Notes for Lecture 14

pn Junction, Electrostatics

14.1 What/Where is a pn Junction?

A pn junction device is usually made with one type of crystal (say Si) doped differently at the two ends. Let us assume that this is the case[1]. How the dopant atoms are distributed in the crystal depends on the manufacturing process – one can have a "step junction" (epitaxial growth) or a "graded junction" (ion implantation)[2].

So, assume that we have a pn junction. What we mean here is that the dopant density N_D and the acceptor density N_A are now functions of the position, and on one side N_A dominates (p) and on another side N_D dominates (n). To simplify the matters, we will assume that N_D and N_A is dependent only on one spatial coordinate, x.

For a pn junction, let us take the coordinate system so that the left side is the p part, and the right side is the n part. Thus, we mean that

$$N_A \gg N_D \qquad \text{for } x \to -\infty \tag{14.1}$$

$$N_D \gg N_A \qquad \text{for } x \to \infty \tag{14.2}$$

Of course, $x \to \pm\infty$ is a mathematical idealization. In actuality, $-\infty$ means the left end of the device, and ∞ means the right end of the device.

[1] This is the case of a "homo-junction." If this assumption is broken for a pn junction, then one is dealing with a "hetro-junction."

[2] Please do not think that you can simply join two crystals of p type and n type and make a pn junction. That is *not* how a pn junction is made, while it is conceptually helpful to think in such a way for a step junction.

How $N_A(x \to -\infty)$ approaches $N_A(x \to \infty)$ and how $N_D(x \to -\infty)$ approaches $N_D(x \to \infty)$ will depend on what kind of junction that we make. But it is clear that the material will be changing from a p type to an n type. By the general argument of band bending (LN 10), we recognize that E_F, which must be constant throughout the device, must cross E_i, the intrinsic Fermi level. Right at the point where this happens, p and n are exactly balanced. We might call this point the junction point, and assign $x = 0$.

However, the more common definition of the junction is the **metallurgical junction**. Here, the junction point is defined as:

$$x = 0 \quad \text{where} \quad N_A = N_D \qquad \text{(metallurgical junction)} \tag{14.3}$$

These two definitions are not consistent in general. Arguably, the first definition ($E_F = E_i$) may be viewed as better due to its being more logical. However, it has a problem: E_F and E_i both have temperature dependences and that their temperature dependences are not identical (LN 10), thus the junction point would be T dependent. For this reason, the metallurgical junction is the better definition. However, notice that the physics near the junction is still best understood in terms of the first definition. That is, $p \approx n$ near the junction because[3] $E_F \approx E_i$. Importantly, this is the physical basis for the starting point assumption, $\rho = e(N_D - N_A)$ in the depletion region, of the depletion approximation. Therefore one could argue that at least initially the two definitions are consistent within the depletion approximation, while they deviate from each other as the calculation proceeds (cf. the discussion at the end of Section 14.4).

14.2 Depletion Approximation

The depletion approximation makes it simple to calculate the essential electro-statistics of a pn junction. Recall that the depletion approximation means that the net charge density of the crystal ρ is given by $e(N_D - N_A)$ in the depletion region. It is necessary for us to define the spatial extent of the depletion region. That region is defined by two numbers x_p and x_n: x_p is the extent of the depletion region in the p part, while x_n is the extent of the depletion region in the n part. These length scales are primarily driven by the diffusion process.

Outside the depletion region, the crystal is charge neutral and is just like a separate p type crystal or a separate n type crystal except for an overall potential energy offset. Inside the depletion region, there *is* net charge in the crystal. Note that in the previous

[3] One might say that $p \approx n$ follows from the equations of the slide 3 of LN 10. This is incorrect. The underlying assumption there was a charge neutrality, which is not valid locally near the junction!

lecture and in HW 7 we learn that the majority carriers ensure the charge neutrality, for a common n type or p type semiconductor. Note that the depletion region is where the semiconductor basically acts like an intrinsic semiconductor (or a "compensated" semiconductor in the case of a graded junction, cf. p60 of the textbook). Therefore, the charge neutrality cannot be realized in the depletion region.

Here, then, the precise statement of the depletion approximation is given. The net charge density of the crystal is given by

$$\rho = e(N_D - N_A) \qquad -x_p < x < x_n \tag{14.4}$$

$$= 0 \qquad \text{otherwise} \tag{14.5}$$

where we continue to work under the assumption that all donors and acceptors are ionized ($N_D^+ = N_D$ and $N_A^- = N_A$).

The regions where the net charge density is not zero is called the **depletion region** or the **space charge region**.

The depletion approximation is the main assumption that we apply for our analysis here.

14.3 Gauss's Law

The Gauss law is the following:

$$\nabla \cdot \vec{E} = \frac{\rho}{K_s \epsilon_0} \tag{14.6}$$

Here, we encounter the divergence operator again, $\nabla \cdot \vec{E} \equiv \partial E_x/\partial x + \partial E_y/\partial y + \partial E_z/\partial z$. For other symbols, ϵ_0 is the vacuum permittivity, and K_s is the static dielectric constant (I used the symbol ε to mean the same thing)[4]. One note is that the dielectric constant is generally a function of the frequency of the light, and here we are concerned with the static situation (i.e. zero frequency or the DC limit), and thus the subscript s.

While the Gauss law involves three dimensional differential calculus, as presented

[4]Note that the dimension of $K_s\epsilon_0$ is charge / (voltage · length), i.e. capacitance per length. Its value for Si is 1.045 pF/cm.

above, our interest here is a one dimensional problem. This simplifies things greatly.

$$\frac{dE}{dx} = \frac{\rho}{K_s \epsilon_0} \tag{14.7}$$

$$E(x) = E(x_0) + \int_{x_0}^{x} dx' \, \frac{\rho(x')}{K_s \epsilon_0} \tag{14.8}$$

It would be helpful, but not essential, if you had some exposure to Gauss's law already and did some basic examples. If not, do not worry – the *pn* junction problem currently in our hand is actually an excellent basic example of applying the Gauss law!

14.4 Step Junction

By a step junction, we mean a *pn* junction made by joining a p type crystal and an n type crystal (two two crystals are usually the same kind of crystal, such as Si, doped differently) at a sharp interface. The epitaxial methods such as MBE (molecular beam epitaxy) or MOCVD (metal-organic chemical vapor deposition) techniques are used to make a step junction.

The depletion approximation for a step junction is then:

$$\rho(x) = \begin{cases} -eN_A & -x_p < x < 0 \\ eN_D & 0 < x < x_n \\ 0 & \text{otherwise} \end{cases} \tag{14.9}$$

The first thing to realize is that

$$N_A x_p = N_D x_n \tag{14.10}$$

This is quite fundamental. It means that the charge is conserved globally. This equation corresponds to the **global charge conservation/neutrality**.

The second thing to realize is that the right end of the crystal ($x = \infty$) will be indistinguishable from a lone n type device and that the left end of the crystal ($x = -\infty$) will be indistinguishable from a lone p type device. Why? This is because of the following reason. As the p part and the n part "mingle with each other" by exchanging charge carriers at the junction, a net local charge density develops, as modeled above. However, at distances much greater than x_p, x_n, it is clear that the

junction will exert no electrostatic force, as the junction will look like an object with no net charge from a large distance[5] That is,

$$E = 0 \quad \text{at the right end } (x = \infty) \text{ and the left end } (x = -\infty) \tag{14.11}$$

So, the values of the potential function at left end and at right end are constants. What is crucial to notice is that, while a constant potential is not significant, here we are dealing with two constants, and the difference between two constants *is* important physically! The difference between the potential function at right end and the potential function at left end is what we call the **built-in potential** or the **contact potential**. The book uses the symbol, V_{bi}, for it, and we will do the same here.

The third thing to realize is that the electron system and the hole system are separately in their own equilibrium – there is no net current in each channel.

$$J_n = e\mu_n nE + eD_n \frac{dn}{dx} = 0 \tag{14.12}$$

$$J_p = e\mu_p pE - eD_p \frac{dp}{dx} = 0 \tag{14.13}$$

Using the Einstein relation[6] ($D/\mu = k_B T/e$), we get

$$enE + k_B T \frac{dn}{dx} = 0$$

$$epE - k_B T \frac{dp}{dx} = 0$$

which leads to

$$E = -\frac{k_B T}{e} \frac{d\ln n}{dx} = \frac{k_B T}{e} \frac{d\ln p}{dx} \tag{14.14}$$

It should not be surprising that $d\ln n = -d\ln p$. That can be derived from $np = n_i^2$. Using this result, note that the built-in potential can be obtained as $V_{bi} = \int_{-\infty}^{\infty} -E dx$. The result is

$$V_{bi} = \frac{k_B T}{e} \ln\left(\frac{n(\infty)}{n(-\infty)}\right) = \frac{k_B T}{e} \ln\left(\frac{p(-\infty)}{p(\infty)}\right) \tag{14.15}$$

[5]Of course, the dipole moment can and does exist for that object but the electrostatic force will be vanishingly small at a large distance. Moreover, our one dimensional approximation actually means that, at a large distance, the junction can be viewed as a pair of positively and negatively charged sheets of charge, which can be shown, from Gauss's law, to produce no electric field at all, except within the gap between the two sheets of charges.

[6]You may realize that the Einstein relation was *derived* previously from these equilibrium conditions! In that derivation, the formula $n = n_i \exp(\beta(E_F - E_i))$ and $p = p_i \exp(\beta(E_i - E_F))$ were used as inputs, with $dE_i/dx = eE$. Here, we could have used the same inputs to derive V_{bi}.

Using either expression, the following result follows (e.g. $n(\infty) = N_D$ and $n(-\infty) = n_i^2/p(-\infty) = n_i^2/N_A$):

$$V_{bi} = \frac{k_B T}{e} \ln\left(\frac{N_D N_A}{n_i^2}\right) \tag{14.16}$$

As is evident from this formula, the built-in potential is an *intrinsic* property of a *pn* junction, determined completely by the dopant densities and the temperature. The contact potential or the built-in potential is the "total band bending amount" that results due to the matching of the Fermi level.

Now, all we need to do is to integrate $\rho(x)$ twice. Using Gauss's law and Eq. 14.11, we get

$$E(x) = \begin{cases} 0 & x < -x_p \\ (K_s\epsilon_0)^{-1}(-e)N_A(x + x_p) & -x_p < x < 0 \\ (K_s\epsilon_0)^{-1}(-eN_A x_p + eN_D x) = (K_s\epsilon_0)^{-1}eN_D(x - x_n) & 0 < x < x_n \\ 0 & x > x_n \end{cases} \tag{14.17}$$

Notice that the electric field $E(x)$ is a continuous function of x, while $\rho(x)$ was a discontinuous function of x. Integration make it nice. As long as $\rho(x)$ remains finite, $E(x)$ is a continuous function[7].

Excellent! Now, we can figure out the electrostatic potential function V as in $E = -dV/dx$. All we need to do is to integrate E and negate it.

$$V(x) = \begin{cases} 0 & x < -x_p \\ (K_s\epsilon_0)^{-1}eN_A\frac{(x+x_p)^2}{2} & -x_p < x < 0 \\ (K_s\epsilon_0)^{-1}e\left(N_A\frac{x_p^2}{2} + N_D\frac{x_n^2}{2} - N_D\frac{(x-x_n)^2}{2}\right) & 0 < x < x_n \\ (K_s\epsilon_0)^{-1}e\left(N_A\frac{x_p^2}{2} + N_D\frac{x_n^2}{2}\right) & x > x_n \end{cases} \tag{14.18}$$

Here we defined, arbitrarily, $V(x \to -\infty)$ as zero. Independent of that arbitrary convention, we have

$$V_{bi} \equiv V(\infty) - V(-\infty) \qquad \text{(only with zero bias!)} \tag{14.19}$$

$$= (K_s\epsilon_0)^{-1}\frac{e(N_A x_p^2 + N_D x_n^2)}{2} \tag{14.20}$$

In this equation only x_p and x_n are unknowns since V_{bi} is already expressed in terms of fundamental material parameters in Eq. 14.16.

[7] If $\rho(x)$ is singular at some point of x, with a "Dirac-delta function" singularity, representing a finite *surface* charge, then $E(x)$ will become discontinuous at that point.

Now, we are ready to solve for x_n and x_p to complete the solution. Combining Eqs. 14.10, and 14.20, one can solve for two unknowns x_n and x_p. The mathematics, which is solving a quadratic equation, is basic. Here, we simply state the results.

$$x_p = \sqrt{\frac{2K_s\epsilon_0}{e} \frac{N_D}{N_A(N_A+N_D)} V_{bi}} \tag{14.21}$$

$$x_n = \sqrt{\frac{2K_s\epsilon_0}{e} \frac{N_A}{N_D(N_A+N_D)} V_{bi}} \tag{14.22}$$

$$W \equiv x_p + x_n = \sqrt{\frac{2K_s\epsilon_0}{e} \left(\frac{N_A+N_D}{N_AN_D}\right) V_{bi}} \tag{14.23}$$

Here, W is the total width of the depletion region. Note that for the more highly doped side contributes less to W. This makes sense since the primary cause of the depletion region is the diffusion of the majority carriers to the other side, where they turn into minority carriers and will be subjected to eh-pair annihilating interactions (Section 12.6).

Fig. T5.9 is a nice summary of the pn junction electrostatics.

Obtaining these solutions gives us a good feeling. On the other hand, if it gives a "how can it be this easy?" kind of feeling, then you are quite right also. Why? Well, the solutions that we obtained are *not* self-consistent, although we are kind of aware of this inconsistency. Notice that we started from $\rho(x)$ and then ended up with $V(x)$. With $V(x)$ we have the full knowledge of how the band bends, since each of $E_i(x)$, $E_c(x)$ and $E_v(x)$ is basically $(-e)V(x)$ plus a constant (which is different, of course, depending on which of E_i, E_c, E_v we are dealing with). So, knowing $V(x)$, we have, in turn, the full knowledge of $E_i(x), E_c(x), E_v(x)$. We can also figure out what the value of E_F is, since that can be determined at one end of the device, i.e. in any uniform device, as explained in LN 10, and once it is known on one end it applies everywhere else in the device due to the equilibrium condition. E_F can be determined from N_D, N_A and T. What does all this mean? It means that we can now calculate *exactly* what $p(x)$ and $n(x)$ are by locally applying $n = n_i \exp(\beta(E_F - E_i))$ and $p = p_i \exp(-\beta(E_F - E_i))$. When we calculate $p(x)$ and $n(x)$ and then calculate the net charge $\rho(x) = e(p - n + N_D - N_A)$, we will for sure find that $\rho(x)$ is not as simple as we assumed above in Eq. 14.9. It will still feature a discontinuity at $x = 0$, but it will no longer feature discontinuities at $x = x_n$ and $x = -x_p$. This is because the progression from an intrinsic semiconductor behavior (at the junction) to a normal p type (left) or n type (right) behavior should show a rounded crossover behavior rather than an abrupt step like change. This rounded crossover behavior would introduce a correction to our starting point assumption, making our solution self-inconsistent. For the purpose of this course, such a correction term can be deemed

small enough to be ignored. However, what would we do if we wish to ensure the self-consistency of our solution? Here is a systematic approach: an "iterative method" or a "perturbation method" that can assure the self consistency starting from our solutions above[8]. (1) Assume $\rho^{(0)}(x)$, like our depletion approximation formula. (2) Calculate the output $V^{(0)}(x)$ using $\rho^{(0)}(x)$ as input. As shown above, this involves integrating $\rho^{(0)}(x)$ twice. We call $\rho^{(0)}(x)$ and $V^{(0)}(x)$ as the "zero-th order solution." (3) From $V^{(0)}(x)$, calculate a new input $\rho^{(1)}(x)$, as outlined in this paragraph. (4) Then, by integrating $\rho^{(1)}(x)$ twice, obtain $V^{(1)}(x)$. $\rho^{(1)}(x)$ and $V^{(1)}(x)$ are our "first order solution." (5) Repeat steps (3,4) to get the n-th solution $\rho^{(n)}(x)$ and $V^{(n)}(x)$ from the $(n-1)$-th order solution $\rho^{(n-1)}(x)$ and $V^{(n-1)}(x)$ with $n = 2, 3$ and on and on. When $\rho^{(n)}(x)$ and $\rho^{(n-1)}(x)$ are sufficiently close to each other, the self-consistency of the solution is attained and the iteration can be stopped.

14.5 Step Junction, Biased

Consider a step junction that is biased (Figure T5.10). When the voltage applied to the p end is higher, we call it a "forward bias" and use a positive value for the net bias voltage V_A. The opposite case is the "reverse bias." The applied voltage V_A will have to drop somewhere in the device. A constant current flows. Where would the voltage drop the most? Considering the pn junction device as a locally resistive device, with resistance changing as a function of position, it is obvious that the more resistive part will have the largest drop of the voltage. Recall that the conductivity is given by $\mu_n en + \mu_p ep$. So, the resistivity will be the maximum when n and p are both very small. That is the junction region, where the device approximately acts like an intrinsic semiconductor. Outside this junction, the resistance is low and when the current is small, the voltage drop there is negligible. Therefore, a biased junction device can be approximated as an unbiased junction device with the built-in voltage modified to $V_{bi} \to V_{bi} - V_A$! For a forward bias, the potential difference between the p part and the n part is decreases. For a reverse bias, the potential difference increases, i.e. the band bending becomes greater.

Here is a summary of what changes are to be made when a bias V_A is applied, within this idealized "low-current" approximation.

1. Eqs. 14.21–14.23 will remain valid if and only if V_{bi} is changed to $V_{bi} - V_A$.

[8]Regardless of the actual methods of getting the output from the input and calculating the new input from the output, the general algorithmic structure outlined here is of general importance for science or engineering problems.

2. Eqs. 14.9,14.10,14.17,14.18 will remain valid if and only if the new values of x_p and x_n, obtained from the previous step, are taken into account.

3. Eqs. 14.11 remains valid.

4. In a strict sense, Eqs. 14.12–14.14 are not valid any more, since **the system is no longer in equilibrium when an external bias is applied**. Instead, the electron system or the hole system will each exhibit a finite current. Also, the concept of the Fermi level is not applicable any more, near the junction. **However, we shall see in the next two lectures that Eqs. 14.12–14.14 gain important approximate validity in a controlled approximation scheme in the biased case.**

5. Eqs. 14.15 and 14.16 remains valid as is (*no* changing V_{bi} to $V_{bi} - V_A$!). These formulae could in fact be taken as the more generalized *definition* of V_{bi}.

6. Eqs. 14.19,14.20 will remain valid if and only if V_{bi} is changed to $V_{bi} - V_A$ and the new values of x_p and x_n are used. The modified form of Eq. 14.19, $V_{bi} - V_A = V(\infty) - V(-\infty)$ is the *definition* of what we mean when we say "we apply a bias V_A to the device."

Here is a qualitative summary of what happens. The extent of the depletion regions scales like $\sqrt{V_{bi} - V_A}$, becoming narrower for a forward bias and becoming wider for a reverse bias. As the net charge density remains the same, it follows that the total charge accumulated on either side of the junction increases for a reverse bias and decreases for a forward bias. Last but far from the least, now there is a current, which is basically what this device is all about.