4 Describe in your own words what drift current is.

6.5 A semiconductor has an electron mobility of $1000cm^2/Vs$ and an effective mass $m^* = 0.5m_0$ (where m_0 is the actual mass of an electron). What is the average time between scattering events in the crystal? What is the average scattering rate? (Be careful with units here.) What is scattering and what causes electron scattering in semiconductors?

6.6 In a silicon bar uniformly doped with 1×10^{15} phosphorus atoms per cm^3 and 1×10^{18} boron atoms per cm^3.

(a) Calculate the mobile electron and hole concentrations for this bar. What do you notice about the mobile hole concentration compared to the acceptor doping? How about the mobile electron concentration compared to the donor doping.

(b) Now, if we apply an electric field in the +x-direction with a magnitude of $10^3 V/cm$ to the silicon bar, what are the electron and hole drift current densities? What is the total current density in the bar? Remember charge neutrality will exist in this uniformly doped bar. Use $\mu_n = 1000cm^2/Vs$ and $\mu_p = 500cm^2/Vs$ for electron and hole mobilities respectively.

(c) If the bar is $100um$ long and has a square cross-section that is $10um \times 10um$, calculate the resistance of the bar in ohms. (Be careful with units here)

6.7 Describe in your own words what diffusion current is.

6.8 The hole concentration in a piece of silicon is given by:

$$p(x) = 10^{14}(1 - \exp(-x/3\mu m))cm^{-3}$$

What is the hole concentration at $x = 1um$? What is the hole diffusion current density at $x = 1um$? Take the diffusion coefficient for holes to be $D_p = 2.5cm^2/s$.

Chapter 7

Non-Uniform Doping and the Built-In Electric Field

7.1 Introduction

In this chapter we find out that in semiconductors when the doping varies with position, internal electric fields and internal electrostatic potentials arise. This occurs because of the interplay of both drift and diffusion currents that exist in non-uniformly doped semiconductors.

7.2 Built in Electric Field: Balance of Drift and Diffusion Currents

Assume we have two different separate semiconductor blocks: block A and block B as shown in the top of Figure 7.1. Any semiconductor will do, but let's assume that the semiconductor is silicon, since that is the one most common and comprises almost all electronics. Now, let's assume block A is doped N-type with doping N_{D1}, and block B is also doped N-type with a different donor concentration say N_{D2}. Also, let's assume that $N_{D1} > N_{D2}$, so block A has higher doping than block B, and thus block A has more mobile electrons than block B. (Recall, the mobile electrons are in the semiconductor conduction band in energy space.) The blocks are not connected and have no net charge. Each block will have no net charge because within each the number of negatively charged mobile electrons will be the same as the number of positively charged ionized donor atoms. In other words, in

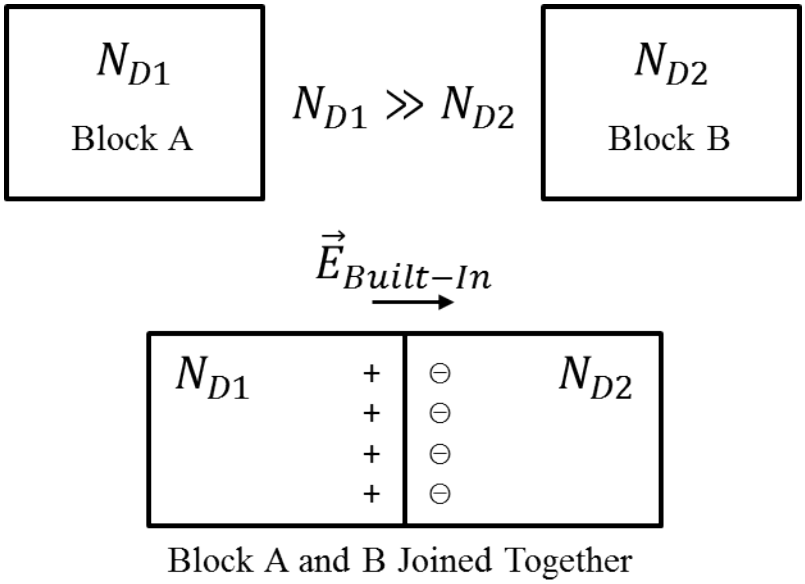


Figure 7.1: Block A and block B be are both doped N-type, the doping in block A is greater concentration than that of block B. At the bottom, the two blocks are brought together in intimate contact. Mobile electrons from block A diffuse to block B. This causes the block A side to become positively charged with respect to the block B side and a built-in electric field arises pointing from block A to block B. The circled negative symbols represent mobile electrons and the plus signs represent fixed ionized donor atoms that a stuck in the semiconductor lattice.

block A the number of mobile electrons will be approximatey equal to N_{D1} and the number of ionized donors N_D^+ will also equal N_{D1}, thus the sum of positive and negative charges will be zero.

Now, let's connect the blocks together to form a junction between block A and block B. What will happen? Well, since block A has higher concentration of mobile electrons than block B, the mobile electrons from block A will diffuse to block B as a result of random motion. Now, block A will wind up being positively charged since it started out as neutral and then lost some of its electrons, and block B will be negatively charged since it has gained some electrons. Now, since block A is now positively charged, and block B is now negatively charged, an electric field will arise that points from block A to block B. This is called the **Built-In Electric Field** and is very important in semiconductor operation. This built-in electric field will then act on the electrons that have diffused over from A to B, and start pulling them back from block B back to block A. This flow from B to A due to the electric field that has arisen is drfit current; and the flow from A to B is diffusion current. Now, if there is no applied voltage to the blocks, then there will be no net current, and the flow due to diffusion will be equal and opposite to the flow due to drift. The charge separation and the built-in field are shown in the bottom of Figure 7.1.

In summary, when we have a region of high N-type doping concentration next to a region of low N-type doping concentration in equilibrium, then we will wind up with a built in electric field (and a built in potential), and the drift current will be equal and opposite to the diffusion current.

7.2.1 Derivation of Built-In Potential

To derive the expression for the relationship between built in potential and the differences in electron concentration, we start by considering a bar of semiconductor material. One side of the bar is doped with N_{D1} and has n_1 mobile electron concentration, and $N_{D1} \approx n_1$. The other side is doped with N_{D2}, and $N_{D2} \approx n_2$. Now, the bar is open circuited and thus has no net current flowing. We know the total electron current, already given by equation (7.1), has both drift and diffusion components as follows:

$$J_{nT} = -q\mu_n n\frac{d\phi}{dx} + qD_n\frac{dn}{dx} \tag{7.1}$$

Since the bar is open circuited no net current will flow so $J_{nT} = 0$, and we have

$$0 = -q\mu_n n\frac{d\phi}{dx} + qD_n\frac{dn}{dx} \tag{7.2}$$

Now recall the Einstein relation that relates diffusivity to mobility that we derived previously:

$$D_n = \frac{KT}{q}\mu_n = V_T\mu_n, \tag{7.3}$$

where V_T is the thermal voltage which is equal to $0.026V$ at $300K$. (not to be confused with the thermal velocity v_t), Substituting for the diffusion coefficient we have:

$$0 = -q\mu_n n\frac{d\phi}{dx} + qV_T\mu_n\frac{dn}{dx} \tag{7.4}$$

Cancel μ_n and q from both sides, rearrange and separate variables:

$$\frac{1}{V_T}\frac{d\phi}{dx} = \frac{1}{n}\frac{dn}{dx} \tag{7.5}$$

Now integrating both sides from x_1 to x_2.

$$\frac{1}{V_T}\int_{\phi(x_1)}^{\phi(x_2)} d\phi = \int_{n(x_1)}^{n(x_2)} \frac{1}{n}dn \tag{7.6}$$

Completing the integration and now applying the limits gives:

$$\frac{1}{V_T}[\phi(x_2) - \phi(x_1)] = \ln[n(x_2)] - \ln[n(x_1)] \tag{7.7}$$

Multiplying both sides by -1, writing the difference of logs as a fraction, defining $n(x_1)$ as n_1, and similarly for the other variables, we have:

$$\frac{1}{V_T}[\phi_1 - \phi_2] = \ln\frac{n_1}{n_2} \tag{7.8}$$

We can also write equation (7.8) in terms of an exponential:

$$\boxed{n_1 = n_2\, e^{\frac{[\phi_1-\phi_2]}{V_T}}} \tag{7.9}$$

Where $[\phi_1 - \phi_2]$ is the built-in potential between points x_1 and x_2.

Equation (7.9) is very important. It says that when we have a difference in electron concentration at two different points in the semiconductor, then we will also have a difference in potential between those points. It also says that there is an exponential relationship between potential differences and carrier concentrations inside the semiconductor.

Now, if we make the following definition for the built in potential of this n_1/n_2 junction:

$$\phi_{BI} = \phi_1 - \phi_2 \tag{7.10}$$

And we make the approximations that $n_1 = N_{D1}$ and $n_2 = N_{D2}$, we can come up with the built in potential as a function of doping concentration. Since the manufacturer of the material will know and provides the doping concentration, we can determine the built in potential as:

$$\boxed{\phi_{BI} = V_T\, \ln\frac{N_{D1}}{N_{D2}}} \tag{7.11}$$

Example 7.1:
If a semiconductor bar is doped with donors on one side with $N_{D1} = 1 \times 10^{18}/cm^3$ and donors on the other side with $N_{D2} = 1 \times 10^{15}/cm^3$, calculate the buil-in voltage.

$$\phi_{BI} = 0.026 \ln\left(\frac{10^{18}}{10^{15}}\right) = 0.18V$$

The actual smoothly-varying potential profile inside of this example N-N junction is shown in Figure 7.2. For $x < 0$ the doping is N_{D1} and for $x > 0$ the doping is N_{D2}. We can confirm from this picture that the built in potential is indeed $0.18V$ by subtracting the potential on the left side from the potential on the right side:

$$0.48V - 0.3V = 0.18V$$

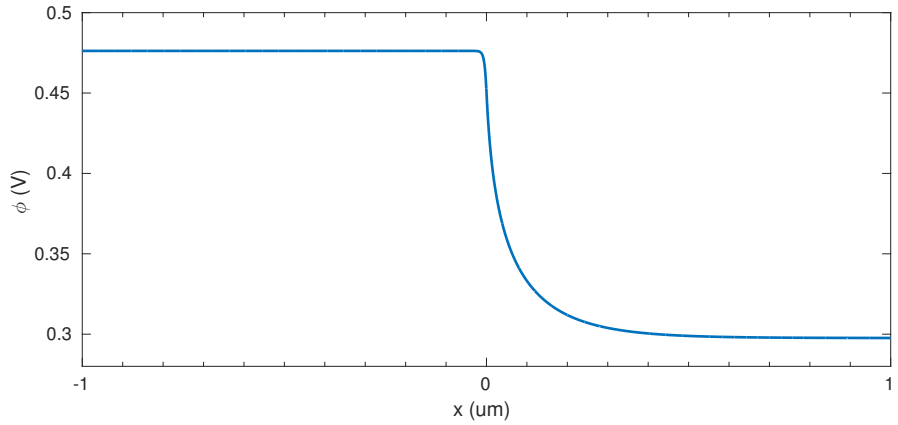


Figure 7.2: Potential profile inside N-N junction formed with $N_{D1} = 1 \times 10^{18}/cm^3$ $(x < 0)$ and $N_{D2} = 1 \times 10^{15}/cm^3$ $(x > 0)$

Note that we have here come up with the built-in voltage for an **n/n** junction. A totally analogous derivation can be applied to find the built-in voltage for a **p/p** junction. Even more important, we will find in the next chapter an analogous expression for the ubiquitous **PN** junction.

7.2.2 The Reference Potential

As we learned early on in our science education, electrostatic potential must be with respect to a reference voltage. Otherwise, we always need to speak in terms of potential differences. In semiconductors, we typically set a reference potential

which is correlated to the intrinsic carrier concentration of a semiconductor. To understand this better let's refer back to equation (7.9). Now, let's replace n_2 with n_i, the intrinsic electron concentration, and equation (7.9) is written as

$$n_1 = n_i \, e^{\frac{[\phi_1 - \phi_i]}{V_T}} \tag{7.12}$$

Now, let's make the definition that at any point in the semiconductor where the mobile electron concentration is intrinsic or n_i at equilibrium, then the potential at that point will be $\phi_i = 0V$. This defines a reference or zero potential at equilibrium. In other words, in equilibrium at any point inside the semiconductor where the mobile electron concentration is n_i, the potential at that point will be $0V$. We can then write an equilibrium expression that relates any carrier concentration at a specific point to the absolute potential at that same point as follows:

$$\boxed{n(x) = n_i \, e^{\frac{\phi(x)}{V_T}}} \tag{7.13}$$

Since at equilibrium the expression $n(x)p(x) = n_i^2$ always holds, then we can express the local hole concentration $p(x)$ to the potential at that point wth a negative in the exponential as follows:

$$\boxed{p(x) = n_i \, e^{\frac{-\phi(x)}{V_T}}} \tag{7.14}$$

By taking the natural log of both sides in Equations 7.13 and 7.14, we can also write the potential in a semiconductor in equilibrium in terms of the ratio of the mobile carrier concentration at a particular position to the intrinsic concentration:

$$\boxed{\phi(x) = V_T \ln \frac{n(x)}{n_i} = -V_T \ln \frac{p(x)}{n_i}} \tag{7.15}$$

Example 7.2:
If we have a silicon bar that is doped at the left end of the bar with $N_D = 10^{17}/cm^3$ donors, and at the right end the bar is doped with $N_A = 10^{16}/cm^3$ acceptors. Then we can make the approximation that the left end of the bar has an electron concentration of $10^{17}/cm^3$, and the right end of the bar has a hole concentration of $10^{16}/cm^3$ Also, we can then call the potential at the left end ϕ_n and the potential on the right end ϕ_p, where:

$$\phi_n = V_T \ln \frac{N_D}{n_i} = 0.026 \ln \left(\frac{10^{17}}{10^{10}} \right) = 0.42V \tag{7.16}$$

and

$$\phi_p = -V_T \ln \frac{N_A}{n_i} = -0.026 \ln \left(\frac{10^{16}}{10^{10}} \right) = -0.36V \tag{7.17}$$

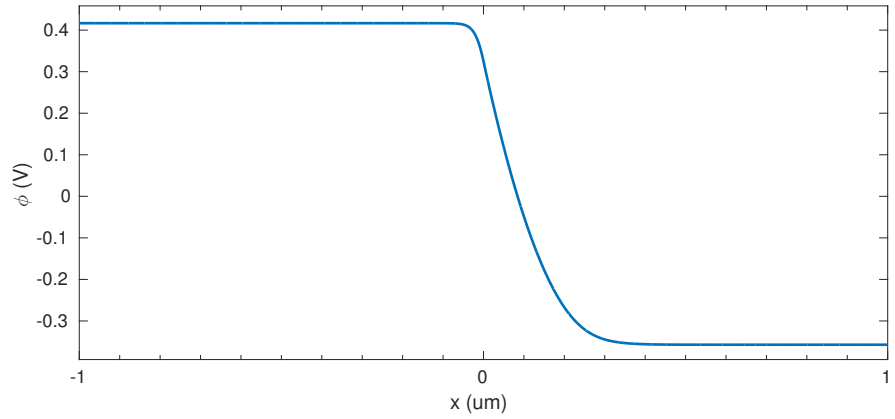


Figure 7.3: Potential profile inside the PN junction formed with $N_D = 10^{17}/cm^3$ $(x < 0)$ and $N_A = 10^{16}/cm^3$ $(x > 0)$

Figure 7.3 shows the potential profile inside this example PN junction. The n-doped side far from the junction has constant potential equal to ϕ_n and the p-doped side equal to ϕ_p. Near the junction in a region called the **depletion region**, the potential makes a smooth transition between these values. Also, please note that when we have equilibrium, regions where the material is doped n-type the potential is positive, and the potential is negative where the bar is doped p-type. (Recall, in equilibrium we do not have any outside energy sources applied to the bar, like no applied voltages.)

Relationship between Potential and Fermi Level

It turns out that the difference between the actual Fermi level in a semiconductor and the intrinsic Fermi level also describes the electrostatic potential for mobile carriers in the material. If you go on and study more semiconductor physics, you will most likely run into analyses that involve the Fermi level, because it becomes a very useful tool for the analysis of devices that contain more than one material (not just silicon for example). We therefore introduce here some key concepts that relate to the Fermi level that will help if future encounters with device physics wind up involving the Fermi level.

In Chapter 5 we introduced the concept of the Fermi Level as a way describing the concentration of electrons and holes in the valence and conduction bands in equilibrium. Recall that the equilibrium electron concentration written in terms of the Fermi as the following.

$$n = N_C \exp[-(\mathcal{E}_C - \mathcal{E}_F)/KT] \tag{7.18}$$

And the intrinsic carrier concentration is given by

$$n_i = N_C \exp[-(\mathcal{E}_C - \mathcal{E}_{Fi})/KT] \tag{7.19}$$

Taking the natural log of the ratio between n and n_i and then multplying by $\frac{KT}{q}$ gives the following relationship:

$$\frac{KT}{q}\ln\frac{n(x)}{n_i} = V_T \ln\frac{n(x)}{n_i} = \frac{\mathcal{E}_F - \mathcal{E}_{Fi}(x)}{q} \tag{7.20}$$

Comparing Equations 7.20 and 7.15 we see that in equilibrium, we can express the internal potential in a semiconductor as the difference between the Fermi level and intrinsic Fermi level divided by the magnitude of the electronic charge:

$$\boxed{\phi(x) = \frac{\mathcal{E}_F - \mathcal{E}_{Fi}(x)}{q}} \tag{7.21}$$

This is a fundamental relationship which connects the description of semiconductors in terms of the Fermi level to the description used in this text which is in terms of the electrostatic potential.

7.2.3 Poisson Equation

One of the most basic equations from electrostatics that often arises in semiconductors is the Poisson equation. From your electromagnetics classes, you probably recall that the Poisson equation relates the charge to the potential, and it is often thought of as another way of stating Gausses law. In general the Poisson equation is as follows:

$$\nabla^2\phi = -\frac{\rho}{\epsilon} \tag{7.22}$$

Where ρ is the charge concentration.

As we have seen before, in semiconductors the charge concentration at any point is composed of the contribution from the mobile electron n, the mobile hole p, the ionized acceptor N_A^-, and the ionized donor N_D^+ concentrations at that point, respectively. This gives the following general form for the Poisson equation in semiconductors:

$$\nabla^2\phi = -\frac{q}{\epsilon}(p - n + N_D^+ - N_A^-) \tag{7.23}$$

Equilibrium Poisson Equation: When the semiconductor is in equilibrium condition (no applied voltage, etc.), the electron and hole concentrations are written using equations (7.13) and (7.14), respectively. This gives the following form for the Poisson equation in for semiconductors in equilibrium:

$$\nabla^2\phi = -\frac{q}{\epsilon}(n_i\, e^{\frac{-\phi}{V_T}} - n_i\, e^{\frac{\phi}{V_T}} + N_D^+ - N_A^-) \tag{7.24}$$

In one dimension equation (7.24) is expressed as:

$$\frac{d^2\phi(x)}{dx^2} = -\frac{q}{\epsilon}[n_i\, e^{\frac{-\phi(x)}{V_T}} - n_i\, e^{\frac{\phi(x)}{V_T}} + N_D^+(x) - N_A^-(x)] \tag{7.25}$$

Note that in equation (7.25) we have explicitly included the x dependence to remind us that all the quantities in the Poisson equation are typically functions of position.

Example 7.3:
Using the current equation and the definition of carrier concentration with respect to the reference potential, show that both the electron current density and the hole current density are both identically zero at any point in a semiconductor in equilibrium.
Start with the Einstein relation $\mu_n V_T = D_n$, the current equation for electrons (7.1) and the expression for mobile electron concentration at equilibrium:

$$J_{nT}(x) = -q\mu_n n(x)\frac{d\phi(x)}{dx} + qD_n\frac{dn(x)}{dx}$$

$$n(x) = n_i\, e^{\frac{\phi(x)}{V_T}}$$

Evaluating the derivative from the diffusion current term:

$$\frac{dn}{dx} = \frac{d}{dx}(n_i e^{\frac{\phi(x)}{V_T}}) = \frac{n_i\, e^{\frac{\phi(x)}{V_T}}}{V_T}\frac{d\phi(x)}{dx} = \frac{n(x)}{V_T}\frac{d\phi(x)}{dx} \tag{7.26}$$

Substituting this for the derivative of concentration in the current eqution and using the Einstein relation gives:

$$J_{nT}(x) = -q\mu_n n(x)\frac{d\phi(x)}{dx} + q\mu_n n(x)\frac{d\phi(x)}{dx} = 0 \tag{7.27}$$

This is an important example because it verifies mathematically that at equilibrium, the drift and the diffusion currents for mobile electrons are equal in magnitude and opposite in direction and then cancel each other to yield zero net current. A similar treatment can be provided for holes which is left for an excersize in the problems at then end of this chapter.

7.3 Problems

7.1 In words, qualitatively describe how the built in potential would arise in a non-uniformly doped P-type Silicon bar. Assume that the right end of the bar has higher acceptor (N_A) doping than the left side of the bar.

7.2 Derive the expression for the hole diffusion current density across a plane by considering the flow of holes across it.

7.3 Using the current equation for holes, derive the expression relating built-in voltage and hole concentration at equilibrium ($J_p = 0$) that results from the balance of drift and diffusion currents in a non-uniformly doped silicon bar.

7.4 One half of a silicon bar is doped with $10^{19} cm^{-3}$ of donor atoms, and the other half is doped with $10^{15} cm^{-3}$ donors, what is the built-in potential inside this bar. Assume all the donors are ionized.

7.5 If the electron concentration as a function of position (z) in a $100um$ long silicon bar at equilibrium is found to be:

$$n(z) = n_b \left[1 + \frac{z}{L}\right]$$

(a) Graph the potential as a function of position in the bar from $z = 0$ to $100um$. Let $n_b = 1 \times 10^{15} cm^{-3}$, and $L = 1\mu m$

(b) Graph the base-10 log of the hole concentration as a function of position along the bar.

(Note equilibrium means the total current in the bar is zero.)

7.6 A semiconductor device relies on various specific regions having a excess of electrons (heavily n-doped) and other regions having an excess of holes (heavily p-doped) in order to function properly.

(a) Explain what would happen to the functionality of this device at high temperatures (below melting temp.)? Why? Include things like 'more/less n/p-type' and 'intrinsic' in your response.

(b) What material property would you want to change to improve this issue?

7.7 The electrostatic potential in a semiconductor is found to be:

$$\phi(x) = \frac{qN_D}{2\epsilon}x^2 \quad Volts$$

Where N_D is the ionized donor concentration. Note that the units for electrostatic potential are Volts.

(a) Obtain an expression for the electric field in terms of the parameters given in the above expression for potential. Include units in your answer.

(b) Obtain an expression for the charge density. Include units in your answer.

7.8 Using the current equation, the Einstein relation, and the definition of carrier concentration with respect to the reference potential, show that both the electron current density and the hole current density are both identically zero at any point in a nonuniformly doped semiconductor in equilibrium.

Chapter 8

The PN Junction Part I

8.1 Introduction

A PN junction is probably the most fundamental structure in electronics. A PN junction is the intimate contact of an N-type and a P-type semiconductor. This structure forms a diode, which we know allows electric current to flow in one direction but not the other. The PN junction structure forms a diode, but it is also a critical component of many other devices including the BJT and the ubiquitous MOSFET. A diode is composed mainly of a PN junction. The governing equation for the PN junction diode is:

$$I_D = I_o \left[e^{V_A/V_T} - 1 \right] \tag{8.1}$$

Where I_o is called the saturation current and it is determined by the doping and material parameters. V_A is the applied voltage. It is a positive number when the plus side is of V_A is applied to the P-side of the junction. V_T is the thermal voltage KT/q.

In this chapter and the next, we will learn how a PN junction works and how it gives rise to the governing equation (8.1). The block diagram of a PN junction and the circuit symbol are shown in Figure 8.1. In the figure, the N-side is on the left and the P-side is on the right. (In this chapter, we will use this convention of having the N-side on the left and P-side on the right. I find that it makes the PN junction analysis somewhat easier.)

8.2 Built-In Potential of a PN Junction

Since a PN junction is a nonuniformly doped semiconductor, it will have a built in potential ϕ_o. Like we found in the previous chapter, the built in potential

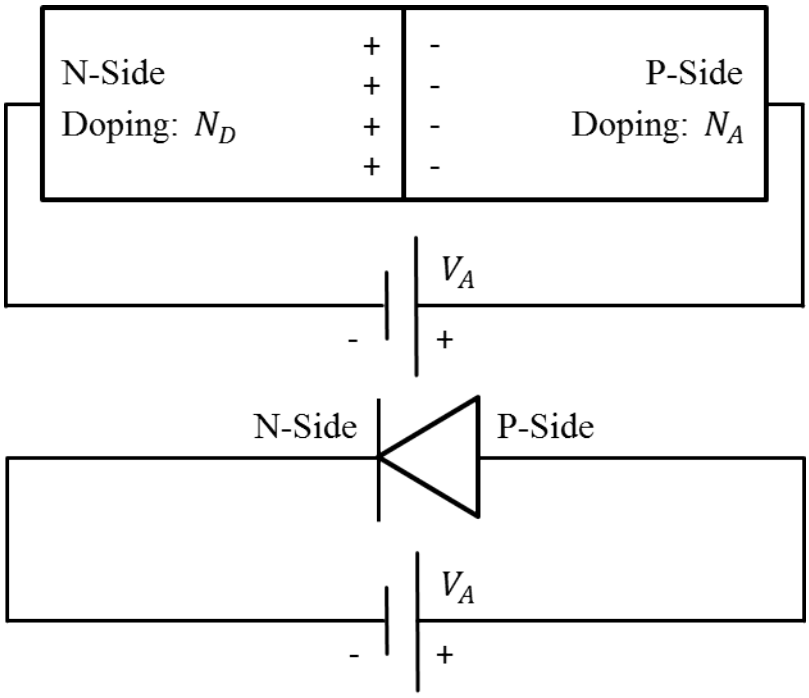


Figure 8.1: PN Junction Block Diagram. The N-side is on the left doped with donors and the P-side is on the right doped with acceptors.The lower figure shows the circuit symbol for the PN junction with N-side and P-side indicated. The applied voltage V_A is positive as indicated because the positive terminal is connected to the P-Side.

is given by:

$$\phi_o = V_T \ \ln\left(\frac{n_n}{n_p}\right) \tag{8.2}$$

Where n_n is the mobile electron concentration on the N-side and n_p is the concentration of mobile electrons on the P-side. Now it is an excellent approximation to say that the mobile electron concentration on the N-side is equal to the donor concentration ($n_n = N_D$), and the mobile hole concentration on the P-side is equal to the acceptor concentration ($p_p = N_A$). Now if we invoke the equilibrium condition that $np = n_i^2$, then n_p the electron concentration on the P-side is $n_p = n_i^2/N_A$. Substituting into equation (8.2) we obtain the following well known equation for the built in potential of a PN junction:

$$\boxed{\phi_o = V_T \ \ln\left(\frac{N_D N_A}{n_i^2}\right)} \tag{8.3}$$

This same result is obtained if we use the ratio of hole concentrations instead of taking the ratio of electrons in Equation 8.2. However, the ratio is flipped due to the opposite charge of the holes i.e. $\phi_o = V_T \ln\left(\frac{p_p}{p_n}\right)$.

Built-in Potential versus Temperature

It is important to note that the ϕ_o depends on absolute temperature in two basic ways. First of all $V_T = K_b T/q$ which is obviously directly proportional to temperature. In addition, the intrinsic carrier concentration n_i increases with increasing temperature as was described in Chapter 5. The increase in the intrinsic concentration is the dominant effect here, so as temperature increases, the built in potential will typically decrease in semiconductors.

8.3 PN Junction Operation: Qualitative

The operation of a PN junction is similar to that of the N1/N2 junction we studied in the previous chapter. The PN junction and the N1/N2 junction both have a built-in potential. However, the fact that we have an N-type material in contact with a P-type material gives rise rectification. More specifically, a PN junction will rectify electrical current, but an N1/N2 junction will not. The rectification properties of the PN junction can easily be seen from the plot of the diode's current-voltage relationship (Equation 8.1) shown in Figure 8.2. When the forward voltage bias is applied (positive), substantial current is able to flow through the diode, however in reverse bias, the current is greatly suppressed and only a small leakage current I_o can make it through. The leakage current is generally many orders of magnitude less than the forward current so we say the diode acts as a rectifier.

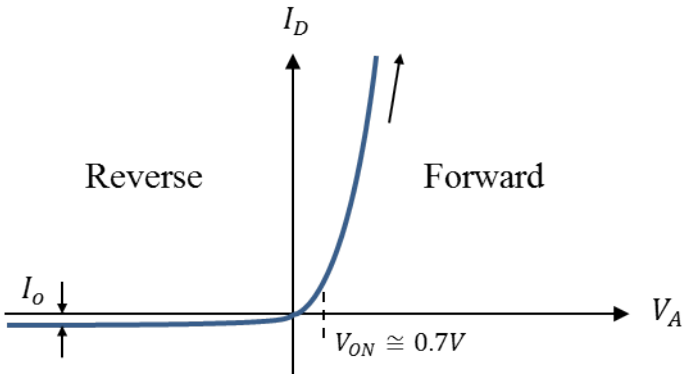


Figure 8.2: Plot of PN Junction current equation. The 'turn-on' or 'knee' voltage V_{ON} for a typical Si diode is around $0.7V$. Leakage current is typically in the micro-amp regime, whereas the forward current typically ranges from milli-amps to amps.

8.3.1 Equilibrium, No Net Current

The top illustration in Figure 8.3 shows the PN junction at equilibrium. On the N-side we have a very large concentration of mobile electrons. On the P-side the concentration of mobile electrons is many orders of magnitude smaller. Thus, electrons from the N-side will diffuse to the P-side in response to the concentration gradient. Since the N-side loses electrons it will become positively charged, and since the P-side gains electrons it will take on a negative charge. As electrons continue to diffuse from N-side to P-side, an electric field will arise due to the separation of charge. This built-in electric field will point from the N-side to the P-side since the N-side is now positive with respect to the P-side. This built-in field will then act to pull mobile electrons back to the N-side. Since there is no applied bias to the PN junction there will be no net electron current, so the electron diffusion from N to P will be equal and opposite to the drift flow from P to N, which has resulted from the built-in field. A similar set of events will occur for mobile holes. In other words, holes will diffuse from P to N and the built-in electric field will cause holes to drift back from N to P giving rise to zero net hole current. Thus, at equilibrium in a PN junction electron drift current will be equal and opposite to electron diffusion current so that the total electron current will be zero. Likewise, hole drift and diffusion currents are equal and opposite as well, so the total hole current will also be zero. So for electrons we have:

$$J_{nT} = 0 = q\mu_n n E_o + qD_n \frac{dn}{dx} \tag{8.4}$$

And for hole current we have:

$$J_{pT} = 0 = q\mu_p p E_o - qD_p \frac{dp}{dx} \tag{8.5}$$

Recall that the first terms on the right hand side of equations (8.4) and (8.5)

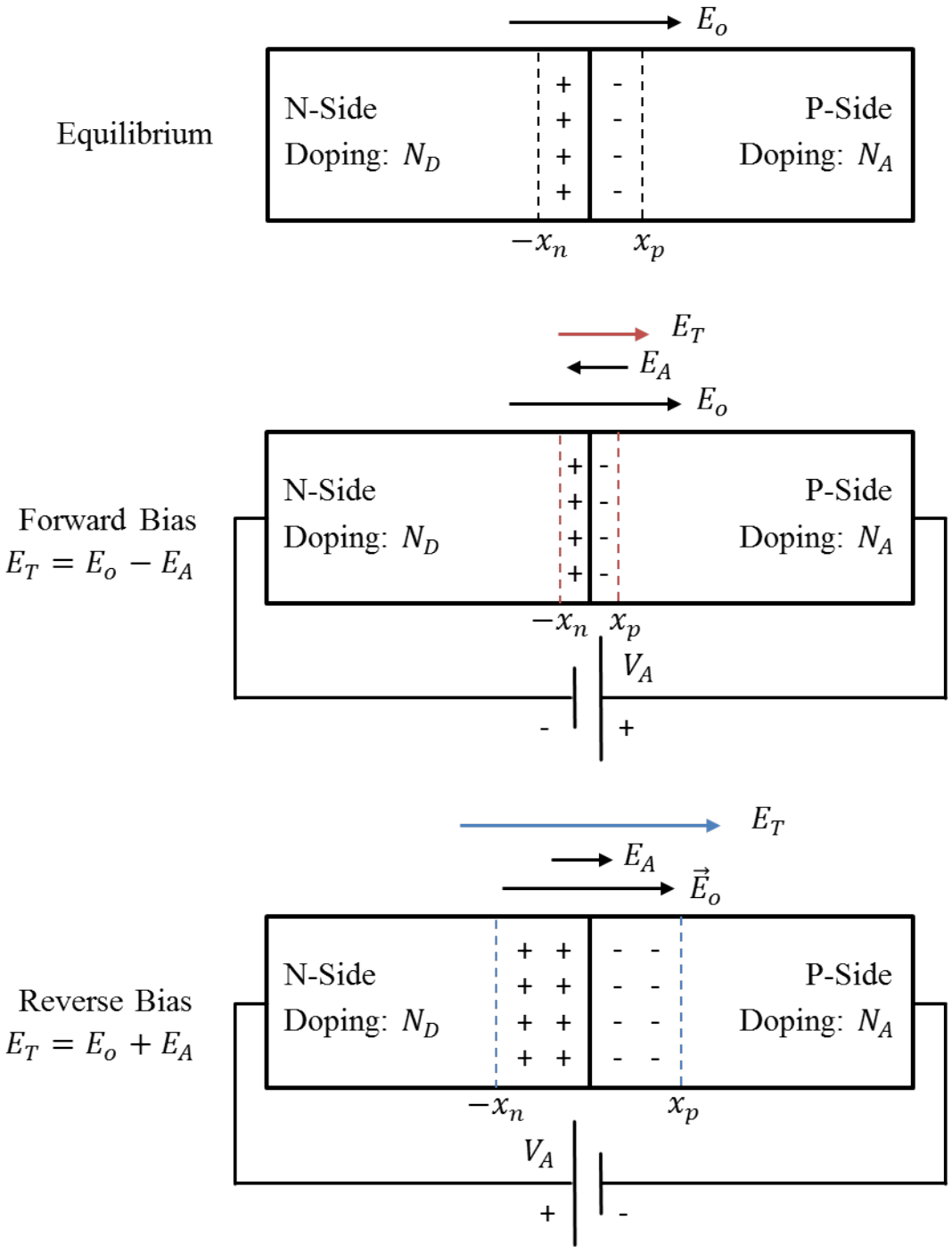


Figure 8.3: PN Junction: Top Equilibrium; Middle Forward Bias; Bottom Reverse Bias. The N-Side is on left, the P-Side is on the right of each figure. The depletion region is between $-x_n$ and x_p. Outside the depletion region are the quasi neutral N-side and P-side bulk regions.

are drift currents and the second terms are the diffusion currents, and E_o is the built-in electric field, which points from N to P.

Depletion Region: Since most of the mobile charge has diffused away, the region near the junction will be largely void of mobile electrons and holes. Thus, this region will consist mainly of positively charged ionized donors on the N-side and negatively charged ionized acceptors on the P-side. Since this region is mainly void of mobile charges, it is called the **Depletion Region**. Furthermore, the internal electric field and the internal electrostatic potential will be dropped across this depletion region, for equilibrium, forward bias and reverse bias as well. The depletion region is shown as the charged area between $-x_n$ and x_p in Figure 8.5.

8.3.2 Forward Bias

In forward bias, we apply a positive external voltage to the P-side with respect to N-side. This is shown in the middle illustration in Figure 8.3. The external voltage gives rise to a field component E_A inside the device that points from P to N. Now the total field E_T in the device is now:

$$E_T = E_o - E_A, \qquad E_T < E_o \tag{8.6}$$

So the applied field has the effect of reducing the total field because it points in the opposite direction as E_o. Now, it is important to keep in mind that the total field E_T still points in the same direction as E_o, but tends to be less than E_o.

Now, since the electric field is now smaller across the depletion region, the drift current components are decreased. Thus, electron drift current is now less than the electron diffusion current and a net electron current will flow. This net electron current is largely due to diffusion. Similarly, the applied field also reduces the hole drift current and a net hole current will now flow as well, which is largely diffusion current in nature.

So that is how forward bias works. In summary, we apply an external potential that reduces the internal field, thereby reducing the drift current. Furthermore, the diffusion current has not changed too much from equilibrium. Thus diffusion is now greater than drift and a net current will flow, which is given by:

$$I_D = I_o \, e^{V_A/V_T} \tag{8.7}$$

8.3.3 Reverse Bias

In reverse bias, we apply a positive voltage to the N-side with respect to the P-side. This is shown on the bottom of Figure 8.3. The added bias is in the same direction as the built-in field, and thus adds to it. So the total field across the junction is now greater than the built in field or:

$$E_T = E_o + E_A, \qquad E_T > E_o \tag{8.8}$$

At first this seems like it would significantly increase the drift component, but it does not. In fact hardly any reverse current flows at all. The reason is that there are relatively very few electrons on the P-side, and therefore, there are very few electrons available for drift current from P to N. For significant drift current to flow, you need both field and carriers, but the carriers that would respond to the enhanced field are minority carriers, and thus very low in concentration.

More specifically, the total electric field in Reverse Bias is mainly contained within the depletion region and it points from N to P. Furthermore, the net electron current in reverse bias is due largely to the small number of electrons in the P-side near the junction that can be pulled over by the field. Similarly, the net hole current consists of the small number of minority holes in the N-side that are pushed over to the P-side by the electric field. Thus this enhanced field only increases the drift current from its equilibrium level by a very small amount. Furthermore, the diffusion current does not really change too much from the equilibrium case. Thus, in reverse bias, the drift current becomes only slightly larger than its equilibrium value, and drift and diffusion still balance except for a very slightly increased drift component. Therefore, a very very small current flows into the N-side and out of the P-side in reverse bias. This reverse bias current is many orders of magnitude smaller than the forward bias current at typical values of V_A, and is approximately given by:

$$I_D = -I_o \tag{8.9}$$

(Note that I_D in reverse bias will typically be a little larger than $-I_o$, but the reason for this will be saved for later.)

8.3.4 Requirements for Rectification

A rectifier is an electronic device that only allows current to flow in one direction. As discussed, a PN junction allows current to flow in one direction but, except for a very small leakage current, not the other. The PN junction is the most ubiquitous rectifier in electronics. We found in Chapter 7 that non-uniform doping using the same gives rise to a built-in potential. In the current chapter we find that non-uniform doping that consists of two different dopant species (donors and acceptors) gives rise to not only a built-in potential but also a depletion region. For a device to be able to rectify two physical attributes are required;

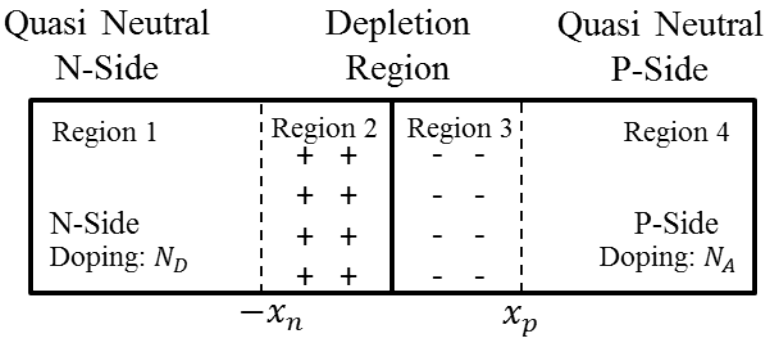


Figure 8.4: Depletion region and Quasi Neutal regions of PN Junction

Requirements for Rectification

1. **A Built in Potential**
2. **A Depletion Region**

This important result is at the heart of modern semiconductor electronics.

8.4 Electric Field, Potential and the Depletion Approximation

To analyze the electric field, the electrostatic potential and the charge distribution in a PN junction we use the **Depletion Approximation**. The depletion approximation says that there is a small region on either side of the PN junction that has very few (or is depleted of) mobile electrons or holes compared to the concentration of ionized donors and acceptors, and this region is therefore electrically charged. Furthermore, all areas outside of the depletion region are approximated as charge neutral. As a result all the electric field in the PN junction is contained within this depletion region. Furthermore, the electrostatic potential is totally dropped across the depletion region. Outside the depletion region, the electric field is approximated to be zero, and the electrostatic potential is approximated to be constant. Overall, the depletion approximation is very accurate for most applications, especially when the PN junction is in equilibrium.

In this text, when we use the depletion approximation, it is convenient to divide the PN junction into four regions, as shown in Figure 8.4.

- **Region 1: Quasi Neutral N-Type**: This is the N-side of the PN junction that is doped with donors. Here the concentration of mobile electrons is provided by the donor atoms which become ionized (Donor atoms in Si

are typically Phosphorous, a column 5 element), The concentration of mobile electrons is essentially equal to the ionized donor concentration so the Quasi Neutral N-side of the PN junction is charge neutral. This region extends from the metal contact to the N-side of the depletion region. In this quasi neutral N-side, electrons are the majority carriers, and holes are the minority carriers. This large region is also called the N-type bulk region in the PN junction.

- **Region 2: N-Side of the Depletion Region**: This narrow region extends from the edge of the N-side of the depletion region to the edge of junction with the P-Side. Since this narrow region is largely void of mobile electrons and holes, it has a charge concentration that is due to the ionized donors, so it has positive charge density of $\rho = +qN_D$, where N_D is the donor concentration.

- **Region 3: P-Side of the Depletion Region**: This is analogous to Region 2, but it is on the P-side of the depletion region. This narrow region extends from the edge of the junction with the N-Side to the edge of the P-side quasi-neutral region. Since this narrow region is largely void of mobile electrons and holes, it has a charge concentration that is due to the ionized acceptors, so it has negative charge density of $\rho = -qN_A$, where N_A is the acceptor concentration.

- **Region 4: Quasi Neutral P-Type**: This is the P-side of the PN junction that is doped with acceptors. Here the concentration of mobile holes is provided by the acceptor atoms which become ionized (Acceptor atoms in Si are typically Boron, a column 3 element), The concentration of mobile holes is essentially equal to the ionized acceptor concentration so the Quasi Neutral P-side of the PN junction is charge neutral. This region extends from the P-side of the depletion region to the P-side contact. In this quasi neutral P-side, holes are the majority carriers, and electrons are the minority carriers. This large region is also called the P-type bulk region in the PN junction.

8.4.1 Doping and Charge Summary in PN Junction Regions

The doping and charge summaries are illustrated in Figure 8.4. Below is a further description.

Region 1: Quasi Neutral N-Side ($x \leq -x_n$):
Doping = N_D = Donor Concentration
N_A=0.
$n \approx N_D$
$p \approx \frac{n_i^2}{n}$ (extremely small and taken as negligible)
$\rho = q(p - n + N_D) \approx 0$

Region 2: Depletion Region N-Side ($-x_n \leq x \leq 0$):
Doping = N_D = Donor Concentration
$N_A = 0$
$n << N_D$
$p << N_D$
n, p are negligible compared with N_D
$\rho = +q(p - n + N_D^+) \approx +qN_D$
Region 2 is largely depleted of mobile carriers. Hence, it is called N-side of the Depletion Region

Region 3: Depletion Region P-Side ($0 \leq x \leq x_p$):
Doping = N_A = Acceptor Concentration
$N_D = 0$
$n << N_A$
$p << N_A$
n, p are negligible compared with N_A
$\rho = +q(p - n - N_A^-) \approx -qN_A^-$
R3 is largely depleted of mobile carriers. Hence, it is called P-side of the Depletion Region.

Region 4: Quasi Neutral P-Side ($x \geq x_p$):
Doping = N_A = Acceptor Concentration
$N_D = 0$
$p \approx N_A$
$n \approx \frac{n_i^2}{p}$ (very small and taken as negligible)
So, $\rho = +q(p - n - N_A^-) \approx 0$

8.4.2 PN Junction Electric Field and Potential Distribution Using the Depletion Approximation

We will now derive the expressions for the built-in electric field and the built-in electrostatic potential as a function of position in the PN junction. We will do this by solving the Poisson equation for regions 1 through 4 within the framework of the very accurate depletion approximation. First of all, let's rewrite the Poisson equation for semiconductors in one dimension:

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon}(p - n + N_D^+ - N_A^-) \tag{8.10}$$

And also recall that the electric field is the negative gradient of the potential which in one dimension reduces to the simple derivative:

$$E = -\frac{d\phi}{dx} \tag{8.11}$$

Region 1: Quasi-Neutral N-Side

Let's remind ourselves the depletion approximation says that all the potential drops across the depletion region, and the electric field is totally contained within the depletion region. Keeping this in mind let's solve the Poisson equation for Region 1, which is charge neutral:

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon}(p - n + N_D^+) = 0 \tag{8.12}$$

The solution to 8.12 says that the potential in the bulk N-side is either linear or constant. Since the depletion approximation says that the potential is totally dropped across the depletion region (Regions 2 and 3), the potential in this region is constant. Furthermore, using equation (7.13), we find that the constant potential is:

$$\phi_n = V_T \ln \frac{N_D}{n_i} \tag{8.13}$$

Furthermore, since the potential is constant, then the electric field $E = 0$ in this bulk N-Side. The constant potential and the zero field in the bulk N and P regions are illustrated in Figure 8.5.

Region 4: Quasi-Neutral P-Side

The analysis for the bulk P-Side is analogous to that of the quasi-neutral bulk N-side. In the bulk P-side we have charge neutrality so the Poisson equation becomes

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon}(p - n - N_A^-) = 0 \tag{8.14}$$

The solution to 8.14 says that the potential in the bulk P-side is either linear or constant. Since the depletion approximation says that the potential is totally dropped across the depletion region (Regions 2 and 3), the potential in this region is constant. Furthermore, using equation (7.14), we find that the constant potential is:

$$\phi_p = -V_T \ln \frac{N_A}{n_i} \tag{8.15}$$

Furthermore, since the potential is constant, then the electric field $E = 0$ in this bulk P-Side. It is also important to notice that using n_i as the point of reference potential, then the potential in the P-Side turns out to be a negative number. See Figure 8.5.

Region 2, N-Side of Depletion Region

The N-side of the depletion region is doped with donors, and has negligible mobile electrons and holes. Thus the Poisson equation for this region is:

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon}N_D^+ \tag{8.16}$$

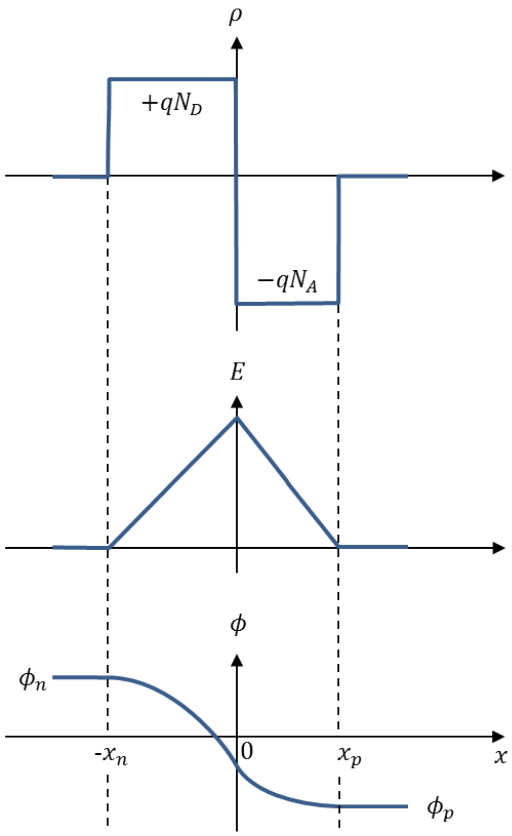


Figure 8.5: PN Junction in Equilibrium: Top is the Charge Density; Middle is the Built-In Electric Field; Bottom is the Built-In Electrostatic Potential.The N-Side is on left, the P-Side is on the right of each figure. The depletion region is between $-x_n$ and x_p. Outside the depletion region are the quasi neutral N-side and P-side bulk regions.

Integrating and remembering that the field is the negative of the derivative of the potential we obtain

$$-\frac{d\phi(x)}{dx} = E(x) = \frac{q}{\epsilon}N_D x - c \tag{8.17}$$

Now apply boundary condition that the field must go to zero at $-x_n$ and we can solve for the constant of integration.

$E(-x_n) = 0 = \frac{-q}{\epsilon}N_D x_n - c$. Solving for c gives $c = \frac{-q}{\epsilon}N_D x_n$. Substituting this expression for c back into equation (8.17) gives the following expression for the electric field as a function of position in the N-side of the depletion region:

$$\boxed{E(x) = \frac{q}{\epsilon}N_D(x + x_n) \qquad (-x_n \leq x \leq 0)} \tag{8.18}$$

Now find the potential $\phi(x)$. We start with equation (8.18), writing the field as the negative derivative of the potential we have:

$$\frac{d\phi}{dx} = \frac{-q}{\epsilon}N_D(x + x_n) \tag{8.19}$$

Now integrating both sides from the edge of the depletion region $-x_n$ to some point x in the depletion region:

$$\int_{\phi(-x_n)}^{\phi(x)} d\phi = \frac{-q}{\epsilon} N_D \int_{-x_n}^{x} (x + x_n) dx \tag{8.20}$$

which gives:

$$\phi(x) - \phi(-x_n) = \frac{-q}{2\epsilon} N_D (x + x_n)^2 \tag{8.21}$$

Rearranging gives the expression for the potential as a function of position in the N-side of the depletion region.

$$\boxed{\phi(x) = \phi_n - \frac{q}{2\epsilon} N_D (x + x_n)^2} \tag{8.22}$$

where $\phi_n = \phi(-x_n) = V_T \ln \frac{N_D}{n_i}$. See Figure 8.5.

Region 3, P-Side of the Depletion Region

The analysis of the field and potential on the P-Side of the depletion region is analogous to that of the N-side depletion region. As in the N-side, the concentration of mobile electrons and holes is negligible compared with the ionized dopant concentration. But here, the dopants are acceptors so the charge concentration is negative and is given as $-qN_A$. Starting with this charge concentration we find the electric field and the electrostatic potential by integration of the Poisson equation. This gives the following:

$$\boxed{E(x) = \frac{q}{\epsilon} N_A (x_p - x) \qquad (0 \leq x \leq x_p)} \tag{8.23}$$

$$\boxed{\phi(x) = \frac{q}{2\epsilon} N_A (x_p - x)^2 + \phi_p \qquad (0 \leq x \leq x_p)} \tag{8.24}$$

where, $\phi_p = \phi(x_p) = -V_T \ln \frac{N_A}{n_i}$. Figure 8.5 illustrates the built-in electric field and the built-in electrostatic potential.

Depletion Region Length

Once we have the field and the built in potential, it is straightforward to obtain the lengths of the depletion region on each side:

$$x_n = \sqrt{\frac{2\epsilon(\phi_0 - V_A)}{q} \left[\frac{N_A}{N_D(N_A + N_D)}\right]} \tag{8.25}$$

$$x_p = \sqrt{\frac{2\epsilon(\phi_0 - V_A)}{q} \left[\frac{N_D}{N_A(N_A + N_D)}\right]} \tag{8.26}$$

Where V_A is the applied voltage across the PN junction, as shown in Figure 8.1. V_A is positive for forward bias and negative for reverse bias. Also, for this expression, there are no restrictions on the values of V_A for reverse bias, but for forward bias V_A must be less than the built-in potential or ($V_A < \phi_0$ for FB).

The entire depletion region width can easily be calculated from $W_D = x_n + x_p$:

$$W_D = \sqrt{\frac{2\epsilon(\phi_0 - V_A)}{q}\left[\frac{N_A + N_D}{N_A N_D}\right]} \tag{8.27}$$

Additionally, the depletion lengths x_n and x_p can be determined from the width:

$$x_n = \frac{N_A}{N_A + N_D} W_D \tag{8.28}$$

$$x_p = \frac{N_D}{N_A + N_D} W_D \tag{8.29}$$

Example 8.1:
A silicon PN junction is doped on one side with the acceptor boron at a concentration of $1 \times 10^{17}/cm^3$ and the other side is doped with the donor phosphorous at a concentration of $1 \times 10^{16}/cm^3$. Assume all dopant atoms are ionized. Calculate the built in potential ϕ_0, the length of the depletion region region W_D and the peak electric field at room temperature. Also calculate the value of the electrostatic potential at the junction.

$$\phi_0 = 0.026 \ln\left[\frac{(1 \times 10^{16})(1 \times 10^{17})}{(1 \times 10^{10})^2}\right] = 0.77V$$

$$W_D = \sqrt{\frac{2 \times 11.7 \times 8.85 \times 10^{-14} \times 0.77}{1.6 \times 10^{-19}}\left[\frac{(1 \times 10^{16}) + (1 \times 10^{17})}{(1 \times 10^{16})(1 \times 10^{17})}\right]}$$

$$W_D = 3.4 \times 10^{-5} cm = 0.33 \mu m$$

To calculate the peak electric field we note that the field is continuous and is maximum at the junction:

$$E_{max} = E(0) = \frac{q}{\epsilon} N_D x_n = \frac{q}{\epsilon} N_A x_p$$

Using Equations 8.25 and 8.26 for x_n and x_p and multiplying by the corresponding constants gives:

$$E_{max} = E(0) = \sqrt{\frac{2q\phi_0}{\epsilon}\frac{N_A N_D}{N_A + N_D}} = \sqrt{\frac{2 \times 1.6 \times 10^{-19} \times 0.77}{11.7 \times 8.85 \times 10^{-14}}\frac{(10^{17})(10^{16})}{(10^{17}) + (10^{16})}}$$

$$E_{max} = E(0) = 4.65 \times 10^4 V/cm$$

To calculate the value of the electrostatic potential at the junction we can use either one of equations (8.22) or (8.24) and evaluate at $\phi(0)$.

$$\phi(0) = \frac{q}{2\epsilon} N_A x_p^2 + \phi_p = \phi_n - \frac{q}{2\epsilon} N_D x_n^2 = -0.35V$$

8.4.3 PN Junction Depletion Approximation Comparison to Numerical Solution and Effects of Bias

In the previous sections we have used the depletion approximation to calculate reasonably simple closed form solutions of the charge density, field, and potential everywhere inside the PN junction. The key assumption in this approximation was that an abrupt depletion region surrounding the junction forms, and is completely devoid of electrons and holes. In the real world, however, the concentration of

electrons and holes does not suddenly become zero inside the depletion region and instead there is a transition period at the depletion region edges where the concentration changes continuously from its equilibrium value in the quasi-neutral region to a small value inside the depletion region. The validity of the depletion approximation relies on the fact that this change is quite rapid due to the exponential dependence of the electron and hole concentrations on the potential (Equations 7.13 and 7.14). Since this fact is true, the charge density inside the depletion region is well approximated by a square-wave function. A comparison of the depletion approximation solution to a full numerical solution of the Poisson equation in a PN junction is provided in Figure 8.6. The full numerical solution shown is a self-consistent solution to the Poisson equation in one dimension (Equation 7.25) for all values of x performed without breaking the space up into different regions.

This solution is valid for the PN junction with no applied bias. The derivation of the solution to the Poisson equation for a PN junction under applied bias can be carried out in the same manner as we have provided for the case of no applied bias. The only change is that that we must add the applied potential to the boundary conditions i.e. the potential value in the quasi-neutral regions. When applying forward bias of V_A, the positive potential is applied to the p region of the junction and we will take the n region to be our ground. Following this, our constant potentials inside the quasi-neutral regions will become $\phi_p + V_A$ and $phi_n + 0$ respectively. The applied potential also affects the location of the depletion region edges, as given in Equations 8.25 and 8.26. Forward bias shrinks the depletion region and reverse bias increases the length of the depletion region. The consequence of these changed bias conditions are shown in Figure 8.7, where the equilibrium, reverse, and forward bias conditions are shown.

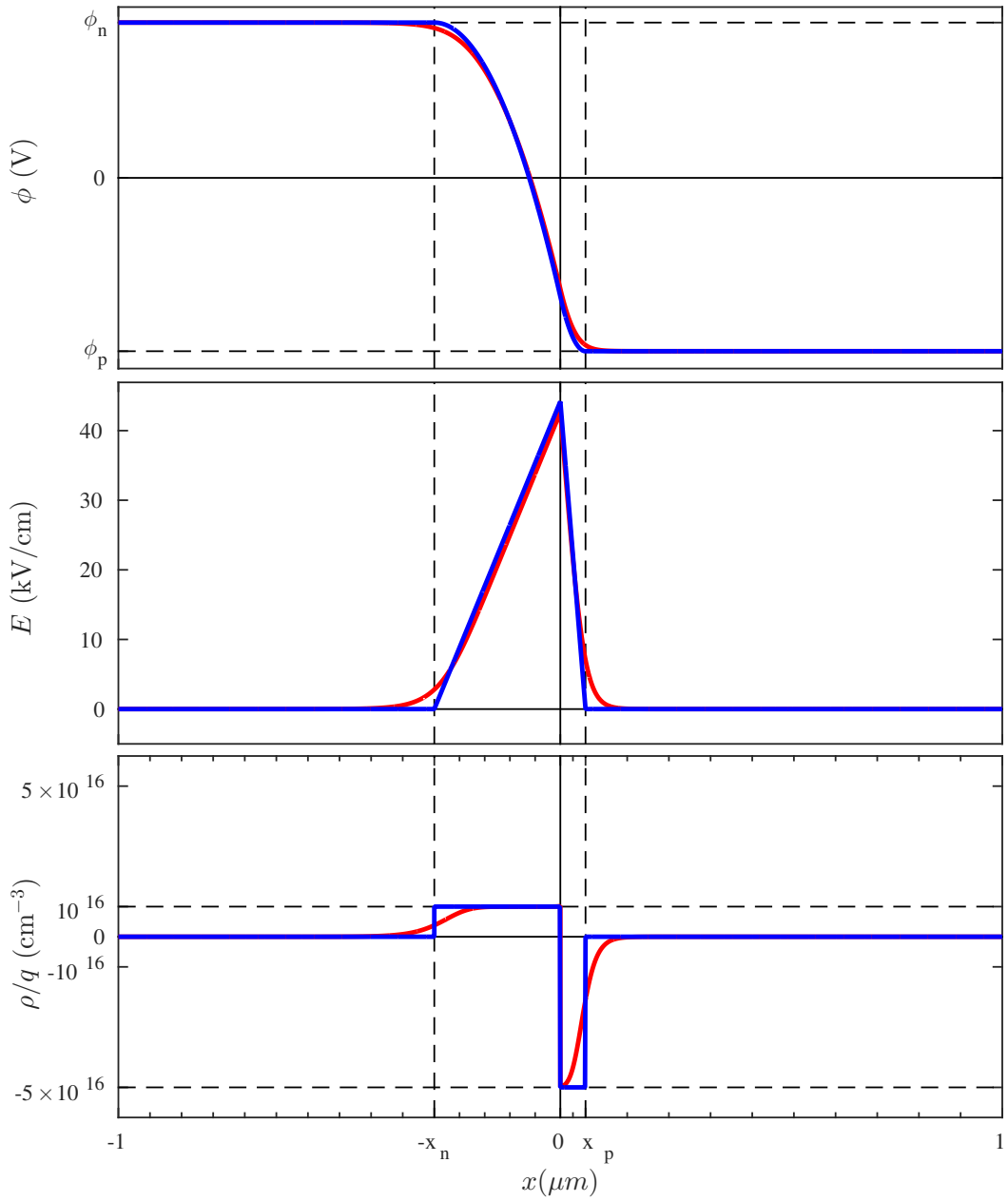


Figure 8.6: Solutions to Poisson equation in PN junction doped with $N_D = 10^{16}/cm^3$ ($x < 0$) and $N_A = 5 \times 10^{16}/cm^3$ ($x > 0$). Numerical solution to the full equation is in red. Depletion approximation solution is in blue.

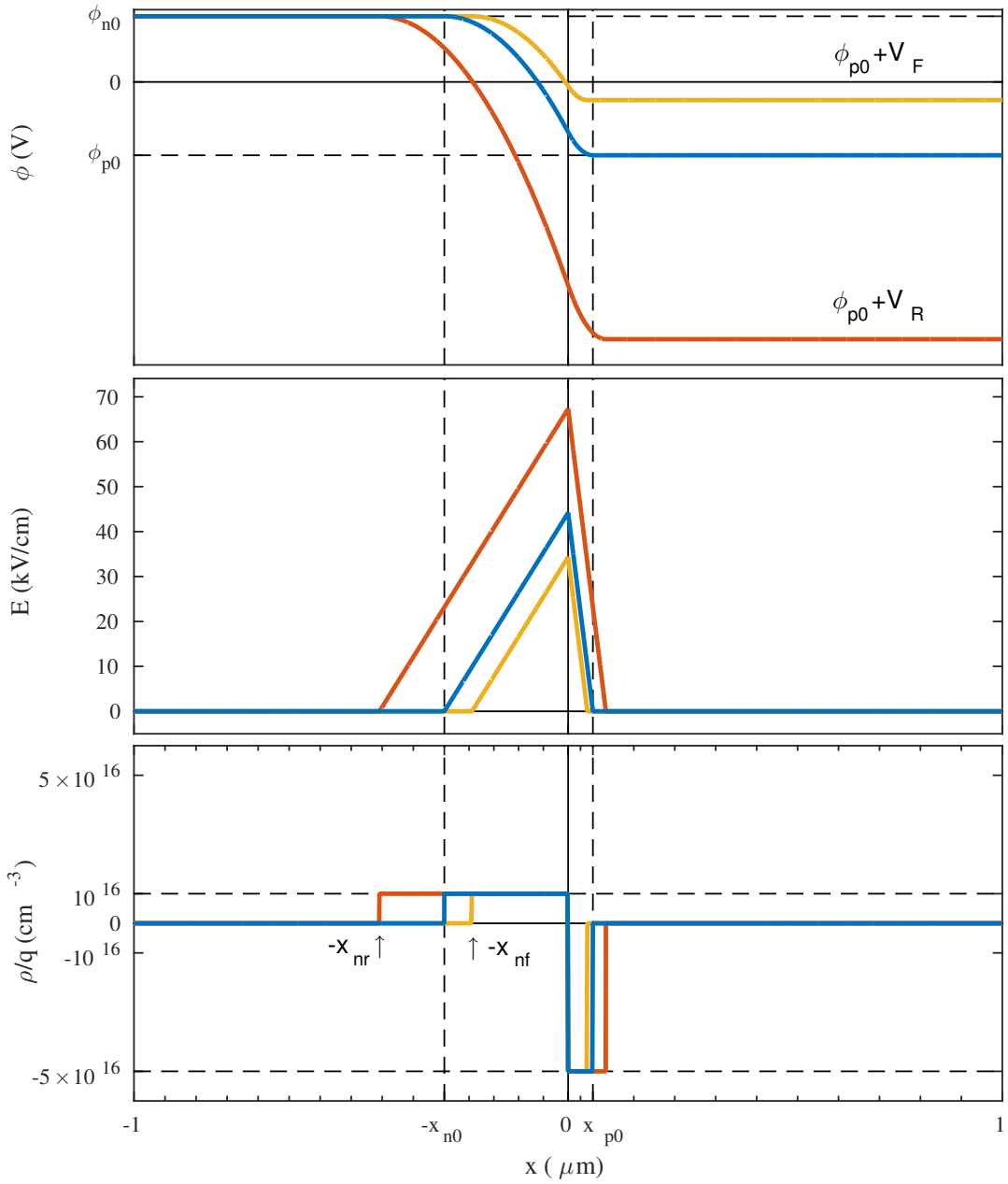


Figure 8.7: Depletion approximation solutions to Poisson equation in PN junction doped with $N_D = 10^{16}/cm^3$ $(x < 0)$ and $N_A = 5 \times 10^{16}/cm^3$ $(x > 0)$. Equilibrium solution is shown in blue, forward bias is shows in yellow, and reverse bias is shown in red. The reverse bias voltage $V_R = -1V$ and the forward bias voltage $V_F = 0.3V$.

Example 8.2:
Based on what you learned in this and the several preceding chapters, answer the following questions in a sentence of two:

1) How do you connect a power supply to forward or reverse bias the PN junction in Example 8.1?

To forward bias the junction, connect the positive potential to the boron doped (p) side, and the negative potential to the phosphorous (n) side. For reverse bias, apply the potential in the opposite manner.

2) Why is the electric field smaller during forward bias than in equilibrium or reverse bias, as evidenced by Figure 8.7?

In forward bias, the field due to the biasing potential is oriented opposite to that of the built-in field resulting from the built-in potential. These fields add together to result in a net field which is smaller, but still oriented the same direction as the original built-in field. Remember, the forward bias potential will be **less** than the built-in potential so the sign of the electric field will not reverse. In reverse bias, the fields will add to create a larger net field.

3) Why doesn't the drift current increase significantly in reverse bias even though the internal field increases as shown in Figure 8.7?

The drift current doesn't increase in reverse bias because despite the field magnitude, there are very few electrons in the depletion region. Because the field increases but the carrier concentration is still low, the drift current only increases by a very small amount.

8.5 Problems

8.1 If $I_o = 1 \times 10^{-15} A$, graph the diode equation from $-1 \leq V_A \leq 0.8$ volts. Indicate on the graph the regions of forward and reverse bias. Assume the temperature is $27^o C$, or room temperature.

8.2 PN Junction in Equilibrium Qualitative Operation

(a) What are the four basic current components in a PN junction.

(b) What is meant when we say a 'PN junction is in equilibrium'?

(c) In equilibrium, what are the relative magnitudes and directions of each of the four current components?

8.3 PN Junction Forward Bias Qualitative Operation

(a) How do we connect a battery to forward bias a PN junction? Indicate which pole of the battery connects to the P side or N side.

(b) When you forward bias a PN junction, what happens to the drift current for each carrier? What happens to the diffusion current for each carrier?

(c) Use your answers to the previous two questions to summarize the operation of a PN junction diode for forward bias operation in a few sentences.

8.4 PN Junction Reverse Bias Qualitative Operation

(a) How do we connect a battery to reverse bias a PN junction? Indicate which pole of the battery connects to the P side or N side.

(b) When you reverse bias a PN junction, what happens to the drift current for each carrier? What happens to the diffusion current for each carrier? Do any of the current components change significantly?

(c) Use your answers to the previous two questions to summarize the operation of a PN junction diode for reverse bias operation in a few sentences.

8.5 Derive the expressions for the electric field and electrostatic potential versus position in the P-side of the depletion region for a PN junction. Graph the electric field and electrostatic potential for the entire PN (or NP) junction.

8.6 Derive the expressions for the lengths of the depletion regions x_n and x_p.

8.7 If a PN junction is doped with $N_D = 10^{16} cm^{-3}$ and $N_A = 10^{17} cm^{-3}$:

(a) Graph the electric field as a function of position and the built-in potential as a function of position with zero applied voltage. Include the field and potential in the quasi neutral regions on the P and N sides, as well as in the depletion region near the junction.

(b) Determine the length of the depletion region length on the n-side and the p-side, respectively.

(c) Give the expression for boundary value potentials, ϕ_n and ϕ_p.

8.8 In an abrupt PN junction, where is the field strength at its maximum? What happens to the maximum and the depletion width if you increase both the acceptor and donor doping concentrations? Why?

8.9 Derive an expression for the maximum of the electric field in an abrupt PN junction under no bias as a function of **only**: the doping concentrations N_A and N_D, the intrinsic carrier concentration n_i, the dielectric constant of the semiconductor ϵ, the Boltzmann constant K and the temperature T.

8.10 If one side of the PN junction is doped more heavily than the other, comment on how the field inside the depletion region is affected.

8.11 An abrupt PN junction is created with $N_A = 10^{17} cm^{-3}$ and $N_D = 10^{18} cm^{-3}$. For a certain bias condition you know that the depletion length on the p-side extends $1um$ from the junction. Only using charge neutrality, calculate the depletion region length on the n-side. Why can we calculate the depletion length this way?

8.12 What are the key assumptions made in the depletion approximation? What do these assumptions mean physically? How might things be different in the real world?

8.13 We know from Gauss's Law and the definition of the electric field in one dimension that the integral of the charge density gives us the electric field, and the integral of the electric field gives us the minus potential. As a result, given the abrupt PN junction charge density shown in the first row of Figure 8.8 below, we can deduce the triangular electric field profile and quadratic (minus) potential profile. For charge densities A), B), and C), match their corresponding electric field and minus potential profiles.

8.14 Built-in Potential versus Temperature: Calculate and graph the built-in potential versus Temperature for the silicon PN junction that has $N_D = 10^{16} cm^{-3}$ and $N_A = 10^{16} cm^{-3}$ for temperature ranging between $0^o C$ and $200^o C$. You will need to also calculate the intrinsic carrier concentration as a function of temperature to do this problem, which was discussed in Chapter 5.

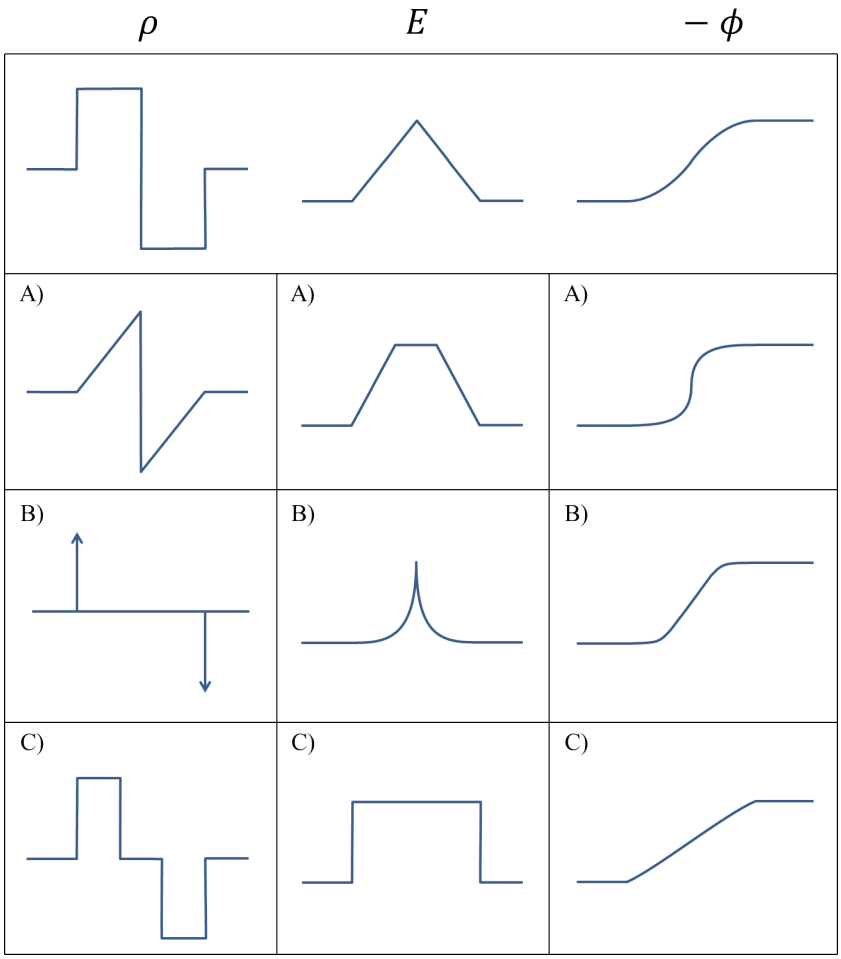


Figure 8.8: Plots of charge density ρ (first column), electric field E (second column), and minus potential $-\phi$ (third column) for Question 8.13. The first row shows a properly matched set based on the governing physical equations.

8.15 A silicon PN junction is doped on one side with $10^{16}/cm^3$ boron and the other side with $10^{16}/cm^3$ phosphorous. The depletion approximation says that the concentration of mobile carrier is negligible compared to that of the ionized dopant atoms in the vicinity of the PN junction. Using the exponential expressions relating potential and carrier concentration in equilibrium, along with the calculated electrostatic potential from the depletion approximation, calculate and graph the mobile electron and hole concentrations along the depletion region. Compare your result of the mobile carrier concentration with that of the ionized dopant concentration. Calculate and graph the percent error in the depletion approximation as a function of position along the depletion region.

Chapter 9

PN Junction Part II

9.1 Introduction

In this chapter we will derive the diode equation which is one of the most fundamental equations in electronics:

$$I_D = I_0(e^{V_A/V_T} - 1) \tag{9.1}$$

9.2 Continuity Equations

First we derive continuity equation for electrons which is a partial differential equation in space and time. In three spatial dimensions it is given as:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \vec{J_n} + G_n - R_n \tag{9.2}$$

In one spatial dimension the continuity equation for electrons is:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J}{\partial x} + G_n - R_n \tag{9.3}$$

The continuity equation says that the change in the number of electrons per unit time in a small volume element is equal to the flow of electrons into the volume element, minus the flow of electrons out of the volume element, plus any generation of electrons minus any recombination of electrons within the volume. This is illustrated in Figure 9.1, and expressed mathematically as follows:

$dVolume = (\Delta x A)$

$$\frac{\partial n}{\partial t}\Delta x A = \frac{1}{-q}(J_n(x) - J_n(x + \Delta x))A + (G_n - R_n)\Delta x A \tag{9.4}$$

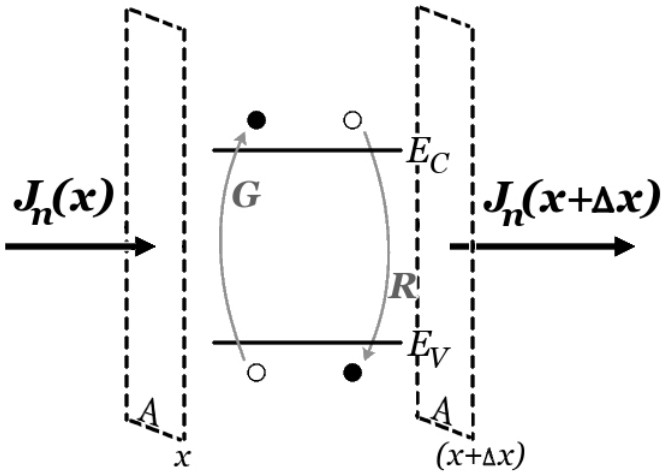


Figure 9.1: Illustration of physical meaning of Continuity Equation.

Expanding $J_n(x + \Delta x)$ in a Taylor series and canceling the cross-sectional area A from all terms gives:

$$\frac{\partial n}{\partial t}\Delta x = \frac{1}{-q}(J_n(x) - J_n(x) - \frac{\partial J_n}{\partial x}\Delta x) + (G_n - R_n)\Delta x \tag{9.5}$$

Canceling the Δx, the $J_n(x)$ terms and the negative sign gives the continuity equation for electons in one spatial dimension and time:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} + (G_n - R_n) \tag{9.6}$$

Following an analogous procedure for holes will give the following continuity equation for holes:

$$\frac{\partial p}{\partial t} = \frac{-1}{q}\frac{\partial J_p}{\partial x} + (G_p - R_p) \tag{9.7}$$

9.2.1 Generation and Recombination

The term $-R_n$ stands for the loss of electrons from the conduction band by their recombining with holes in the valence band. Physically, what is actually happening is that a conduction band electron falls back into the valence band. As a result, there is one less electron in the conduction band, and one less hole in the valence band. The units for R_n are concentration/time or $1/cm^3\,sec$. There are various mechanisms of recombination, which students and engineers will study later in their careers. For the standard electrical operation of semiconductor devices, like diodes, MOSFETs and BJTs, the most common type of recombination is called

Shockley-Read-Hall recombination, which describes recombination of electrons with holes being facilitated by defect traps in semiconductor.

The term G_n in the continuity equation reflects the generation of electrons from the valence band into the conduction band. This will occur thermally when the temperature rises, or when electrons absorp photons with sufficient energy. It can also occur when electrons in the conduction band get so much energy by an applied electric field that they collide with electrons in the valence band and cause them to jump to the conduction band. This is called impact-ionization and it generates both an electron and a hole. These are mechanisms for further study later in the students' careers.

During standard diode operation, the main phenomenon occuring is the recombination of electron holes pairs in the bulk or quasi-neutral regions of the PN junction. For these circumstances, it turns out that the generation/recombination term can be given as:

$$G_n - R_n = -\frac{\Delta n}{\tau_n} \tag{9.8}$$

Where τ is the recombination lifetime. (Note that the τ_n here is different from the one in the formula for mobility, ,which is the mean time between collisions.) There will be an analogous expression for the recombination of holes in the hole-continuity equation.

Example 9.1:
A steady state direct electron flow is entering into one end of a region of a semiconductor device that is $50\mu m$ long with a cross-sectional area of $100\mu m \times 100\mu m$. Measurements find that the electron current is totally absorbed in this region by electron-hole recombination. In other words, electrons flows into the region but none flow out. We also find that the mobile electron concentration increases uniformly in the region by $\Delta n = 1\times10^{14}/cm^3$ in order to accommodate the recombination process, and the recombination lifetime $\tau_n = 1\times10^{-6}s$. At the other end of the region hole current is flowing in and it too is totally absorbed by recombining with electrons that are entering from the other end. How much electron current is flowing through one side of the region, and how much hole current is flowing through the other side?

Start with the continuity equation for electrons:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} + (G_n - R_n)$$

DC Steady State current means the concentration doesn't change with time:

$$\frac{\partial n}{\partial t} = 0$$

We know that the total generation/recombination term must be negative since we only have recombination:

$$G_n - R_n = -\frac{\Delta n}{\tau_n}$$

Plugging these values into the steady state continuity equation:

$$\frac{1}{q}\frac{dJ_n}{dx} = \frac{\Delta n}{\tau_n}$$

Since Δn is constant, this equation can be directly integrated to get the current density. We can multiply by the area A to get the total current component.

$$I_n = A\frac{q\Delta n}{\tau_n}\int_0^{50\mu m} dx = (100\mu m)^2\frac{1.6\times10^{-19}C\cdot10^{14}cm^{-3}\cdot50\mu m}{10^{-6}s} = 8\times10^{-6}A$$

We know that all holes also recombine in this region, so from symmetry, the same amount of hole current must flow **in** to the other side.

$$I_p = I_n$$

9.3 Fundamental Semiconductor Equations:

Now that we have derived the continuity equations, we have seen in this course the five coupled fundamental semiconductor equations. These fundamental semiconductor equations are to semiconductors as Maxwell's equation are to Electricity and Magnetism, and as Schrodinger's equation is to Quantum Mechanics. We list the five fundamental semiconductor equations below:

Semiconductor Equations

$$\frac{\partial^2\phi}{\partial x^2} = -\frac{q}{\epsilon}(p - n + N_D^+ - N_A^-) \quad \text{Poisson Equation} \tag{9.9}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} + G_n - R_n \quad \text{Electron Continuity Equation} \tag{9.10}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} + G_p - R_p \quad \text{Hole Continuity Equation} \tag{9.11}$$

$$J_n = -qn\mu_n\frac{\partial \phi}{\partial x} + qD_n\frac{\partial n}{\partial x} \quad \text{Electron Current Equation} \tag{9.12}$$

$$J_p = -qp\mu_p\frac{\partial \phi}{\partial x} - qD_p\frac{\partial p}{\partial x} \quad \text{Hole Current Equation} \tag{9.13}$$

We have already used some of these equations previously in this course. For example, we used the current equations to obtain the built-in electric field in an N1/N2 junction and a PN junction in equilibrium. Also, we used the Poisson equation to describe the electric field and the electrostatic potential in the Depletion region of a PN junction. To perform a detailed, comprehensive device analysis (of a diode, BJT, MOSFET, etc.) or to design a device, these equations are solved fully on a computer, all at once, self-consistently, with specific device structure (doping, geometry, etc.) as input. The result gives a comprehensive description of the operation of the device including the current-voltage characteristics, the MOSFET threshold voltage, BJT Beta, etc.

In this course we will continue to use these equations, but in an approximate decoupled way to obtain analytical expressions for the device operation. In this chapter we will use them to derive the current equation for a PN junction. More specifically, we will use mainly the current and continuity equations.

9.4 Derivation of Diode Equation

Before starting our derivation let's specify the following notation which we will use in the derivation.

N-Side:

n_{n0} = Electron equilibrium concentration in N-Side bulk.
p_{n0} = Hole equilibrium concentration in N-Side bulk
p_n = Non-Equilibrium Hole concentration on N-side,

P-side:

p_{p0} = Equilibrium hole concentration in P-Side bulk.
n_{p0} = Electron equilibrium concentration in P-Side bulk.
n_p = Non Equilibrium Electron concentration on P-side

Minority Carriers: Electrons on the P-Side, and Holes on the N-Side.
Majority Carriers: Electrons on the N-Side, and Holes on the P-Side.

Electron Diffusion Length: L_n: Distance electron diffuses in P-type material before it recombines with a hole ($L_n = \sqrt{D_n \, \tau_n}$).
Hole Diffusion Length: L_p: Distance hole diffuses in N-type material before it recombines with an electron ($L_p = \sqrt{D_p \, \tau_p}$).

Equilibrium: Recall, the equilibrium concentrations are when there is no applied voltage and no current flowing. In this section, we will not be in equilibrium because we will apply a voltage and obtain the resulting current.

General Approach

Recall that when we apply a forward bias voltage, built in field and the built-in potential decrease. This allows diffusion current to dominate over drift current. Our analysis will be based largely on obtaining the diffusion current that results after we apply a forward bias. To do so, we will solve the current and continuity equations for minority carriers. This will first give us the electron concentration in the P-side and the hole concentration in the N-side. When we get these non-equilibrium concentrations, we will utilize them in the diffusion current expression to get the minority carrier current flows. Finally, we will use continuity (KCL) to get the total current which will give the diode equation.

Electron Current in Quasi-Neutral P-Side

To find the current we will start by finding the electron concentration and the electron current in the bulk P-side. In other words, we will find the minority carrier concentration and current in what we called 'Region 4' in the previous chapter. To do this, we will start with the Poisson (9.9), Electron Continuity (9.10), and Electron Current (9.12) equations from above, and solve them self-consistently for the potential $\phi(x)$, the electron concentration $n(x)$, and the electron current $J_n(x)$ throughout the bulk P-side.

First of all, the Poisson equation solution is easy. Since we make the very good approximation that the field is contained in the depletion region, then the potential in the quasi neutral P-side is constant and the field $E = 0$. This leaves us with only the continuity and current equations to solve.

Now, let's further make our lives easier by looking only for a steady state or time independent solution. We can then set $\frac{\partial n}{\partial t} = 0$. This leaves us with the following forms of the Electron Current and Electron Continuity Equations in the quasi-neutral P-region.

$$0 = \frac{1}{q}\frac{dJ_n}{dx} + G_n - R_n \tag{9.14}$$

$$J_n = qD_n\frac{dn}{dx} \tag{9.15}$$

Note that the current equation now only has a diffusion component because the field is zero and hence no drift.

We now have two equations and two unknowns (J_n and n). Now we reduce to one equation by substituting the expression for J_n from the diffusion current into the continuity equation. This leaves only the continuity equation and one unknown which is $n(x)$:

$$0 = \frac{1}{q}\frac{d}{dx}(qD_n\frac{dn}{dx}) + G_n - R_n \tag{9.16}$$

Now, let's re-write $n(x)$ as the sum of the excess electron concentration resulting from the applied bias $\Delta n(x)$ and the constant equilibrium concentration:

$$n(x) = \Delta n(x) + n_{p0} \tag{9.17}$$

The Generation / Recombination term, which is the rate in which the excess electrons recombine with holes in the P-side, can be approximated as:

$$G_n - R_n = \frac{-\Delta n}{\tau} \tag{9.18}$$

Remember, when an electron and hole recombine, we lose both of these mobile carriers. Now substituting for $n(x)$ and $G_n - R_n$ into equation (9.16) gives

$$0 = D_n\frac{d^2\Delta n}{dx^2} - \frac{\Delta n}{\tau} \tag{9.19}$$

Note that $\frac{d(\Delta n + n_{p0})}{dx} = \frac{d\Delta n}{dx}$ since n_{p0} is a constant. Equation (9.19) is a simple second order homogeneous differential equation with constant coefficients which has the general solution:

$$\Delta n(x) = Ae^{x/L_n} + Be^{-x/L_n} \tag{9.20}$$

where, $L_n^2 = D_n\tau$. The parameter L_n is called the electron diffusion length, and it is the average distance that the electron will diffuse in the P-side before it recombines with a hole.

Boundary Conditions

Now apply boundary conditions at $x = \infty$ and $x = x_p$, to get A and B. First, $\Delta n(\infty) = 0$, so $A = 0$. This leaves

$$\Delta n(x) = Be^{-x/L_n} \tag{9.21}$$

Now apply boundary condition at $x = x_p$. To find the concentration at the beginning of the P-bulk region (Depletion region edge), we first recall built-in potential relations from the previous chapter:

$$n_{p0} = n_{n0}e^{-\phi_0/V_T} \tag{9.22}$$

Where ϕ_0 is the Built-in potential.

Now in addition to the built-in potential, we also have an external or applied potential. So we will add this additional potential to ϕ_0. This gives the electron concentration at the edge of the P-side while including the effect of the added voltage V_A:

$$n_p(x_p) = n_{n0}e^{(-\phi_0+V_A)/V_T} \tag{9.23}$$

Now, we write in terms of the excess concentration at the boundary by subtracting off the equilibrium value:

$$\Delta n(x_p) = n_p(x_p) - n_{p0} = n_{n0}e^{-\phi_0/V_T}(e^{V_A/V_T} - 1) \tag{9.24}$$

Using equation (9.22) gives:

$$\boxed{\Delta n(x_p) = n_{p0}(e^{V_A/V_t} - 1)} \tag{9.25}$$

Note that equation is very important and intuitive. It says that the extra electron concentration at the boundary of the P-side is simply the equilibrium concentration multiplied by the exponential factor that contains the applied bias V_A. In other words, equation (9.25) tells us a great deal about the effect of the applied bias.

Now continuing to find B:

$$\Delta n(x_p) = n_{p0}(e^{V_A/V_T} - 1) = Be^{-x_p/L_n} \tag{9.26}$$

Solving for B gives:

$$B = \Delta n(x_p)e^{x_p/L_n} = n_{p0}(e^{V_A/V_T} - 1)e^{x_p/L_n} \tag{9.27}$$

Finally, substituting for B into equation (9.21) gives:

$$\boxed{\Delta n(x) = n_{p0}(e^{V_A/V_T} - 1)e^{(x_p-x)/L_n}} \tag{9.28}$$

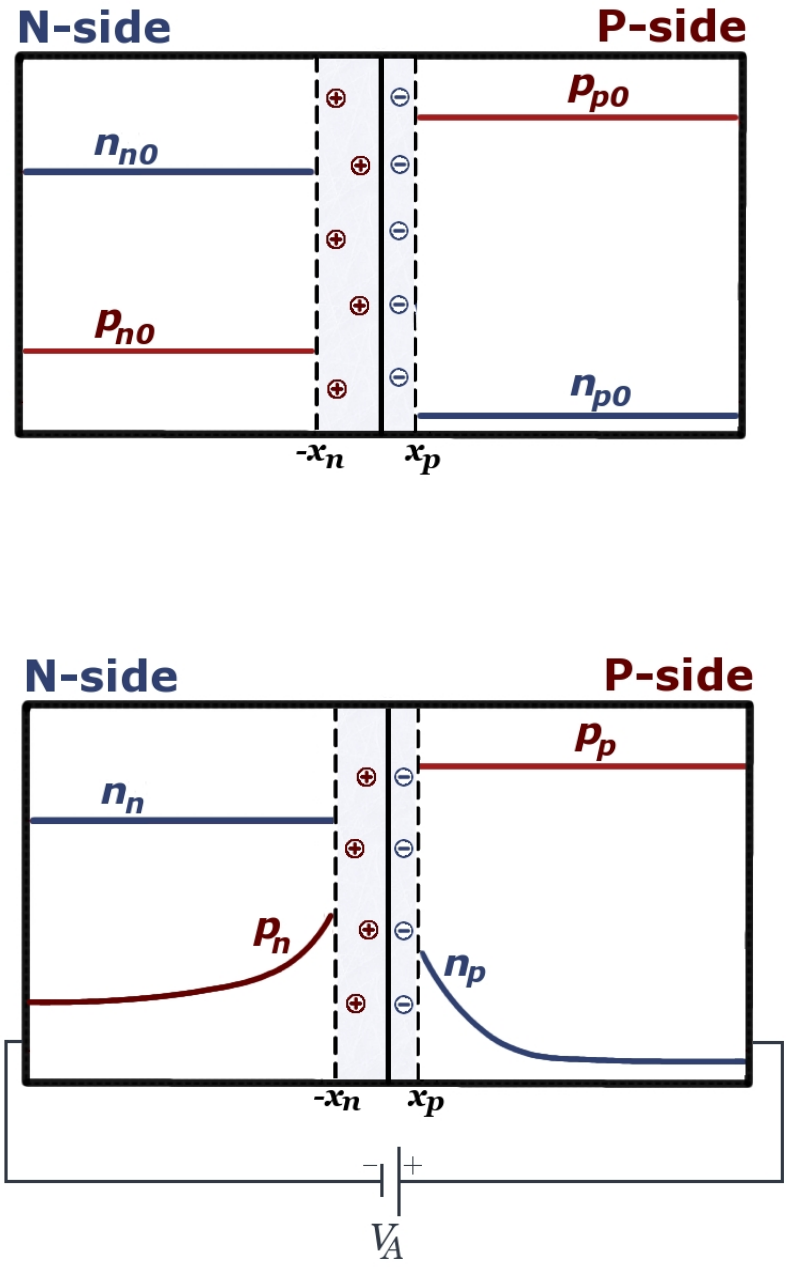


Figure 9.2: Top figure illustrates the electron and hole concentrations in equilibrium. The bottom figure illustrates the carrier concentration when a forward bias is applied to the PN junction. The figure is not really to scale since the majority concentration in each region is many orders of magnitude larger than the minority concentration. Thus the scale is generally logarithmic. 165

Equation (9.28) is important. It gives the excess electron concentration (concentration in excess of the equilibrium) on the P-side as a function of position after applying external voltage V_A to the PN junction. As a check, note that if we don't apply an external voltage, then V_A would be zero and there would also be zero excess p-side electrons.

Electron Current on P-Side

Now that we have the electron concentration as a function of position, we can easily substitute it back into equation for diffusion current:

$$J_n(x) = qD_n \frac{dn(x)}{dx} \tag{9.29}$$

Substituting $n(x)$ from (9.28) into (9.29) and then taking the derivative of the concentration with respect to x and multiplying by q and diffusivity D_n gives the following final expression for electron current density in the quasi-neutral p-side:

$$\boxed{J_{n_p}(x) = -\frac{qD_n n_{p0}}{L_n}(e^{V_A/V_T} - 1)e^{(x_p - x)/L_n}; \quad (x \geq x_p)} \tag{9.30}$$

Excess Hole Concentration and Hole Current on N-Side

Performing a totally analogous procedure for the minority carrier holes, we get the following final expression for excess hole concentration versus position on the N-side:

$$\boxed{\Delta p(x) = p_{n0}(e^{V_A/V_T} - 1)e^{(x_n + x)/L_p}} \tag{9.31}$$

Substituting the hole concentration into the expression for diffusion current:

$$J_p(x) = -qD_p \frac{dp(x)}{dx} \tag{9.32}$$

We obtain the following final expression for hole current density in the N-side quasi neutral region:

$$\boxed{J_{p_n}(x) = -\frac{qD_p p_{n0}}{L_p}(e^{V_A/V_T} - 1)e^{(x_n + x)/L_p}; \quad (x \leq -x_n)} \tag{9.33}$$

The electron and hole current densities, as well as the total current density, under forward bias are illustrated in Figure 9.3.

Current in Depletion Region:

To find currents J_n, J_p in the depletion region , we utilize he very good approximation that there is negligible generation or recombination of mobile electrons or holes in the depletion region because the depletion region being thin, the electrons and holes do not find enough time to recombine. So $(G_n - R_n)$ and $(G_p - R_p)$ are

both taken as zero in this region. So the continuity equation for electrons and holes in the depletion region both simplify to:

$$\frac{dJ_n}{dx} = 0 \tag{9.34}$$

$$\frac{dJ_p}{dx} = 0 \tag{9.35}$$

Thus, J_n = constant and J_p = constant for $(-x_n \leq x \leq x_p)$. Furthermore, the constants are equal to the currents at the edges of the depletion region, which we have already determined above. So the values of the electron and hole currents throughout the depletion region are given as the known constants:

$$J_n(x_p) = -\frac{qD_n n_{p0}}{L_n}(e^{V_A/V_T} - 1) \qquad (-x_n \leq x \leq x_p) \tag{9.36}$$

$$J_p(-x_n) = -\frac{qD_p p_{n0}}{L_p}(e^{V_A/V_T} - 1) \qquad (-x_n \leq x \leq x_p) \tag{9.37}$$

Furthermore, since the current components are constant in the depletion region, we also have $J_n(x_p) = J_n(-x_n)$ and $J_p(-x_n) = J_p(x_p)$. See Figure 9.3 to get a picture of the constant currents in the depletion region.

Total Diode Current

Using Kirchoff's Current Law (which is also a form of the continuity equation), we know that the total current J_{Tot} in the PN junction must constant. Furthermore we also know that the total current at any point is the sum of the electron current and the hole current:

$$J_{Tot} = J_n(x) + J_p(x) = constant \tag{9.38}$$

We know the total electron current and the total hole current at every point in the depletion region. So, if we add the depletion region currents given by equations (9.36) and (9.37) we will get the total current density:

$$J_{Tot} = J_n(x_p) + J_p(-x_n) \tag{9.39}$$

$$\boxed{J_{Tot} = \left[\frac{qD_p p_{n0}}{L_p} + \frac{qD_n n_{p0}}{L_n}\right](e^{V_A/V_T} - 1)} \tag{9.40}$$

or

$$J_{Tot} = J_0(e^{V_A/V_T} - 1) \tag{9.41}$$

where

$$J_0 = \left[\frac{qD_p n_i^2}{N_D L_p} + \frac{qD_n n_i^2}{N_A L_n}\right] \tag{9.42}$$

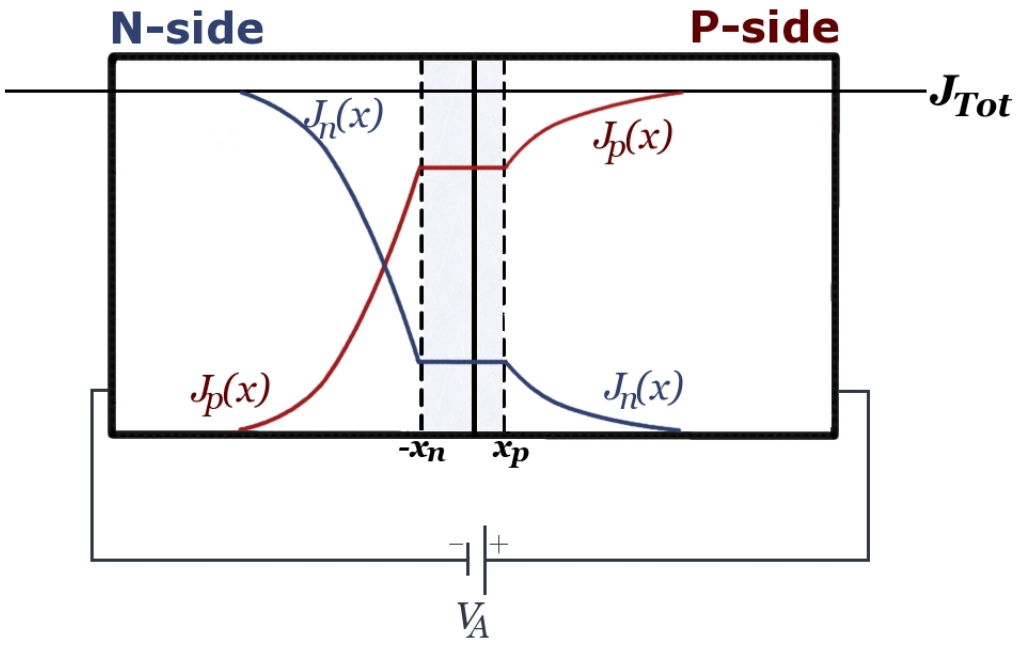


Figure 9.3: The electron, hole and total currents throughout the PN junction under forward bias.

Where we have used $n_{p0} = \frac{n_i^2}{N_A}$ and $p_{n0} = \frac{n_i^2}{N_D}$. In our equation for total current density the negative sign has been left off because we are mainly interested in **magnitude** of the current.

Recall the J_{Tot} is actually the current density, so in order to get the current, we need to multiply by the cross sectional area of the diode or $I_o = AJ_o$ so that we obtain:

$$\boxed{I_D = I_0(e^{V_A/V_T} - 1)} \tag{9.43}$$

Equation (9.43) is called the **Diode Equation** which we presented in the beginning of this chapter. Here we have derived the diode equation, which is one of the most fundamental equations in solid state electronics.

Example 9.2:
If the N-side of a Silicon PN junction is doped with $10^{17}/cm^3$ donor phosphorous atoms, and the P-side of the junction is doped with $10^{16}/cm^3$ acceptor boron atoms, calculate the total current in the diode at a forward bias of $V_A = 0.6V$. Take the cross sectional area to be $A_C = (100um)^2$. Also, what percentage of the total current across the junction at $x = 0$ is due to electron current and hole current, respectively?

In order to evaluate the diode current equation we need to find values for the diffusion coefficients D_n and D_p and diffusion lengths L_n and L_p.

$$D_n = \mu_n \frac{KT}{q} = 1400cm^2/Vs \cdot 0.026V = 36cm^2/Vs$$
$$D_p = \mu_p \frac{KT}{q} = 450cm^2/Vs \cdot 0.026V = 12cm^2/Vs$$
$$L_n = \sqrt{D_n \tau_n} = \sqrt{36cm^2/Vs \cdot 4 \times 10^-6s} = 120um$$
$$L_p = \sqrt{D_p \tau_p} = \sqrt{12cm^2/Vs \cdot 2 \times 10^-6s} = 49um$$

Plugging these values into the diode current equation:

$$I_D = A_C q n_i^2 \left(\frac{D_p}{N_D L_p} + \frac{D_n}{N_A L_n} \right) \left(\exp\left(\frac{V_A}{V_T} \right) - 1 \right) = 5.5 \times 10^{-6} A$$

To find the percentage of the current components at $x = 0$, we use the constant value of J_n or J_p inside the depletion region. In this case, we'll use the equation for holes, but either equation could be used.

Because the current components inside the depletion region are constant due to our assumption that there is no generation or recombination here:

$$J_p(0) = J_p(-x_n) = \frac{q D_p n_i^2}{N_D L_p} (e^{V_A/V_T} - 1) = 4.13 \times 10^{-3} A/cm^2$$

Taking the ratio of the hole current to the total gives us the hole current fraction:

$$p\% = \frac{41.3A/m^2 \cdot (100um)^2}{5.5 \times 10^{-6} A} = 7.5\%$$
$$n\% = 100\% - p\% = 92.5\%$$

Majority Carrier Currents

While we have found the total diode current and the minority carrier currents,

we can also now find the majority carrier currents. Since the total current is constant, and we know the minority carrier current in each region, we can just subtract the minority current from the total current to get the majority carrier current. In other words, the majority carrier or electron current density in the N-side is:

$$J_{n_n}(x) = J_{Tot} - J_{p_n}(x) \tag{9.44}$$

and the hole current density in the P-side is:

$$J_{p_p}(x) = J_{Tot} - J_{n_p}(x) \tag{9.45}$$

Where J_{Tot}, $J_{p_n}(x)$ and $J_{n_p}(x)$ are given by equations (9.41), (9.33) and (9.30), respectively. Figure 9.3 illustrates the majority and minority currents throughout the forward biased PN junction.

Section Summary

In this section we derived the diode equation which is one of the most important parts of semiconductor device physics. The main result indicates that the current in forward bias is exponentially dependent on the applied voltage V_A. Furthermore, if we look at equation (9.40), we see that, in addition to the applied voltage, the current is given by material parameters: Diffusivity ($D_n,\ D_p$), Diffusion Lengths ($L_n,\ L_p$), intrinsic carrier concentration n_i, thermal voltage $V_T = \frac{KT}{q}$, as well as the doping concentrations N_A and N_D. Note, electronic devices are engineered by carefully choosing the geometry and doping concentrations while using equations like (9.40) as guidelines.

Example 9.3:
If the N-side of a silicon PN junction is doped with $10^{17}/cm^3$ phosphorous atoms, and the P-side is doped with $10^{17}/cm^3$ boron atoms, calculate the built-in potential. What are are the electron and hole concentrations at the edges of the quasi-neutral regions $-x_n$ and x_p in equilibrium?

$$\phi_o = V_T \ln \frac{N_D N_A}{n_i^2} = 0.026 \ln \frac{10^{17}10^{17}}{(10^{10})^2} = 0.84V$$

$$n(-x_n) = N_D = 10^{17}/cm^3$$

$$n(x_p) = \frac{n_i^2}{N_A} = \frac{(10^{10})^2}{10^{17}} = 10^3/cm^3$$

$$p(x_p) = N_A = 10^{17}/cm^3$$

$$p(-x_n) = \frac{n_i^2}{N_D} = \frac{(10^{10})^2}{10^{17}} = 10^3/cm^3$$

Now, suppose a positive voltage of $V_A = 0.6V$ is applied to between the ends of the junction with the positive terminal connected to the P-side so it is forward biased. What are the electron and hole concentrations now at the edges of the quasi-neutral regions?

$$n(x_p) = n(-x_n)e^{(-\phi_o+V_A)/V_T} = N_D e^{(-0.84+0.60)/0.026} = 9.8 \times 10^{12}/cm^3$$

$$p(-x_n) = p(x_p)e^{(-\phi_o+V_A)/V_T} = N_A e^{(-0.84+0.60)/0.026} = 9.8 \times 10^{12}/cm^3$$

9.5 Diode Capacitances

Associated with a PN junction are also intrinsic capacitors which limit the speed at which a diode can respond. The capacitors are in parallel with the ideal diode, and are illustrated in Figure 9.4. As shown in the figure, there are two basic types of capacitances in a PN junction:

1. C_j : Depletion region or junction capacitance.
2. C_{Diff} : Diffusion capacitance: mainly in forward bias. The diffusion capacitance can be further identified as a diffusion capacitance due to excess electrons (C_{Diff_n}) or a diffusion capacitance due to excess holes (C_{Diff_p}).

In general, we will be looking for the small signal capacitance which is given by:

$$C = \left|\frac{dQ_n}{dV_A}\right| = \left|\frac{dQ_p}{dV_A}\right| \qquad Farads/area \tag{9.46}$$

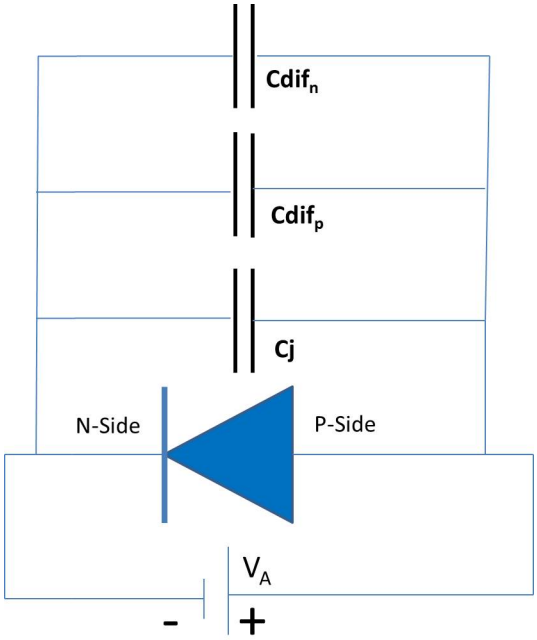


Figure 9.4: The intrinsic capacitors that are present in a PN junction diode. There are two types of capacitors: the junction capacitance and the diffusion capacitances. There is an N-type diffusion capacitor and a P-type diffusion capacitor.

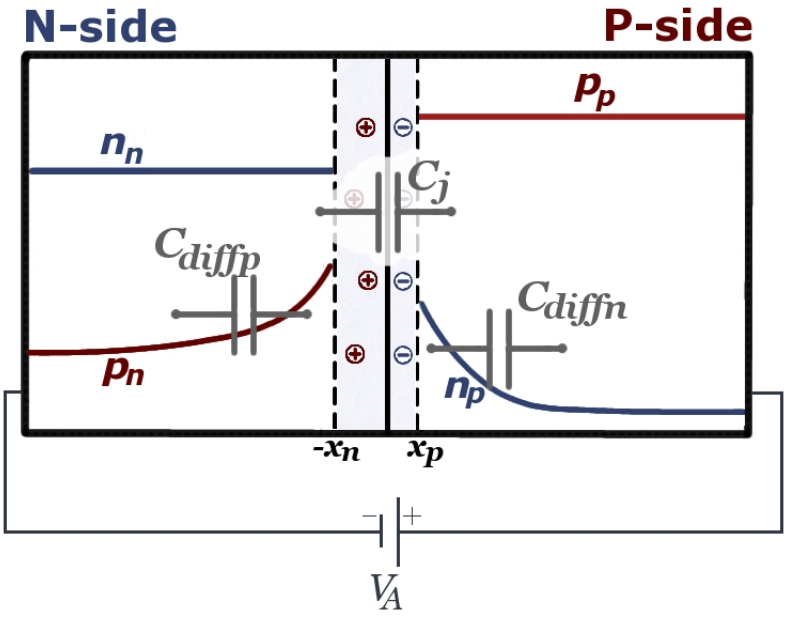


Figure 9.5: A cartoon-like drawing illustrates the sources of the junction and diffusion capacitances by placing the capacitor circuit symbol over the regions from which the capacitances arise.

Junction Capacitance

The junction capacitance comes from the charge in the depletion region as illustrated in Figure 9.5.The positive charge in the depletion region is: $Q_n = qN_D x_n$. We derive junction capacitance as follows:

$$C_j = \left|\frac{dQ_n}{dV_A}\right| = \left|\frac{dQ_n}{dx_n}\frac{dx_n}{dV_A}\right| = \left|qN_D\frac{dx_n}{dV_A}\right| \tag{9.47}$$

$$\frac{dx_n}{dV_A} = -\sqrt{\frac{\epsilon}{2q(\phi_0 - V_A)}[\frac{N_A}{N_D(N_A + N_D)}]} \tag{9.48}$$

$$\boxed{C_j = \sqrt{\frac{q\epsilon}{2(\phi_0 - V_A)}[\frac{N_D N_A}{(N_A + N_D)}]}} \tag{9.49}$$

Diffusion Capacitance:

Since it takes time to set up the excess carrier concentrations Δn and Δp after applying a forward bias, this gives rise to a charge and a capacitance:

$$C_{Diff_n} = \frac{dQ_{Diff_n}}{dV_A} \tag{9.50}$$

The charge associated with the diffusion capacitance due to electrons is the from the total excess electron concentration on the P-side during forward bias. This total charge is obtained from integrating the excess minority electron concentration throughout the quasi-neutral P-side.

$$Q_{Diff_n} = q \int_{x_p}^{\infty} \Delta n(x) dx \tag{9.51}$$

Substituting equation (9.28) for $\Delta n(x)$ and integrating yields:

$$Q_{Diff_n} = L_n q n_{p0} (e^{V_A/V_T} - 1) \tag{9.52}$$

Now differentiating with respect to the applied voltage gives the diffusion capacitance:

$$\boxed{C_{Diff_n} = \frac{dQ_{Diff_n}}{dV_A} = \frac{L_n q n_{p0}}{V_T} (e^{V_A/V_T})} \tag{9.53}$$

A similar process for excess minority holes gives:

$$\boxed{C_{Diff_p} = \frac{dQ_{dDiff_p}}{dV_A} = \frac{L_p q p_{n0}}{V_T} (e^{V_A/V_T})} \tag{9.54}$$

It is important to note that these capacitances are in units of Farads per unit area. To get the capacitor value for a specific diode, one has to multiply by the cross sectional area of the device. Figure 9.5 illustrates the source of the junction and diffusion capacitances by placing the capacitance circuit symbol over the related regions in the PN junction.

9.6 Problems

9.1 What is meant by the diffusion lengths L_n and L_p? Using the Einstein relation and the values in Table C.2, calculate typical diffusion lengths for electrons and holes in silicon.

9.2 A Silicon PN junction is doped with $N_D = 5 \times 10^{16} cm^{-3}$ and $N_A = 10^{17} cm^{-3}$, and has a cross-section area of $100um \times 100um$. Assume that the diode can be considered to be infinitely long. Obtain material parameters from Table C.2 A forward bias of 0.3V is applied to the diode at $300^{\circ}K$.

(a) Calculate and graph the minority carrier concentration as a function of position for this device.

(b) Calculate and graph the electron and hole currents as a function of position throughout the entire device, as well as the total current (Give $I_n(x)$ and $I_p(x)$ everywhere.)

(c) Graph the diode current versus applied voltage for this diode for $-1.0V < V_A < 0.7V$.

9.3 (a) Calculate the built in potential for the diode in the previous problem.

(b) Now, without doing any calculations, decide what is larger, x_n or x_p. Explain how you arrived at your answer.

9.4 Starting from the current and continuity equations for holes, derive the equation for excess minority hole concentration and minority carrier hole current for a PN junction under bias.

9.5 What are the boundary conditions $\Delta n(x_p)$ and $\Delta p(-x_n)$ for the excess minority carrier concentrations used when solving the current continuity equations to obtain the excess minority carrier concentrations: $\Delta n(x)$ and $\Delta p(x)$? Explain qualitatively how they come about and derive them algebraically.

9.6 Explain qualitatively what is meant by the depletion approximation. Describe the different regions of the PN junction under this approximate method of viewing the device.

9.7 A Silicon PN junction is doped with $N_D = 1 \times 10^{17} cm^{-3}$ and $N_A = 10^{17} cm^{-3}$, and has a cross-section area of $100um \times 100um$. Assume that the diode can be considered to be infinitely long. Obtain material parameters from Table C.2. Graph the diode current for V_A between -1.0V and 0.7V at $T = 27^oC$. Graph the diode current for the same voltages again, but this time for $T = 200^oC$.

9.8 A clever device engineer decided to add strain the Silicon lattice which slightly changed the spacing between atoms. This caused the curvature of the valence band to double. How would this affect the diode current in the preceding problem. Provide a qualitative explanation and a numerical value.

9.9 If the diode in problem (2) above has a reverse bias of $V_A = -1V$:

(a) Calculate the capacitance. Note that under reverse bias only the junction capacitance is important.

(b) Sketch the small signal frequency response of this diode if it is driven by a small AC signal with source resistance of 50Ω while also under this -1V reverse bias. (Ignore any diode resistance.)

9.10 If the diode in problem (2) above has a forward bias of $V_A = 0.3V$: Calculate the total capacitance of this diode at the given bias. (Note that under forward bias, the diode will have both junction and diffusion capacitance.)

Chapter 10

BJT: Bipolar Junction Transistor

10.1 Introduction

The bipolar junction transistor (BJT) is a three-terminal device which is mainly configured with external resistors and capacitors to be used as an amplifier in analog electronics. A BJT is typically described as a current-controlled current source where the current flowing between the emitter (E) and collector (C) terminals is controlled by the much smaller current flowing into or out of the base (B) terminal. In addition, it can also be described as a voltage controlled current source where the the current flowing between the emitter (E) and collector (C) terminals is controlled by the voltage applied between the base (B) and emitter (E) terminals. Since a linear variation in the B-E voltage gives rise to an exponential variation in the C-E current, the BJT has applications as a high gain amplifier.

The BJT exists in two varieties: NPN and PNP. An NPN BJT is composed of an N-type emitter, connected to the P-type base, which in turn is connected to an N-type collector. The emitter-base junction is an NP junction and the base-collector junction is a PN junction.

For the NPN BJT, the N-type doping in the emitter is typically a hundred times larger than the P-type doping in the base. Furthermore, the width of the base is typically very narrow compared to the recombination length for electrons and holes for reasons that will be explained later in the chapter.

10.2 BJT Modes of Operation and Basic Current Relationships

The BJT has four basic modes of operation:

1. Forward Active (Analog Applications)
2. Reverse Active (Not typically used)
3. Saturation (Digital Applications)
4. Cutoff (Digital Applications)

10.2.1 Forward Active Mode Basic Current Relations

The Forward Active Mode is by far the most commonly used of BJTs. It is typically used to construct voltage amplifiers. We will therefore focus on this mode of operation in this text. For a BJT to be biased in the forward active mode, the terminal voltages have to have the following relationships:

$$\begin{array}{ll} \text{NPN:} & V_C > V_B > V_E \\ \text{PNP:} & V_E > V_B > V_C \end{array}$$

Writing the KCL for currents in the BJT (See Fig. 10.1):

$$I_C + I_B = I_E \tag{10.1}$$

We here define β, the current gain:

$$I_C = \beta I_B \tag{10.2}$$

where the value of β varies, but it is typically greater than 100:

$$\boxed{\beta > 100} \tag{10.3}$$

The value of β depends on the structural design of the BJT. We will show how to calculate it later on in the chapter. Using Eqn. 10.2 in Eqn. 10.1 we can rewrite:

$$(\beta + 1)\frac{I_C}{\beta} = I_E \tag{10.4}$$

$$\Rightarrow \frac{\beta}{\beta + 1} I_E = I_C \tag{10.5}$$

Defining α as the ratio

$$\boxed{\alpha = \frac{\beta}{\beta + 1}} \tag{10.6}$$

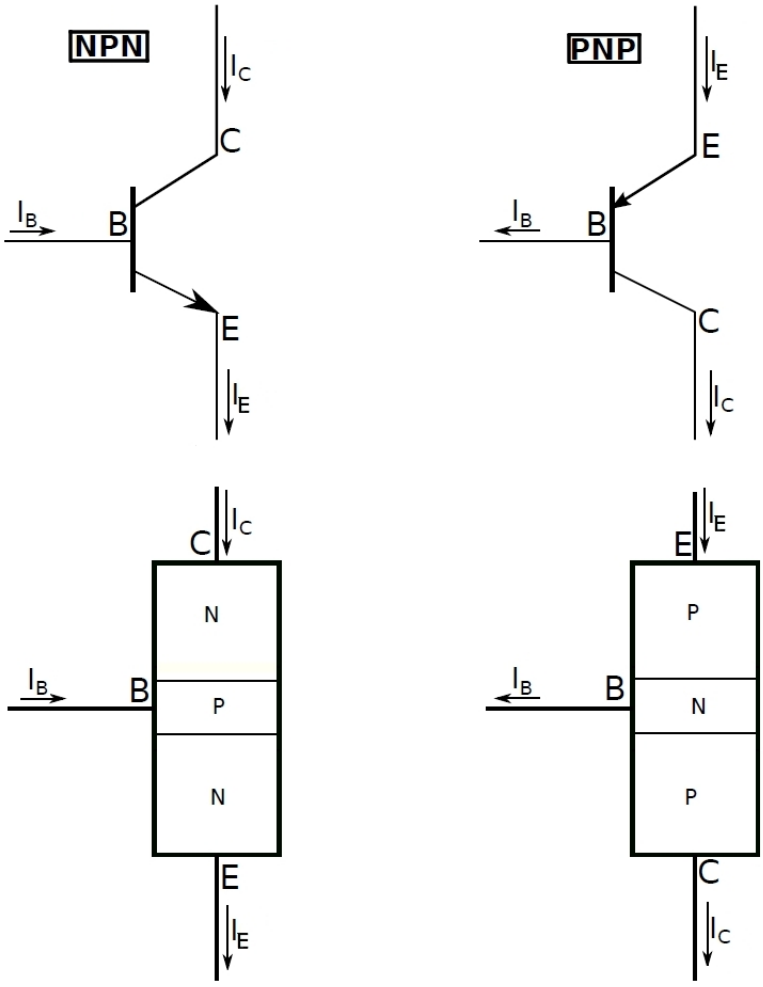


Figure 10.1: A BJT is a three terminal device with emitter, collector and base terminals. An NPN BJT has an n-type collector and emitter, and a thin, lightly doped p-type layer as the base in between. A PNP BJT has a p-type collector and emitter, and a thin, lightly doped n-type layer forms the base.

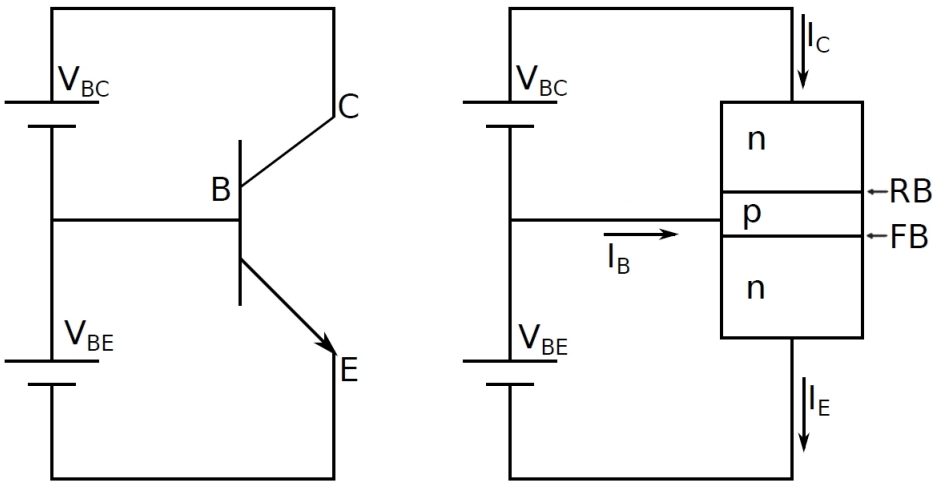


Figure 10.2: NPN BJT biased in the forward active mode of operation. On the left is the circuit symbol and on the right is an illustration of the BJT structure with external forward bias connections.

we obtain

$$\alpha I_E = I_C \,. \tag{10.7}$$

We can also write I_E in terms of I_B and β:

$$\beta I_B + I_B = I_E \tag{10.8}$$

$$\Rightarrow (\beta + 1) I_B = I_E \tag{10.9}$$

Typically,

$$\boxed{0.99 < \alpha < 1.00} \tag{10.10}$$

The basic IV characteristics under normal operation are shown in Figures 10.3 and 10.4. In forward active mode, we often think of the BJT current I_C as being independent of V_{CE} because these current lines are approximately flat. This, however, is an idealized situation and in reality, there is a slight increase in the current as a function of V_{CE} in the forward action region. Increasing either the base current I_B or the base-emitter bias voltage V_{BE} will increase the current through the collector.

10.3 Physical BJT Operation in Forward Active

Forward active is the standard mode of operation for BJTs when they are used in analog electronics applications (amplifiers etc.). For forward-active operation, the

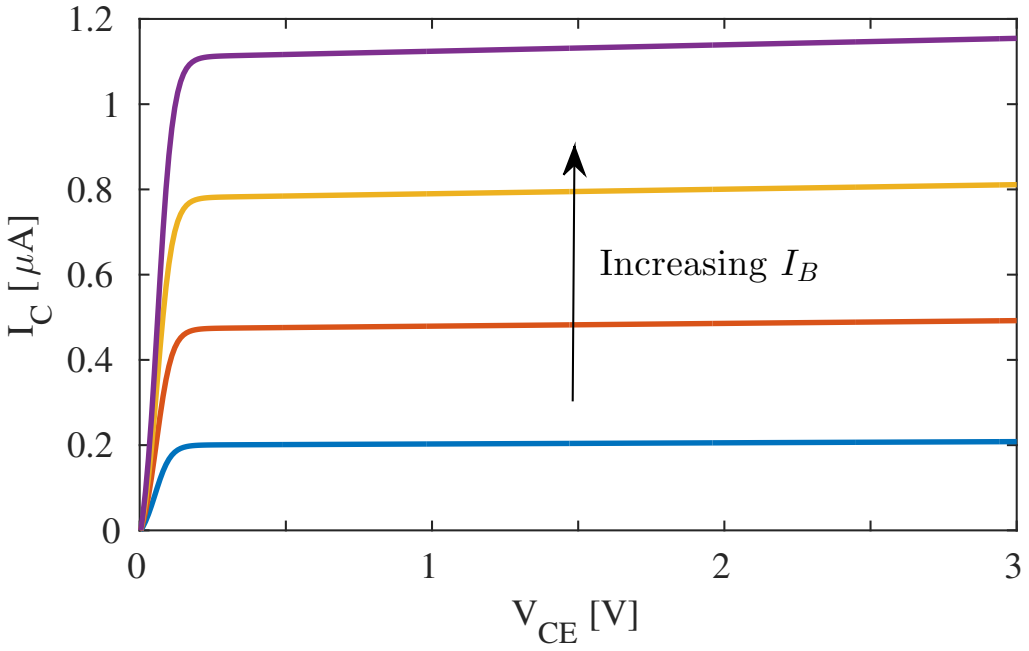


Figure 10.3: BJT IV curves with linearly increasing Base Current I_B from 5nA to 20nA

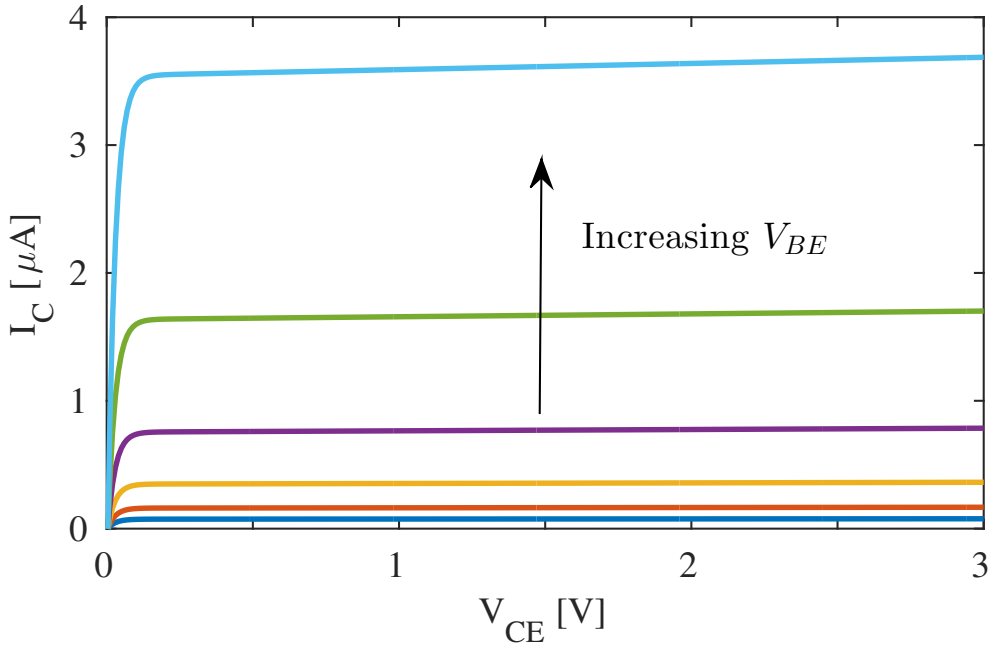


Figure 10.4: BJT IV curves with linearly increasing Base Voltage V_{BE} from 0.4V to 0.5V

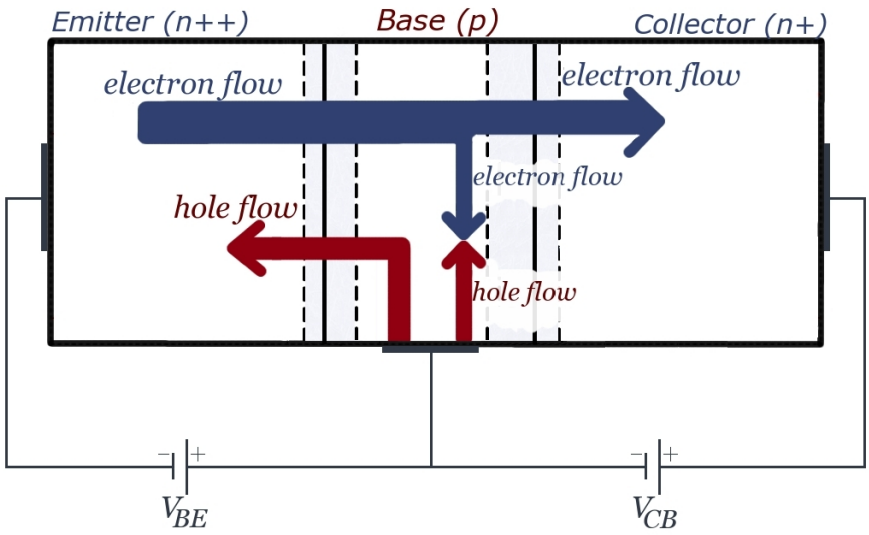


Figure 10.5: Flow of electrons and holes in Forward Biased NPN BJT.

emitter-base junction is forward-biased and the base-collector junction is reverse-biased, as shown in Figure 10.2 for an NPN transistor. To understand BJT operation and how current gain β comes about it is important to keep in mind the following two physical characteristics of BJT structure:

1. **Emitter and Base Doping:** The magnitude of donor doping in the N-type emitter is much larger than the magnitude of the acceptor doping in the P-type base. (This refers to NPN configuration.) (In general doping in emitter is much much larger than doping in base.)

2. **Base Width:** The base is very narrow so that the recombination length for electrons L_n is much larger than the base width or ($W_B << L_n$).

10.3.1 Electron and Hole Flow and Currents for NPN BJT

In an NPN BJT, under forward bias conditions, electrons enter the base from the emitter, and then flow across the base due to diffusion. However, before most of these electrons have the time to recombine in the base, they reach the end of the base and are then pulled into the collector by the large electric field, from the reverse biased collector base junction, that points from the collector into the base. In this way electrons flow from the emitter to the collector and give rise to collector current.

Since the base-emitter junction is forward biased, we also have holes diffusing from the base into the emitter. This gives rise to one of the components of base current. However, since the emitter is much more highly doped than the base, the

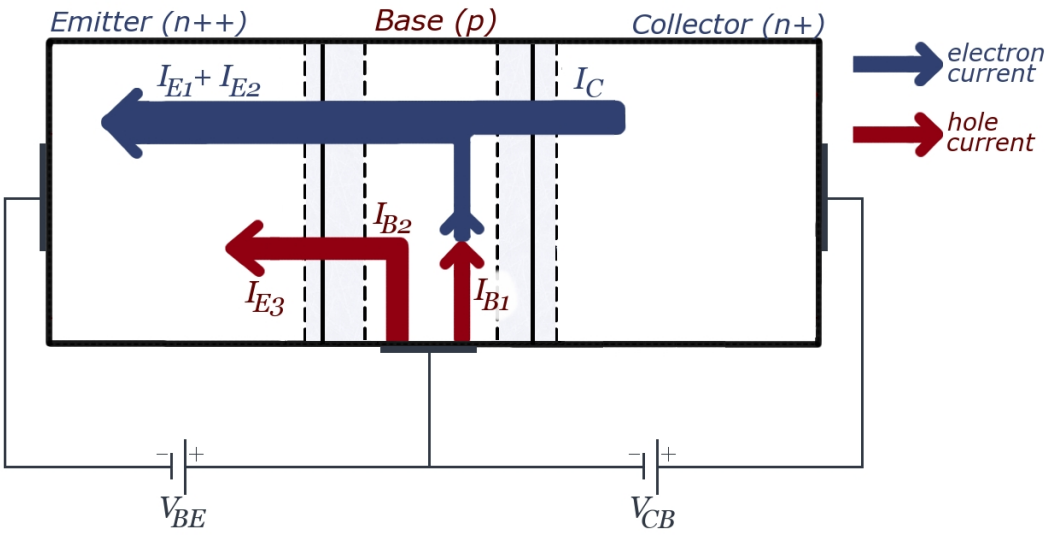


Figure 10.6: Current components in the NPN BJT.

flow of holes into the emitter is much much less than the flow of electrons into the base. Thus, the magnitude of the current from the base into the emitter is much less than the total emitter current and thus also much less than the collector current. We describe this again in a little more detail below.

The flow of electrons and holes, and the resulting electron and hole current components are illustrated in Figures 10.5 and 10.6, respectively.

Emitter Current: I_E has three components, which are also the sum of the collector and base currents.

$$I_E = I_{E1} + I_{E2} + I_{E3} \tag{10.11}$$

Which can also be written as:

$$I_E = I_C + I_{B1} + I_{B2} \tag{10.12}$$

- $I_{E1} = I_C$ which is composed of electrons flowing from emitter into base and then to collector.
- I_{E2} is composed of electrons flowing into base that then recombine with holes in base, $(I_{E2} = I_{B1})$.
- I_{E3} is composed of holes flowing from base to emitter due to forward biased emitter-base junction $(I_{E3} = I_{B2})$.

Base Currents I_{B1} and I_{B2}:

Base Current I_B has two components:

$$I_B = I_{B1} + I_{B2} \tag{10.13}$$

- I_{B1} is due to holes in the base that recombine with electrons that have diffused into the base from the emitter.

- I_{B2} is due to holes that diffuse from the base into the emitter across the forward biased base emitter junction.

Since the emitter is much more highly doped than the base, most of the current across the base-emitter junction is due to the electrons flowing from the emitter to the base, not from the holes flowing from the base to the emitter. Furthermore, since the base is so short, very few electrons and holes recombine there. In summary, since the base is doped at a very low level compared to the emitter, and there is very little recombination in the base, the total base current is very small compared to the total emitter current. So the current ratio I_E/I_B is very large ($I_E/I_B = (\beta + 1)$).

Collector Current: I_C

The collector current is composed of electrons that have flowed from the N-type emitter, through the base and then into the collector.

- $I_C = I_{E1}$

Since the Base-Emitter junction is forward biased, electrons diffuse from the emitter into the base due to standard PN junction operation. However, since the base is very narrow, most of the electrons do not recombine with holes in the P-type base, but instead diffuse across the base toward the collector. They are then swept into the collector by the large electric field in the reverse-biased base-collector junction. Once these electrons are in the collector, they will eventually exit the collector wire contact and be manifested as the collector current.

10.4 Derivation of BJT Current - Voltage Relationships

To derive the Current - Voltage (I-V) relationships for a BJT, we will utilize an approach that is very similar to the one we followed for a PN junction. This makes sense since a BJT is essentially composed of two PN junctions. This approach uses the semiconductor equations, especially the current and continuity equations, to determine the carrier concentrations. We then use the carrier concentrations to obtain the current. We will start by deriving the expression for the collector current as a function of the base-emitter voltage.

10.4.1 Collector Current

The collector current I_C is composed of electrons that diffuse from the emitter, across the forward biased base-emitter junction, which then continue to diffuse across the base region, and are then pulled into the collector by the strong electric field across the reverse biased base-collector junction. To quantify this current, we will start by solving the current and continuity equations for electrons (minority carriers) in the P-type base.

$$\frac{\partial n_p}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} + G_n - R_n \tag{10.14}$$

$$J_n = -qn_p\mu_n\frac{\partial \phi}{\partial x} + qD_n\frac{\partial n}{\partial x} \tag{10.15}$$

where n_p is the minority carrier concentration in the base.

We can now apply the conditions in the base to simplify the above continuity and current equations. First of all, we will assume steady state or DC operation so $\frac{\partial n_p}{\partial t} = 0$. Also, since the base is very thin, we will neglect the generation/recombination term so $G_n - R_n = 0$. Finally, since the P-type base is quasi-neutral, the electric field is very small so we will neglect the drift component of the current. The current and continuity equations are thus greatly simplified to the following:

$$0 = \frac{1}{q}\frac{\partial J_n}{\partial x} \tag{10.16}$$

$$J_n = qD_n\frac{\partial n_p}{\partial x} \tag{10.17}$$

Recall from our derivation of the diode current that the electron concentration can be rewritten as an excess component plus an equilibrium component (Equation 9.17).

$$n_p(x) = \Delta n_p(x) + n_{po} \tag{10.18}$$

The concentration gradient in the electron current equation now only operates on the position-dependent excess component $\Delta n_p(x)$ because the equilibrium component has a derivative of 0:

$$J_n = qD_n\frac{\partial \Delta n_p}{\partial x} \tag{10.19}$$

We now have two equations and two unknowns (J_n and Δn_p).

Minority Carrier Concentration n_p in Base: Now substituting for J_n gives the following equation for n_p only:

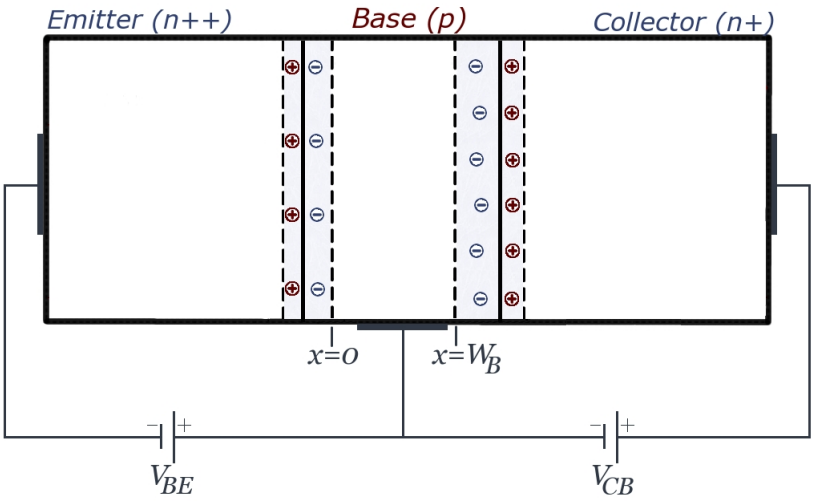


Figure 10.7: Cross-Section of BJT showing coordinates at boundaries of base.

$$\frac{d^2\Delta n_p(x)}{dx^2} = 0 \tag{10.20}$$

Integrating two times gives:

$$\frac{d\Delta n_p(x)}{dx} = A \tag{10.21}$$

$$\Delta n_p(x) = Ax + B \tag{10.22}$$

Boundary Conditions: The solution to the continuity equation indicates that the electron concentration distribution in the base varies linearly as they diffuse across the base region. To obtain the values of the constants A and B we need to apply boundary conditions in a way that is similar to what we did for the PN junction. First of all, we set up a coordinate system as shown in Figure 10.7. Let $x = 0$ at the edge base-emitter depletion region, and we let $x = W_B$ at the edge of the base-collector depletion region.

BC 1: at $x = 0$: At the base-edge of the emitter-base junction, the injected electron concentration depends exponentially on the base-emitter bias, and can be calculated with the same method as was used while deriving currents for a forward-biased PN-junction transistor (see equation (9.25) from PN junction chapter). Thus, the boundary condition at $x = 0$ is:

$$n_p(0) = n_{po}\exp(\frac{V_{BE}}{V_T}) \tag{10.23}$$

$$\Delta n_p(0) = n_{po}\left(\exp(\frac{V_{BE}}{V_T}) - 1\right) \tag{10.24}$$

Substituting for this value for $x = 0$ in equation (10.22) gives us the value of the constant B:

$$\Delta n_p(0) = A \cdot 0 + B = B = n_{po}\left(\exp(\frac{V_{BE}}{V_T}) - 1\right) \tag{10.25}$$

BC 2: at $x = W_B$: On the collector side, the electrons which have moved to the junction edge are immediately swept away by the depletion region electric field, and therefore the excess minority concentration at this edge may be taken as zero or:

$$\Delta n_p(W_B) = 0 \tag{10.26}$$

Substituting this value into equation (10.22), and the value of B obtained above, gives the following for the constant A:

$$\Delta n_p(W_B) = 0 = AW_B + B \tag{10.27}$$

Or

$$A = \frac{-B}{W_B} = \frac{-\Delta n_p(0)}{W_B} = -\frac{n_{po}}{W_B}\exp(\frac{V_{BE}}{V_T}) \tag{10.28}$$

Substituting for A and B into equation (10.22) gives the following expression for the excess electron concentration in the base in forward active mode:

$$\boxed{\Delta n_p(x) = n_{po}\left(\exp(\frac{V_{BE}}{V_T}) - 1\right)(1 - \frac{x}{W_B})} \tag{10.29}$$

Therefore the excess electron concentration in the base decays linearly from the emitter junction edge to the collector junction edge - valid when the base is thin compared to the diffusion length. Figure 10.8 shows the calculated concentration of electrons in the base during forward active. The figure also shows the hole concentrations in the emitter and collector for forward active.

Current:

Now that we have the minority carrier concentration in the base as a function of position, we can substitute it into the current equation to get the electron current density in the base. So, substituting $\Delta n_p(x)$ from equation (10.29) into equation (10.19), and then taking the derivative as indicated, we get the following expression for the electron current density in the base, which is equal to the current density of electrons that flows into the collector.

$$J_n = -\frac{qD_n}{W_B}n_{po}\left(\exp(\frac{V_{BE}}{V_T}) - 1\right) \tag{10.30}$$

The collector current can then be found by using this current density. Assuming a cross-sectional area of A_c for the transistor and using $n_{po} = n_i^2/N_A$, we obtain the following expression for the **Collector Current**:

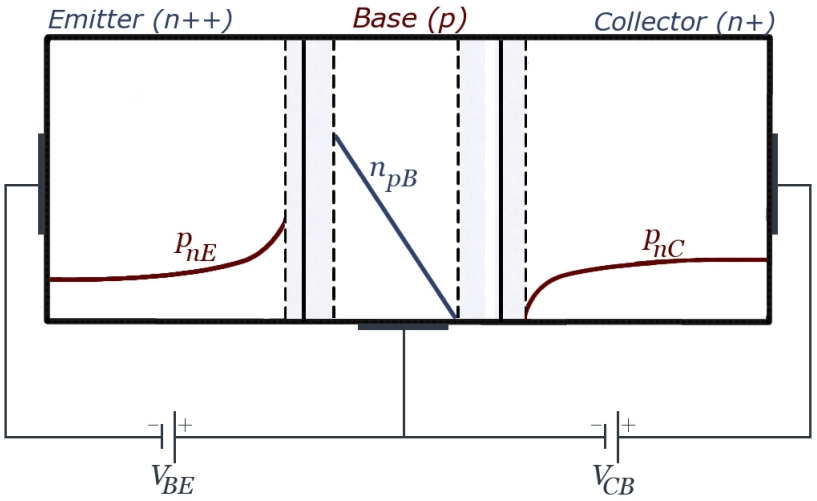


Figure 10.8: Cross-Section of BJT showing minority carrier concentrations during forward active operation.

$$I_C = \frac{qA_cD_nn_i^2}{W_BN_A}\left(\exp(\frac{V_{BE}}{V_T}) - 1\right) \tag{10.31}$$

We will assume the BJT is operating in forward active mode, meaning $\exp(\frac{V_{BE}}{V_T}) \gg 1$ so a more tractable form of the equation is obtained:

$$\boxed{I_C = \frac{qA_cD_nn_i^2}{W_BN_A}\exp(\frac{V_{BE}}{V_T})} \tag{10.32}$$

Defining the transistor saturation current I_{C0} as a constant dependent on the transistor structural elements, such that $I_{C0} = qA_cD_nn_{po}/W_B$,

$$\boxed{I_C = I_{C0}\exp(\frac{V_{BE}}{V_T})} \tag{10.33}$$

Example 10.1:
A silicon BJT has the following doping profile and transport parameters: Emitter Doping is $N_D = 10^{19} cm^{-3}$; Base doping is $N_A = 10^{17} cm^{-3}$; Collector Doping is $N_D = 10^{16} cm^{-3}$; Base Width = $1.0 \mu m$; Electron lifetime in base is $\tau_b = 10^{-6} s$; Hole lifetime in emitter is $\tau_p = 10^{-7} s$; Electron diffusivity $D_n = 20 cm^2/s$; Hole diffusivity $D_p = 10 cm^2/s$ and BJT cross-sectional area $100\mu m \times 100\mu m$.
If the base-emitter voltage $V_{BE} = 0.6V$, and the base-collector junction is reverse biased, calculate the following:

1. The electron concentration in the base as a function of position.

Recall, from the solution of the current and continuity equations we know that the excess electron concentration drops approximately linearly across the base:

$$\Delta n_{pB}(x) = n_{po}\left(\exp(\frac{V_{BE}}{V_T}) - 1\right)\left(1 - \frac{x}{W_B}\right)$$

And the total electron concentration becomes:

$$\begin{aligned} n_{pB}(x) &= \frac{(10^{10})^2}{(10^{17})}\left[\left(\exp(\frac{0.6}{0.026}) - 1\right)\left(1 - \frac{x}{1um}\right) + 1\right] \\ n_{pB}(x) &= 1.05 \times 10^{13}\left(1 - \frac{x}{1um}\right) \end{aligned}$$

2. The total number of electrons in the base.

Integrating $n_{pB}(x)$ from 0 to W_B and multiplying by the BJT area A:

$$\begin{aligned} A \cdot \int_0^{W_B} n_{pB}(x)dx &= A \cdot \int_0^{W_B} \frac{n_i^2}{N_A}\left[\left(\exp(\frac{V_{BE}}{V_T}) - 1\right)\left(1 - \frac{x}{W_B}\right) + 1\right] dx \\ &= A \cdot 5.26 \times 10^{12} cm^{-3} \cdot W_B = 5.26 \times 10^4 electrons \end{aligned}$$

3. The collector current.

$$\begin{aligned} I_C &= \frac{qA_c D_n n_i^2}{W_B N_A}\left(\exp(\frac{V_{BE}}{V_T}) - 1\right) \\ I_C &= \frac{1.6 \times 10^{-19} C \cdot (100um)^2 \cdot 20 cm^2/s \cdot (10^{10} cm^{-3})^2}{1um \cdot (10^{17} cm^{-3})}\left(\exp(\frac{0.6}{0.026}) - 1\right) \\ I_C &= 3.4 \times 10^{-5} A \end{aligned}$$

10.4.2 Base Currents

The base current I_B has two components: $I_B = I_{B1} + I_{B2}$.

Base Current Component I_{B1}

As discussed previously, I_{B1} is due to electron/hole recombination in the base, and depends on the amount of minority carriers in this region. To get the total amount of current due to recombination of electrons and holes in the narrow base we again start with the continuity equation.

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} + G_p - R_p \tag{10.34}$$

As we did previously, we look only for steady state solutions, so $\frac{\partial p}{\partial t} = 0$. Just as we did for the PN junction, we approximate the generation/recombination term as follows:

$$G_p - R_p \approx -\frac{p(x) - p_o}{\tau_b} = -\frac{\Delta p(x)}{\tau_b} \tag{10.35}$$

Where τ_b is the carrier recombination lifetime in the base, and p is the hole concentration in the base during forward active operation, and p_o is the equilibrium hole concentration. The continuity equation for holes in the base then becomes:

$$0 = -\frac{1}{q}\frac{\partial J_p}{\partial x} - \frac{\Delta p(x)}{\tau_b} \tag{10.36}$$

Now, let's separate variables and integrate over the base to obtain the total current due to recombination:

$$\int_0^{W_B} dJ_p = -q\int_0^{W_B} \frac{\Delta p(x)}{\tau_b} dx \tag{10.37}$$

Evaluating the left hand side says that the total current density due to recombination in the base is:

$$J_p = -q\int_0^{W_B} \frac{\Delta p(x)}{\tau_b} dx \tag{10.38}$$

Observing equation (10.38) we have one equation and two unknowns (J_p and p). This difficulty can be overcome by realizing that the recombination rate for holes in the base must be the same as the recombination rate for electrons in the base since electrons recombine with holes. Furthermore, we can readily calculate the recombination rate for electrons because we have already calculated the electron concentration in the previous section as given by equation (10.29). In other words:

$$\frac{\Delta p(x)}{\tau_b} = \frac{\Delta n_p(x)}{\tau_b} = \frac{n_{po}\left(\exp(\frac{V_{BE}}{V_T}) - 1\right)(1 - \frac{x}{W_B})}{\tau_b} \tag{10.39}$$

Again, because the BJT is operating in forward active, we can neglect the '-1' term in equation 10.39. Now, substituting the right hand side of the above equation into (10.38) gives:

$$J_p = -q \int_0^{W_B} \frac{n_{po} \exp(\frac{V_{BE}}{V_T})(1 - \frac{x}{W_B})}{\tau_b} dx \tag{10.40}$$

Performing the integration and multiplying by the cross-sectional area of the base A_c gives the following expression for the base current component due to recombination of electrons, injected from the emitter, with holes in the base:

$$\boxed{I_{B1} = \frac{1}{2} \frac{q A_c n_{po} W_B}{\tau_b} \exp(\frac{V_{BE}}{V_T})} \tag{10.41}$$

where τ_b is the recombination rate in the base.

Example 10.2:
For the BJT and bias conditions in Example 10.1, calculate the following:

1) The electron current from the emitter that is due to recombination in the base.

$$\begin{aligned} I_{E2} = I_{B1} &= \frac{1}{2}\frac{qA_c n_{po} W_B}{\tau_b}\exp(\frac{V_{BE}}{V_T}) \\ I_{B1} &= \frac{1.6\times10^{-19}C\cdot(100um)^2\cdot(10^{10}cm^{-3})^2\cdot 1um}{2\cdot(10^{17}cm^{-3})\cdot(10^{-6}s)}\exp(\frac{0.6}{0.026}) \\ I_{B1} &= 8.4\times10^{-9}A \end{aligned}$$

2) The emitter current component, and collector current that is due to electrons from the emitter that then diffuse through the narrow base and get swept into the collector by the reverse biased base-collector junction.

$$\begin{aligned} I_{E1} = I_C &= \frac{qA_c D_n n_i^2}{W_B N_A}\exp(\frac{V_{BE}}{V_T}) \\ I_C &= \frac{1.6\times10^{-19}C\cdot(100um)^2\cdot 20cm^2/s\cdot(10^{10}cm^{-3})^2}{(1um)\cdot(10^{17}cm^{-3})}\exp(\frac{0.6}{0.026}) \\ I_C &= 3.4\times10^{-5}A \end{aligned}$$

3) The hole current that enters the base electrode due to recombination with electrons.

An equal amount of hole current enters the base from the external connection as there is electron current entering from the emitter because these currents are due to recombination of electrons with holes.

$$I_{B1} = 8.4\times10^{-9}A$$

4) The ratio components of the emitter current that are due to electrons from the emitter that enter the base to the component of hole current in the base that is due to recombination with electrons.

$$\frac{I_{E1}+I_{E2}}{I_{B1}} = \frac{I_C + I_{B1}}{I_{B1}} = \frac{3.4\times10^{-5}+8.4\times10^{-9}}{8.4\times10^{-9}} = 4049$$

Base Current Component I_{B2}

I_{B2} is due to holes that diffused from the base into the emitter across the forward biased base-emitter junction. This current is totally analogous to the hole current in a PN junction that diffuses from the P-side into the N-side under forward bias. Recall from the previous chapter that this current was obtain by solving the

current and continuity equations for minority carriers. We will not go through this entire lengthy derivation again, but simply use the results from the PN junction described in the previous chapter. So, as obtained in Chapter 9, the current due to holes at the junction edge is given by equation (9.37):

$$J_p(-x_n) = -\frac{qD_p p_{n0}}{L_p}(e^{V_A/V_T} - 1) \tag{10.42}$$

Now, substituting V_{BE} for the applied voltage, neglecting '-1' term compared to the exponential, and multiplying by the transistor cross-sectional area A_c, we obtain the component of the base current due to holes diffusing from the P-base to the N-emitter:

$$\boxed{I_{B2} = \frac{qA_c D_p p_{n0}}{L_p}\, e^{V_{BE}/V_T}} \tag{10.43}$$

where D_p and L_p are hole diffusion constant and recombination length in the n-type emitter, respectively.

Example 10.3:
For the BJT and bias conditions in Examples 10.1 and 10.2, calculate the following:

1) The base current that is due to hole diffusion from the base into the emitter.

$$\begin{aligned} I_{E2} = I_{B2} &= \frac{qA_c D_p p_{n0}}{L_p}\exp(\frac{V_{BE}}{V_T}) \\ I_{B2} &= \frac{1.6\times10^{-19}C\cdot(100um)^2\cdot 10cm^2/s\cdot(10^{10}cm^{-3})^2}{(10^{19}cm^{-3})\cdot(\sqrt{10cm^2/s\cdot 10^{-7}s})}\exp(\frac{0.6}{0.026}) \\ I_{B2} &= 1.7\times10^{-8}A \end{aligned}$$

2) The ratio of the total emitter current to the component of base current that is due to hole current diffusing from the base into the emitter.

$$\begin{aligned} \frac{I_{E1}+I_{E2}+I_{E3}}{I_{B2}} = \frac{I_C+I_{B1}+I_{B2}}{I_{B2}} &= \frac{3.4\times10^{-5}+8.4\times10^{-9}+1.7\times10^{-8}}{1.7\times10^{-8}} \\ \frac{I_C+I_{B1}+I_{B2}}{I_{B2}} &= 2001 \end{aligned}$$

Total Base Current

The total base current I_B can be found by summing the two components up.

$$\boxed{I_B = I_{B1}+I_{B2} = \left[\frac{1}{2}\frac{qAW_B}{\tau_b}\frac{n_i^2}{N_A} + \frac{qAD_p}{L_p}\frac{n_i^2}{N_D}\right]\exp(\frac{V_{BE}}{V_T})} \tag{10.44}$$

Where the values for the minority carrier equilibrium concentrations have be written in terms of the doping: $n_{p0} = n_i^2/N_A$ and $p_{n0} = n_i^2/N_D$. For compactness, we will typically write

$$\boxed{I_B = I_{B0} \exp(\frac{V_{BE}}{V_T})} \tag{10.45}$$

where

$$I_{B0} = \left[\frac{1}{2}\frac{qAW_B}{\tau_b}\frac{n_i^2}{N_A} + \frac{qAD_p}{L_p}\frac{n_i^2}{N_D}\right] \tag{10.46}$$

10.4.3 Current Gain β

Since the result for total base current I_B depends directly on $\exp(V_{BE}/V_T)$, just like I_C does, the two may be stated in terms of each other. This leads us to a definition of β, the current gain, given earlier in the chapter as I_C/I_B. Substituting for I_C from equation (10.32) and I_B from equation (10.44), we obtain the following expression for β:

$$\beta = \frac{I_C}{I_B} \tag{10.47}$$

$$\beta = \frac{\frac{qAD_n n_{po}}{W_B} \exp(\frac{V_{BE}}{V_T})}{(\frac{1}{2}\frac{qAn_{po}W_B}{\tau_b} + \frac{qAD_p}{L_p}\frac{n_i^2}{N_D}) \exp(\frac{V_{BE}}{V_T})} \tag{10.48}$$

or

$$\boxed{\beta = \frac{1}{\frac{W_B^2}{2\tau_b D_n} + \frac{D_p}{D_n}\frac{W_B}{L_p}\frac{N_A}{N_D}}} \tag{10.49}$$

Since there is very little base current compared to the electron current that travels from the emitter to the collector, the ratio I_C/I_B is large. Device design decisions to optimize β include keeping the base region very narrow and doping the emitter heavily compared to the base. Remember, N_D in the equation for β is the **emitter** doping.

Example 10.4:
For the BJT and bias conditions in Examples 10.1-10.3, calculate the value of β by taking the ratio of the collector electron current to the total base hole current.

The base current is comprised of two components: I_{B1} which is due to holes recombining with electrons in the base, and I_{B2} which is due to the forward biased emitter-base junction.

$$\begin{aligned} I_{B1} = & \frac{1}{2}\frac{qA_c n_{po} W_B}{\tau_b} \exp(\frac{V_{BE}}{V_T}) = 8.4 \times 10^{-9} A \\ I_{B2} = & \frac{qA_c D_p p_{n0}}{L_p} \exp(\frac{V_{BE}}{V_T}) = 1.7 \times 10^{-8} A \\ I_C = & \frac{qA_c D_n n_i^2}{W_B N_A} \exp(\frac{V_{BE}}{V_T}) = 3.4 \times 10^{-5} A \\ \beta = & \frac{I_C}{I_{B1}+I_{B2}} = \frac{3.4\times10^{-5}A}{8.4\times10^{-9}A+1.7\times10^{-8}A} = 1339 \end{aligned}$$

You should get the same value for beta (within rounding error) by plugging in the appropriate values directly into Equation 10.49.

10.5 BJT I-V Characteristics

We can now describe the current-voltage characteristics of the BJT in terms of I_C vs V_{CE}, with V_{BE} as the parameter. Note that we can rewrite $V_{CE} = V_{CB} + V_{BE}$ We have described so far that

$$I_C = \beta I_B \tag{10.50}$$
$$I_E = (\beta + 1) I_B \tag{10.51}$$
$$I_B = I_{B0} e^{V_{BE}/V_T} \tag{10.52}$$

I_C ideally does not depend on V_{CE}. It can be described as either a voltage or a current dependent current source, which is possible since it is assumed to be not load-dependent. Which is to say, whatever load is attached to the collector, the current profile remains flat.

10.5.1 Early Voltage

As a matter of fact, this idea of a perfectly-horizontal I_C vs V_{CE} relationship, with I_C as an ideal dependent source, is an idealized description. I_C actually has a slight slope with respect to V_{CE}. As the collector bias is raised with respect to the

base, causing the base-collector junction to be deeper in reverse-bias, the depletion region between the base and the collector gets wider. This effectively causes the base width W_B to shrink, which increases the current flow. This phenomenon is called the **Early Effect**.

Recall the ideal description for I_C:

$$I_C = \frac{A_c q D_n n_i^2}{W_B N_A} e^{V_{BE}/V_T} \tag{10.53}$$

Including the effects of the Early voltage V_A:

$$I_C = \frac{A_c q D_n n_i^2}{W_B N_A} e^{V_{BE}/V_T} [1 + \frac{V_{CE}}{V_A}] \tag{10.54}$$

V_A is called the **Early Voltage**. It is a measure of the how non-ideal the BJT is in acting as a current source. V_A is typically very large compared to V_{CE}. Therefore the Early effect does not have a major impact, but it is important nonetheless, as it gives rise to the small-signal parameter r_0, the output resistance.

The small dependence of collector current I_C on V_{CE}, the Early Effect and the Early Voltage are shown in Figure 10.9.

10.6 Small Signal Parameters

When we use BJTs to construct amplifiers, we typically make circuits that amplify small variations of input signals to obtain larger variations of signals at the output. To quantify these variations, we typically define small signal parameters for BJTs. These small signal parameters are the first order terms in Taylor series expansions of the BJT current-voltage relations when currents and/or voltages are varied.

The small (first order) variation in collector current I_C in response to small variations in V_{BE} and V_{CE} are as follows:

$$\Delta I_C = \frac{\partial I_C}{\partial V_{BE}} \Delta V_{BE} + \frac{\partial I_C}{\partial V_{CE}} \Delta V_{CE} \tag{10.55}$$

Now if we define the small signal transconductance as: $g_m = \frac{\partial I_C}{\partial V_{BE}}$,
And the small signal output conductance as: $g_o = \frac{\partial I_C}{\partial V_C}$,
We can rewrite equation (10.55) as follows:

$$\Delta I_C = g_m \Delta V_{BE} + g_o \Delta V_{CE} \tag{10.56}$$

It is often more convenient to express the variation of the collector current that results from its dependence on V_{CE} as an output resistance. So we define the output resistance parameter r_o as the reciprocal of the output conductance:

$$r_o = 1/g_o \tag{10.57}$$

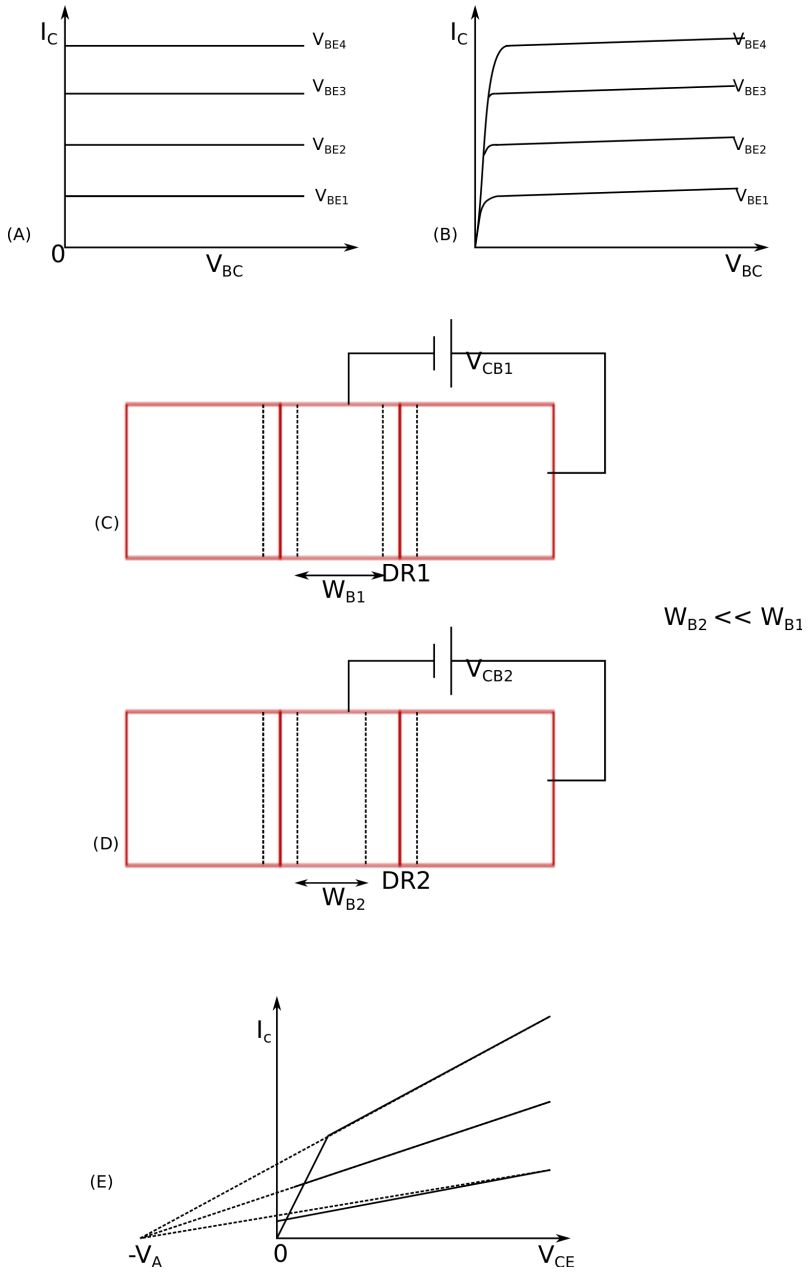


Figure 10.9: A. (Top left) Ideal IV curves: The current value is determined by V_{BE} and remains constant with changing V_{CE}. B. (Top right) Actual IV curves: The current increases with increasing V_{CE}. This is caused by the Early effect, as described in the text. C. The BJT cross-section in the forward active region. W_{B1} is the base region width for a given V_{CE}. D. Higher V_{CE} causes the base-collector junction depletion region to widen and the effective base width to decrease, to $W_{B2} << W_{B1}$. This results in an increase in the collector current with higher V_{CE}, which is the **Early effect**. E: All the IV curves converge to V_A, called the **Early Voltage**.

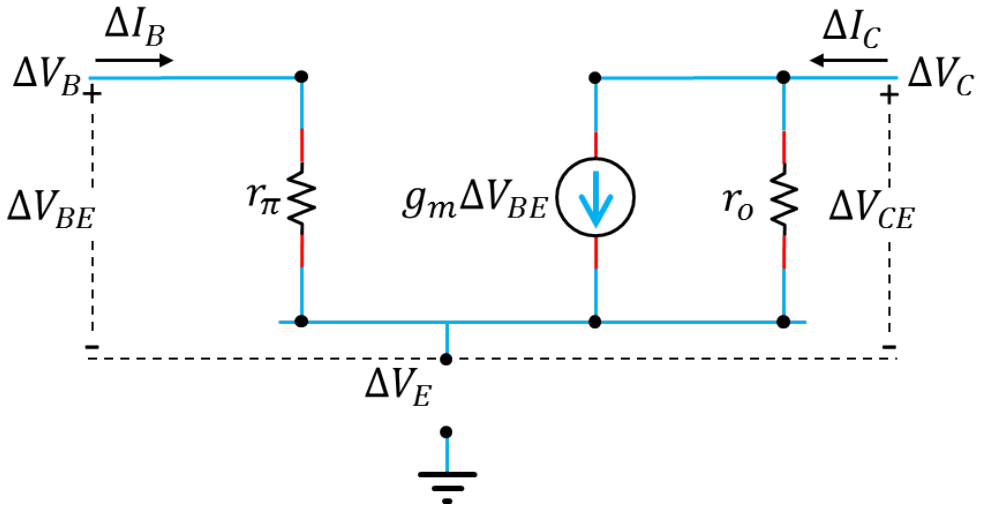


Figure 10.10: Basic small signal model for BJT.

And equation (10.55) can now be written as:

$$\Delta I_C = g_m \Delta V_{BE} + \frac{1}{r_o} \Delta V_{CE} \tag{10.58}$$

We also typically want to calculate the variation in the base current that results from a small variation in the base-emitter voltage. Again applying Taylor series to first order gives:

$$\Delta I_B = \frac{\partial I_B}{\partial V_{BE}} \Delta V_{BE} \tag{10.59}$$

Now we define a small signal input conductance as $g_\pi = \frac{\partial I_B}{\partial V_{BE}}$. And it is often convenient to write the conductance in terms of a resistance such that $\frac{1}{g_\pi} = r_\pi$. We can then write the change in base current due to a small change in base-emitter voltage as:

$$\Delta I_B = \frac{1}{r_\pi} \Delta V_{BE} \tag{10.60}$$

We can now use these relationships to construct the small signal equivalent circuit for a BJT amplifier. Figure 10.10 shows a standard BJT small signal model equivalent circuit obtained using the methods described in this section.

10.7 BJT Capacitances

Associated with the BJT are intrinsic capacitors that play a very large role in determining the operation speed of a BJT. The intrinsic capacitors affect the BJT frequency response and switching speed. The intrinsic capacitors are analogous to

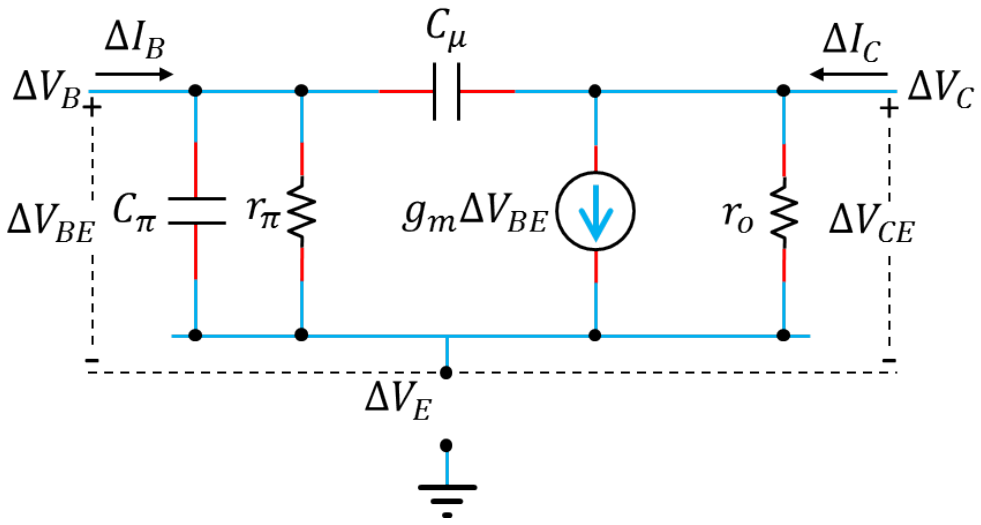


Figure 10.11: Basic small signal model for BJT.

those associated with the PN junctions we studied previously. In the small signal model there are typically two major capacitors: C_π and C_μ, which are associated with Base-Emitter junction and the Base-Collector junction, respectively. The capacitance C_π is composed of the parallel combination of the Base-Emitter junction capacitance and the diffusion capacitance due to minority charge in the base. Expressions for these types of capacitors are analogous to those in equations (9.49) and (9.53), respectively. The capacitor C_μ is due to the depletion region of the reverse biased Base-Collector junction. It is thus a depletion capacitance and has the form of equation (9.49).

10.8 Problems

10.1 Sketch and label the cross-sections of an NPN and a PNP BJT.

10.2 Describe in your own words how a BJT operates in forward active mode. Use words that include drift, diffusion, built-in fields, reverse bias, forward bias, short base, diffusion length, etc.

10.3 A silicon BJT has the structure and parameters below. Calculate β and α; Assume forward active operation.

(a) Base Width = $1.0 \mu m$

(b) Electron lifetime in base is $1 \times 10^{-6} s$

(c) Base doping is $N_A = 10^{17} cm^{-3}$

(d) Emitter Doping is $N_D = 2 \times 10^{19} cm^{-3}$

(e) Collector Doping is $N_D = 10^{18} cm^{-3}$

(f) Electron diffusivity $D_n = 20 cm^2/s$

(g) Hole diffusivity $D_p = 10 cm^2/s$

(h) BJT cross-sectional area $100\mu m \times 100\mu m$

10.4 For the BJT with parameters above, graph the collector current versus the collector-emitter voltage for the voltage range $0 < V_{CE} < 5$ and three value of $V_{BE} = (0.4, 0.5, 0.6)V$.

10.5 For the BJT with the structure given in problem 3, calculate C_π and C_μ if $V_{BE} = 0.6V$ and $V_{CE} = 2V$.

10.6 Describe the key attributes of a BJT design that give rise to a large value of β.

10.7 By solving the relevant semiconductor equations, derive the current gain β, and the current components for the base, emitter and collector for a PNP BJT in forward active mode. (Remember, for PNP, forward active means $V_E > V_B > V_C$.)

10.8 By taking the ratio of the sum of the base current components to the sum of the emitter current components, and using the relationships between the various components of base and emitter currents, show that $I_E/I_B = \beta + 1$.

Chapter 11

MOSFET: Metal Oxide Semiconductor Field Effect Transistor

11.1 Introduction

The MOSFET is the most common transistor of all. It has both analog and digital applications. Almost all computer chips have MOSFETs as their basic building blocks. So MOSFETs can act as digital switches. MOSFETs can also be configured to operate as analog amplifiers. The MOSFET is usually considered to be a three terminal device. The three terminals are the gate, source and drain. (The MOSFET also has a fourth terminal called the body, but the body is usually shorted to the source or grounded, it thus plays a secondary role so we won't worry about it too much in this text.) Although there are many nuances, the general operation is that a voltage applied to the gate terminal controls the current that flows between the drain and the source. One of the key features of the MOSFET is the gate oxide. This is a very thin insulating layer that separates the gate terminal from the rest of the device. Because of this oxide, when you apply a DC voltage to the gate, no gate current will flow. Thus the current between the source and the drain is largely controlled by the electric field that arises due to the gate voltage. This is where the MOSFET gets its name as a 'field-effect transistor'. This is in contrast to a BJT where there is a DC base current and a very specific ratio between the base and collector currents).

In a modern digital CPU chip, there can be as many as a billion MOSFET

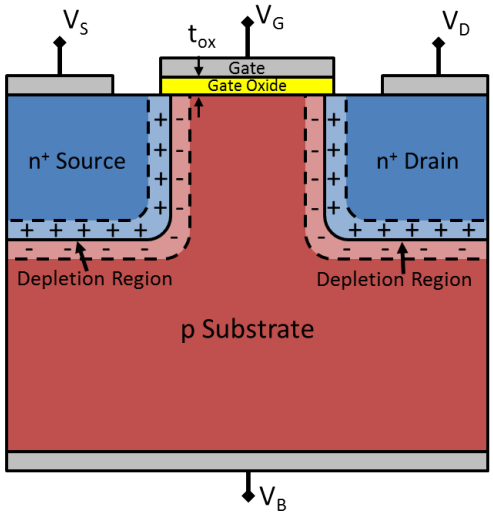


Figure 11.1: Cross Section of N-MOSFET.

transistors. In analog circuits, the number of MOSFETs is much smaller, usually numbering between ten and one thousand. As you can surmise by the number of MOSFETs on a computer chip, MOSFETs can be incredibly small. Modern processing technology has allowed engineers to fabricate MOSFETs with typical sizes of less than a few tens of nanometers.

11.2 MOSFET Structure and Circuit Symbol

Figure 11.1 shows the physical two-dimensional cross-section of an N-Channel MOSFET. The figure shows the N-type Source and Drain regions, the P-type Substrate, the thin Silicon-Dioxide layer and the Gate. The thickness of the oxide layer is only a few nanometers, and is designated as t_{ox}. Figure 11.2 shows a cross-section of a P-Channel MOSFET. As shown in the figure, the doping of the P-Channel MOSFET is opposite that of the N-Channel device. An important thing to notice is that there are PN junctions formed between the source and the substrate and the drain and the substrate. Also, the doping in the source and drain is typically 100 to 1000 times greater than the doping in the substrate. Thus, as shown in the figures, we designate the source and drains with a (+) sign to indicate that they are highly doped.

In Figure 11.3 three dimensional illustrations of N and P channel MOSFETs are shown. The key attribute to take from the 3D figures is the gate width 'W' and the gate length L, which are important parameters for setting the device current.

Their circuit symbols and corresponding physical contacts and orientation are shown in Figure 11.4.

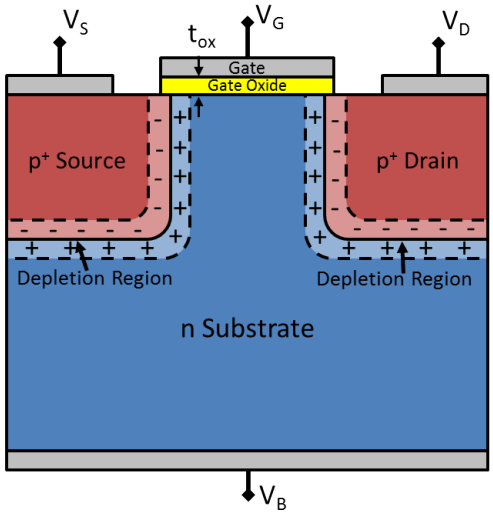


Figure 11.2: Cross Section of P-MOSFET.

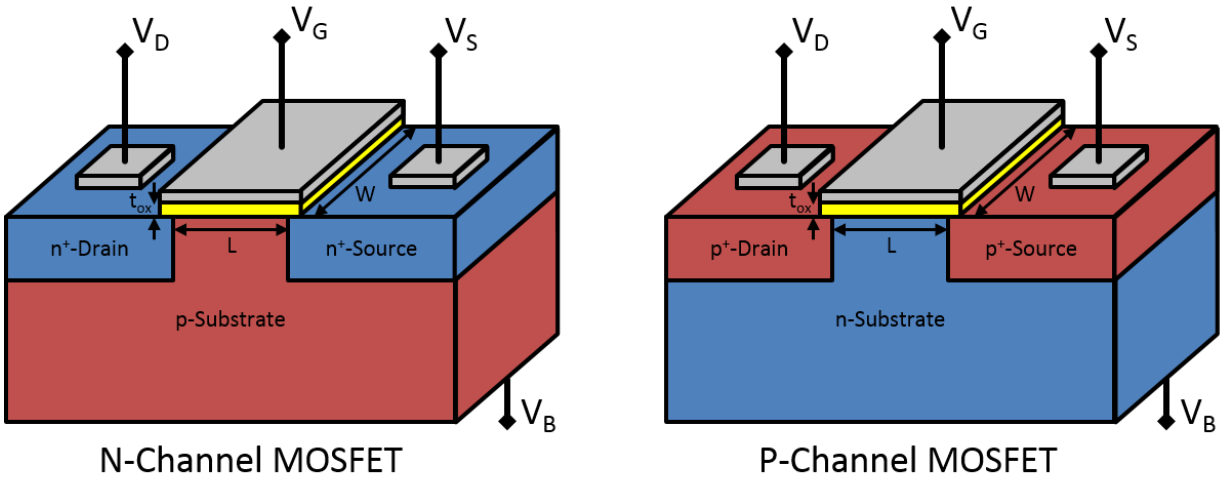


Figure 11.3: 3-Dimensional MOSFET Diagrams

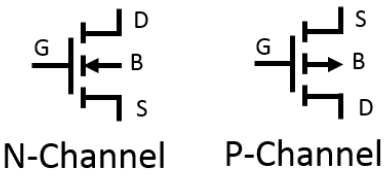


Figure 11.4: N-MOSFET and P-MOSFET Circuit Symbols

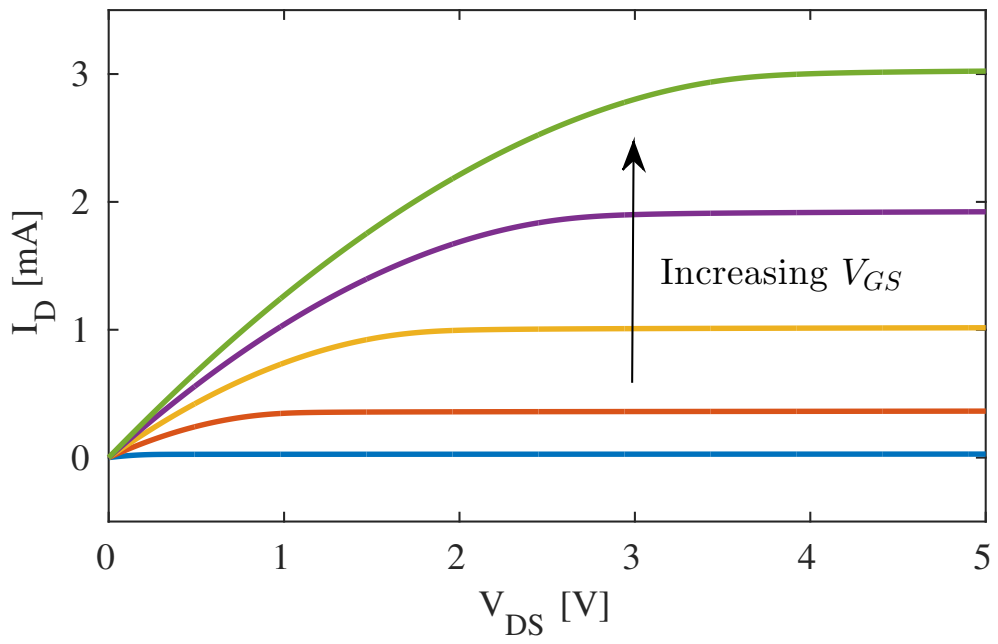


Figure 11.5: MOSFET IV curves with linearly increasing Gate Voltage V_{GS} from 1V to 5V

11.3 Qualitative MOSFET Operation

In this section our discussion will be for N-Channel MOSFETs. A totally analogous description is applicable to P-Channel MOSFETs, but the voltage polarities would be opposite. In an N-Channel MOSFET, the current that flows into the the drain and out of the source is largely controlled by the voltage applied to the gate. Since the current flow between two terminals (the source and drain) is controlled by a voltage applied to a third terminal (the gate), we refer to the MOSFET as a voltage controlled current source. Like a BJT, the MOSFET IV characteristics in Figure 11.5 show that the MOSFET can act as a constant current source under certain bias conditions. Figure 11.6 shows the effect that the gate voltage V_{GS} has in controlling the current as well as the location of the turn-on or **Threshold Voltage** V_{TH} - below which current does not readily flow.

11.3.1 Channel Formation and Threshold Voltage

Now, let's describe this qualitative operation in a little more detail. First of all refer to Figure 11.7. In this figure the NMOS transistor has the source grounded and the body grounded. A positive voltage is then applied to the gate, and another positive voltage is applied to the drain. The positive potential applied between the gate and source (V_{GS}) acts to forward bias the source-substrate (N-P) region near

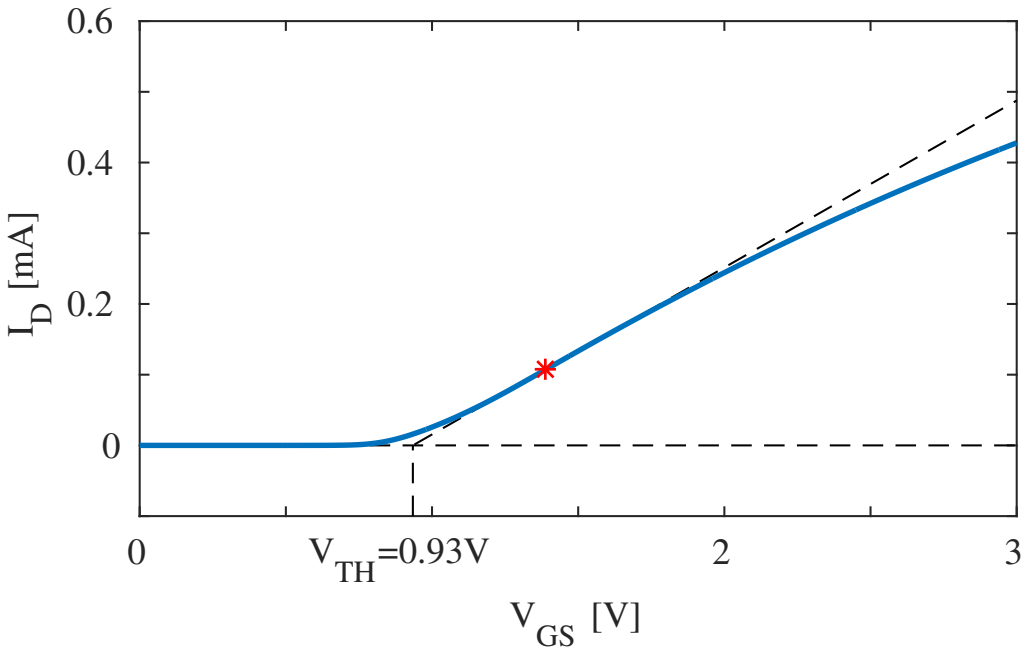


Figure 11.6: MOSFET IV curves with constant Drain Voltage V_{DS} of 0.5V. The red star indicates the inflection point of the curve which is used in the extraction of the threshold voltage V_{TH}

oxide at the top of the device. The electric field arising from the gate-to-source voltage points in the direction that is opposite the built-in field due to the source-substrate PN junction. This reduces the total electric field and allows electrons from the source to diffuse into the substrate region at the top of the source near the gate oxide. Now, these electrons that diffuse out of the source are also pulled up against the insulating oxide by the gate field. However, since the oxide is an insulator, the electrons from the source cannot enter the gate electrode but will gather under the oxide, resulting in a large concentration of mobile electrons in the P-substrate right under the gate oxide. This large concentration of electrons under the oxide is called the channel of mobile electrons, or just the **channel**. The value of the gate-source voltage (V_{GS}) necessary to create a channel is called the **Threshold Voltage** or V_{TH}. Figure 11.8 emphasizes the channel under the oxide by showing the mobile electrons as circled negative charges.

11.3.2 MOSFET Current Flow

Once a channel is formed, current can flow between the source and the drain. If a voltage is now applied to the drain, with the source still grounded, then the field that results from this drain-source voltage (V_{DS}), acts to pull the electrons that are in the channel into the drain and eventually into the drain contact wire. This

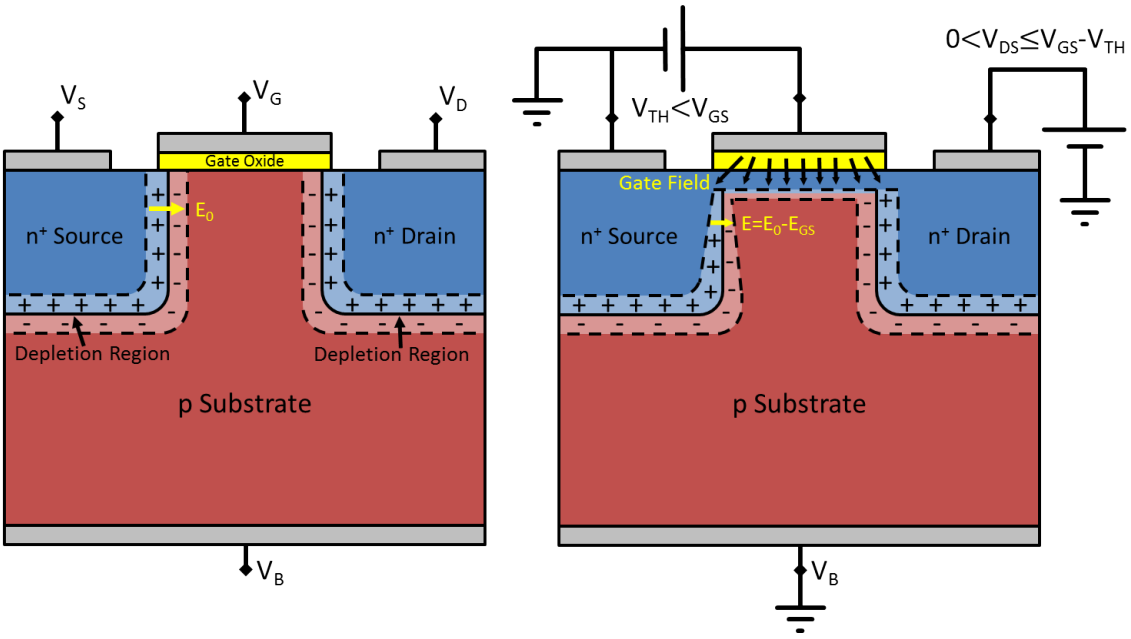


Figure 11.7: Cross Section of N-MOSFET. showing mobile electrons in channel. The picture shows that the gate voltage is greater than the threshold voltage so a channel has been formed. The figure also shows the depletion regions around the source and drain regions

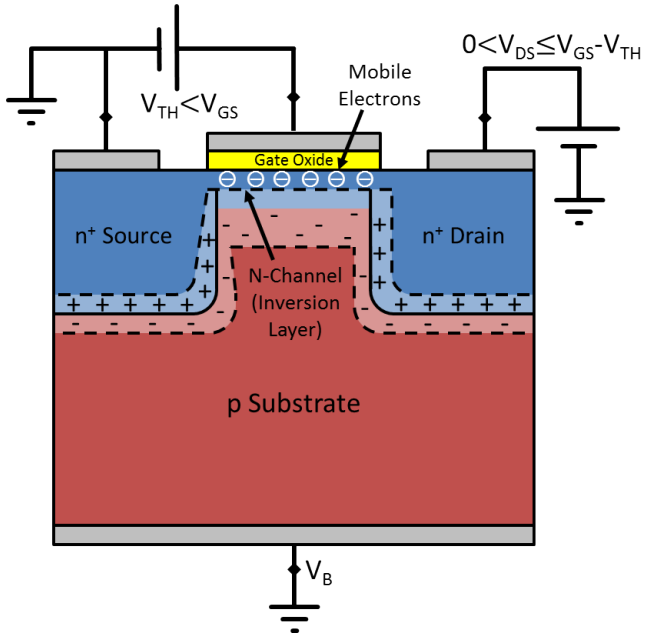


Figure 11.8: Cross Section of N-MOSFET showing the formation of the channel due to the forward bias of the source junction near the surface due to the gate field. Once the gate voltage is greater than the threshold voltage, the channel forms because the gate field pulls electrons (which can now leave the source) up to the gate oxide.

electron flow out of the drain contact then manifests itself as the drain current (I_D), which flows into the drain. In the steady state, these gate and drain voltages are kept fixed, and electrons will continue to flow out of the source, into the channel, and into the drain. This flow manifests itself as an electrical current that flows into the drain and out of the source. See Figure 11.7 for illustrations.

11.3.3 Cutoff

It is important to understand, that if there is no channel formed, there will not be a drain-to-source current in the MOSFET. To understand this consider the following. The voltage on the drain is a polarity that reverse biases the drain-substrate PN junction. Typically, virtually no current can flow in a reverse biased PN junction. So, unless there is a channel that supplies electrons to the drain, no drain current will flow if a drain voltage is applied. When there is no gate-source voltage, or if it is below a certain threshold, then there will be no channel formed, and there will be no drain-source current, even when a voltage is applied to the drain. Under these conditions the MOSFET is said to be in **Cutoff**. Thus, when the gate-source voltage is below the threshold voltage, a channel is not created, and no significant drain current can flow. (If you go on to study devices and electronics more, you will find that this low gate voltage condition is called subthreshold, but we will not consider this in here.) Figure 11.9 illustrates the situation in cutoff. The positive and negative charges are fixed (not mobile so they are not circled) and are due to fixed ionized dopant atoms. There is not a channel of mobile charges so no current will flow.

11.4 Current-Voltage Characteristics and Equations

For an N-Channel MOSFET, when $V_{GS} \geq V_{TH}$ and when $V_{DS} > 0$, drain current I_D will flow. The characteristic curves of I_D *vs.* V_{DS} for several different values of V_{GS} are shown in Figure 11.10. The figure shows that for a given curve (specific value of V_{GS}), the drain current first depends linearly on V_{DS}, then takes on a quadratic character, and then finally I_D becomes relatively constant even as V_{DS} increases.

11.4.1 Linear Region: $(V_{GS} - V_{TH}) \geq V_{DS}$:

We call the region of relatively low V_{DS}, where I_{DS} increases rapidly with increasing V_{DS} the linear or triode region. The relationship between drain current

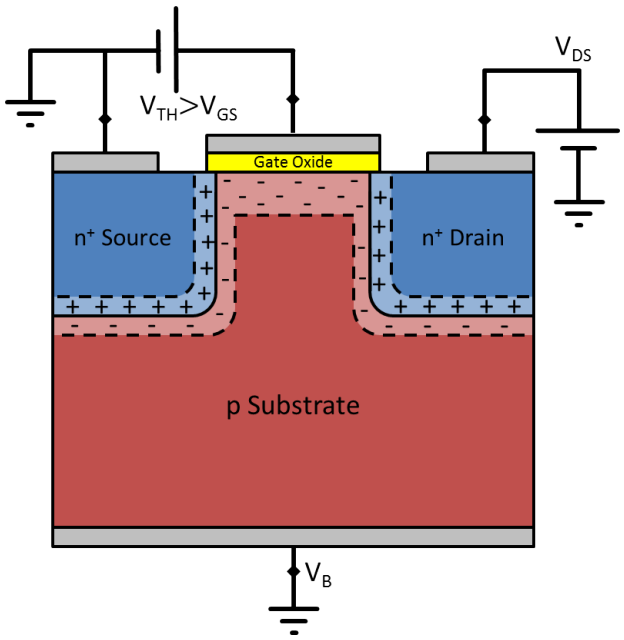


Figure 11.9: Cross-Section of an N-MOSFET in Cutoff region. Note that there is no channel formed because since it is assumed that any applied gate voltage would be less than the threshold voltage. The charges illustrated represent the charged depletion regions that are formed under the gate as around the source-substrate and drain-substrate PN junctions. The charges are fixed because they are from ionized dopants, not from mobile electrons or holes.

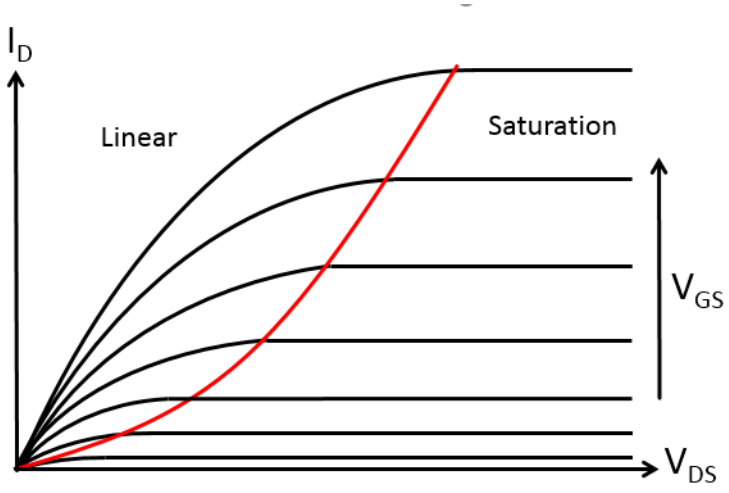


Figure 11.10: Current-Voltage Characteristic of an N-MOSFET. Each curve is a plot of V_{DS} vs. I_D at a constant value of V_{GS}. Each curve is for a different value of V_{GS}

.

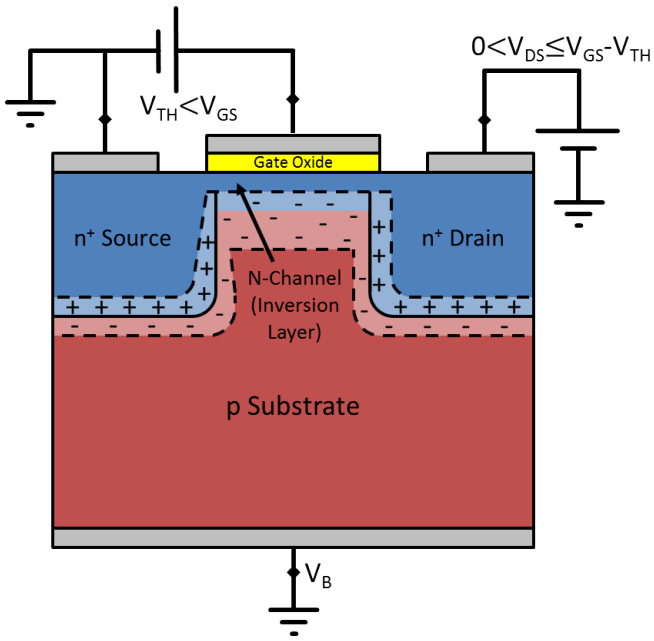


Figure 11.11: Illustration of the electron channel under formed on the gate oxide in the linear region of operation. The channel is also called the inversion layer since the mobile carriers at the top of the P-Substrate at the oxide interface are now electrons

and gate-source voltage and drain-source voltage is given as follows:

$$\boxed{I_D = \mu C_{ox} \frac{W}{L} \left[(V_{GS} - V_{TH}) V_{DS} - \frac{V_{DS}^2}{2} \right]}, \qquad (V_{GS} - V_{TH}) \geq V_{DS} \qquad (11.1)$$

In this region of operation, the quantity of the gate-source voltage minus the threshold voltage, is greater than the drain source voltage, or $(V_{GS} - V_{TH} \geq V_{DS})$.

Under these voltage conditions, the electron channel underneath the gate oxide is fairly uniform between the source and the drain. This is illustrated in Figures 11.11 and 11.12.

In equations (11.1) and (11.2) we have the following parameters and variables:
μ is the electron mobility;
L is the gate length;
W is the gate width;
C_{ox} is the gate oxide capacitance per unit area; $C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$;
ϵ_{ox} and t_{ox} are the gate oxide permittivity and gate oxide thickness;
V_{TH} is the threshold voltage;
V_{GS} is the gate-source voltage;
V_{DS} is the drain-source voltage;

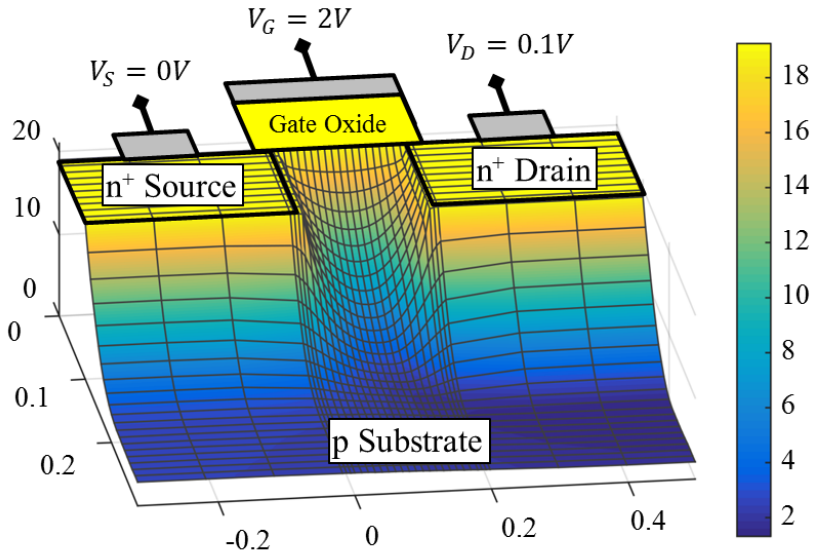


Figure 11.12: Simulated electron concentration within an N-MOSFET biased in Linear region. The colored surface indicates the log_10 of electron density in units of cm^{-3}. The electron concentration at the semiconductor surface (channel) is high and uniform from source to drain.

11.4.2 Saturation Region: $(V_{GS} - V_{TH}) \leq V_{DS}$:

We call the region of relatively high V_{DS}, where I_{DS} stays essentially constant with increasing V_{DS} the saturation region. The relationship between drain current and gate-source voltage and drain-source voltage in the saturation region is given as follows:

$$\boxed{I_D = \frac{\mu C_{ox} W}{2L}(V_{GS} - V_{TH})^2}, \qquad (V_{GS} - V_{TH}) \leq V_{DS} \tag{11.2}$$

In the saturation region of operation the channel is not uniform under the gate oxide. The channel concentration of electrons is much higher at the source side of the channel than at the drain side. In fact at the drain side the electron concentration is sufficiently low that it can be considered to be a depletion-type region, and the channel is said to be 'pinched-off' near the drain. Figures 11.13 and 11.14 illustrates this condition.

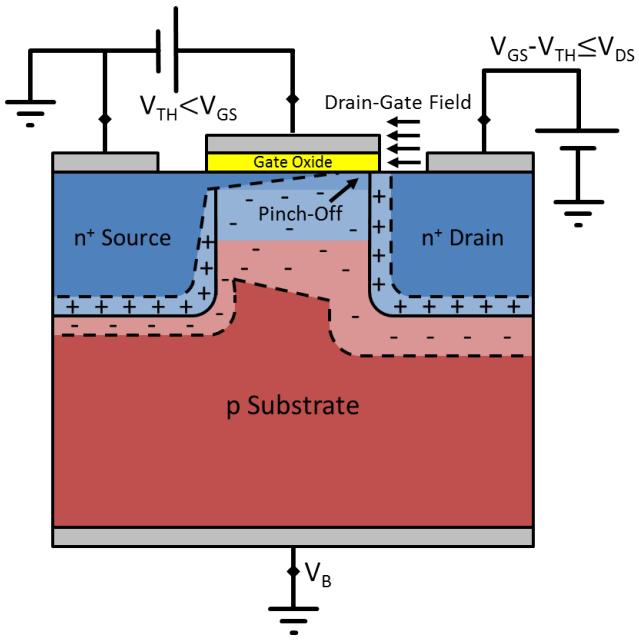


Figure 11.13: Cross Section of N-MOSFET biased in Saturation. The channel has two distinct regions. Near the source the channel has a very high concentration of mobile electrons that are pulled against the oxide. Near the drain, the channel is not so tightly pulled up to the gate oxide but is spread out more vertically and has much lower concentration but is much thicker. The figure actually shows a contour plot of the channel. The darker triangular region has very high mobile electron density while the lighter more rectangular region has low mobile electron density

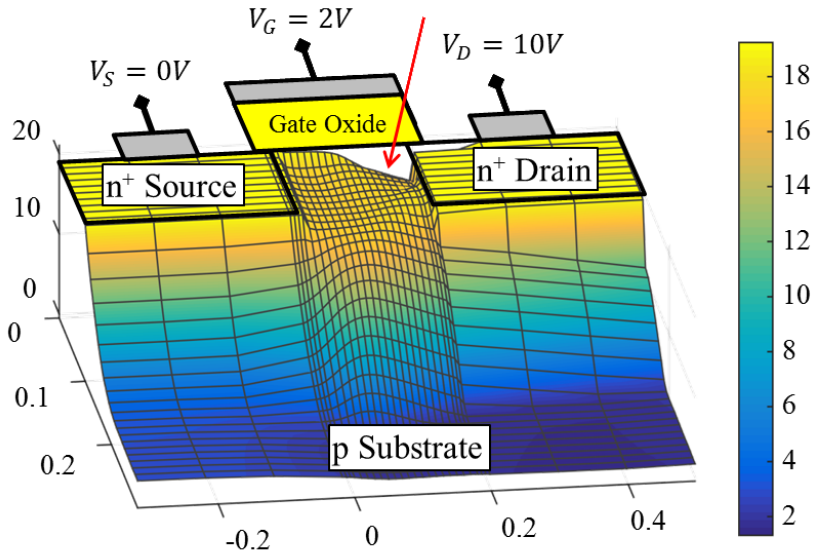


Figure 11.14: Simulated electron concentration within an N-MOSFET biased in Saturation. The colored surface indicates the log_{10} of electron density in units of cm^{-3}. The red arrow indicates the significant drop in electron concentration at the semiconductor surface (channel) near the drain, meaning the device in pinched off

Example 11.5:
A silicon N-MOSFET has channel with width $W = 2\mu m$ and length $L = 0.5\mu m$ with a gate oxide thickness of $t_{ox} = 20nm$ made of SiO_2 ($\epsilon_r = 3.9$).

Calculate C_{ox} (capacitance per unit area) as well as the total capacitance of the gate oxide.

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{3.9\epsilon_0}{20nm} = \frac{3.9 \cdot 8.854 \times 10^{-14} F/cm}{2 \times 10^{-6}} = 1.7 \times 10^{-7} F/cm^2$$

To calculate the total gate capacitance multiply the capacitance per area by the total gate area covering the channel.

$$C_{Gate} = C_{ox} \cdot A = C_{ox} \cdot W \cdot L =$$
$$= (1.7 \times 10^{-7} \frac{F}{cm^2}) \cdot (2 \times 10^{-4} cm) \cdot (5 \times 10^{-5} cm) = 1.7 \times 10^{-15} F$$

Example 11.6:
The same N-MOSFET in Example 11.5 is biased with $V_{GS} = 5V$ and $V_{DS} = 1V$. The threshold voltage is $V_{TH} = 0.8V$. Take the mobility to be $\mu = 300 cm^2/Vs$. This mobility is generally lower than the bulk mobility due to defects at the semiconductor-oxide interface.

What region is this device operation in (Cutoff, Linear, Saturation)? Calculate the drain current at this bias point.

Because $V_{GS} \geq V_{TH} \Rightarrow 5V > 0.8V$ we know the device is not in cutoff and since $(V_{GS} - V_{TH}) \geq V_{DS} \Rightarrow (5V - 0.8V) \geq 1V$ we know that the MOSFET is operating in the linear region. Apply the linear region current formula to calculate the drain current:

$$I_D = \mu C_{ox} \frac{W}{L} \left[(V_{GS} - V_{TH}) V_{DS} - \frac{V_{DS}^2}{2} \right] =$$
$$= \left(300 \frac{cm^2}{Vs} \right) \left(1.7 \times 10^{-7} \frac{F}{cm^2} \right) \left(\frac{2\mu m}{0.5\mu m} \right) \left[(5V - 0.8V) \cdot (1V) - \frac{(1V)^2}{2} \right]$$
$$= 7.5 \times 10^{-4} A$$

Example 11.7:
The same N-MOSFET in Examples 11.5 and 11.6 is now biased with $V_{GS} = 5V$ and $V_{DS} = 10V$.

Confirm that this device is operating in the saturation region and calculate the new drain current at this bias point.

Because $(V_{GS} - V_{TH}) \leq V_{DS} \Rightarrow (5V - 0.8V) \leq 10V$ the MOSFET is indeed operating in the saturation region. Applying the appropriate current formula to calculate the drain current:

$$\begin{aligned} I_D &= \frac{\mu C_{ox} W}{2L}(V_{GS} - V_{TH})^2 = \\ &= \frac{\left(300\frac{cm^2}{Vs}\right)\left(1.7 \times 10^{-7}\frac{F}{cm^2}\right) 2\mu m}{2 * 0.5\mu m}(5V - 0.8V)^2 \\ &= 1.8 \times 10^{-3} A \end{aligned}$$

11.5 Simplified Derivation of Current-Voltage Relations: Applicable for Low V_{DS}

In this section we will derive a simple or zero order version of the MOSFET current-voltage relationship. The derivation gives a good intuition of MOSFET operation, and the result is applicable to very low drain-source voltages. To obtain this relation for current versus voltage at very low V_{DS}, we recall that a MOSFET is similar to a capacitor. The polysilicon gate acts like the top plate of the capacitor, and the channel of mobile electrons acts like the bottom plate. The voltage between the gate and the source V_{GS} mainly controls the concentration of mobile electrons in the channel. The larger V_{GS}, the higher the channel electron concentration. Once the channel is formed, the voltage between the drain and the source V_{DS} acts to pull the channel electrons from the source into the drain, which then gives rise to drain current I_D. To get an expression for the drain current I_D we start with the drift term of the current equation that we have seen in previous chapters:

$$J = q\mu nE, \qquad I = AJ, \qquad I = Aq\mu nE \tag{11.3}$$

Where J is the drift current density and A is the cross-section area that the current density flows through. Now, let's work with drain current and we get:

$$I_D = A_c qn\mu E = -A_c qn\mu \frac{d\phi}{dx} = -Wt_{ch} qn\mu \frac{d\phi}{dx} \tag{11.4}$$

Where I_D is the drain current; $A_c = Wt_{ch}$ is the cross sectional area of the channel where W is the channel width and t_{ch} is the channel thickness; n is the concentration

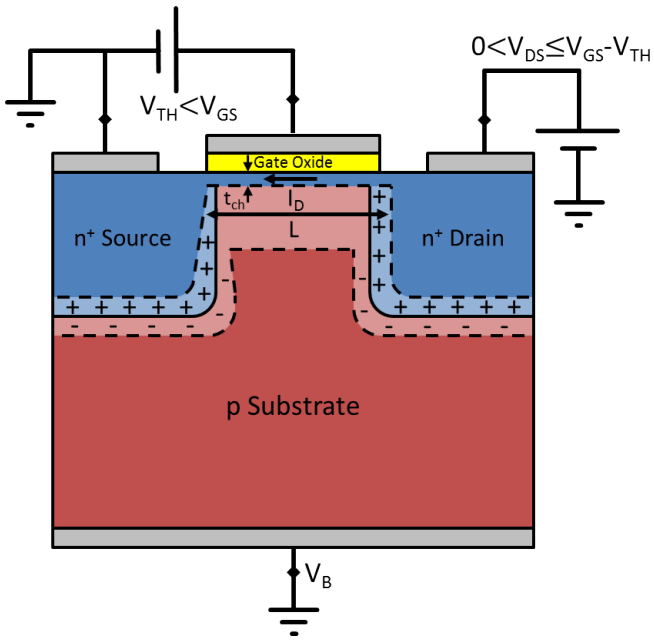


Figure 11.15: Cross Section of N-MOSFET illustrating channel length (L) and thickness (t_{ch})

of mobile electrons in the channel; μ is the electron mobility; ϕ is the electrostatic potential; and $-d\phi/dx$ is the electric field E. (See Figure 11.15). The goal now is to express the above variables in terms of known parameters. First of all, let's approximate the derivative of the potential in terms of the voltage V_{DS} and the length of the gate L, which are both known.

$$\frac{d\phi}{dx} = \frac{V_{DS}}{L} \tag{11.5}$$

Now we need to express the channel electron concentration in terms of known quantities and parameters as well. To do this we recall that the MOSFET is like a capacitor and recall that the charge on a capacitor is $Q = CV$. Since the channel can be thought of as the charge on the lower plate of a capacitor we can say the following:

$$Q_{channel} = Q_{mobile} + Q_{fixed} = C_{ox}V_{GS} \tag{11.6}$$

Where $Q_{channel}$ is the areal charge concentration in the channel, which is composed of the mobile electron charge Q_{mobile}, and the fixed background charge Q_{fixed} that is due to ionized acceptors. (The units of the areal charge is $coulombs/cm^2$.)

However, only the mobile charge contributes to the current, so we subtract the threshold voltage V_{TH} from the gate-source voltage to get Q_{mobile}.

$$Q_{mobile} = C_{ox}(V_{GS} - V_{TH}) = qnt_{ch} \tag{11.7}$$

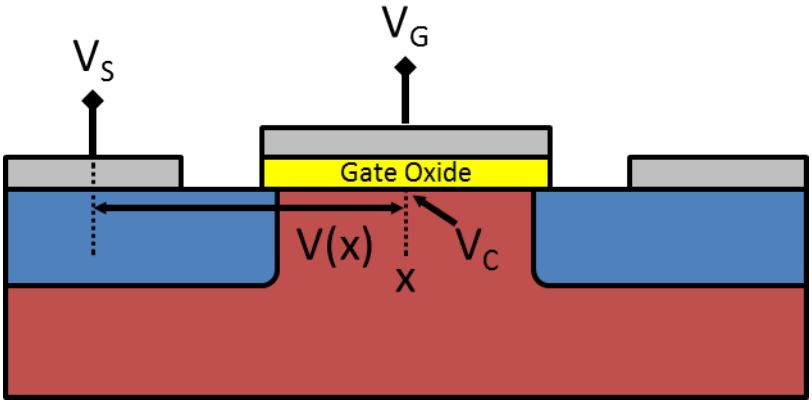


Figure 11.16: Cross section of MOSFET showing the V(x) for Current-Voltage (I_D *vs.* V_{GS}, V_{DS}) derivation.

Note that we have to multiply the three dimensional channel concentration n by the channel thickness t_{ch} to convert to charge per unit area. Now we substitute the above for $\frac{d\phi}{dx}$ and qnt_{ch} into equation (11.4), and we obtain the following expression for drain current:

$$\boxed{I_D = \frac{W\mu C_{ox}}{L}(V_{GS} - V_{TH})V_{DS}}, \qquad V_{DS} << (V_{GS} - V_{TH}) \tag{11.8}$$

Now it is important not to forget that the expression for MOSFET drain current I_D given by equation (11.8) is only only valid for very small values of V_{DS}. It is also worth noting that more complete expression for current in the linear region given by equation (11.1) reduces to equation (11.8) for very small values of V_{DS}. This is because when V_{DS} is much less than one, V_{DS}^2 in equation (11.1) becomes negligible. Finally, we need to mention that the negative sign in equation (11.8) has not been written explicitly because we're focusing on current magnitude and not worrying about direction at this particular time.

11.6 Derivation of Drain Current versus Gate and Drain Voltages for Linear Region: $(V_{GS} - V_{TH} \geq V_{DS})$

Here we will derive equation (11.1) for operation in the linear region. This will follow from the previous derivation, but we will account for the fact that the electron concentration in the channel is not constant, but will be a function of position or $n = n(x)$. The geometry for this derivation is provided in Figure 11.16. Like before, we start with the standard expression for electron drift current density:

$$J = -q\mu n\frac{d\phi}{dx} \tag{11.9}$$

Now we multiply by the channel cross-sectional area, explicitly indicate the position dependence on the carrier concentration $n(x)$, and write the potential or voltage using the letter V instead of ϕ to obtain an expression for drain current in the channel which is similar to (11.4) in the previous section:

$$I_D = A_c J = A_c q\mu n(x)\frac{d\phi}{dx} = W t_{ch} q\mu n(x)\frac{d\phi}{dx} \tag{11.10}$$

Where the parameters have the same meaning as before, except now $n(x)$ is now a function of position. We now express the charge density at the point x in the channel as the product of the oxide capacitance and the voltage across the capacitor a point x:

$$Q(x)_{channel} = Q(x)_{mobile} + Q(x)_{fixed} = C_{ox}(V_G - V(x)) \tag{11.11}$$

Where V(x) is the potential at the point in the channel at the location where the charge is $Q(x)$. Also, recall that C_{ox} is the gate capacitance per unit area and $Q(x)_{channel}$, $Q(x)_{mobile}$ and $Q(x)_{fixed}$ are all areal charge which is typically expressed as $Coulombs/cm^2$

Now, we are only interested in the mobile charge in the channel since that gives to the current, so we subtract the threshold voltage to obtain the mobile charge only:

$$Q(x)_{mobile} = C_{ox}(V_G - V(x) - V_{TH}) \tag{11.12}$$

We now express $Q(x)_{mobile}$ in terms of the mobile charge concentration as follows:

$$qn(x)t_{ch} = Q(x)_{mobile} = C_{ox}(V_G - V(x) - V_{TH}) \tag{11.13}$$

Note that we have multiplied the carrier concentration, which is per unit volume, by the channel thickness to get the charge per unit area at the x coordinate under the gate. Now, we have $n(x)$ in terms of device parameters so let's substitute them for $n(x)$ in equation (11.10).

$$I_D = W\mu[C_{ox}(V_G - V(x) - V_{TH})]\frac{dV}{dx} \tag{11.14}$$

Where we have also replaced the symbol ϕ we use for potential with voltage symbol V, which of course has the same meaning. Now, using KVL, we can re-write the quantity $(V_G - V(x))$ in terms of gate and source voltage as indicated in Figure 11.16.

$$(V_G - V(x)) = V_G - V_S - V(x) + V_S = V_{GS} + V_S - V(x) \tag{11.15}$$

Where we have used the standard notation that the difference between the two potentials can be written as $(V_G - V_S) = V_{GS}$ Substituting for $(V_G - V(x))$ in equation (11.14) and rearranging gives

$$I_D = W\mu C_{ox}[(V_{GS} - V_{TH}) + V_S - V(x)]\frac{dV}{dx} \tag{11.16}$$

Multiplying both sides by dx and integrating from source to drain gives:

$$\int_{x_S}^{x_D} I_D dx = \int_{V(x_S)}^{V(x_D)} W\mu C_{ox}[(V_{GS} - V_{TH}) + V_S - V(x)]dV \tag{11.17}$$

Performing the indicated integration gives:

$$I_D\, x\Big|_{x_S}^{x_D} = W\mu C_{ox}\left[(V_{GS} - V_{TH})V + V_S V - \frac{V^2}{2}\right]\Bigg|_{V_S}^{V_D} \tag{11.18}$$

Substituting in the limits of integration and knowing that the gate length is $L = x_D - x_S$, we obtain the following:

$$I_D\, L = W\mu C_{ox}\left[(V_{GS} - V_{TH})V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{11.19}$$

Note that I_D is factored out of the integral since the total drain current must be constant since we have to satisfy KCL. Now we divide both sides by the gate length L, and we obtain the final expression for drain current versus gate, drain and source voltages that we gave earlier in Section 11.4.1 for MOSFET operation in the linear region:

$$\boxed{I_D = \frac{W\mu C_{ox}}{L}\left[(V_{GS} - V_{TH})V_{DS} - \frac{V_{DS}^2}{2}\right]}, \quad (V_{GS} - V_{TH}) \geq V_{DS} \tag{11.20}$$

Equation (11.20) says that the current going into the drain of an N-Channel MOSFET is determined by the values of V_{DS} and V_{GS} that you apply to the device. Since the drain current I_D depends on the terminal voltages, the device is called a voltage controlled current source. Equation (11.20) is applicable only in the linear (also called the triode) region of operation. Also, it is worth noting that for small values of V_{DS}, the quadratic $V_{DS}^2/2$ term is negligible, and equation (11.20) will reduce to equation (11.8) of the previous section.

11.7 Derivation of Drain Current versus Gate and Drain Voltages for Saturation Region: $(V_{GS} - V_{TH} \leq V_{DS})$

The derivation for current in the saturation region of operation is very simple. If we recall that for values of V_{DS} which are greater than the quantity $(V_{GS} - V_{TH})$

the drain current does not change when we continue to increase V_{DS}. (We will find later that this is an approximation, but we will not worry about that now.) Thus, to find the current in the saturation region we simply equate V_{DS} and $(V_{GS} - V_{TH})$ and substitute this into equation (11.20). In other words we do the following:

$$I_D = \frac{W\mu C_{ox}}{L}\left[(V_{GS} - V_{TH})(V_{GS} - V_{TH}) - \frac{(V_{GS} - V_{TH})^2}{2}\right] \tag{11.21}$$

Doing the algebra gives rise to the expression that we gave earlier in Section 11.4.2 for drain current versus gate-source voltage for N-channel MOSFETs in the saturation region.

$$\boxed{I_D = \frac{W\mu C_{ox}}{2L}(V_{GS} - V_{TH})^2}, \quad (V_{GS} - V_{TH}) \leq V_{DS} \tag{11.22}$$

It is important to observe that as equation (11.22) indicates, current in the saturation region only depends on V_{GS}, and is independent of V_{DS}. Since I_D is independent of V_{DS} the MOSFET in operating in saturation is an excellent voltage controlled current source, and thus we use MOSFETs operating in saturation for designing amplifiers. (As mentioned previously, this result will be slighly modified later to account for channel length modulation, which gives rise to a very small dependence of I_D on V_{DS} in saturation).

11.8 The MOS Capacitor and Threshold Voltage

As a first step to obtaining the expression for MOSFET threshold voltage V_{TH}, we start with the MOS capacitor. The derivation of the expression for the threshold voltage V_{TH} is based on the electrostatic analysis of the MOS capacitor. The MOS capacitor is a MOSFET without the source and drain (Figure 11.17). In other words, it is a three layer device. The top layer is the metal or highly doped polysilcon gate; The middle layer is a very thin insulating oxide (SiO_2); and the third layer is composed of a semiconductor, which is almost always silicon. Figure 11.18 shows the typical MOS capacitor structure. As shown in the figure, the polysilicon gate is so highly doped that is acts like a metal; the oxide is very thin with t_{ox} typically being on the order of nanometers. The semiconductor layer this MOS capacitor is P-type and is typically several microns thick.

11.8.1 Electric Field, Potential and Charge in a MOS Capacitor

In Figure 11.19 we show the charge charge distribution of the MOS capacitor when both the gate and the P-substrate layers are both grounded. We see a very very thin positively charged depletion layer on the N-type polysilicon gate, and a thicker negatively charged depletion layer in the P-type material under the oxide.

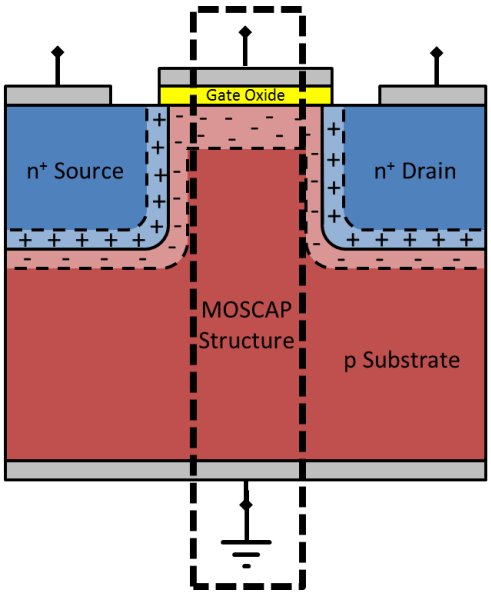


Figure 11.17: The MOS Capacitor within the full MOSFET

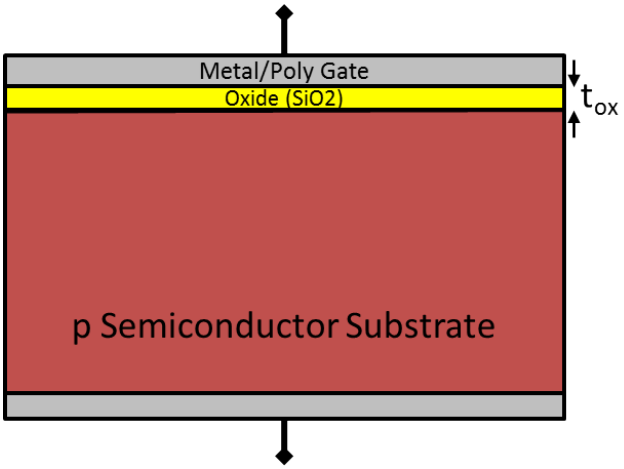


Figure 11.18: Cross section of MOS Capacitor

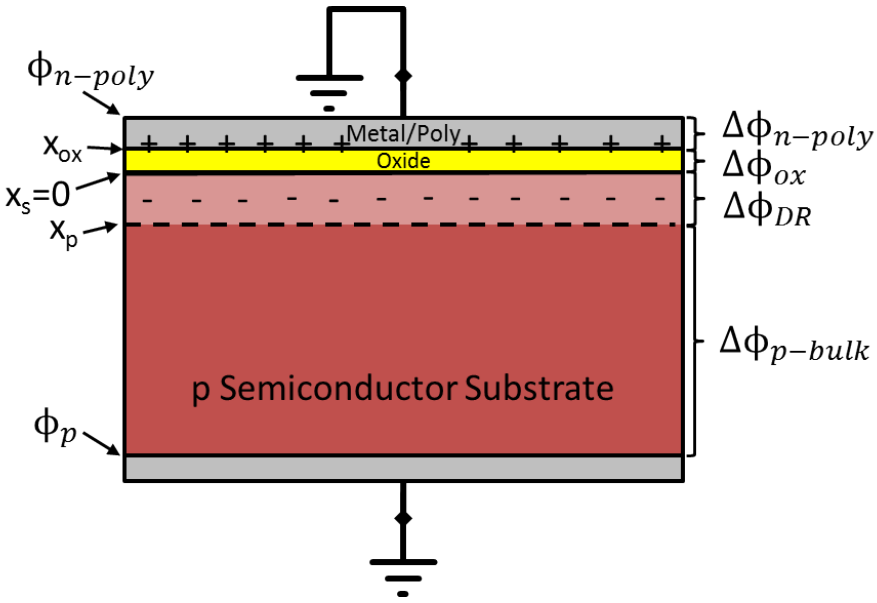


Figure 11.19: MOS Capacitor for derivation of threshold voltage V_{TH}

This electrostatic layer is similar to that in a PN junction, except in the MOS capacitor there is an insulating oxide between the N and P type materials.

Since there is not a direct contact between the N and P layers, it may seem curious as to how this charge re-arrangement occurs. The answer is that the large concentration of mobile electrons in the N-type polysilicon will flow through ground to the P-type substrate region where the concentration of mobile electrons is very very small. In other words electrons will flow from a region of high concentration to a region of low concentration resulting in a displacement of charge. Also, holes will flow from the P-substrate through ground to the N-type polysilicon gate. This will give rise to a very thin positively charged surface layer of ionized donors on the N-type polysilicon and a region of negatively charged ionized acceptor ions in the P-type region under the oxide. In other words, after this rearrangement of charge a depletion region will arise in the P-type silicon under the oxide and in the the N-type polysilicon gate. This is illustrated in Figure 11.19. One key point about this depletion region is that it is virtually entirely in the P-type region under the gate. Since the N-type doping in the polysilicon is several orders of magnitude greater than the P-type doping in the substrate, the depletion region in the N-type polysilicon is negligibly thin. As a result of this transfer of charge, we wind up with a built-in potential across the MOS capacitor, which is analogous to that of a PN junction. The magnitude of this built in potential ϕ_o is given by:

$$\phi_o = \phi_{n-poly} - \phi_p = V_T \ln\left(\frac{N_{Dpoly} N_A}{n_i^2}\right) \tag{11.23}$$

Where the parameters have their usual meanings: ϕ_o is the built-in potential, N_{Dpoly} is the N-type doping concentration of the polysilicon gate, N_A is the P-type doping

concentration in the silicon substrate under the gate oxide, n_i is the silicon intrinsic mobile carrier concentration, and V_T is the thermal voltage $= KT/q$.

Example 11.8:
A silicon N-MOSFET has a substrate doping of $N_A = 10^{17} cm^{-3}$ and has a gate made of heavily doped polysilicon $N_{D-poly} = 5 \times 10^{19} cm^{-3}$. Calculate the built-in potential of this device.

The built-in potential of the MOSFET is the difference between the potential at the polysilicon gate and the potential deep in the p-type substrate.

$$\phi_{n-poly} = V_T \ln(N_{Dpoly}/n_i)$$
$$\phi_p = -V_T \ln(N_A/n_i)$$
$$\phi_o = \phi_{n-poly} - \phi_p = V_T \ln(\frac{N_{Dpoly} N_A}{n_i^2}) = 0.026V \ln(\frac{5 \times 10^{19} \cdot 10^{17}}{10^{20}}) = 1V$$

11.8.2 Applying the Depletion Approximation to Calculate the Field and Potential

To calculate the electric field and the electrostatic potential in the MOS capacitor, we solve the Poisson equation while applying the Depletion Approximation in a way very similar to what we did for the PN junction in Section 8.4 of Chapter 8. To do so we will divide the MOS capacitor into three regions and calculate the field and potential distribution in each region using the analysis below.

Region 1: The Polysilicon Gate

The polysilicon gate has extremely high doping so that it acts like a metal. This causes there to be no charge density, so from Poisson's equation we know:

$$\frac{d^2\phi}{dx^2} = 0 \tag{11.24}$$

In addition, since the polysilicon acts like a metal, there is no potential drop across it ($\Delta\phi_{n-poly} = 0$), so the electrostatic potential in this region is constant ($\phi(x) = Constant$) and has the value due to the built-in potential. Because the potential is a constant, the electric field will therefore be zero in this region.

$$E_{poly} = 0 \tag{11.25}$$

$$\phi = V_T \, ln \frac{N_{Dpoly}}{n_i} \equiv \phi_{n-poly} \tag{11.26}$$

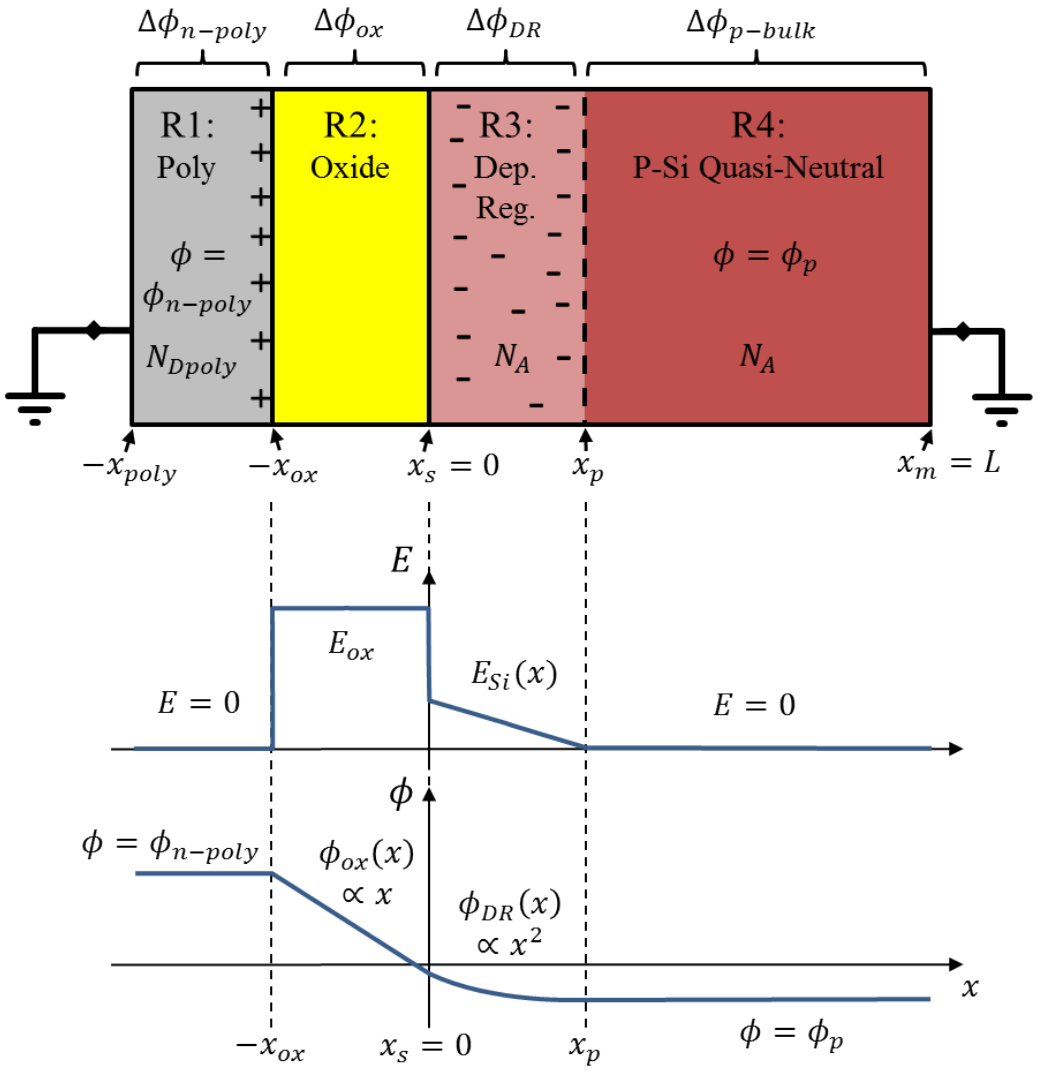


Figure 11.20: Depletion approximation solutions to electric field and potential inside a p-type MOS Capacitor

Region 3: The P-Type Silicon Substrate

We will now work out of order and focus on Region 3, which is the P-type depletion under the gate oxide. To calculate the electric field in the P region near the gate oxide, we will use the depletion approximation. To do this we solve the Poisson equation in this depletion region. Starting with the general Poisson equation for semiconductor we have

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon_{Si}}(p - n + N_D^+ - N_A^-) \qquad (0 \leq x \leq x_p) \tag{11.27}$$

where $x_s = 0$ is the coordinate of the Si/SiO2 surface or interface, and x_p is the end of the depletion region in the p-substrate as shown in Figure 11.19.

Now directly following the work done in Chapter 8 for calculating the internal field and potential, 11.27 reduces to the following since the doping is by acceptors, and the mobile electron and hole concentrations are negligible.

$$\frac{d^2\phi}{dx^2} = \frac{q}{\epsilon_{Si}}(N_A^-) \qquad (0 \leq x \leq x_p) \tag{11.28}$$

Integrating once we obtain:

$$\frac{d\phi}{dx} = \frac{q}{\epsilon_{Si}} N_A x + Constant \tag{11.29}$$

To derive an expression for the electric field in the depletion region directly under the gate oxide recall that $E = -d\phi/dx$. Applying the boundary condition that the electric field is zero in the quasi neutral P-type substrate (starting at x_p) we can solve for the constant to obtain:

$$\frac{d\phi}{dx}|_{x_p} = 0 \tag{11.30}$$

$$\boxed{E(x) = -\frac{d\phi}{dx} = \frac{qN_A}{\epsilon_{Si}}(x_p - x) \qquad (0 \leq x \leq x_p)} \tag{11.31}$$

Following the method in Section 8.4, we now integrate again and obtain the expression for electrostatic potential in this region as a function of position

$$\int_{\phi(x)}^{\phi(x_p)} d\phi = \frac{qN_A}{\epsilon_{Si}} \int_x^{x_p} (x - x_p)dx \tag{11.32}$$

$$\phi(x_p) - \phi(x) = \frac{-qN_A}{2\epsilon_{Si}}(x - x_p)^2 \tag{11.33}$$

At the depletion region edge x_p, the potential must equal its constant value in the quasi-neutral region $\phi(x_p) = \phi_p$ used in the definition of the built-in potential:

$$\boxed{\phi(x) = \frac{q}{2\epsilon_{Si}} N_A(x_p - x)^2 + \phi_p \qquad (0 \leq x \leq x_p)} \tag{11.34}$$

where,

$$\phi_p = \phi(x_p) = -V_T \, ln \frac{N_A}{n_i} \tag{11.35}$$

Figure 8.5 illustrates the built-in electric field and the built-in electrostatic potential.

Region 2: The Gate Oxide

Since there is no charge in the insulating gate oxide, the solution of the Poisson equation becomes:

$$\frac{d^2\phi}{dx^2} = 0 \qquad (-x_{ox} \leq x \leq 0) \tag{11.36}$$

By integrating, this immediately gives

$$\frac{d\phi}{dx} = Constant \qquad (-x_{ox} \leq x \leq 0) \tag{11.37}$$

And since the electric field $E = -d\phi/dx$, it means that the electric field in the oxide is a constant as well.

$$E_{ox} = Constant \tag{11.38}$$

To determine the value of that constant we apply Gauss' Law which requires that at the oxide/semiconductor interface the product of the electric field and the permittivity must be continuous or:

$$\epsilon_{ox} E_{ox}(0) = \epsilon_{Si} E_{Si}(0) \tag{11.39}$$

We have already obtained the expression for the electric field at the surface in the previous section. Evaluating equation 11.31 at the surface ($x = 0$) and doing a little algebra we obtain an expression for the oxide field in terms of known parameters:

$$\boxed{E_{ox} = \frac{\epsilon_{Si}}{\epsilon_{ox}} E_{Si}(0) = \frac{q}{\epsilon_{ox}} N_A (x_p - 0)} \tag{11.40}$$

Solving for the potential by integrating the negative of the field:

$$\frac{d\phi}{dx} = -E \tag{11.41}$$

$$\int_{\phi(-x_{ox})}^{\phi(x)} d\phi = -E_{ox} \int_{-x_{ox}}^{x} dx \tag{11.42}$$

$$\phi(x) - \phi(-x_{ox}) = -E_{ox}(x + x_{ox}) \tag{11.43}$$

From our results in Region 1, we apply the potential boundary condition:

$$\phi(-x_{ox}) = \phi_{n-poly} \tag{11.44}$$

$$\boxed{\phi(x) = \phi_{n-poly} - E_{ox}(x + x_{ox})} \tag{11.45}$$

We can see from this that the potential drops linearly across the oxide. Calculating the drop across the oxide $\Delta\phi_{ox}$ is just

$$\Delta\phi_{ox} = E_{ox}(0 + x_{ox}) = E_{ox}t_{ox} \tag{11.46}$$

Where $t_{ox} = x_{ox}$ is the thickness of the gate oxide.

The complete electric field and potential we have derived are plotted in Figure 11.20.

11.8.3 Threshold Voltage V_{TH} Derivation

Now that we did the electrostatic analysis of the MOS capacitor, we will use the results to help as part of our threshold voltage derivation. The threshold voltage V_{TH} is the value of gate-source voltage that is necessary to invert the surface from P-type to N-type. In other words, the value of V_{GS} that is necessary to change the surface potential $\phi_s \equiv \phi(x = 0)$ equal to $-\phi_p$. Here we will derive an expression for this voltage in terms of the applied gate voltage and known parameters using a MOS capacitor. The approach will be to start off with the equilibrium condition, and then add an additional gate voltage V_G to obtain the final expression.

If we consider the MOS capacitor in Figure 11.19, we know that the built in potential for this structure is the difference between built in potential from the N-poly doping of the gate and the built-in potential from the P doping of the substrate:

$$\phi_o = \phi_{n-poly} - \phi_p \tag{11.47}$$

Where ϕ_{n-poly} and ϕ_p are known from the doping as given by equations (11.26 and (11.35), respectively.

Now that we know the built in potential, the set of steps will be to determine how the built in potential is distributed over the MOS capacitor:

$$\phi_o = \Delta\phi_{n-poly} + \Delta\phi_{ox} + \Delta\phi_{DR} + \Delta\phi_{p-bulk} \tag{11.48}$$

$\Delta\phi_{n-poly} = 0$, since the n-poly gate is doped very highly it acts like a metal and thus there is negligble voltage drop across it.

$\Delta\phi_{p-bulk} = 0$, since the p-bulk is far from the oxide it is charge neutral ($p \approxeq N_A$), so there is negligible electric field and thus there is a negligible voltage drop across it.

Therefore, all the built-in potential is dropped across the gate oxide and the depletion region in the p-type silicon under the gate or:

$$\phi_o = \Delta\phi_{ox} + \Delta\phi_{DR} \tag{11.49}$$

Now let's determine how the built-in voltage is distributed over the gate oxide and the depletion region as indicated in Figure 11.19.

Gate Oxide Voltage Drop

In this section we'll find an expression for the voltage drop across the gate oxide in terms of known parameters. We found earlier in this chapter that the electric field in the oxide is constant. We also found the voltage drop across the oxide as the product of the field and the oxide thickness (Equation 11.46):

$$\Delta\phi_{ox} = E_{ox}t_{ox} \tag{11.50}$$

We now need to determine E_{ox} in terms of known parameters. Since we found an expression for the field in the semiconductor earlier in the chapter, we employ electrostatics to write the oxide field in terms of the semiconductor field at the interface. Electrostatics says that the product of the permittivity and the electric field must be continuous at the SiO_2/Si interface as we showed earlier in equation (11.39) :

$$\epsilon_{ox}E_{ox}(0) = \epsilon_{Si}E_{Si}(0) \tag{11.51}$$

Now, find $E_{ox}(0)$ from $E_{Si}(0)$ using the expression for the electric field in the silicon depletion region (11.31) that we derived in the previous section:

$$E_{Si}(0) = \frac{q}{\epsilon_{Si}}N_A(x_p - 0) \tag{11.52}$$

$$E_{ox}(0) = \frac{\epsilon_{Si}E_{Si}(0)}{\epsilon_{ox}} = \frac{qN_Ax_p}{\epsilon_{ox}} \tag{11.53}$$

Now, we find x_p in terms of ϕ using our solution to the potential in the depletion region (Equation 11.34):

$$\phi(0) - \phi_p = \frac{qN_A}{2\epsilon_{Si}}(x_p - 0)^2 \tag{11.54}$$

$$x_p = \left[\frac{2\epsilon_{Si}}{qN_A}(\phi(0) - \phi_p)\right]^{\frac{1}{2}} \tag{11.55}$$

Now that we have an expression for x_p we can finally evaluate Equation 11.50 for $\Delta\phi_{ox}$ in terms of known physical quantities (except for $\phi(0)$ which we will discuss later). Substituting Equation 11.55 into the oxide field from equation (11.53) then into equation (11.50) we get the potential drop across the oxide:

$$\Delta\phi_{ox} = E_{ox}t_{ox} = \frac{qN_Ax_p}{\epsilon_{ox}}t_{ox} \tag{11.56}$$

$$\Delta\phi_{ox} = \frac{qN_At_{ox}}{\epsilon_{ox}}\left[\frac{2\epsilon_{Si}}{qN_A}(\phi(0) - \phi_p)\right]^{\frac{1}{2}} \tag{11.57}$$

Making the substitution for the oxide capacitance (per unit area) $C_{ox} = \epsilon_{ox}/t_{ox}$ and defining the surface potential $\phi_s \equiv \phi(0)$:

$$\Delta\phi_{ox} = \frac{1}{C_{ox}} \left[2qN_A\epsilon_{Si}(\phi_s - \phi_p)\right]^{\frac{1}{2}} \tag{11.58}$$

We now use equations (11.58), (11.49) and (11.61) to express the entire built-in potential as the sum of the drop across the oxide and the drop across the semiconductor depletion region:

$$\boxed{\phi_o = \frac{1}{C_{ox}} \left[2q\,\epsilon_{Si}\,N_A(\phi_s - \phi_p)\right]^{\frac{1}{2}} + (\phi_s - \phi_p)} \tag{11.59}$$

Where we have written the potential $\phi(0)$ in the simplified form of ϕ_s. Now it is important to note that the built in potential on the left hand side is known from the doping and given by equation (11.23), and all terms on the right hand sides of equation (11.59) are known except for the surface potential ϕ_s. So, the surface potential could be readily determined by solving for it in equation (11.59).

Finding Threshold Voltage

Depletion Region Voltage Drop

The voltage drop across the depletion region is given as a function of position in equation (11.34). Here, we are interested in the total potential drop across the depletion region which we can calculate as the difference between the potential at the semiconductor surface $\phi(0)$ and the potential at the edge of the depletion region $\phi(x_p)$:

$$\Delta\phi_{DR} = \phi(0) - \phi(x_p) \tag{11.60}$$

Or using the more compact notation:

$$\Delta\phi_{DR} = \phi_s - \phi_p \tag{11.61}$$

Where $\phi(x_p)$ has its usual definition of the constant bulk potential determined by the doping as in equation (11.35).

We will now use equation (11.59) to find the MOSFET threshold voltage. Let's start by adding the gate voltage V_G to the LHS of the equation and moving the built in voltage to the RHS:

$$V_G = \frac{1}{C_{ox}} \left[2q\,\epsilon_{Si}\,N_A(\phi_s(V_G) - \phi_p)\right]^{\frac{1}{2}} + (\phi_s(V_G) - \phi_p) - \phi_o \tag{11.62}$$

It is important to note that the surface potential does not have the same value as it did in equation (11.59) because it is a function of the gate voltage V_G. We now introduce the definition of the threshold voltage as the following:

The MOSFET Threshold Voltage V_{TH} is amount of gate voltage necessary so that the surface potential is equal to the magnitude of the bulk potential but opposite in sign. In other words:

$$\boxed{V_{TH} = V_G|_{\phi_s=-\phi_p}} \tag{11.63}$$

Using the definition (11.63) in equation (11.62). we substitute $-\phi_p$ for ϕ_s and arrive at the following expression for threshold voltage in terms of known parameters:

$$V_{TH} = \frac{1}{C_{ox}}\left[2q\,\epsilon_{Si}\,N_A(-2\phi_p)\right]^{\frac{1}{2}} - (2\phi_p) - \phi_o \tag{11.64}$$

Now, since the built-in potential ϕ_p in the p-substrate is always less than zero, we can write equation (11.64) using absolute values and arrive at the final expression for the MOSFET threshold voltage:

$$\boxed{V_{TH} = \frac{1}{C_{ox}}\left[4q\,\epsilon_{Si}\,N_A|\phi_p|\,\right]^{\frac{1}{2}} + 2|\phi_p| - \phi_o} \tag{11.65}$$

Before ending this section, we remind ourselves that it is not at all arbitrary that we define threshold voltage by equation (11.63). It has this definition because when the surface potential is equal and opposite the bulk potential, it means that the concentration of electrons at the surface is equal in magnitude to the concentration of holes in the substrate. Thus, the surface has been inverted from a region where the mobile carriers were holes, to a region where the mobile carriers are electrons. Thus, the surface has been inverted and the channel of conducting electrons (or the n-channel) has formed.

Example 11.9:
The same N-MOSFET as in 11.8 has a gate oxide made of SiO_2 which is $30nm$ thick. Calculate its threshold voltage.

First we need to find ϕ_p and C_{ox}:

$$\phi_p = V_T \ln(N_A/n_i) = -0.026V \ln(10^{17}/10^{10}) = -0.42V$$
$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{3.9\epsilon_0}{30nm} = \frac{3.9 \cdot 8.85 \times 10^{-14} F/cm}{3 \times 10^{-6} cm} = 1.15 \times 10^{-7} F/cm^2$$

Then apply the formula for threshold voltage using our physical parameters and the values previously calculated:

$$V_{TH} = \frac{1}{C_{ox}}\left[4q\,\epsilon_{Si}\,N_A|\phi_p|\,\right]^{\frac{1}{2}} + 2|\phi_p| - \phi_o$$
$$= \frac{\left[4(1.6 \times 10^{-19})(11.7 \cdot 8.85 \times 10^{-14})(10^{17})(0.42)\right]^{\frac{1}{2}}}{1.15 \times 10^{-7}} + 2(0.42) - 1 = 1.3V$$

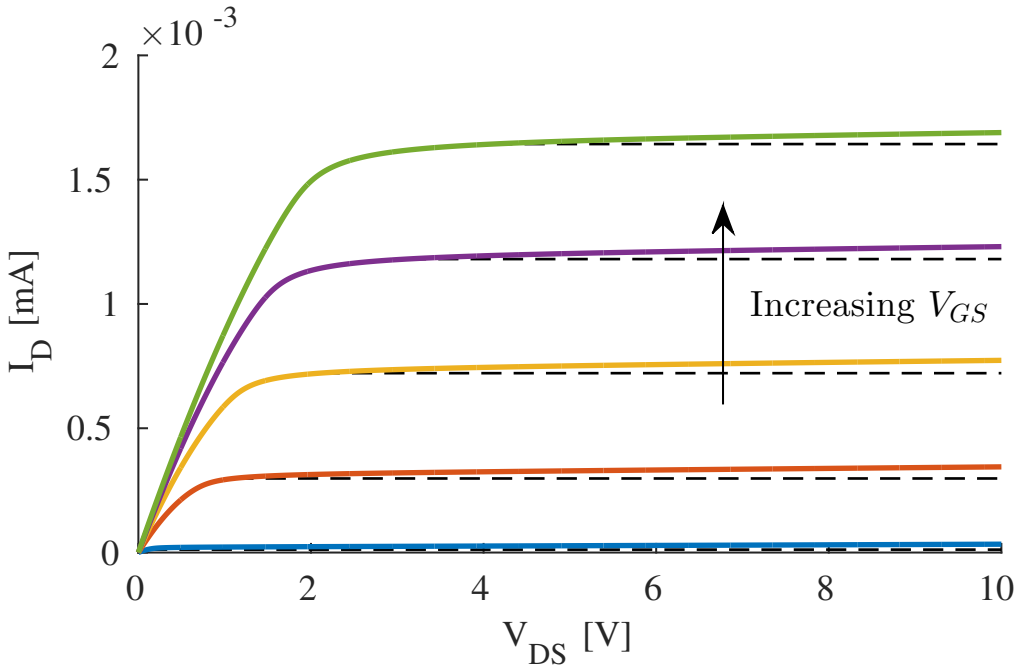


Figure 11.21: MOSFET IV curves with linearly increasing Gate Voltage V_{GS} from 1V to 5V. When the channel length modulation parameter λ is not included, the saturation current remains constant with respect to V_{DS} (shown as dashed black lines).

11.9 Small Signal Analog Model

Earlier in this chapter we derive the DC drain current MOSFET in saturation, and found it is given by:

$$I_D = \frac{\mu C_{ox} W}{2L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{11.66}$$

Where the parameter λ is the channel length modulation factor and it is typically small, so that the term $(\lambda V_{DS} << 1)$, and λV_{DS} represents a small correction. This factor results in a slight V_{DS} dependency in the drain current for a MOSFET operating in the saturation region, shown in Figure 11.21.

We now find the small signal current. The drain current depends strongly on V_{GS} and a little on V_{DS}, so the total small signal drain current is obtained by the total differential with respect to both: ΔI_D for a given ΔV_{GS} and ΔV_{DS}

$$\Delta I_D = \frac{\partial I_D}{\partial V_{GS}}\Delta V_{GS} + \frac{\partial I_D}{\partial V_{DS}}\Delta V_{DS} \tag{11.67}$$

We now define: $\frac{\partial I_D}{\partial V_{GS}} = g_m$ to be the Small Signal Transconductance.
Taking derivative of I_D in saturation with respect to V_{GS} and substituting:

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = \frac{\mu C_{ox} W}{L}(V_{GS} - V_{TH}) \tag{11.68}$$

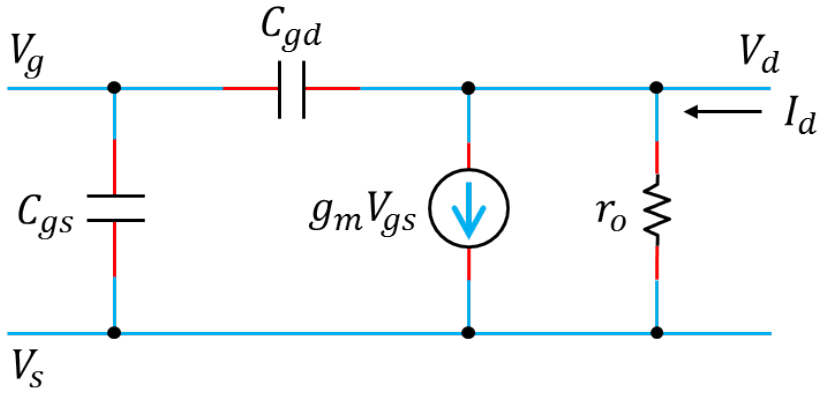


Figure 11.22: Small Signal Equivalent Circuit of N-MOSFET.

Or, written in terms of the drain current we get:

$$g_m = \sqrt{\frac{2\mu C_{ox} W I_D}{L}} \tag{11.69}$$

Next we define the output conductance $g_o = \frac{\partial I_D}{\partial V_{DS}}$. Performing the indicated differentiation gives:

$$g_o = \frac{\mu C_{ox} W}{2L}(V_{GS} - V_{TH})^2 \lambda = \lambda I_D \tag{11.70}$$

Finally, when we work with small signal equivalent circuits we typically use output resistance r_o which is just the reciprocal of the ouput conductance:

$$r_o = \frac{1}{g_o} \tag{11.71}$$

In Figure 11.22 we show the small signal N-MOSFET equivalent circuit. The capacitors in the equivalent circuit come from C_{ox}. In general the capacitors are given as follows:

$$C_{gs} = \frac{2}{3} C_{ox} L W \tag{11.72}$$

and

$$C_{gd} = C_{ox} L_d W \tag{11.73}$$

Where L_d is the small distance that the gate oxide overlaps the drain, and it is call the lateral diffusion length. Typically, $C_{gs} >> C_{gd}$. However, C_{gd} is a Miller capacitor so its effect is increased for high gain amplifiers. The physical locations of these capacitors inside the MOSFET are shown in Figure 11.23.

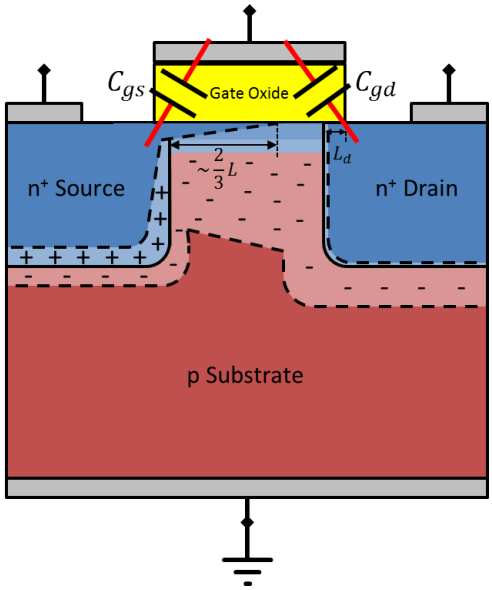


Figure 11.23: Small Signal Equivalent Circuit of N-MOSFET.

Example 11.10:
The N-MOSFET in Example 11.5 is biased such that it has drain current $I_D = 3 \times 10^{-3} A$ flowing. Calculate the small signal transconductance g_m and the output resistance r_o assuming $\lambda = 100 V^{-1}$.

$$g_m = \sqrt{\frac{2\mu C_{ox} W I_D}{L}} = \sqrt{\frac{2(300)(1.7 \times 10^{-7})(2)(3 \times 10^{-3})}{0.5}} = 1.1 \times 10^{-3} \Omega^{-1}$$

$$g_o = \lambda I_D = 3 \times 10^{-5} \Omega^{-1}$$

$$r_o = \frac{1}{g_o} = 33.3 k\Omega$$

11.10 Problems

11.1 (a) Draw two cross-sections of an N-MOSFET. Label the source, drain, gate, body, oxide, substrate, and indicate the type of doping in each region. Draw one cross-section with zero gate voltage, and another when gate voltage is large enough to establish a channel.

(b) What is meant by the inversion layer, how is it formed? What are the mobile carriers in the channel for a N-Channel MOSET and a P-Channel MOSFET.

(c) Qualitatively, explain how and why the gate voltage affects the mobile

carrier concentration in the channel and the drain current.

11.2 Describe qualitatively how a MOSFET works.

11.3 Why is the DC input resistance of a MOSFET equal to infinity. Why is a MOSFET a field effect device, while a BJT is not.

11.4 Describe in a sentence or two what is meant by the threshold voltage in a MOSFET. Why do we say that inversion is reached when the surface potential is equal to the negative of the substrate potential?

11.5 An N-Channel MOSFET has $N_A = 10^{17}cm^{-3}$, $N_{D-poly} = 10^{19}cm^{-3}$, $Tox = 10nm$, $L = 0.5um$ and $W = 20.0um$. Let mobility $\mu_n = 500cm^2/Vs$. (Note that the electron mobility by the MOSFET surface is lower than in the bulk.)

(a) Calculate the Threshold Voltage for this MOSFET.

(b) Graph carefully the family of curves for the MOSFET above for $V_{GS} = 1, 2$ and $3V$, and $V_{DS} = 0$ to $5V$ for each value of V_{GS}.

(c) Calculate the width of the depletion region under the gate if the surface potential is zero volts.

(d) Calculate the electric field in the oxide when the surface potential is zero volts.

11.6 Derive the expression for Id vs. V_{GS} for extremely low values of V_{DS}, including all of the steps. Also, point out where the channel charge and the electric field parallel to the channel is hidden the in the equation.

11.7 Explain the difference between the two regions of operation of a MOSFET: Linear and Saturation. Include the words pinch-off and depletion region in your answer.

11.8 Derive the expressions for the small signal output resistance and the transconductance.

11.9 Calculate values for and draw the small signal equivalent circuit of an N-MOSFET in Problem 11.5 above when $V_{GS} = 2V$ and $V_{DS} = 3V$. Take the channel length modulation parameter $\lambda = 0.05/V$. Include C_{gs} and C_{ds}. Assume lateral diffusion $L_d = 0.05um$.

11.10 Derive the expression for threshold voltage, including all steps. Explain each step with a short sentence.

11.11 Derive the expression for the depletion region length x_p in terms of the built in potential, the substrate doping N_A, oxide thickness t_{ox}, q, and the material permittivities.

11.12 Starting with the Poisson equation, show that the electric field in the oxide E_{ox} is constant and given by: $E_{ox} = \frac{qN_A x_p}{\epsilon_{ox}}$, where x_p is the thickness of the depletion region in the semiconductor.

Appendix A

Quantum Mechanics

A.1 Finite Well

As mentioned in Section 3.5.2 algebraic derivation is lengthy, however here we will present the final solutions for computing the allowed energies. Equations A.1 through A.4 can be obtained by solving for the wavefunction coefficients using Equations 3.55 through 3.56 with boundary conditions set by Equations 3.58 and 3.62.

$$\text{Odd Solutions :} \qquad \tan\left(\frac{k_n L}{2}\right) = \sqrt{\left(\frac{\alpha}{k_n}\right)^2 - 1} \qquad n = 1, 3, 5, \ldots \tag{A.1}$$

$$\text{Even Solutions :} \qquad -\cot\left(\frac{k_n L}{2}\right) = \sqrt{\left(\frac{\alpha}{k_n}\right)^2 - 1} \qquad n = 2, 4, 6, \ldots \tag{A.2}$$

$$\alpha = \frac{\sqrt{2mV_o}}{\hbar} \tag{A.3}$$

$$\mathcal{E}_n = \frac{\hbar^2 k_n^2}{2m} \tag{A.4}$$

From this new set of equations, the energy solutions can be calculated numerically or graphically (Figure A.1) with relative ease by solving for the allowed k_n values and converting these to their corresponding $\mathcal{E}_n$.

The right hand side of equations A.1 and A.2 is plotted in black in Figure A.1, along with the tangent (red) and the negative cotangent (blue) functions. Where these lines cross are the discrete energy solutions for the bound states in a finite well with $V_o = 5eV$ and $L = 2nm$. A finite well will have a finite number of bound states that 'fit' into it before the particle becomes free. We can tell that there are a finite number of states graphically because even though the red and blue lines will continue appearing out towards infinite energy, eventually the quantity $\sqrt{\left(\frac{\alpha}{k}\right)^2 - 1}$

in black will touch the x axis at $\mathcal{E} = V_o$ before it becomes imaginary at larger energies. This limits the number of bound solutions and indicates that all bound states must have energy less than the well height.

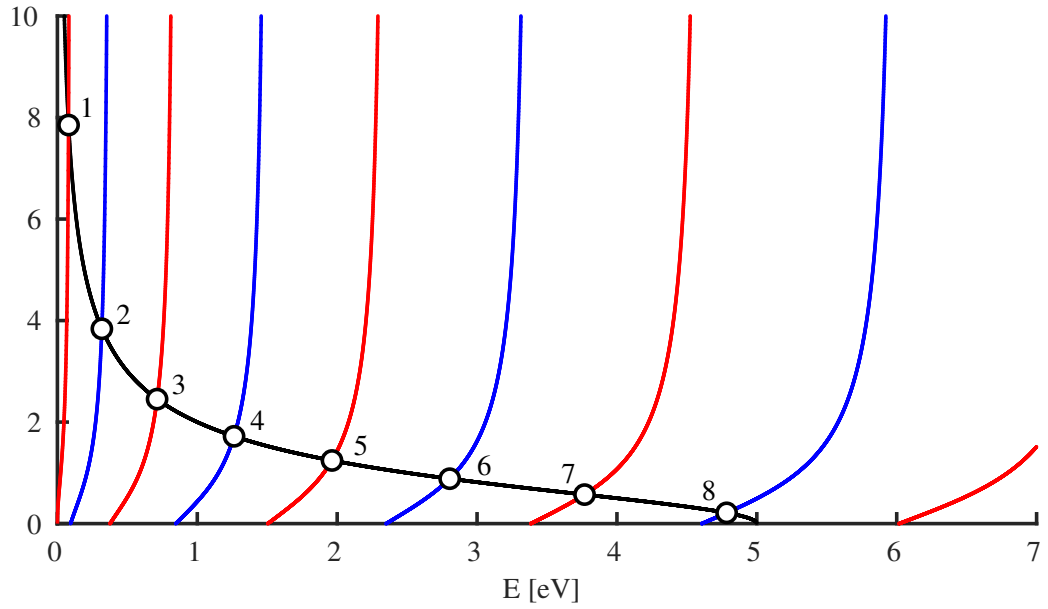


Figure A.1: Finite well energy solutions are circled for a well with $V_o = 5eV$ and $L = 2nm$. The black line is the quantity $\sqrt{\left(\frac{\alpha}{k}\right)^2 - 1}$, the red lines are $\tan\left(\frac{kL}{2}\right)$ and the blue lines are $-\cot\left(\frac{kL}{2}\right)$. The x value where the lines intersect gives the allowed energy of that state.

For the example finite well in Figure A.1, the first six allowed energies are compared to those in an infinite well of the same length below:

	1	2	3	4	5	6
Finite:	0.0795 eV	0.3175 eV	0.7128 eV	1.2626 eV	1.9625 eV	2.8038 eV
Infinite:	0.0940 eV	0.3761 eV	0.8462 eV	1.5043 eV	2.3505 eV	3.3848 eV

Example A.1:
You wish to create a finite well system which has the same ground state energy as an infinite well system. If the infinite well system has $L = 1nm$, what would the length need to be for the finite well if it's barriers are $V_o = 3eV$?

Start by finding the ground state energy of the infinite well which we will use to compare the final solution:

$$\mathcal{E}_1 = \frac{\pi^2\hbar^2 1^2}{2mL^2} = \frac{\pi^2(1.054\times10^{-34}Js)^2 1^2}{2\cdot 9.1\times10^{-28}kg\cdot(1nm)^2} = 0.376eV$$

Next, get k_1 from E_1 (or obtain it directly with $k_n = n\pi/L$):

$$k_1 = \sqrt{2m\mathcal{E}_1}/\hbar = 3.14\times10^9 m^{-1}$$

Evaluate α:

$$\alpha = \frac{\sqrt{2mV_o}}{\hbar} = \frac{\sqrt{2\cdot 9.1\times10^{-28}kg\cdot 3eV}}{1.054\times10^{-34}Js} = 8.87\times10^9 m^{-1}$$

Then, for the ground state of the finite well, we must evaluate the solution using the odd equation (A.1) since the ground state has $n = 1$.

$$\tan\left(\frac{k_n L}{2}\right) = \sqrt{\left(\frac{\alpha}{k_n}\right)^2 - 1}$$

$$L = \frac{2}{k_1}\tan^{-1}\left(\sqrt{\left(\frac{\alpha}{k_1}\right)^2 - 1}\right) = \frac{2}{k_1}\tan^{-1}\left(\sqrt{\left(\frac{\alpha}{k_1}\right)^2 - 1}\right) = \mathbf{0.77nm}$$

So the finite well must be slightly contracted compared to the infinite well. This result makes sense because for the same L, the finite well energies are slightly lower so to compensate, the well must shrink because in the infinite well solution $\mathcal{E} \propto 1/L^2$.

A.2 Finite Barrier Tunneling

The transmission coefficient determines the probability of an incident electron passing through the region of space occupied by a potential barrier - whether that be emission over the 'top' (when $\mathcal{E} > V_B$) or 'through' in the case of tunneling (when $\mathcal{E} < V_B$). Here, $\mathcal{E}$ is the energy of the incoming electron and V_B is the energy of the rectangular barrier with length L. Transmission is determined by dividing the probability density of the electron transmitted through the barrier $|F|^2$ by that of

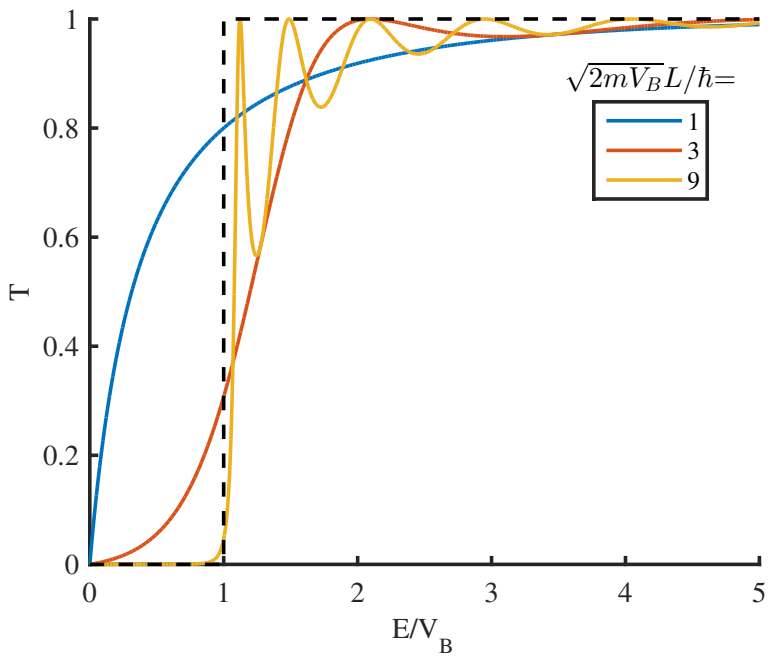


Figure A.2: Quantum mechanical transmission through a finite barrier. As the term $\sqrt{2mV_B}L/\hbar$ increases, the transmission approaches the classical limit (black dashed line).

the incident electron $|A|^2$.

$$T = \frac{|F|^2}{|A|^2} = \begin{cases} \left(1 + \dfrac{V_B^2 \sinh^2(K_{II}L)}{4\mathcal{E}(V_B - \mathcal{E})}\right)^{-1}, & V_B > \mathcal{E} \quad K_{II} = \sqrt{2m(V_B - \mathcal{E})}/\hbar \\ \left(1 + \dfrac{V_B^2 \sin^2(K_{II}L)}{4\mathcal{E}(\mathcal{E} - V_B)}\right)^{-1}, & V_B < \mathcal{E} \quad K_{II} = \sqrt{2m(\mathcal{E} - V_B)}/\hbar \\ \left(1 + \dfrac{mL^2 V_B}{2\hbar^2}\right)^{-1}, & V_B = \mathcal{E} \end{cases} \tag{A.5}$$

From this result, it is straightforward to recover the reflection coefficient using the relation $T + R = 1$. Immediately we can see a drastic difference in the quantum mechanical transmission probability compared to what we would expect classically. In a classical system, for all energies less than the barrier height we would expect to see identically zero transmission, and for all energies greater than the barrier we would expect to see perfect transmission. Interestingly, the quantum mechanically determined transmission probability approaches this classical limit as the barrier height and width increase, shown in Figure A.2 as the black dashed line. Please note that these formulas are only valid when the left and right sides of the barrier are at equal potentials, otherwise a factor of the ratio of K_I and K_{III} must be

included.

Example A.2:
A potential barrier of $V_B = 2eV$ is $1nm$ thick. Calculate the transmission probability for an electron incident with energy $1eV$. If the incident energy is instead $2.5eV$, what is the new transmission probability?

Since the first electron has less energy than the barrier, we use first equation in A.5:

$$K_{II} = \sqrt{2m(V_B - \mathcal{E})}/\hbar = \frac{\sqrt{2 \cdot 9.1 \times 10^{-31} kg \cdot (2eV - 1eV)}}{1.055 \times 10^{-34} Js} = 5.12nm^{-1}$$

$$T(1eV) = \left(1 + \frac{V_B^2 \sinh^2(K_{II}L)}{4\mathcal{E}(V_B - \mathcal{E})}\right)^{-1} = \left(1 + \frac{(2eV)^2 \sinh^2(5.12nm^{-1} \cdot 1nm)}{4 \cdot 1eV \cdot (2eV - 1eV)}\right)^{-1}$$
$$= \mathbf{1.42 \times 10^{-4}}$$

Barrier transmission is extremely low in this case because our T vs $\mathcal{E}/V_B$ looks similar to the yellow curve in Figure A.2 because our barrier has $\sqrt{2mV_B}L/\hbar = 7.25$. For the yellow curve, transmission close to zero for all $E/V_B < 1$ which includes our specific case of $E/V_B = 0.5$.

For the higher energy electron:

$$K_{II} = \sqrt{2m(\mathcal{E} - V_B)}/\hbar = \frac{\sqrt{2 \cdot 9.1 \times 10^{-31} kg \cdot (2.5eV - 2eV)}}{1.055 \times 10^{-34} Js} = 3.62nm^{-1}$$

$$T(2.5eV) = \left(1 + \frac{V_B^2 \sin^2(K_{II}L)}{4\mathcal{E}(\mathcal{E} - V_B)}\right)^{-1} = \left(1 + \frac{(2eV)^2 \sin^2(5.12nm^{-1} \cdot 1nm)}{4 \cdot 1eV \cdot (2.5eV - 2eV)}\right)^{-1}$$
$$= \mathbf{0.7}$$

Barrier transmission in this case is significant but not perfect.

A.3 Density of States

To derive the density of states in the conduction and valence bands, we assume that the electrons (and holes) are free particles confined to a crystal with side lengths L. From Bloch's theorem, we know that the wavefunction (in 3D) may be expressed as:

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) \tag{A.6}$$

$$k_x L = 2\pi n_x \tag{A.7}$$

$$k_y L = 2\pi n_y \tag{A.8}$$

$$k_z L = 2\pi n_z \tag{A.9}$$

$$n = \ldots, -2, -1, 0, 1, 2, \ldots \tag{A.10}$$

with the stipulation that $\mathbf{k}$ must obey periodic boundary conditions with the crystal boundary. From this, we know that in reciprocal (k) space, there is one state per $(2\pi/L)^3$. We also know the dispersion relation for a free particle:

$$E_{\mathbf{k}} = \frac{\hbar^2 |k|^2}{2m^*} \tag{A.11}$$

From this we can calculate the number of states N within a spherical volume of radius $|k|$. The total number of states must be divided by 8 because to account for the equivalence of $\pm k_{x,y,z}$ values, which represent a phase shift in the wavefunction but are physically the same state. This can be thought of as taking only the octant of Cartesian space where k_x, k_y, k_z are all positive (Figure A.3). To account for the spin degeneracy of the states we multiply by two.

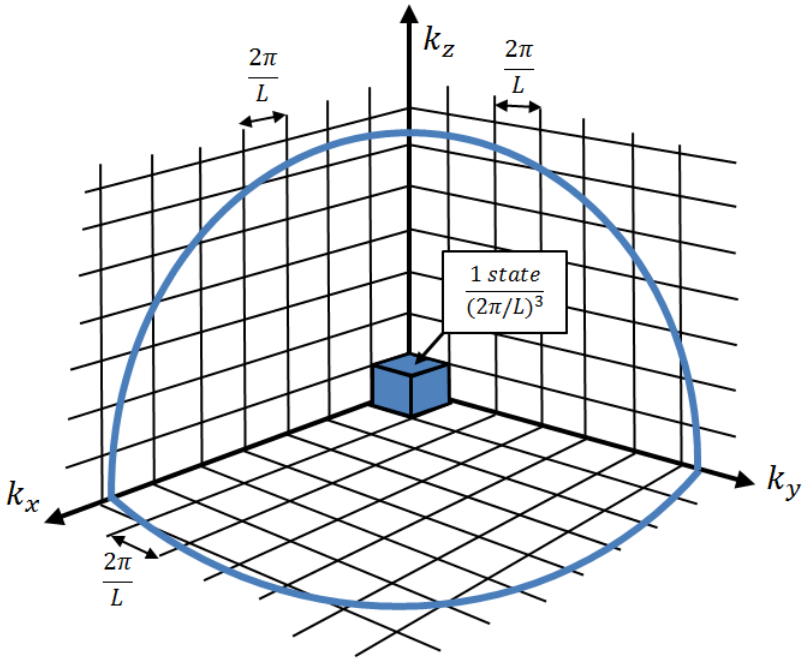


Figure A.3: Counting number of states in k-space that fit within the 8^{th} sphere of a given energy.

$$N = 2\left(\frac{1}{8}\right)\frac{(4/3)\pi|k|^3}{(2\pi/L)^3} = \frac{V}{3\pi^2}|k|^3 = \frac{V}{3\pi^2}\left(\frac{2m^*E}{\hbar^2}\right)^{3/2} \tag{A.12}$$

Here, V is the volume of the crystal. To get the density of states $DoS(E)$ (per volume) we take the derivative of N with respect to E and divide by V:

$$DoS(E) = \frac{dN}{dE} = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E} \tag{A.13}$$

$$DoS(E) = \frac{4\pi(2m^*)^{3/2}}{h^3}\sqrt{E} \tag{A.14}$$

You can see that the density of states has a square root dependence in 3D space, where we are normally operating. Devices which exploit quantum confinement can also be fabricated, which effectively changes the density of states to their analogous 0D, 1D, and 2D forms with different energy dependence ($\delta(E)$, $1/\sqrt{E}$, and constant). Changing the density of states can have many various and interesting consequences to device operation and is often used in conjunction with other quantum effects like tunneling.

Appendix B

Advanced Semiconductor Devices

B.1 JFETs

The Junction Field Effect Transistor is a relatively simple three-terminal device that is mainly composed of a single semiconductor PN junction. Figure xx shows a N-channel Silicon JFET cross section. The JFET has three terminals: the Source, Gate and Drain. As can be seen from the figure, the JFET is largely composed of a thin N-type silicon bar with source and drain contacts at the end with a P-type section in the middle that serves as the gate. In contrast to the MOSFET, the JFET does not have an oxide at the gate, it is just a region of P-type silicon. For N-type JFETs, the typical operation is as follows. A positive voltage is applied from the drain to the source. Since the voltage is higher at the drain contact than the source contact, electrons will flow from the source to the drain, the current goes into the drain and out of the source. The region where the electrons flow is called the channel. Now a voltage is applied to the gate that is negative with respect to both the source and the drain. Therefore the PN junction in the JFET is reverse biased. The larger the magnitude of the negative gate voltage, the more reverse biased the PN junction and the larger the depletion region. As the electrons flow from source to drain they need to flow around the P-region where the channel becomes narrow. As the magnitude of the negative voltage is increased the depletion region increases and the channel becomes more and more narrow and the depletion region extends all the way down across the channel. This limits or saturates the current no matter how much drain voltage you apply, the current stays fixed. The current voltage characteristics of an n-channel JFET are shown in Figure xx. The region where the Id vs Vds curves are horizontal is called saturation, just like in MOSFETs. Algebraic relationships for the drain current as a function of the gate-source and drain-source voltages can be obtained by solving the semiconductor equations, while

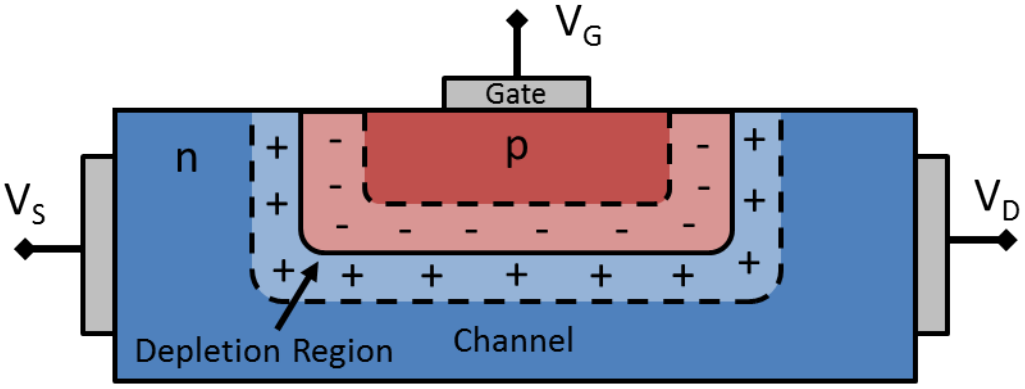


Figure B.1: Cross Section of N-channel JFET.

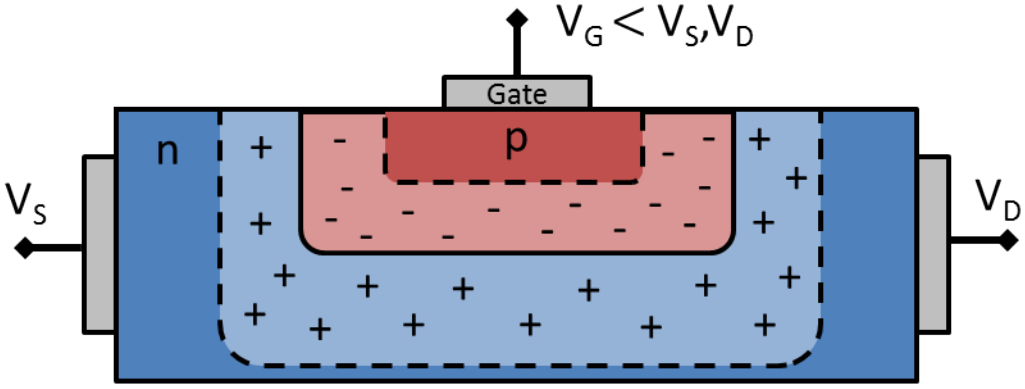


Figure B.2: Cross Section of N-channel JFET showing channel nearly cutoff.

employing appropriate approximations. These current voltage characteristics can be found in other device physics and physical circuit design books. Keeping with the introductory and concise nature of this text, here we present only the common formula of JFET operation the saturation region:

$$I_D = I_{DSS}\left(1 - \frac{V_{GS}}{V_P}\right)^2 \qquad -V_P \leq V_{GS} \leq 0 \tag{B.1}$$

Where I_{DSS} is the maximum current which is obtained when $V_{GS} = 0$.

The benefits of the JFET are that it can carry a significant amount of current. However, it is often complicated and not always practical to use from a circuit perspective because you need to apply negative voltage to the gate and therefore two different voltage polarities are required. Additionally, in contrast to a MOSFET, the reverse biased PN junction at the gate does have some reverse bias current so you do need to supply current to the gate of the device in order to make it work, although not as much as to the base of a BJT.

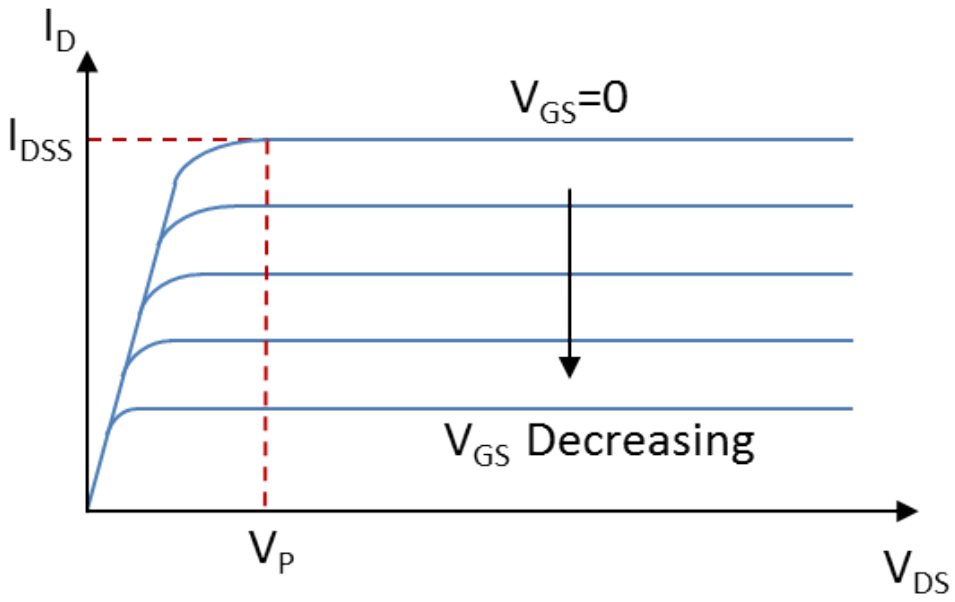


Figure B.3: Current-Voltage characteristics of a JFET.

B.2 Metal-Semiconductor Junctions

B.2.1 Schottky Barrier Diodes and Metal Semiconductor Contacts

In addition to diodes that are formed by semiconductor PN junctions, it is also possible to create a diode or rectifier that is made of a contact between a doped semiconductor and a metal. These kind of rectifiers are called Schottky diodes after Walter Schottky who invented them in the mid twentieth century. The Schottky diode has a similar current voltage relation to a PN junction diode and operates under similar principles. When certain types of metals are brought into contact with a certain semiconductor with specific doping, a built in potential and a depletion region in the semiconductor form. As we said in Chapter 8, requirements for rectification are to have a junction that has both, a depletion region and a built in potential. In addition to the depletion region and the built-in potential, there is an abrupt offset between the band structure of the metal and that of the semiconductor. This offset is another factor that allows particular metal-semiconductor contacts to function as rectifiers.

B.2.2 Schottky Barrier Diodes with Metal and N-Type Semiconductor: Rectifying Contacts

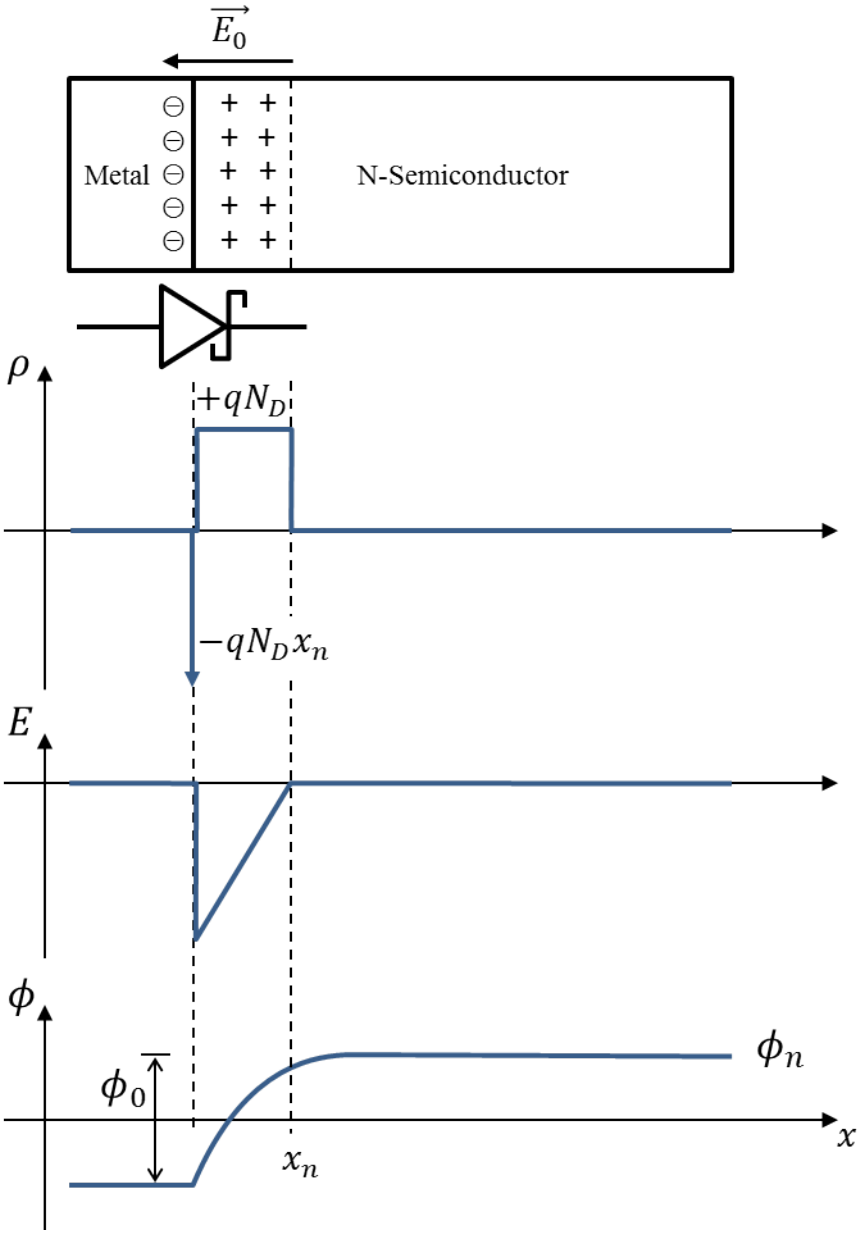


Figure B.4: Block diagram of Schottky structure. Plots show the position-dependent charge density, electric field profile, and potential profile inside the device.

At this point it is worth studying the band structure characteristics of a metal and N-type semiconductor that would form a rectifying contact or a Schottky diode when connected. On the right of Figure B.5 is an illustration of the valence band maximum $\mathcal{E}_V$, conduction band minimum $\mathcal{E}_C$, and the location of the Fermi level

in an N-type semiconductor $\mathcal{E}_{FS}$. Also shown is the zero energy line $\mathcal{E}_{vac}$ which is the energy of an electron after it has been extracted from the material, as well as the electron affinity $q\chi$, which is the energy required to remove an electron that is in the semiconductor conduction band away from the material. On the left of the figure is the Fermi level of the metal and the zero energy line. Also shown is the metal work function $q\Phi_M$, which is the energy required to remove an electron from the surface of the metal.

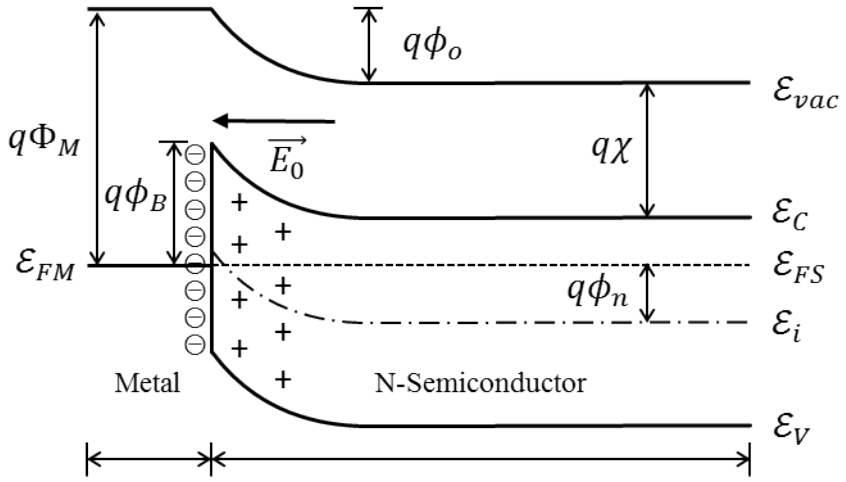


Figure B.5: Energy versus position band structure diagram of an n-type contact in equilibrium. The built-in electric field $\vec{E_0}$ points from the N-semiconductor to the surface of the metal.

Under certain conditions, when N-type silicon is brought into contact with certain metals, electrons from the silicon naturally transfer to the metal. This happens only if the Fermi level of the silicon, that has been doped with donors, is higher than the Fermi level of the metal. In other words, the chemical potential of electrons in the semiconductor is higher than that of the metal, so electrons in the semiconductor will be in higher energy states than those in the metal. Physical systems tend to want to reduce their energy to achieve equilibrium and therefore electrons in the N-type semiconductor will naturally transfer to the metal when they are brought into contact. When this transfer occurs, the region in the semiconductor near the metal becomes depleted of electrons and a depletion layer forms in an analogous manner to what occurs in a PN junction. As a result, the metal takes on more electrons. However, since the metal already has a very high concentration of electrons it does not obtain an accumulation region, but only an extremely thin negatively charged region at its surface near metal-semiconductor interface. Electrons will continue to flow from the semiconductor to the metal and the chemical potential of the semiconductor will decrease while the chemical potential of the metal will increase. Eventually the Fermi levels (or chemical potentials) will become equal and there will no longer be a net flow of electrons between the two materials. At this time a depletion region will

have formed in the semiconductor that contains an electric field that points from the now positively charged N-type semiconductor to the now negatively charged metal surface. The built-in potential is proportional to Fermi level metal subtracted from the Fermi level in the semiconductor or $q\phi_o = \mathcal{E}_{FS} - \mathcal{E}_{FM} = q\Phi_M - (q\chi + E_g/2 - q\phi_n)$.

Equilibrium: The operation of the Schottky Barrier Diode is analogous to that of the PN junction. As mentioned above, we have a positively charged depletion region in the N-type semiconductor and a negative surface charge on the metal and an electric field within the depletion region that points from the N-type semiconductor and terminates on the metal surface. In equilibrium, electrons from the highly doped region deeper in the semiconductor will diffuse toward the depletion region due to the concentration gradient. (Away from the junction and the depletion region, $n_n = N_D$.) However, when these electrons feel the electric field in the depletion region, they are pulled back by the field toward the N-type bulk area. So, in equilibrium, the diffusion current toward the junction is equal and opposite of the drift current from the junction region back into the bulk N-type semiconductor.

Forward Bias: To forward bias this Schottky Diode, we apply a potential to across the semiconductor and the metal as shown in Figure B.6. By applying a negative voltage to the semiconductor with respect to the metal, an internal electric field arises in the depletion region which points in the opposite direction of the built-in field. Thus, just as in the PN junction under forward bias, the total field is now reduced. With the lower electric field, the drift current becomes smaller, but the diffusion current has not appreciably changed because the concentration gradient between the bulk semiconductor and the depletion region at the junction has been maintained. The upshot is that drift current has now been reduced so now there is more diffusion current than drift, and a net current will flow. This net current is composed of electrons flowing from the N-type semiconductor to the metal. Since by convention current flow is in opposite direction to electron flow, current in the forward biased Schottky diode with an N-type semiconductor flows from the metal to the semiconductor.

Reverse Bias: Under forward bias electrons can flow from the semiconductor to the metal. However, under reverse bias now electrons can flow. There are two reasons for this. First, within the semiconductor, if a reverse bias is applied, the field inside the depletion region gets even larger. Thus, any electrons that diffuse from the semiconductor bulk to the depletion region will surely get pulled back into the bulk due to the increased junction field. Secondly, there is a quantum barrier between the metal and the semiconductor for electrons. As shown in Figure B.7, there is a large barrier for electrons from the metal Fermi surface to the conduction

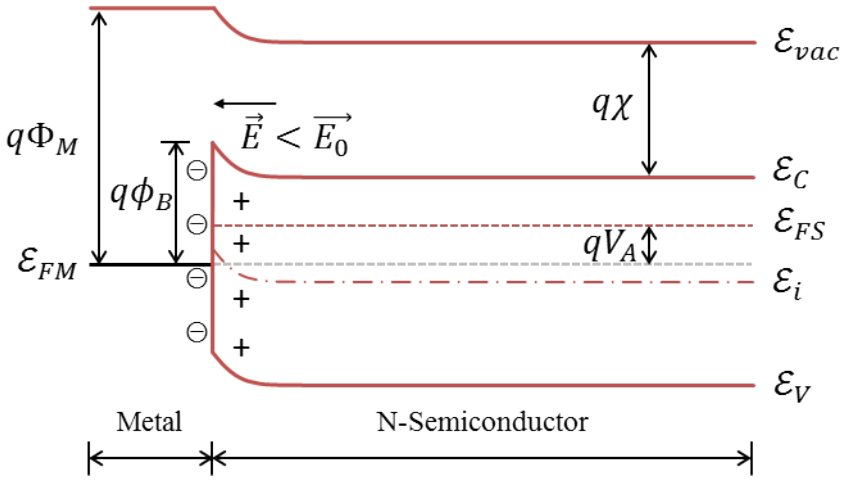


Figure B.6: Energy versus position band structure diagram of an n-type forward biased contact. Applied bias V_A to the metal is positive with respect to the N-semiconductor.

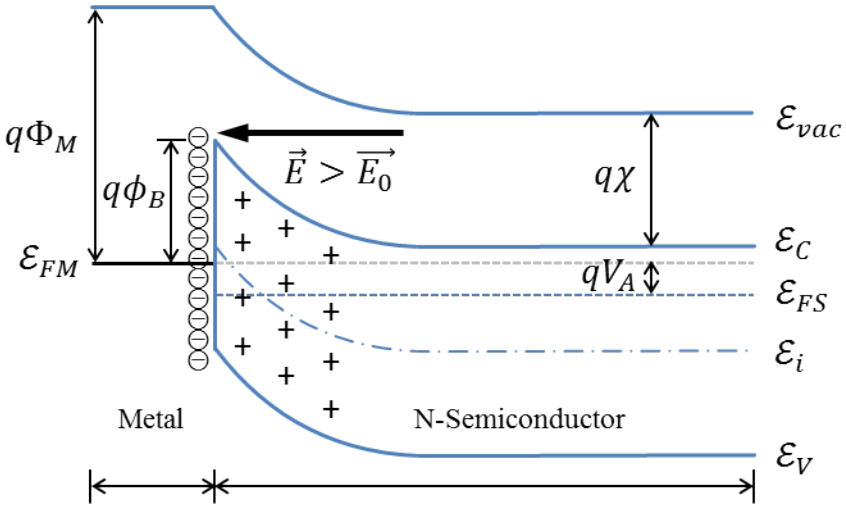


Figure B.7: Energy versus position band structure diagram of an n-type reverse biased contact. Applied bias V_A to the metal is negative with respect to the N-semiconductor.

band in semiconductor. The height of this barrier is ϕ_B, which is given by

$$q\phi_B = q(\Phi_M - \chi) \tag{B.2}$$

Electrons cannot flow from the metal to the semiconductor because of the large barrier $q\phi_B$ between the metal and the conduction band of the semiconductor, even if a negative voltage is applied to the metal with respect to the semiconductor. Therefore, since current is in the opposite direction to electron flow, current cannot flow when a voltage is applied to the Schottky diode that is positive on the semiconductor side with respect to the metal side.

Schottky Diode Equation:

$$J = J_o \left[e^{V_A/V_T} - 1 \right] \tag{B.3}$$

B.2.3 Non-Rectifying Contacts

In the case where the Fermi level is lower in the semiconductor than in the metal, the semiconductor will accumulate electrons instead of creating a depletion region. To balance this charge, the metal forms a very thin layer of positive charge at the interface, causing the field lines to point from the metal to the semiconductor. The metal will not form a depletion region due to the enormous number of electrons it contains. Because the semiconductor is now accumulated, this type of contact is referred to as Non-Rectifying or *ohmic* rather than Schottky. When bias is applied, an ohmic contact will act more similarly to a resistor and will not be rectifying.

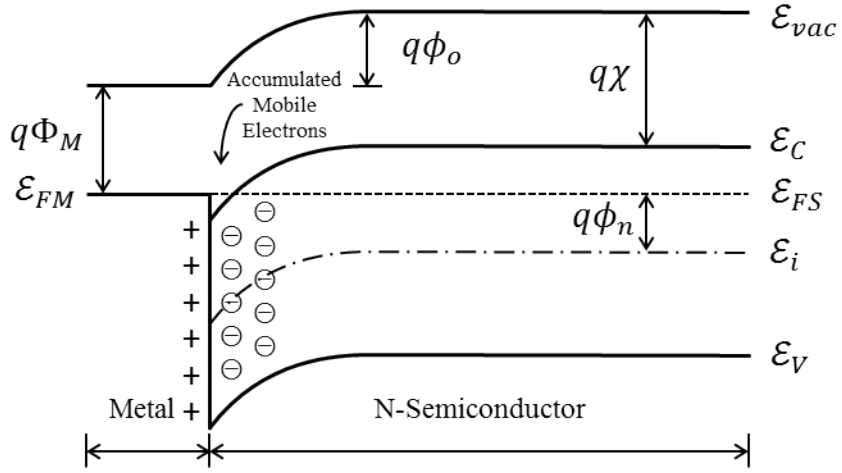


Figure B.8: Energy versus position band structure diagram of an n-type ohmic contact.

B.2.4 Tunneling Contacts

As in the case of the rectifying n-type Schottky contact, an n-type tunneling contact has the Fermi level of the n-type silicon $\mathcal{E}_{FS}$ higher than the Fermi level of the metal $\mathcal{E}_{FM}$. The difference here is that the semiconductor is heavily doped which causes the depletion region that forms at the interface to be extremely thin. As a result, the potential barrier ϕ_B formed by the conduction band is considerably thinner than in the case of the Shottky contact, and the conduction band $\mathcal{E}_C$ is nearly aligned with the Fermi level in the metal $\mathcal{E}_{FM}$. This allows significantly more

electrons to quantum-mechanically tunnel from the metal into the semiconductor and vise-versa so that it passes current both ways about equally. In this way the tunneling contact acts more similarly to an ohmic contact than a Schottky contact.

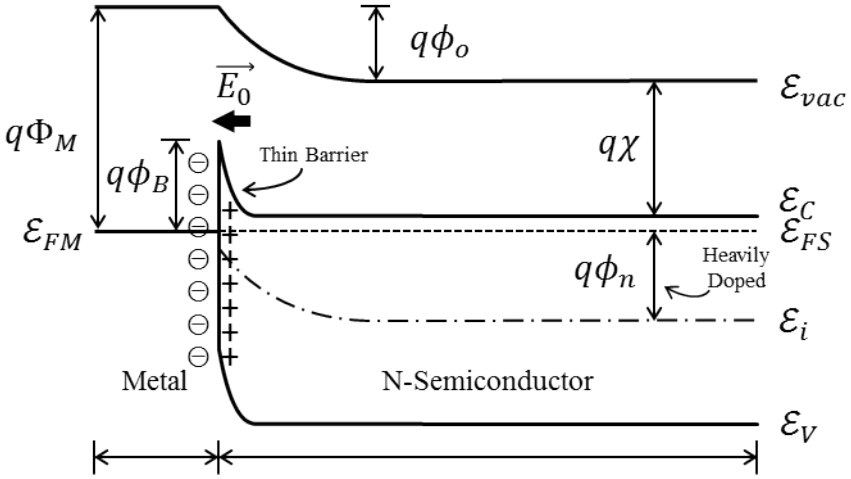


Figure B.9: Energy versus position band structure diagram of an n-type tunneling contact.

B.2.5 Fermi Level in PN Junctions

Just like a Schottky diode, we learned in Chapter 8 that a PN junction has a built-in potential. At the time, we described their operation without the need to introduce the concept of a Fermi level. Just as in the description of the various contact types, we can use the Fermi level to analyze the PN junction and obtain the same built-in potential result. The Fermi level of the P-type semiconductor is lower than the Fermi level of the N-type semiconductor, causing electrons on the N-side to flow to the P-side where they recombine with the excess holes there. This creates the depletion region and built-in field and potential just like in the case of the Schottky contact. Also similarly, the total built-in potential can be determined by the difference in the Fermi levels $q\phi_o = \mathcal{E}_{FN} - \mathcal{E}_{FP} = (q\chi + \mathcal{E}_g/2 + q\,|\phi_p|) - (q\chi + \mathcal{E}_g/2 - q\phi_n) = q(\phi_n + |\phi_p|) = qV_T \ln\left(\frac{N_D N_A}{n_i^2}\right)$. Because the left and right side of the junction are the same material (just doped differently), the effect of the bandgap and affinity cancels out to achieve the result from Chapter 8.

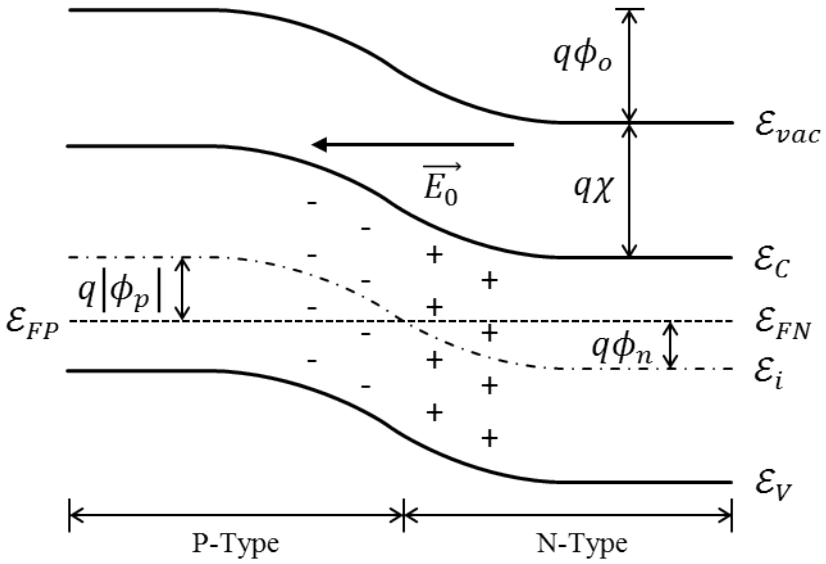


Figure B.10: Energy versus position band structure diagram of a PN junction.

Appendix C

Physical Constants and Material Parameters

Physical Constants

Table C.1: **Most Relevant Physical Constants**

Constant	**Symbol**	**Value**	**Units**
Planck Constant	h	6.626×10^{-34} 4.136×10^{-15}	J·s eV·s
Reduced Planck Constant	$\hbar$	1.055×10^{-34} 6.582×10^{-16}	J·s eV·s
Boltzmann Constant	K	1.381×10^{-23} 8.617×10^{-5}	J·K^{-1} eV·K^{-1}
Electron Charge (magnitude)	q	1.602×10^{-19}	C
Avogadro's Number	A_o	6.0226×10^{23}	mol^{-1}
Electron Mass	m_o	9.109×10^{-31}	kg
Permittivity of Free Space	ϵ_o	8.854×10^{-14}	F·cm^{-1}
Speed of Light	c	2.998×10^{10}	cm·s^{-1}

Material Parameters

Table C.2: **Most Relevant Semiconductor Material Parameters**

Parameter	**Symbol**	**Si**	**Ge**	**GaAs**	**Units**
Atomic Number	Z	14	32	$31-33$	–
Valence	–	4	4	$3-5$	–
Bandgap	E_g	1.12	0.66	1.42	eV
Density	ρ	2.329	5.323	5.32	g·cm^{-3}
Elec. Eff. Mass (DoS)	$m^*_{n,dos}$	1.08	0.56	0.067	m_o
Elec. Eff. Mass (Mobility)	$m^*_{n,cond}$	0.26	0.12	0.067	m_o
Hole Eff. Mass (DoS)	$m^*_{p,dos}$	0.81	0.29	0.47	m_o
Hole Eff. Mass (Mobility)	$m^*_{p,cond}$	0.39	0.21	0.34	m_o
Intrinsic Carrier Conc. ($300K$)	n_i	1×10^{10}	2×10^{13}	2.1×10^{6}	cm^{-3}
Permittivity	ϵ	11.7	16.2	12.9	ϵ_o
Elec. Mobility (max)	μ_n	1400	3900	8500	cm^2/(V·s)
Hole Mobility (max)	μ_p	450	1900	400	cm^2/(V·s)
Elec. Recomb. Lifetime (typ.)	τ_n	4×10^{-6}	$\geq 1\times10^{-3}$	5×10^{-9}	s
Hole Recomb. Lifetime (typ.)	τ_p	2×10^{-6}	$\geq 1\times10^{-3}$	3×10^{-6}	s

Bibliography

[1] Charles Kittel. *Introduction to Solid State Physics*. John Wiley and Sons, 2004.

[2] Bruce C. Reed. *Quantum Mechanics*. Jones and Bartlett, Sudbury, MA, 2008.

[3] Robert Eisberg and Robert Resnick. *Quantum Physics of Atoms, Molecules and Solids*. John Wiley and Sons, 1985.