Fundamentals of Semiconductor Devices

Prof. Digbijoy N. Nath
Department of Electrical Engineering
Indian Institute of Science, Bangalore

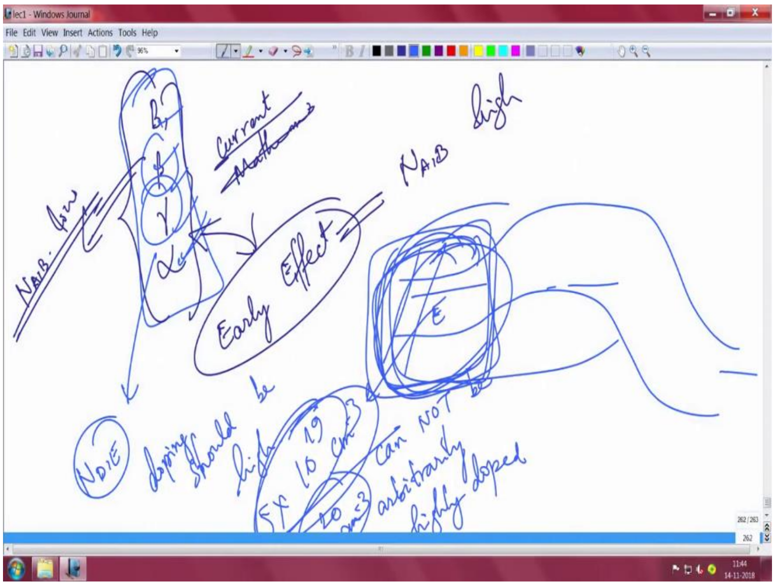
Lecture – 29
Delays in BJT

Welcome back. So, we have been discussing the topic of BJT, we have covered quite a few sub topics within BJT. We have discussed the current transport, the various factors like gain, current transport, the base transport factor, current gain and so on.

If you remember in the last class, we also have discussed about things like Gummel number and Early voltage, which are very important in a BJT, right. We have very thoroughly understood the different components of base current, Emitter collector current, how actually a BJT works and what is it that makes a BJT amplify the signal, ok. It is a current input device and we have more or less covered most of the fundamental things that are required to understand in a BJT. So, what is remaining now in BJT is delay, right. There is something called delay or how fast a BJT can operate, ok. You want to operate a BJT very fast in RF transistor; as an RF transistor for example. So, how fast can it operate? That is something that is remaining.

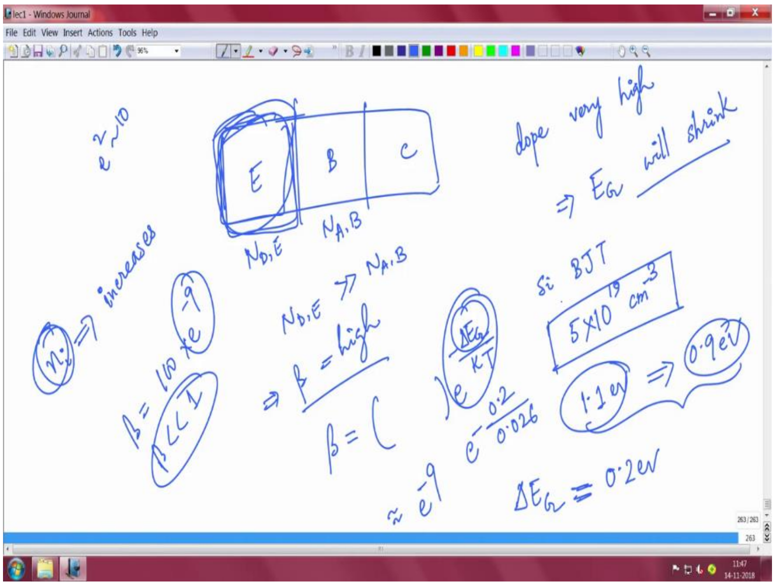
So, in today’s lecture, we will definitely cover delay, ok. What is the speed at which a transistor can operate? A BJT can operate, that is something we will do it from the device point of view, but before that, before that, we will continue slightly from the last slide where we had left in the last class, ok; so we will come to the whiteboard.

(Refer Slide Time: 01:41)



If you recall, in the last class I had told you about the Early voltage and Early effect, and I told you that in general for a BJT to work well or to BJT to you know give you higher gain, the emitter doping should be very high compared to the base doping.

(Refer Slide Time: 01:51)



If you remember, the emitter doping should be, this is emitter doping N_{DE} and this is base doping, N_{AB}. So, the emitter doping should be much higher than the base doping, in order for the, only then the gain beta will be high, right and you will get gain in the device. But, I told you in the last class in the last slide that, the emitter cannot be arbitrarily doped very high, ok.

You cannot dope the emitter arbitrarily very high, because if you recall from one of the earlier lectures, if you dope very high; if you dope a semiconductor very high, then what will happen is that, your band gap will shrink or will reduce - your band gap will reduce, ok, if you dope very high.

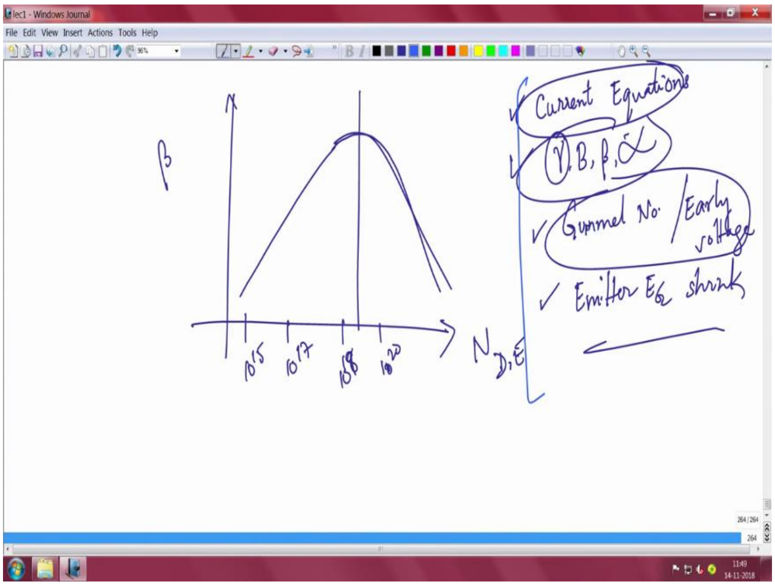
For example, if you take a silicon BJT and your emitter is doped very high, say 5 X $10^{19}/cm^3$, that is an extremely high doping for, you know silicon. So, then what will happen is that your band gap will shrink from 1.1eV to a lower value maybe 0.9 eV or even lower maybe. What will happen is that, when your band gap shrinks, your intrinsic lower concentration n_i will increase, because n_i depends exponentially on the inverse of band gap.

So, if your band gap becomes lower, your intrinsic carrier concentration n_i will increase exponentially. And so, this term n_i, n_i that gets cancelled out in the equations will not get cancelled out now. So, what will happen eventually is that your expression for beta that you have, you have an expression for beta if you recall. That expression will now get reduced by a factor of e to the power minus delta E_G by kT, where delta E_G is the band gap reduction $e^{\frac{-\Delta E_G}{kT}}$.

So, in this case, if the band gap reduces from 0.1, 1.1 eV to point 0.9 eV in the emitter, only in the emitter because your doping is high, then the reduction in band gap is 0.2 eV. So, this term will be, e to the power minus 0.2 at room temperature 0.026 will be like ($e^{\frac{-0.2}{0.026}}$), e to the power minus 9 (e^{-9}). So, whatever you had beta, suppose you had a beta of 100, if you dope it extremely high, will get reduced by e^{-9}) and e^{-9} is a very large number.

So, for example, e^2 you can see e^2 is almost 10. So, e^{-9}, will be you know like 10^5 or so, so this beta will be much much lower than 1. So, your gain has gone for a toss. So, what it means is that you cannot arbitrarily dope the emitter high, because band gap narrowing effect will, what would what, it will lead to reduction in the gain; so essentially what am trying to say is that, initially you know, if you plot the gain beta versus, sorry, if you plot the gain beta versus the emitter doping, N doping in the emitter, then due to your higher emitter doping, your beta increases.

(Refer Slide Time: 04:29)



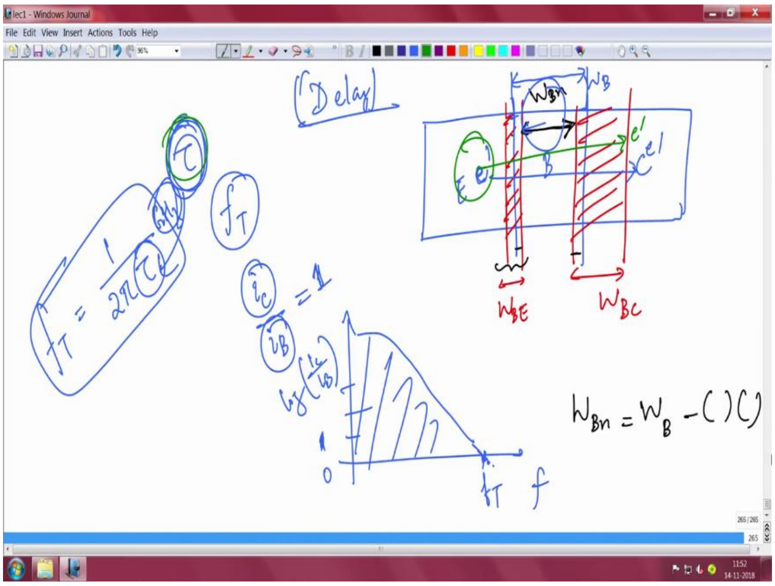
But, at one point, when the emitter’s band gap will start to shrink, then this gain will come down, right. So, you cannot dope, this is doping, right. So, this may be 10^{15}, 10^{17}, 10^{19}, you know and so on, 10^{20}, it cannot I mean it will not even 10^{20}. At 10^{18} or so, it will probably start to come down; so your gain will come down. So, you cannot arbitrarily do that very high, right.

So, these are the things that we have studied, what we have studied? We have studied current; the various current equations and the current transport in a BJT, the base component of the current, emitter injected current, base injected current, recombination current and so on. We have studied things like the current, the base transport factor, the emitter injection efficiency, the beta, the gain, the current transfer ratio alpha, these things we have studied.

We have studied the Gummel number, right. We have studied the Gummel number and the Early voltage, how important they are in the device. We have studied this, you know, like emitter shrinking effect, emitter band gap shrinking effect, right. So, we have studied most of the things that are required to understand the working of the BJT, right.

So, the next thing that the now we have to do is to study the delay, ok; so most of the things are clear. And of course, if in case of any doubts in any of this, either of especially the current equations can be a little bit, you know, confusing if you have, because you have so many different area, you know the types of current there. So, if you have any confusion, you always free to email me or we can discuss offline definitely, right. And you know how BJT amplifies the signal that we also have discussed. Now what remains is basically delay. So, we will now start and finish up delay. So, what does delay mean, ok?

(Refer Slide Time: 06:17)



What does delay mean? Delay essentially is in a way related to how fast the transistor can operate. So, for example, if you talk about a BJT, this is your emitter which is highly doped, this is your base and this is a collector and emitter has a very small depletion region here, right. And the base-collector has a wide depletion region here if you recall, right. I call this W_{BC}, the base collector depletion width, this is very narrow.

This is base-emitter depletion width and of course your, the neutral base width has become now W_{Bn}. This W_{Bn} is the neutral base width, remember, that is equal to the original base width minus the depletion that you subtract, this depletion, the depletion towards the base side. And depletion towards the base side, this depletion, this if you subtract from the original base, you get the neutral base width, ok. The neutral base width essentially is the width of the base after you subtract the depletion that is extending towards the base from both emitter side and collector side, ok.

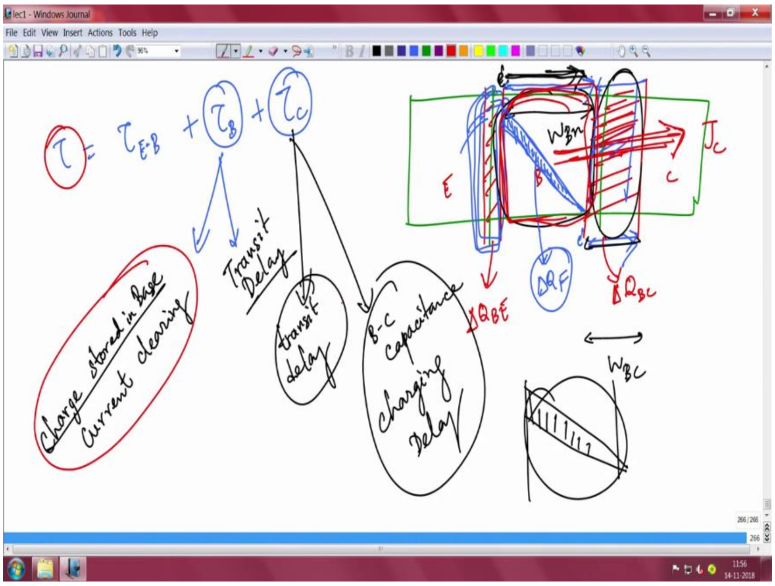
You see this blue, this is the metallurgical width of the base. This point to this point, that is the metallurgical width of the base and that is W_B, you subtract this part and this part and you get W_{Bn}; this is W_{Bn}, alright. So, essentially when an electron moves from emitter to collector, what is the total time it takes; that is called τ and that is the delay. What is the time that electrons take to reach the collector from emitter, ok. That is called the τ, the total time; that is called delay.

If you make this delay, the time shorter, your frequency of operation can be faster. Your transistor will work faster, if the delay is smaller. And there is something called a cut off frequency or f_T; f_T is defined as the frequency at which the current gain is 1, which means, your output current by input current, output current is collector current, input current is base current, that ratio becomes 1; which means your gain is 1. Above this, your gain will be less than 1; there will be no gain. So, gain has to happen only at a frequency less than f_T.

So, if I plotting frequency here and this I am plotting the log of gain here, which is I_C by I_B, then this will fall off at point and in log, in log scale, it is 0 which is 1, log of 1 is 0. So, this point is called f_T; is called current gain cut off frequency. And beyond this, your low, the gain will be less than 1 and log scale, it will be negative.

So, to get some gain, you have to operate in this regime, right. Basically, this is the current gain cut off frequency and f_T is basically given by 1 by 2 pi tau, this tau ($\frac{1}{2\pi\tau}$). So, a frequency will be faster and this is in gigahertz; so hertz. It will be faster when your tau is smaller. So, an electron, when it goes from the emitter side to the collector side; it has to encounter, it will encounter many delays. So, those delays will add up and give you the total delay. So, what are the delays that we should remember?

(Refer Slide Time: 09:15)



So, remember when an electron moves, it will also have to go through capacitors and there will always be capacitive charging and discharging of, you know capacitors, this will add to delay. For example, let me draw good picture here. So, this is again I am drawing the emitter, base, collector; emitter will have a small depletion region here and collector will have a large depletion region here, ok.

So, this capacitance, this is a, this is a depletion region in a forward bias emitter-base junction. This capacitance will store some charge and that will be delta Q_{BE}, base emitter charge

Similarly, this will store some charge, which will be ΔQ_{BC}; that means, every time you are putting a small signal, you are going to modulate this charge, which is ΔQ_{BC}, the charge that is stored here. Of course, you are operating the base, base-emitter of forward bias. So, you will be injecting carriers from electron from the emitter to the base. So, they will decay here like that, the base electrons will in the base, will decay like that which you know almost comes to 0 here; not exactly 0. But every time you change the small signal, essentially, if you are changing the input by a small signal like an AC signal, what is the charge that you actually modulating, ok.

So, every time you are changing the input signal by a small amount in a sinusoidal way, you are going to change the charge that you are injecting into the base also by a small amount here. At this point of course, it has to come to 0, by small amount that you are changing the charge here and that is your ΔQ_F, that is the charge stored in the base because the minority carrier electrons are injected to base. The charge stored in the base also will be modulated by small signal that we refer to as ΔQ_F. So, it is interesting to note that there are many delays that are associated now. So, the delay, the total delay will be made up of, one thing is that you have to charge and discharge this capacitor.

So, there will be delay associated with the emitter-base capacitor, which is this, right. Then the electrons will have to travel through the base, there will be two types of delay there. I will call them the base delay only; I will discuss this later. So, this is the emitter delay that is the emitter-base capacitor charging. As the base, as the electrons travel through base, there are two types of delays that will come that will be clubbed under base delay.

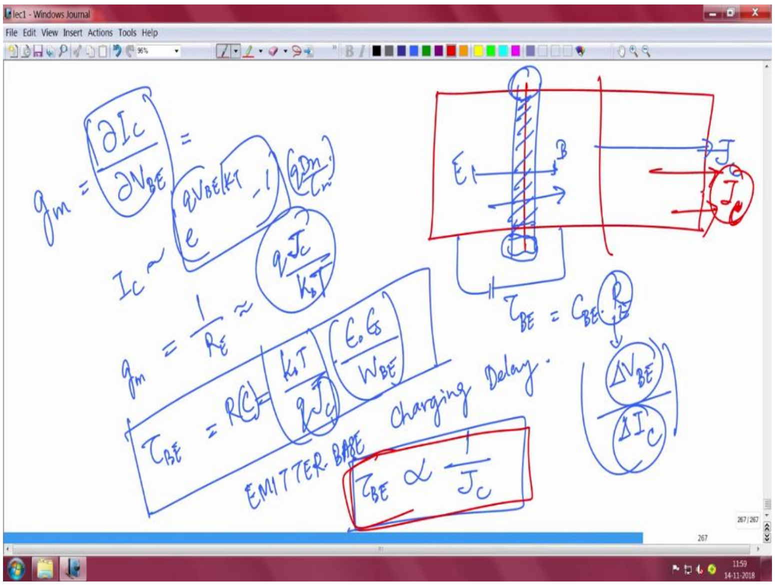
And then there is a collector delay, because there is a charging and discharging of the capacitor and also the electrons have to travel through this distance. So, there is also transit delay as they move across, also there is a charging discharging delay here. So, both of them are clubbed in here. So, the base has two components of delay, right.

And the emitter also has to, the collector also has two components of delay; one is the transit delay, because the electrons have to travel through this W_{BC}, the base-collector depletion, electrons have to physically travel across there. So, that is the transit delay. One is the base-collector capacitance, the capacitance charging delay, because the base collector will have a capacitance, and that, this capacitance will have to charge and discharge.

So, that charging and discharging delay will be there. There will be a transit delay process. Now base has two components one again, one is called transit delay. In the base, transit delay is the fact that electrons have to travel from across the base, electrons have to travel across the base physically. So, there is a time that it will take to cross the base, that is called transit delay, right. And remember the transit delay will be associated only with WBn, the neutral base width that we are talking about.

And then there is another type of base delay and that is the total charge stored in the base that, you see there is a charge that is stored in the base, right. This is the charge that you are modulating, this is the base, this is the charge that is stored in the base, right. There is a charge stored in the base, in the base, divided by the current that is clearing it. There is a current, that is the current that is flowing out, right. There is very much a current that is flowing out; that current is your output current, right. The current that is flowing out is clearing out the charge that is storing in the base, ok. So, the base has some charge and that charge is being cleared out by the collector current. So, that charge clearing out also will add to another delay, ok. It is also another adding term to the delay. So all of these have to be added up to eventually give out the total delay; so now, if you look into each of these delays, right.

(Refer Slide Time: 13:25)



So, what about the base-emitter delay, what about the base-emitter delay, right? So, let me see, you have a small base-emitter forward bias depletion region, this is a very tiny depletion region, but essentially, the delay with the base-emitter junction will be given by whatever the capacitance associated with that is there, which is C_{BE} times your resistance, that is RC delay, if you recall, ok. And what is the resistance that you associate with that. The resistance we associate with that essentially, is your resistance, you can say is actually with respect to the, this is emitter, this is base. So, if you connect this in to positive bias and a forward bias, the change in the base emitter voltage, with respect to the current that is going out which is I_E, but I_E is almost equal to I_C. So, I can say this is I_C, right.

So, this is essentially you know, you know we can say resistor or the inverse of this will be called the transconductance in a way. You know the transconductance, that is the output current with respect to the input voltage, how much is it changing, right? So, in a way, you can say that the output current, the output current is this current, how much is output current changing with the input voltage, input voltage is the voltage here.

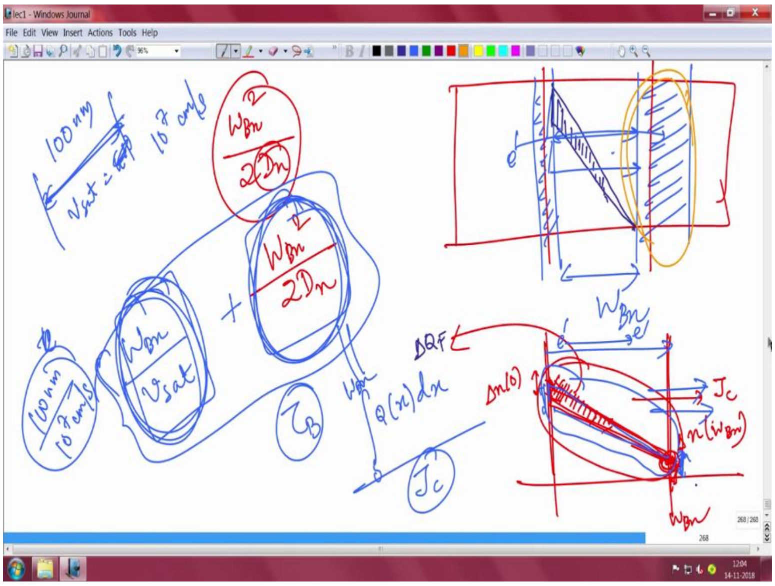
So, τ_{BE} so, this is called transconductance and it gives you, essentially, how much is the tunability of the transistor, how much is the output current changing with respect to the input voltage. And if you recall, I_C actually is given by e to the power qV_{BE} by kT minus 1 with some term like qD_n by L_n and all those things. So, if you do that, if you do this derivation, you will find out that g_m actually is which is 1/R, R_E I can say, is actually given by qJ_C by k T.

$g_m = \frac{qJ_C}{kT}$. So, essentially your g_m is dependent on the current that you are doing, J_C because it is an exponential function.

So, then if I put the delay, τ_{BE}, it will be RC delay and this will be given by, R will be given by the inverse of this, which is kT/q the output current into C, capacitor is the capacitance here which is $\xi_0\xi_S A$, if I normalize with respect to area, then it does not matter, divided by the width - base emitter depletion width, the base- emitter depletion which is very narrow, that is one thing.

And you see this delay, this is the total emitter base charging delay. This is the emitter base charging delay, this is the emitter base charging delay. And this delay is inversely proportional to 1 by J_C, if you recall. What it means is that if you increase the current density, if you increase the current density of the output current, somehow, by either biasing it much at higher voltage, in some way if you increase the output current density, then your delay also will come down and this is a unique situation, because in conventional MOSFETS you will not get this kind of behavior. Only in BJT, you get this behavior that with increasing output current, your delay can come down; it is a very beautiful result actually. But this is basically your emitter delay. And emitter delay could be significant or may or it may not be also significant in some cases. Next is the base delay, there are two components of base delay, I keep telling about, right.

(Refer Slide Time: 17:03)



So, if I look again at the transistor, if I focus on the base, this is an emitter-base depletion region, this is the base-collector depletion region, what remains here, sorry, what remains here is the neutral base width, ok. So, what is the time that you require to supply additional charge. What is the time required to supply additional charge, ΔQ_F, to this quasi neutral base region, you are injecting carriers, they are decaying and they are coming to almost 0 here.

So, this is a charge that you are modulating when you apply a small signal. So, what is the time required to essentially supply the additional charge to the quasi neutral base region, ok? This charge is almost 0, but not exactly 0, right. So, you know the, so essentially for any change that you have here, so, if you look into this condition, again if I am drawing this condition, here. So, suppose I have a carrier profile that is decreasing like that, if I am changing it by small amount delta n at 0 here, then it is coming like that. This is the charge that you are modulating essentially, which I am calling it as ΔQ_F, right. It is the charge, ok, but here it will not come exactly to 0, but it will come to some value, it will not be exactly 0, the reason is there is a velocity saturation.

So, it will not be exactly 0, but there will be some small value, ok. It will be, there will be some small value over which it will modulate and that modulated value will be Δn at this point which is W_{Bn}. What I am trying to say is that because of a finite velocity saturation, it will not, this point will not become 0, there will be some spread there.

So, there is a spread here, you are doing a small signal modulation, you are going to change some charge there. So, there will be a spread there and there will be spread here. So, essentially your there is a charge that is. So, basically, what is the time that you require to charge and discharge the, you know, the base, there will be a base charge and there will be current that is going out. So, how fast is this current getting out this charge, ok? That, that that component is very important in the base delay and that component is basically given by W_{Bn}^2 ok. This is the $W_{Bn}^2/2D_n$, and this is the diffusion coefficient.

So, this component $W_{Bn}^2/2D_n$, it tells you essentially the charge that is stored in the base divided by the current that is clearing it out, that ratio, ok, how fast you are clearing out the charge stored in the base, this is what it tells. There is another base component which is the transit. So, the base electrons are moving across the base, they are going from this point to that point. So, they are covering the base, that is basically at a transit time, that is this distance which is W_{Bn} divided by the velocity with which they move; the maximum velocity they can move is v_{sat}.

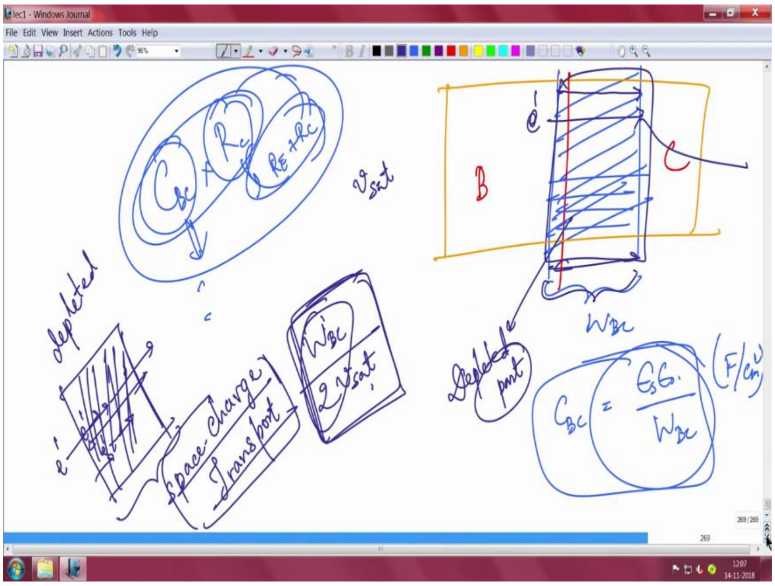
So, this is the total base delay, this is the total base delay. That is the total base delay, ok. So, basically you know you will have a small change in the signal will give you a small perturbation in the charge that you are storing in the base and there will be a small spread here also, because it is a finite velocity, there is not a not an infinite velocity.

So, the collector side you know the carrier density at the collector side also rises there is a small spread here and you need to supply an additional charge that will essentially, you know, make sure that there is a spread here and there is a spread here. How to clear the charge is basically this is the delay associated with that.

And your, the way you calculate this essentially is the total charge stored in the base which is 0 to W_{Bn}, the total charge stored in the base the minority charge, divided by the current that is clearing out J_C, you put the expression for J_C, the current expression, you will get this expression, ok. And this is basically the transit time you have; if you have a base thickness of this 100 nanometer and the velocity of saturation is 10^7, 10^7 cm/s, how fast are the transit electrons moving across that, that is the base delay; this is given by say in this case, 100 nanometer divided by 10^7 cm/s, that will give you this ratio will give you in nanoseconds or pico-second whatever the delay associated with the electrons moving across the base.

So, we now know this base component. Final delay is the collector delay, ok. You see this collector; this collector will charge and discharge. So, there will be two delays associated with that.

(Refer Slide Time: 21:51)



So, I keep telling you that one is that, if I only draw the base collector, for example, if I only draw the base-collector, this is base, this is collector will be small depletion of the base, because the base is doped higher compared to the collector, this is the collector, this is the base-collector depletion region. So, this entire thickness is W_{BC}. So, the capacitance associated with base-collector is $\xi_0\xi_S/W_{BC}$, this is normalized capacitance.

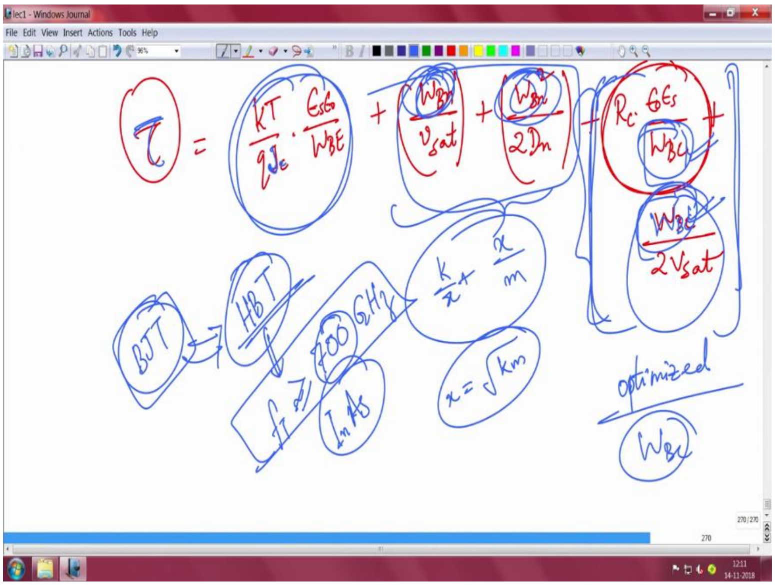
So, this is Farad per centimeter square, right. So, this is your base collector capacitance and the base collector capacitance has to be multiplied by all the parasitic resistances associated with collector. I will call it R_C, that will give you the base collector delay, base collector charging and discharging delay, ok. This is the base charging delay that you will have. And typically this will consist of all the parasitic resistances that you might have, $R_E + R_C$, for example, ok. This will be given in the question; this is the base collector capacitance which is given by this expression.

So, this term, this entire term, gives you the charging and discharging delay of the base collector capacitance, ok. And finally, you have electron that will actually cross across this and will be collected in the collector. So, the electron will take some time to physically cross this depletion region. If this depletion region is W_{BC} and if the electron moves at a velocity of v_{sat}, then this delay should be ideally W_{BC}, this distance divided by the velocity with which they are moving, which is v_{sat}, except that there will be a factor of two here. You know why, because this part is depleted, this is a depleted part. Not unlike the base, this is a unlike the base, this is a depleted part.

And whenever there is a depletion part; that means, there is no free carriers there, whenever there is a depleted part, suppose this is a depleted part, which means there is no carrier concentration almost there. An electron is moving through a depleted part, then the transport is called space charge transport, transport is called space charge transport.

The space charge transport is a type of transport just like you have drift and diffusion that you have learned in the course, if you remember. Similarly, space charge transport is a specific type of transport, not exactly drift diffusion where electric field is dragged by the carriers that are moving in a depleted film. So, the electric field within this is defined by essentially the carriers that are moving across the film and the background carrier is very low, so it is almost depleted. So, in like organic semiconductor or insulators for example, which have very low background concentration like a dielectric, the transport is typically space charge transport, because the carriers that you are injecting will define the field inside. So, in space charge transport, it so happens that, the time taken will be not be the ratio of W_{BC} where the velocity the distance by the velocity, but it will be distance by two times the velocity or half the distance by velocity. So, now we know the various terms that are associated with the delay.

(Refer Slide Time: 24:47)



So, let us write them down. The total delay is given by the base emitter delay, which is your base emitter delay you know is responsible for the charging and discharging of the base emitter capacitor that we had just talked about. So, the base emitter delay will be kT by qI_C into $\xi_0\xi_S$ by W_{BE}.

$$\frac{kT}{qI_C}\frac{\xi_0\xi_S}{W_{BE}}$$

Then there are two components of base delay, one is the base transit time by velocity saturation that is the time taken to cross the base. One is the charging on that clearing up the charge stored in the base, this D_n is the diffusion coefficient, W_{Bn} is the neutral base width; this gives you the time required to clear the charge in the base.

$$\frac{{W_{Bn}}^2}{2D_n} + \frac{W_{Bn}}{v_{sat}}$$

Then there is a base collector capacitance delay $\xi_0\xi_S$ by W_{BC}, this is the base-collector capacitance delay. Finally, there is a transit delay across the base collector depletion region that is given by W_{BC} by $2v_{sat}$. Now, if you recall here, this term will become, you want this to be small.

$$\frac{R_C\xi_0\xi_S}{W_{BC}} + \frac{W_{BC}}{2v_{sat}}$$

So, that the transistor becomes faster. So this term will be small if your base width that the base collector depletion width becomes large. But this part time will become small if the base collector depletion width W_{BC} becomes small. So, here you have to increase the W_{BC}, here you have to reduce W_{BC}. So, there is always, there is always an optimized W_{BC} which is the optimized base collector depletion width across which you will be minimizing this sum, you will be minimizing this sum, ok.

Similarly, you see that base width, the neutral base width has to be small; the thinner the base is, the faster the devices will be, because the delay will come down. And this delay does not matter in many cases, but it can also matter by increasing the J_C very large, you can reduce the delay substantially in which case, only these delays and these delays will matter. These delays will basically come down if your base is thin so, that W_{Bn} is small, this delay will come down with an optimized approach because some, it is like k/x + x/m, this kind of expression So, you have to minimize the expression.

And the minimization will happen at x equal to like km, I guess right $\sqrt{km}$. So, there will be point where you can do that. So, this is the total delay. There could be numericals, you know, that that will ask you to solve these delays and what different kinds of delays will affect the devices. You have to understand that there is a trade off in terms of the base collector capacitance and also this WBn is the neutral base width.

So, you have to subtract out the depletion of from the draws the base from the metallurgical base width, right. So, this is all about delay and you can see that state of the art you know BJT, BJT actually has many drawbacks. So, it is it has an advanced version called HBT. To understand HBT, we have to understand hetero junctions and compound semiconductor, but HBT is hetero junction bipolar transistors which is very similar to BJT except that you use materials of different, different materials actually for emitter and base. But in our record, HBT is an advanced version of HBT.

The cut off frequency f_T greater than 700 gigahertz have been the record, not in silicon, but I guess in Indium Arsenide system, silicon cannot have that kind of fT.

So, there are devices that can actually almost go to 700 or even more than that gigahertz, but again you do not operate the devices at 700 gigahertz, it means that the maximum frequency you can go here is 700 gigahertz. So, that brings us to the conclusion of the ending of this lecture and with that, we basically finish up the BJT. So, the topic of BJT with different, the important parts that we had to discuss are now over. So, what is remaining now is that we will visit BJT very briefly when we discuss about compound semiconductor and hetero junction, because when you have compound semiconductor and hetero junction, then you can overcome some of the problems of BJT like emitter band of shrinking and a low gain, ok. So, that enables you to get high gain despite emitter doping being very high.

So, those things we will just touch base when we talk, when we talk about hetero junctions and compound semiconductors. But as of now, for classical silicon based BJT, whatever we have covered till now is more or less sufficient to understand how the BJTs work. There are more advanced things like Kirk effect which is a high injection effect that we have excluded from the syllabus, that is a non-linear effects sort of thing, but you are most welcome to talk to me about it or discuss in other books.

There is also Ebers-Moll model we did not cover. It is more like understanding the I-V relation with respect to a model that we have, which also we have not covered. But you are free to do that as a self-study reading, but you are always welcome to email me for any questions. So, with that, we will conclude the topics of BJT. And in the next class, we will take up the topics of other things like either compound semiconductor or silicon MOSFET and we will go from there.

Thank you.