Solid State Devices
Dr. S. Karmalkar
Department of Electronics and Communication Engineering
Indian Institute of Technology, Madras
Lecture - 7
Equilibrium Carrier Concentration (Contd…)

Today this is the 7th lecture of the course and the 5th lecture on Equilibrium Carrier Concentration.

(Refer Slide Time: 01:16 - 01:25)

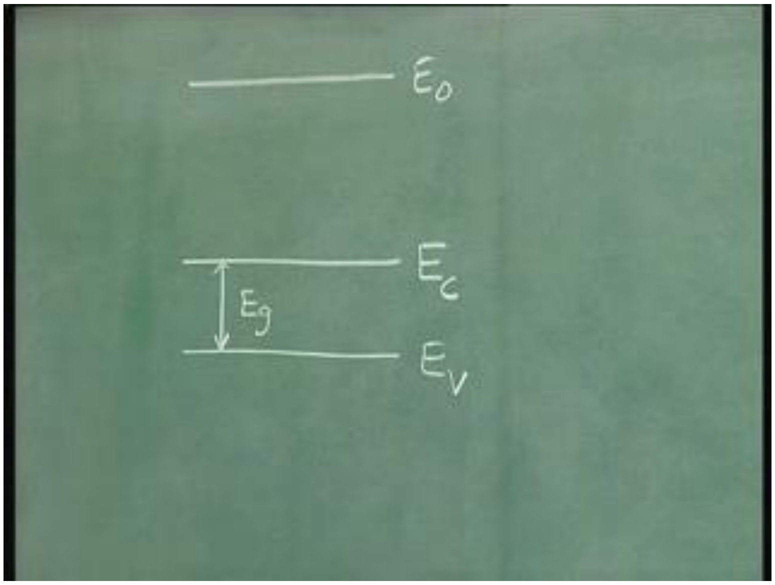
Solid State Devices

Lecture - 7

Equilibrium Carrier Concentration

Yesterday we have started on the energy band diagram or energy band model of the semiconductor which will give us an accurate formula for estimating the carrier concentration at any temperature in an intrinsic semiconductor. Specifically we have discussed the first tenet of this Band Model. That is, what are the allowed states for electrons in a crystal? So the allowed picture is something like this.

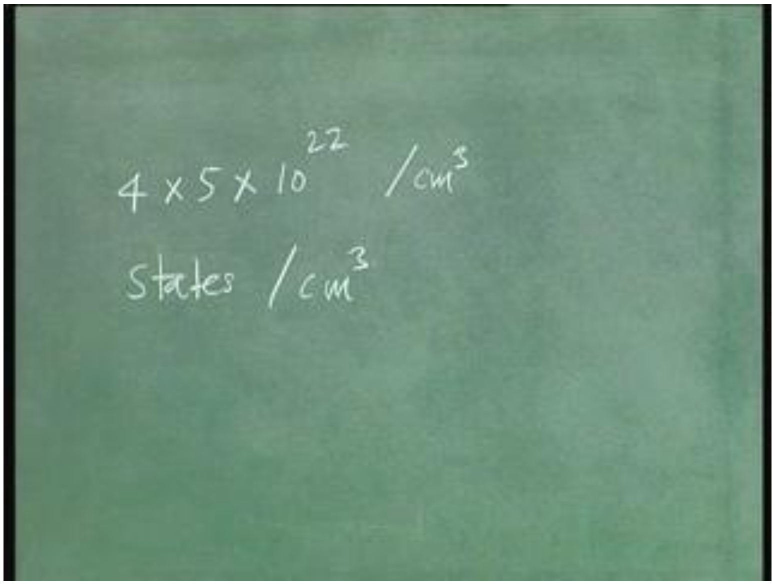
(Refer Slide Time: 01:39)



You have a conduction band separated by a valance band. The bottom of the conduction band is E_C and the top of the valance band is E_V, and here for reference you have the vacuum level. This difference, E_C minus E_V is the energy gap. So electrons are allowed. energy is in this conduction band, energies in this valance band and then below you have other bands.

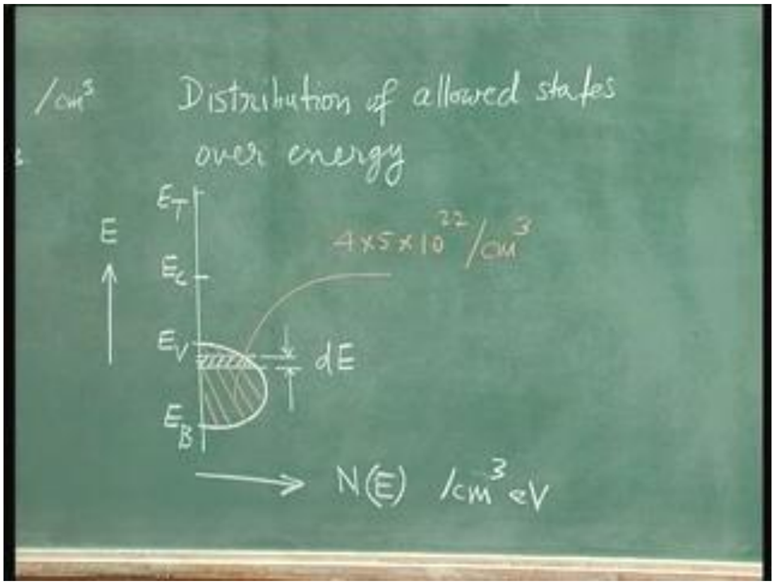
Now the question is what is the distribution of the available states over these allowed energies? Now, you know what the allowed energies are. So, how the allowed states distributed over energy? What is their number? When we look at the valance band and the conduction band, the electrons which are coming here, in these bands, are basically coming from the four valance electrons of each of the atoms in the crystal. In other words, the total number of states corresponding to valance band would be minus 5 into 10 to the power 22 by cm cube into 4 valance electrons corresponding to each of the atoms.

(Refer Slide Time: 02:58)



These are the number of allowed states for electrons. So many states per centimeter cube (states by cm cube).

(Refer Slide Time: 03:38)



Now, we want to see how these are distributed over allowed energies, distribution of allowed states over energy. We will just state the result from Quantum Mechanics. The allowed states are distributed as follows: if we take this y axis to indicate the allowed energies and the x axis to indicate the distribution of the states over energy, N(E), let us say this is E_C and this is E_V. Let us say, the bottom of the valance band is called E_B. Let us assume the top of the conduction band is called E_T. So E_B to E_V this is one allowed band. These 4 into 5 into 10 to the power 22 electrons are equal number of states. It is found that they are distributed as follows: the distribution function looks something like this. What does it mean? This means that if I have to take an energy interval dE, somewhere here, energy interval dE at some energy E, then the number of states I will find in this energy interval would be given by this shaded area.

If I integrate this particular function over this interval which means if I take this total area under this particular function, then that area will be equal to 4 into 5 into 10 to the power 22. So you have a small number of states near the edges of the band. As you move towards the middle of the band you will have a larger number of states within an energy interval. So the area here, this total area is 4 into 5 into 10 to the power 22 by cm cube. If I take one centimeter cube of silicon and I try to see the total number of states they will be this many. And for these states, if I try to see their energy distribution then I will find that this is the energy distribution. So what is the unit of N(E)? If integral N(E) dE, this is N(E) and this is E so you are integrating along this. So if integral N(E) dE has a unit of centimeter per cube (cm cube), it means that N(E) has the unit of per cm cube per eV per energy. This point is very important to note. So the available states are not uniformly distributed over energy, is what we find. Often, when we plot functions, we plot the independent variable on the x axis and the dependent variable on the y axis.

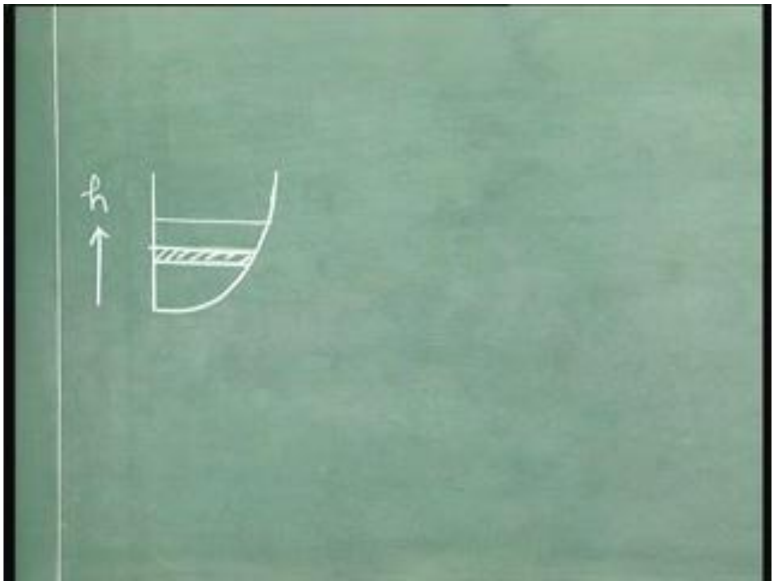
That is, if I were to plot N(E) as a function of E, I would do something like this: N(E) here and E on this axis, which is the normal procedure that we adopt in which case the function would look something like this. This is E_B and this is E_V and this is the area that you are talking about. But here we have plotted the energy on the y axis and N(E) on the x axis.

There are some reasons for following this convention in semiconductor Physics. One is, like you start from here the allowed energy picture, you have shown the energy on the vertical axis, it is useful to align all other diagrams to this diagram. This is one reason why we always choose energy on the vertical axis while plotting N(E). And we will later on see when we plot the Fermi-Dirac Function also we will plot that function on the x axis and energy on the y axis.

Now there are strong reasons for plotting energy on the y axis. This is because intuitively, for us, if you talk of potential energy for example: a body which is at a greater height has higher potential energy so somehow we are used to this kind of a convention when we are talking of energy. So energy is a physical quantity. We are not simply talking of some mathematical entity, it has a physical significance. Higher the location of the body more the potential energy and therefore it makes sense to plot the energy on the y axis.

Location means higher and higher energy. This is the reason why you want to align other diagrams with this allowed energy band diagram and therefore the energy is plotted on the y axis. Let us understand this particular Density of States Function - N(E). What is the meaning of this function? Now, we can take a simple analogy to understand this. Supposing you consider a bucket of water which is something like this and let us also assume that the shape of the bucket is like this.

(Refer Slide Time: 10:59)



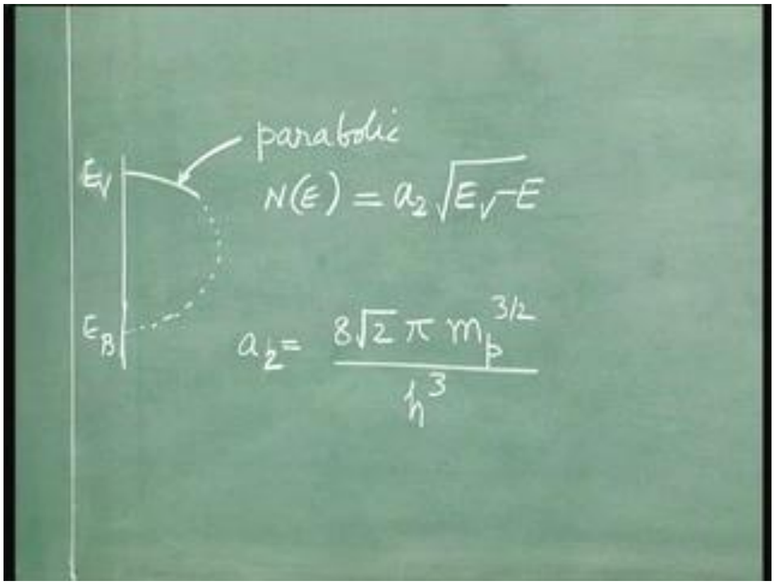
I am assuming a two dimensional situation. Supposing there is a bucket which is like this and you are filling it with water up to some height, now supposing somebody asks you what is the amount of water at any height h, this is height h, supposing I want to know the amount of water at that height, it is easy for us to see that it makes no sense. What is amount of water at any height makes no sense. What makes sense is, what is the amount of water within a interval dh at any h and then it is this, then the amount of water would be that.

Similarly, it makes no sense to ask how many states are available at any energy E. It makes no sense because the states are distributed over energy. This is the approach that is adopted normally, and then we have a large number of states to deal with or a large population. So supposing you have millions of people and you want to talk about their height distribution then it will be very difficult for you to tell at a given height h how many people are there, that is how many people have exactly height h. It is very difficult if you are talking of millions of people. Instead what you would talk about is, within 5 feet and 5 feet 1 inch height interval how many people are there? That is what makes sense.

So, similarly here this is the water in bucket analogy. It makes sense only to talk about water at any height h within an interval dh. So, similarly number of states within an energy interval dE at an energy E. So you take the energy interval dE, around E, only then it makes sense. That is why the unit of N(E) is per centimeter cube per electron volt (by cm cube by eV). I want to emphasize this point because invariably the students forget the electron volt part of this unit and which means they have not understood the physical significance of this function.

We have shown the shape of the function and so on. What is the equation? It is important for us to know the equation. It turns out that you cannot have a single equation for this function. The shape is such that you cannot represent this function accurately with one equation. It turns out that in practice what is of interest is only the function near the edge of the band. So, if you can approximate the shape of the function somewhere here between E_v and E_B that is sufficient for us. And it turns out that a simple parabolic relationship is sufficient for representing the function near the edge of the band.

(Refer Slide Time: 14:48)



$$N(E) = a_2 \sqrt{E_V - E}$$

$$a_2 = \frac{8\sqrt{2}\pi m_p^{3/2}}{h^3}$$

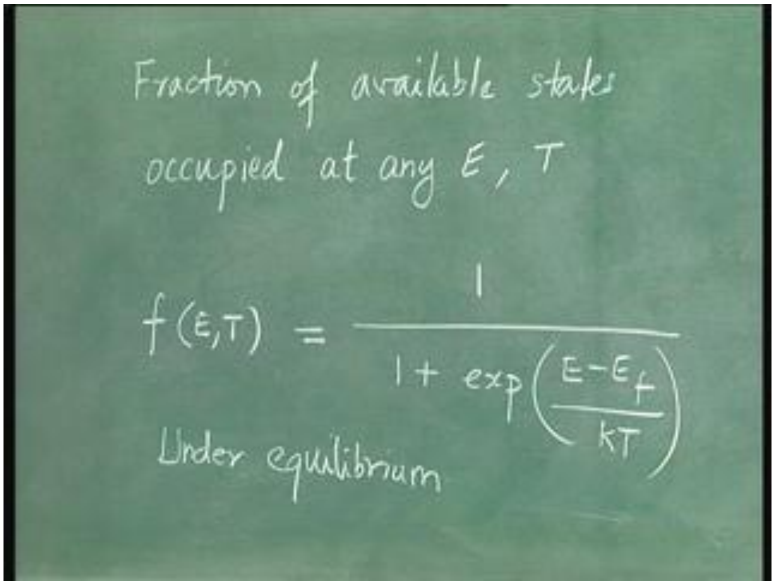
Therefore we can write that function as, I will redraw that portion here, so this is E, this is E_V, this is E_B. If I am interested in this segment, this segment is parabolic, and can be written as N(E) is equal to square root of E_V minus E that is the function.

We will write the function "a" in terms of various physical parameters later on. Let us comeback to the conduction band, the situation in conduction band is similar to that in valance band. I will not plot the exact shape of the conduction band here because that shape is not as simple as this. But again as we have said what is of interest to us is only the function near the band edge and we will see the reason to it later. So if I take the function near the band edge and remaining part I am showing by dotted line, I am not completing it because that shape is not really simple; this part of the function again is a parabolic function.

So I can write this function as constant a into square root of E minus E_C. It turns out that the constant "a" is slightly different for the conduction band than the valance band. So what we will do is, we use a_1 to represent the constant for the conduction band and here a_2 represents the constant associated with the valance band. The constants a_1 and a_2 is equal to (8 square root of 2pi m_p cube by 2) by h cube. So m_p is the effective mass of holes. Therefore for the valance band, the constant a_2 depends on the effective mass of the hole, h is the Planck's constant and the constant a_1 is equal to 8 square root of 2pi m_p is replaced by m_n the effective mass of electrons. For the conduction band the constant depends on effective mass of electrons (m_n). So that is the slight difference between a_1 and a_2: the effective mass of electrons and effective mass of holes.

Now we know how the available states are distributed over energy. Now the final question to be answered before we can estimate the concentration of electrons, free electrons or holes and so on is: if so many states are available what fraction of these states are occupied at any temperature. Once we have that information also then we can put everything together and get a formula.

(Refer Slide Time: 19:24)



Fraction of available states occupied at any E, T

$$f(E,T) = \frac{1}{1 + \exp\left(\frac{E - E_f}{kT}\right)}$$

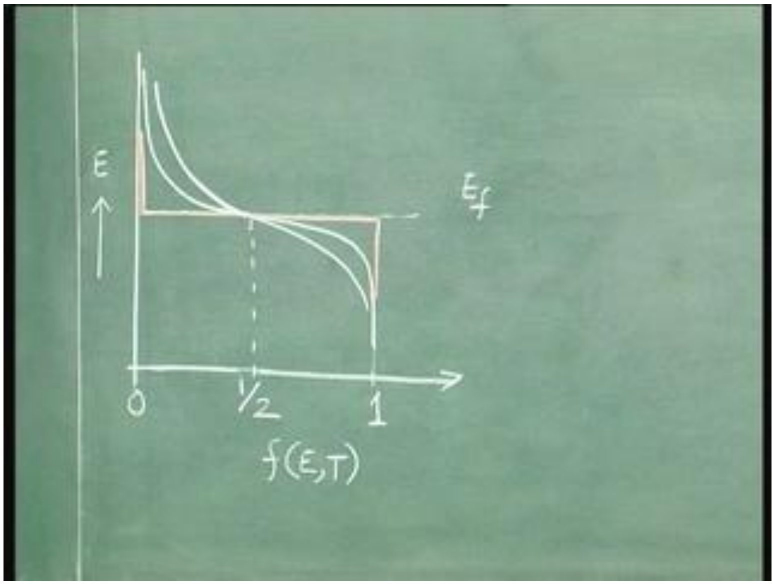
Under equilibrium

We come to the third tenet: that is fraction of available states occupied at any energy and temperature. Remember that whenever we talk of available state at any energy E, it is understood that it is within an interval dE so states available within an energy interval dE at E. Normally within an energy interval dE this phrase is dropped but it is to be always remembered. Supposing there is a state available at any energy interval what fraction is occupied? This is the Fermi-Dirac fraction and this is given by the function f(E, T) is given by 1 by (1 plus exponential (E minus E_f by KT)).

It is extremely important to note that this function or this fraction is derived under equilibrium conditions. What does it mean? It means that you have assumed, while deriving this fraction that if some electrons have been transferred form energy level E_1 to E_2 at some rate then electrons are being transferred from E_2 to E_1 at the same rate; that is a meaning of equilibrium. So using these ideas this fraction has been arrived at.

There are some additional constrains also like: in quantum mechanical system so that no two electrons can be present in any energy and so on. But where is the equilibrium coming from is what I am trying to tell you using this function. Now here this particular symbol E_f stands for Fermi-level, so this is a Fermi-Dirac fraction f(E, T) is equal to [1 by 1 plus exponential (E minus E_f by KT)]. It is useful to see how this fraction looks like when you plot it. So again when we plot, we will choose the energy on the y axis and the function on the x axis.

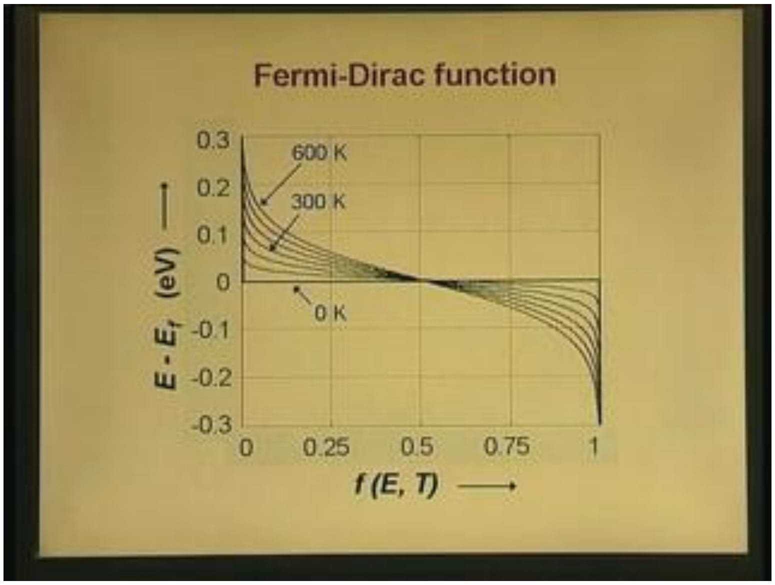
(Refer Slide Time: 22:39)



So this is energy (E) and this is the function (f). So the function would look something like this. If this is E_f for very high energies for E tending to infinity the fraction goes to 0 and let us assume T greater than 0. For E is equal to E_f, this is the exponential of 0, this is 1. The fraction is equal to 1 by 2. So for E very large the fraction is 0. I start from 0 here and as I come down in energy, by the time I come to E_f this value is 1 by 2.

If I proceed further as I go to E minus infinity, this term will go to 0 and the fraction will go to 1. So the maximum value of the fraction is 1. I will get an exactly symmetric segment for any one temperature. This is how the function looks like for any one temperature. How does the function look like for different temperatures? Let us look at it on a graph.

(Refer Slide Time: 24:52)



So you can see here that for higher and higher temperatures, here the curves have been plotted for intervals of 100k starting from 0k. At T is equal to 0 what is happening? Until you come equal to E_f the fraction is 0. That is because, you see here that, if E is greater than E_f and T is equal to 0. $T \rightarrow 0$; this is infinite. No matter of what your E is so long as it is more than E_f. On the other hand, for E is lesser than E_f, for T is equal to 0 it is always minus infinity, so this term is 0. There is an abrupt change so that is what you find here. The function comes down and at E_f it is a constant and then it comes down again for E is lesser than E_f. So you can see that what is being plotted here is E minus E_f. So E minus E_f is equal to 0 so these are Fermi-level location. Then when you increase the temperature in steps of 100k you progressively find the function shifting up.

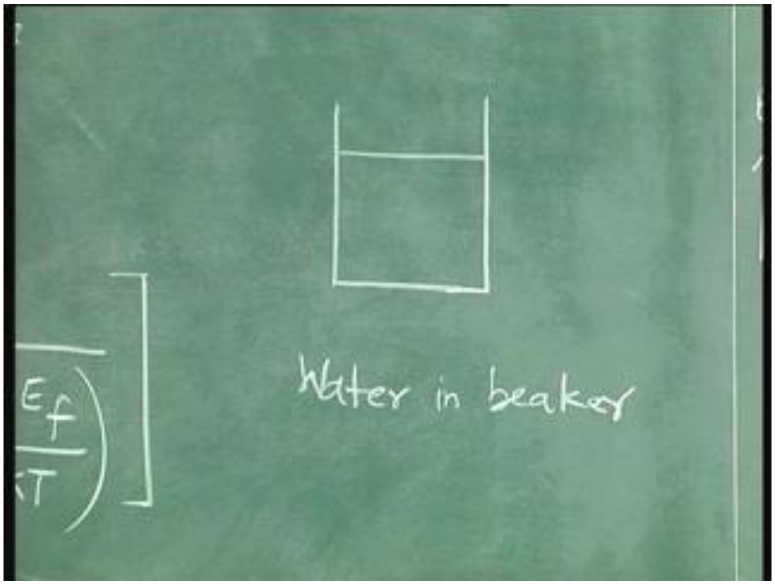
Now what does this function indicate? What does this kind of shape indicate? It is very simple. It means if I have to plot for T is equal to 0k and compare, all that this function says is that electrons from lower energies are moving to higher energies as you increase the temperature.

So, for example, here if you see below E_f, at T is equal to 0 there were electrons and above E_f there were no electrons because even if they were available states, according to this the fraction occupied is 0 which means no electrons are there in those states; the states are unoccupied. When you increase the temperature you can see that the fraction which was 0 now has become non-zero. Whereas here the fraction which was occupied, fraction was 1, it has slightly reduced to less than 1. This means electrons from lower energies are moving to higher energies, that is all what this function shows.

If your function temperature increases further you will have further distortion in this. So, more electrons are moving from lower energies to higher energies. It is important to note that this function does not necessarily say that you have an allowed energy at E. Whether there is an allowed energy state at an energy E is decided by the density of states function and allowed energy levels.

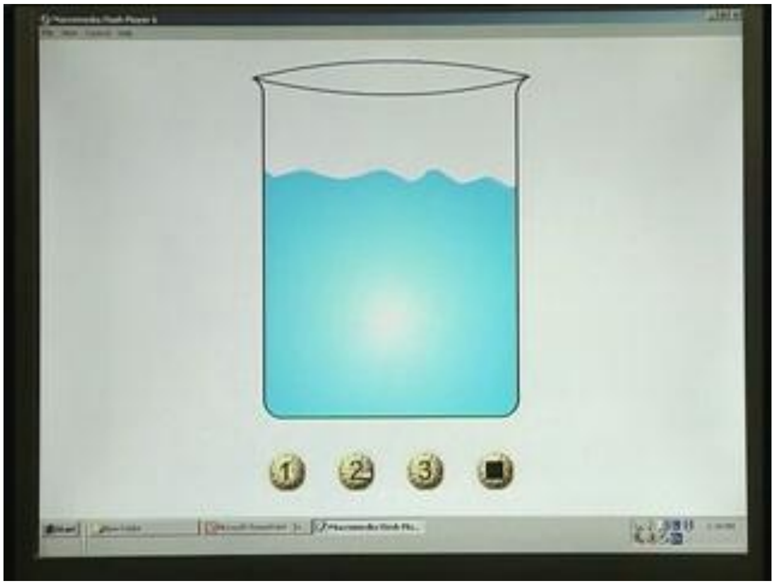
If there is an allowed state and there are states within an energy interval dE around that point then what fraction of this states are occupied is what is given by this function, f(E, T) is equal to [1 by 1 plus exp (E minus E_f by KT)] at any temperature T. So to understand the shape of this function, one can actually again consider an analogy. Again you have water in beaker analogy for this.

(Refer Slide Time: 28:50)



Let us say you have a beaker which is filled with water. Supposing you agitate this beaker and now try to see how the surface of the water in the beaker is, so there is an animation to show this.

(Refer Slide Time: 29:35)



Here the top of the water surface is calm, so there is no agitation. Now when you agitate the beaker you find a disturbance and what is happening there is some amount of water from below the surface is going above the surface in some place, and some other place there is a void from where the water is transferred. You increase the agitation, there is more likelihood and we will find water above the top surface corresponding to the calm situation. And we increase this agitation even more, let me show you the mean surface here and this is the agitated function.

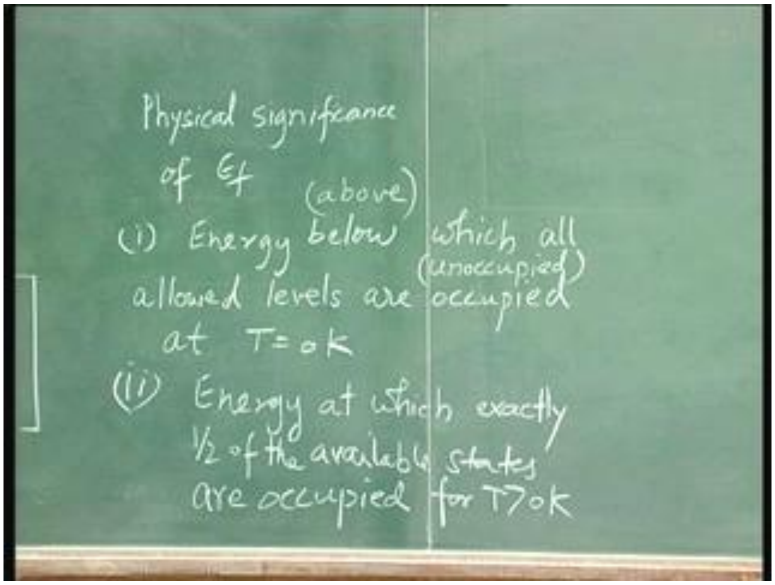
So what we are saying here is, if this is the picture when no agitation is present, when agitation is present it would be something like this; so water from here has gone here. Now when we increase the agitation it could be something like this. If I were to now plot, I replace this energy by height, and instead of the Fermi-level I have called this as Mean Top Surface of water. And what I am plotting here is the fraction of the beaker volume occupied as a function of height and agitation.

Let us say height first because there was E first so E is analogous to height; f stands for fraction of beaker volume occupied. Then in fact you will have a picture as something like this where there is increasing agitation and your plot would look something like this. So this is for zero agitation (the red line is for zero agitation). When there is zero agitation the water surface is like this. Let us say this is height h_0, so above h_0 we have no water and below h_0 the entire beaker volume is occupied with water. So the fraction of beaker volume occupied is 1 below h_0 and above h_0 it is zero. So there is no agitation, the red line is no agitation and that is the difference.

As we increase agitation you find some water above that surface and therefore the fraction of water present is non-zero above the surface. But when you go higher and higher in height the probability of finding water is going to be lesser. If we increase agitation we are likely to find more water above height h. This is an analogy where the water in the beaker is analogous to electrons; the beaker volume is analogous to the allowed energies; the height is analogous to the energy; agitation is analogues to temperature; the fraction of beaker volume is analogous to the fraction of occupied states.

Finally, what is very important is, this mean top surface of water is analogous to the Fermi-level. These kinds of analogies help us to appreciate the physical significance of these ideas which are sometimes very abstract. Now let us focus on this E_f a little bit more. How do we interpret E_f? There are two ways of interpreting E_f. One is in terms of the behavior at T is equal to 0.

(Refer Slide Time: 35:31)



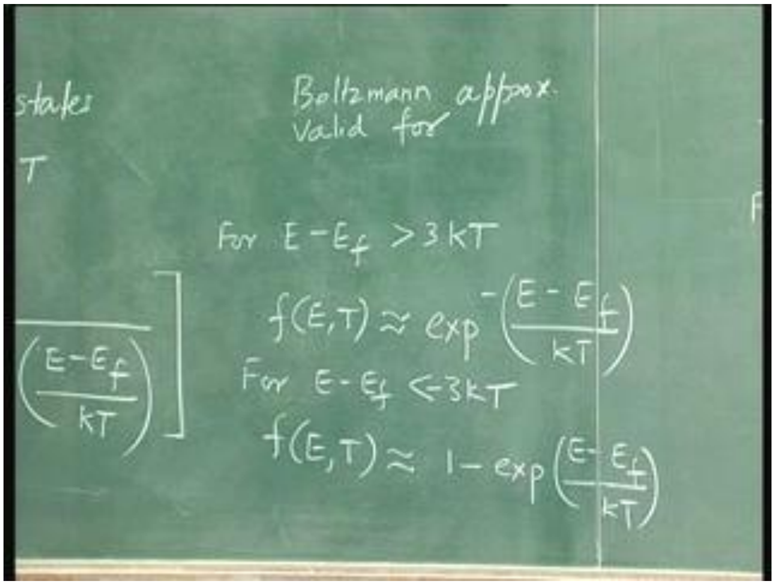
Physical significance
of E_f
(above)
(i) Energy below which all
(unoccupied)
allowed levels are occupied
at $T = 0K$
(ii) Energy at which exactly
$1/2$ of the available states
are occupied for $T > 0K$

In terms of the behavior at T is equal to 0 we can say Physical Significance of E_f. In terms of the T is equal to 0 behavior we can say that it is that level above which no allowed level is occupied and below which all allowed levels are occupied. It is that energy below which all allowed levels are occupied. And I can replace "below" by "above" here and here "occupied" by "unoccupied". So it is an energy above which all allowed levels are unoccupied or none of the allowed levels is occupied. This is one way of interpreting E_f. What is the meaning of E_f at T is equal to 0? This is very important. This interpretation is valid for T is equal to 0.

Let us look at that function. At T is equal to 0, we can see that the function is 0 until you come to E_f and suddenly it changes to 1. Alternately you can interpret in terms of the behavior for T greater than 0. At T greater than 0 if there is an allowed level at E is equal to E_f then exactly half of the number of states which are available will be occupied at E is equal to E_f at T greater than 0. So it is energy at which exactly half of the available states are occupied at T greater than 0.

Now, we will see when we take up of PN junctions, the Fermi-level has other physical significance also. So it is a very important parameter in the analysis of electronic systems under equilibrium conditions. Now we can put our information together namely the allowed energy states, the fraction of states occupied at any energy E, and the density of states function and we can try to derive an equation for the carrier concentration, the intrinsic concentration. Now when we do that or rather before we do that let us discuss one approximation of this function that is often used to derive simple formula and that is called a Boltzmann approximation of this function.

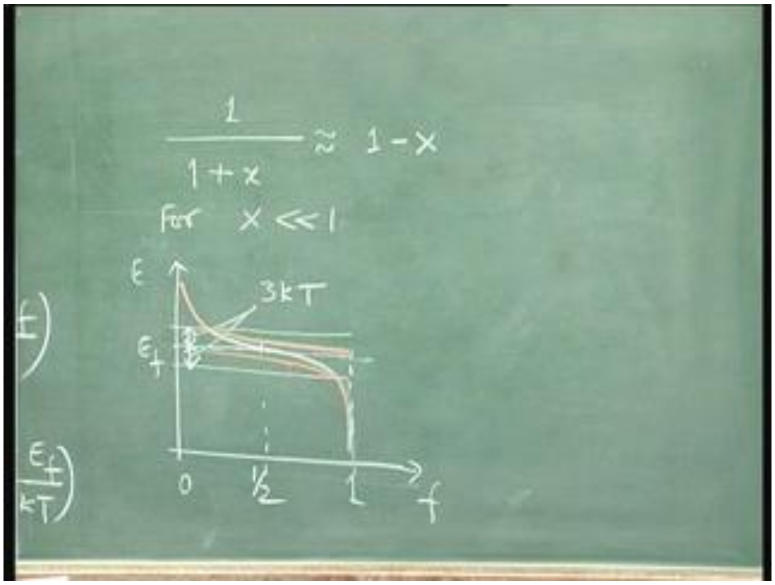
(Refer Slide Time: 39:49)



Boltzmann approximation: This is valid for E minus E_f by KT is greater than 3 or is lesser than minus 3. E minus E_f by KT is either is greater than 3 or less than minus 3. That is E is above E_f by more than 3KT or E is below E_f by more than 3KT. In such a case, this approximation holds. Now, what happens for this case: if E minus E_f by KT is more than 3 then this term [exponential (E minus E_f by KT)] is much larger than 1 and therefore you can neglect 1, and your function will become f(E,T) is approximately equal to exponential minus (E minus E_f by KT) for E minus E_f is greater than 3KT.

Alternately for E minus E_f is lesser than minus 3 KT, if E is greater than E_f by more than 3 KT then this is exponential of minus 3 or exponential of minus of a quantity more than 3 then this quantity becomes much less than 1. If that is a case, you know that 1 by 1 plus x, where x is much less than 1 is (1 minus x).

(Refer Slide Time: 42:11)



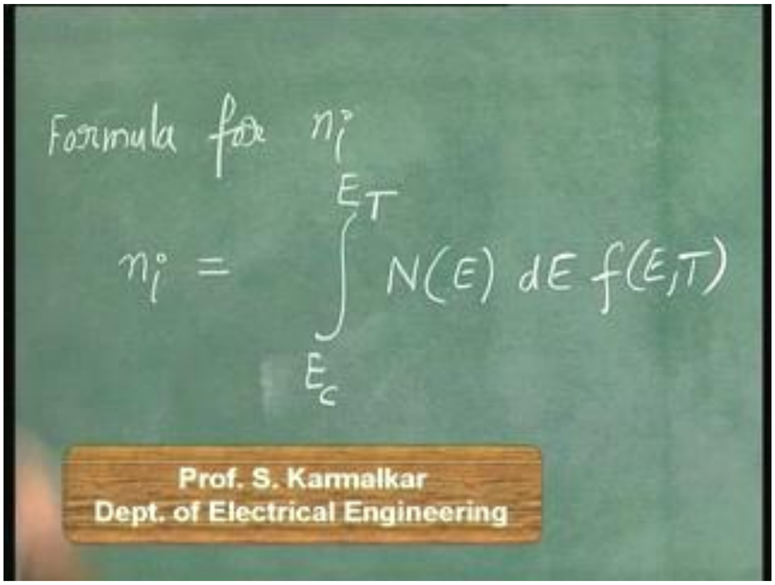
It is 1 by 1 plus x is approximately equal to 1 minus x, for x much less than 1. Using the same approximation we can write f (E, T) is approximately equal to 1 minus exponential (E minus E_f by KT); where, E minus E_f is lesser than minus 3KT that is important. So if you want to compare the Fermi-Dirac function and the Boltzmann distribution function, then how will they look like? It would be something like this. Let us do it for any one temperature, so this is E_f; this is E and this is f, the function f. If this is the Fermi-Dirac function, this is 0 and this is 1, this is 1 by 2.

Now, if you were going to sketch the Boltzmann approximation it would be something like this. It will almost be the same as this function until you come to 3KT above E_f and there it will deviate and it will go like this. The lower segment will be again something like this, so let us mark the interval this is E_f and this is 3KT below E_f, these two intervals are 3KT.

The Fermi-Dirac function and Boltzmann function will almost be the same. It turns out many times or most of the times when we use this function, Fermi-Dirac function, we use it in the range where the Boltzmann approximation is valid. Now one might ask that these are such simple approximations so why does it need Boltzmann to tell you that you can use this. The point is, what we have discussed is, the mathematical method of deriving the Boltzmann function. Now what is the physical significance of that? What exactly has changed? Which assumptions have been modified? There are some assumptions which have been used to derive the Fermi-Dirac function. Which of these functions is modified to derive the Boltzmann function? It turns out that the particular assumption modified is that the Fermi-Dirac function derivation assumes that no two electrons can occupy the same energy whereas Boltzmann derivation assumes that any number of electrons can occupy the same energy. So this is the physical difference between the two.

Now let us discuss the estimation of the concentration. Formula for $N_{i:}$ We can write N_i as the number of electrons in the conduction band at any temperature.

(Refer Slide Time: 47:09)



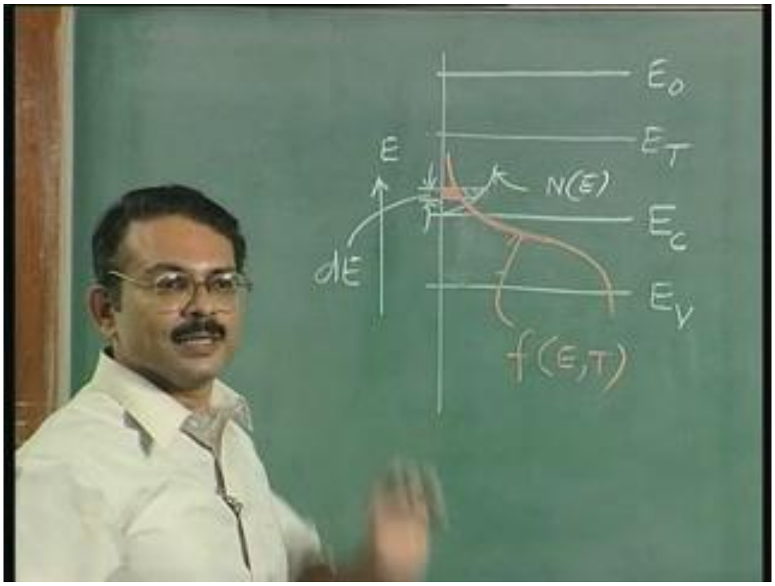
Formula for n_i

$$n_i = \int_{E_C}^{E_T} N(E)\, dE\, f(E,T)$$

Prof. S. Karmalkar
Dept. of Electrical Engineering

This means N_i is equal to integral E_C to E_T. So E_C is this level, bottom of the conduction band and E_T is the top of the conduction band. How many electrons are there? Integral, the available states function. Within energy interval dE at E the number of states available is N(E) into dE. Out of this, f(E, T) is occupied. So you integrate this over the conduction band and you should get the intrinsic concentration.

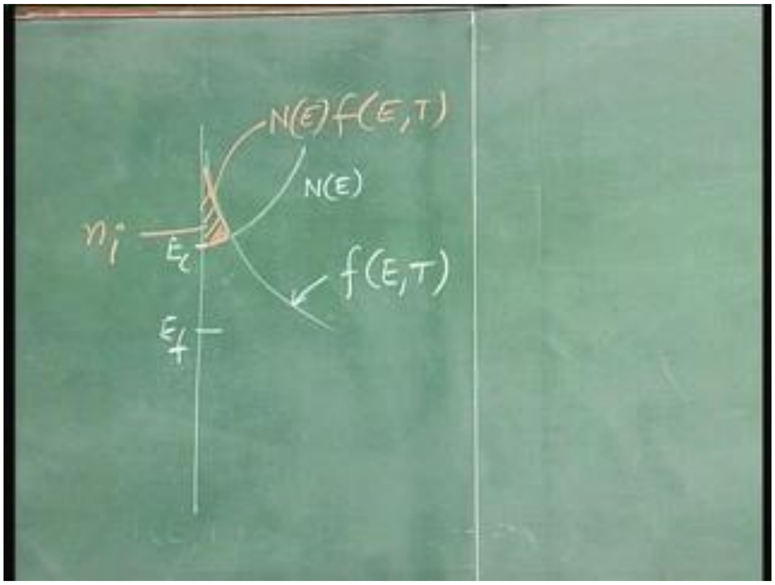
(Refer Slide Time: 48:06)



So graphically if we want to represent this it would be something like this (Refer Slide Time 48:09). So this is the N(E) function. This is N(E) into dE. A fraction of this is occupied because you have a Fermi-Dirac fraction as something like that, so this is f(E). As of now, if I want to sketch this function I should know where the Fermi-level is. I do not know, so I will assume it is somewhere; but I just want to show the fraction function. So here all the three aspects have been shown in this diagram the allowed energy levels, the distribution of available states over energy, and the fraction of state.

These are the available states, multiply this by this fraction which means what is shown in this red portion is the number of electrons in the energy interval dE. So, that shaded area here shows how many electrons are there within energy interval dE because ratio of this area red area to this white area over this interval dE is actually the fraction that is occupied. So it is this term N(E)dE f(E, T) and now you integrate you will get the total number of electrons. How does this function look like N(E) into f(E)? The function N(E) into f(E) looks something like this. It looks something like this and that area is the integral. The boundary of this red, this is N(E) into f(E, T) at any temperature.

(Refer Slide Time: 50:22)

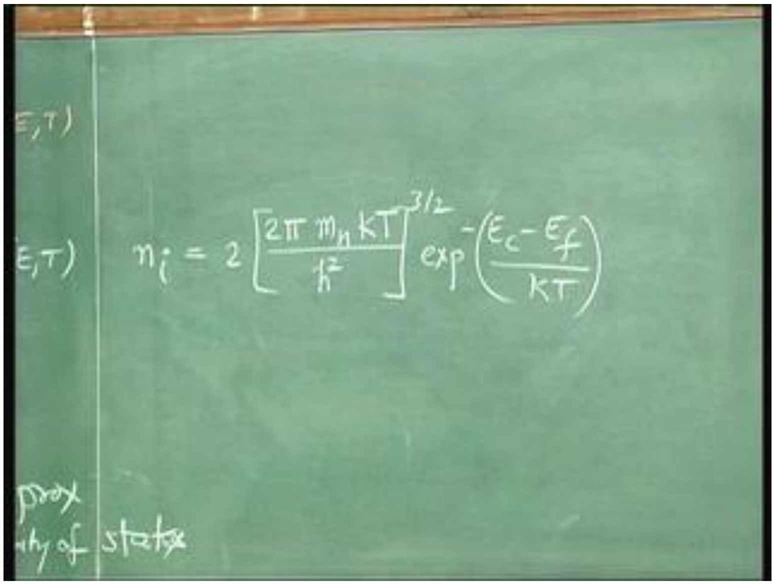


This area is n_i, Fermi-level is somewhere below, this is E_C. Right now we do not know where exactly it is, we have to estimate. You can substitute the various terms and your function will look as something like this, this is equal to integral E_C to E_T, so N(E) is of the form a_1 square root of E minus E_C, f(E) is 1 by 1 plus exponential (E minus E_f by KT) dE. You have to perform this integration. There is no way you can get a simple closed form equation for these general limits E_C and E_T so one uses some approximations.

First thing is the Boltzmann approximation. We use the Boltzmann approximation to this Fermi-Dirac function because one can show that your E_f will be below E_C by more than 3KT. Now you do not know this in advance but from intuitive reasoning one can know this. Normally the way to check such approximations is you make the approximation, derive the result, and from the result you try to see whether your E_f has turned out to be below E_C by more than 3KT. If that is so then your approximation is valid. So you make an approximation, derive the result, then cross check whether the result is in conformity with the assumption or does it contradict the assumption. So the first approximation is the Boltzmann approximation. The next approximation is, you replace E_T by infinity (replace E_T the top of the conduction band by infinity). Why does this work? This works because this function $f(E_T)$ rapidly drops to 0 much before you reach the top of the conduction band. The top of the conduction band will be somewhere here.

So it does not matter whether you are taking the limit up to E_T or to infinity. The area under this is not going to change; it has already dropped to 0 much before you reach E_T. Let us write the two approximations: First E_T turns into infinity; second Boltzmann approximation; third is this formula that we have written already, that is also an approximation that has been used, the Parabolic Density of States Approximation. The actual density of states function is more complex. We are using a parabolic density of states approximation. Under these three approximations you can get a very simple result of this and it looks like this.

(Refer Slide Time: 55:29)



$$n_i = 2\left[\frac{2\pi m_n KT}{h^2}\right]^{3/2} \exp -\left(\frac{E_c - E_f}{KT}\right)$$

It is n_i is equal to 2[2pi m_n KT by h square] whole cube by 2 into exponential of minus (E_C minus E_f by KT) with a negative sign. You are basically substituting for a_1 and of course you are making some substitutions for evaluating the integral. For example, you make a substitution E minus E_C some variable, then your lower limit will become 0, upper limit is already infinity. So how the integral looks like under these approximations? You do it as an exercise. Evaluation of the integral will involve some higher level of Mathematics but at least you can write within the simplest possible form under these approximations integral E_C to E_T a_1 square root of E minus E_C by 1 plus exp (E minus E_f by KT) dE, and then one will find that this is the relation, n_i is equal to 2[2pi m_n KT by h square] whole cube by 2 the exponential of minus (E_C minus E_f by KT). In the next class we shall interpret this formula for n_i physically. We will derive a similar formula for the hole concentration that is p_i and then using the fact that n_i is equal to p_i we will derive an equation for the Fermi-level.