Nanoelectronics: Devices and Materials
Prof. Navakanta Bhat
Centre for Nano Science and Engineering
Indian Institute of Science, Bangalore

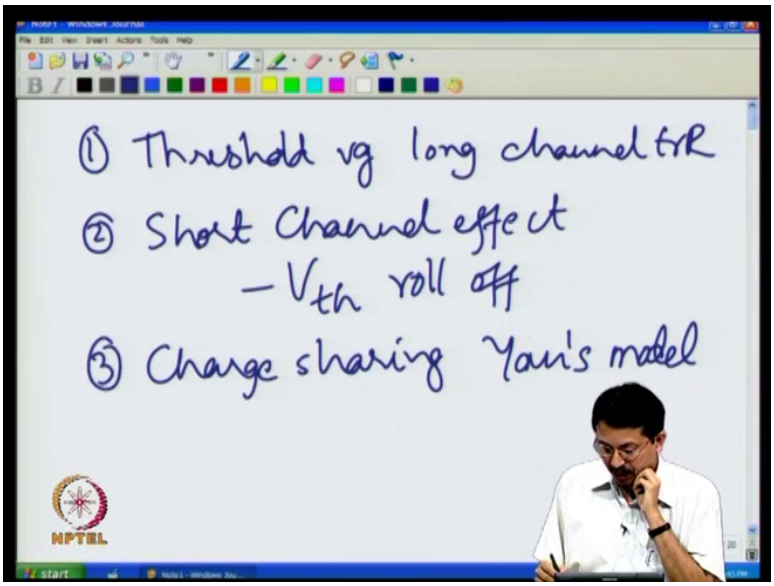
Lecture - 03
Short Channel Effects

So, let us get started. In the last lecture, we had seen you know the different scenarios for scaling and we had also discussed about the non-scaling factors. So, today we will get started with the discussion. Now, that we know we have to scale the c mos technology what are the challenges in scaling, right, what are the some new effects that will come into picture only when we miniature the transistor.

So, the first effect that we are going to talk about is what is called short channel effect. This is a very classic problem, wherein the threshold voltage of the transistor as you know which is a very important metric for MOS transistor that starts rolling off; that is the threshold voltage starts decreasing as you start miniaturising the transistor below maybe a micron or so. Certainly, in the nano meter region you know your threshold voltage will decrease very dramatically.

And to that end, first let us get started with the understanding the threshold voltage in classical transistors.

(Refer Slide Time: 01:29)

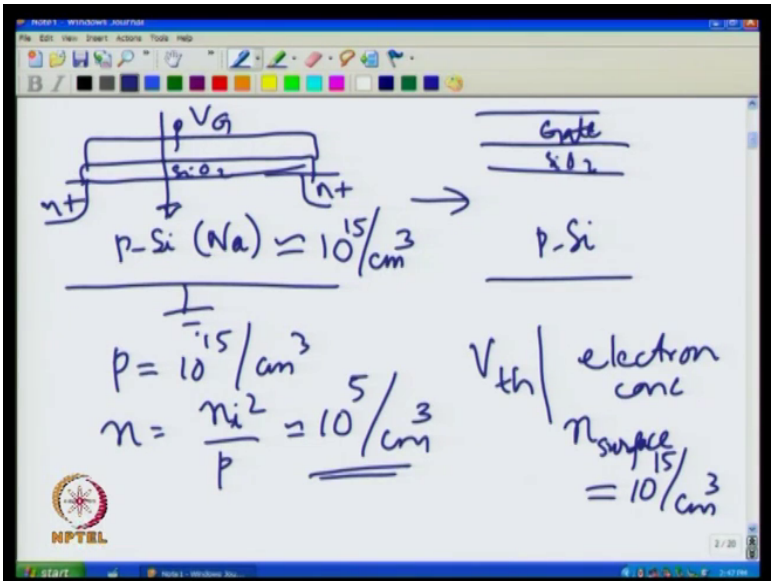


That is what we call long channel transistor. So, we will sort of revisit the threshold voltage equation. We will actually derive the threshold voltage equation for a simple long channel case and then, from there go on to discuss what I called short channel effect. As I mentioned in particular V-th role off, I use this you know a symbol V-th. Do not confuse it with the thermal voltage here. It essentially represents threshold voltage, right.

Now, given that there is a threshold voltage variation, then we will look at a very simple model based on what is called charge sharing, right. This model was first proposed way back in 74 by Yahoo and co-workers. And hence, it is also called Yahoo's model to predict the threshold voltage in miniaturize transistor which is different from the threshold voltage classical equation that you would have seen in a long channel transistor. So, let us set this as the goal for today's lecture.

So, let us first get started with the threshold voltage equation for long channel transistor. So, before that you know we will have to understand some basic aspects here.

(Refer Slide Time: 03:04)



When I am talking about a long channel transistor, you see the source and drains are really placed far apart. The distance between source and drain is huge and then, of course you have the oxide, silicon oxide and your gate electrode, right. This is your structure, right and you apply voltage here, ok and here we are considering n channel transistor and as a result of that we have p silicon substrate here which has a doping concentration acceptor impurities.

Let us say Na for example. This Na may be of the order of 10 to 15 per centimeter cube. So, when we look at this long channel transistor, its suffice is to really look at what is called a one-dimensional picture, right meaning that the source and drain are really so far apart from each other that if you have to just focus on so-called MOs capacitor that is metal oxide semi-conductor is a capacitor, right. You have an insulator in between. Often times we also call this as MIS diode. You see diode is a two terminal device and this also happens to be two terminal device because you have a body or P silicon is one terminal and gate voltage is another terminal, right.

So, all we need to do is only look at a very simple MOS structure which has P silicon substrate and an Si O2 and a gate electrode, right. This is a picture that we are going to focus for the initial part of this lecture.

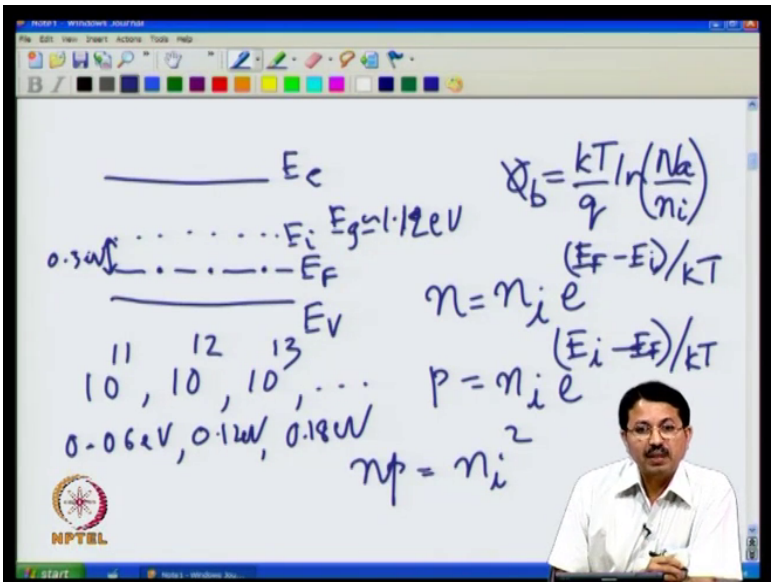
Now, if you recall what is threshold voltage? Threshold voltage is a condition where in we are going to alter the P type doping which essentially holds their majority carriers. If you recall if 10 to 15 is the doping concentration, then my whole concentration is also equal to 10 to 15 per centimeter cube whereas, my electron concentrations remember what is it? It is Na is Ni square by P, right. You want a thermal equilibrium. This relation holds good as you know at room temperature Ni is about 10 to 10 per centimeter cube, right 1.4 10 to 10. Let us approximate it to 10 to 10 per centimeter cube.

So, Ni square is 10 to the 20 and P is 10 to 15 and hence, my N is essentially 10 to 5 centimeter cube, right. They are really minority. Very small concentration of electrons we need to create a electron channel you see in MOS transistor, right. We need to invert the whole population here and convert that into an electron population. We define an inversion voltage V-th, such that electron concentration which is N at surface remember only at the surface. We are only talking of a surface channel device, right. We do not care about what the electron concentration deep in the bulk is.

Electron concentration at the surface should be same as the whole concentration that you began with which is 10 to 15 per centimeter cube. In other words, n surface should go all the way up to 10 to 15 per centimeter cube and that is what we do when we apply a positive gate voltage. We will actually see this picture in a little more detailed by using what is called energy band diagram today, but the fact is simplistic way of explaining that is you apply positive voltage that will attract more electrons to the surface and you

start increasing the electron population at the surface. So, 10 to the 5 electron which was initial concentration starts going up slowly, right and eventually you reach that 10 power 5 and it becomes 10 to the 15. That is the voltage on gate at which we have reached the inversion condition and that is a threshold voltage of the transistor, ok.

(Refer Slide Time: 07:21)



So, that is essentially what we mean by threshold voltage, right. So, now let us look at what is called the energy band picture, right. What we mean by energy band picture is that you know we have silicon which has valence band and conduction band. It has a band gap as you know E g of about 1.12 electron volt at a room temperature and you know we typically locate the middle of this band gap and call it as E i. E i is exactly in between the E c conduction band and the valence band that you have. Now, if it is P type doped semi-conductor as you know, then the Fermi level will be in the lower half of the band gap.

I typically show the Fermi level with this symbol here and the distance between the intrinsic level and the Fermi level is given by what is called a bulk potential. We denote it as phi b which is essentially given by this relation k T over q which is thermal voltage natural log of the doping concentration divided by N r that is the bulk potential. That is essentially difference between the location of Fermi level and the intrinsic level. When you have an intrinsic silicon that is no doping, right your Fermi level will be right on top

of E i. There is no difference between E i and E f, right. That is essentially your intrinsic silicon, right.

Otherwise you know your Fermi level will move along the band gap that you have this is what you will have in P type and in N type. The Fermi level will be on top of E i, right and accordingly, we also indicate that your electron concentration can be represented as you know it has an exponential relationship with respect to the location of the Fermi level, right. K is here Boltzmann constant and t is the absolute temperature, right. So, when your E f is equal to E i, n is equal to N i where E f is below E i. E f minus E i as you know is negative and N will be less than N i and on the other hand, your P is given by N i e to the E i minus E f by k T. In this case, here E i is more than E f. E i minus E f is positive.

Of course, your hole concentration is more than electron concentration and your electron concentration would have gone below N i and as you can very clearly see from these equation, your N p is essentially equal to N i square, right. These two exponent will cancel out, your N p is equal to N i square. So, this is what you have. So, exact location as you know the distance between E i and E v is about 0.56 E v given that the band gap is 1.12 E v.

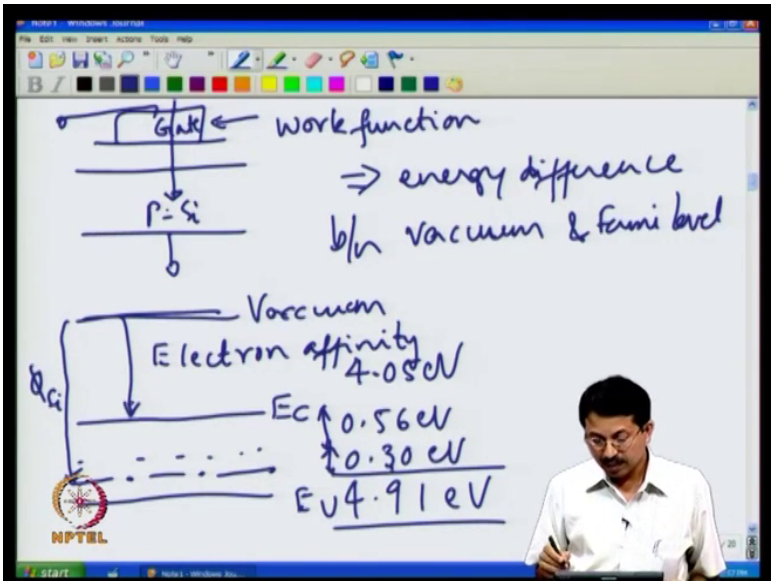
As you start doping the band gap, I mean the Fermi level starts going down, but a very good rule of thumb to remember as far as bulk potential is concerned is if you start increasing the doping concentration above the intrinsic level, remember intrinsic level is 10 to the 10, correct 10 to the 10 per centimeter cube above intrinsic level. If you start increasing the doping concentration by every order of magnitude like 10 to the 11, 10 to the 13 and so on and so forth, right your bulk potential or the distance between E i and E f will increase by about 0.06 electron volt. This is a very good rule. You do a exact calculation and you will find that it is very close.

In other words, for 10 to the 11, this will come down by 0.06 E v for 10 to the 12.06 times 2 because its two orders of magnitude away you see, then it is 0.12 E v and for 10 to the 13, it is 0.18 E v and so and so forth, right.

So, in the previous case as you know we consider the doping concentration which is 10 to the 15, correct. So, for 10 to the 15, the location of the Fermi level will be 0.3 electron volt below intrinsic level because it is five times 0.06 approximately, right which is 0.3.

So, then in the case of particular case that we have looked at P type silicon will be about 0.3 electron that corresponds to a doping level of 10 to the 15 and this is how you will have a picture everywhere whether it is surface or in the bulk to begin with. In other words, you were to consider an energy band diagram going from the gate all the way to oxide and into the substrate, ok.

(Refer Slide Time: 12:28)



We draw energy band diagram again. Let me have that picture for you here. P silicon, this is oxide and this is gate. Let us suppose that we have used the gate, we have a terminology called work function for this gate electrode that you are using. Work function essentially means it is the energy distance. It is the energy difference between the vacuum level and Fermi level.

In other words, if you give so much energy to the electron which is sitting at the Fermi level of any metal, you free that electron out of that metal and bring it to the vacuum, right. So, that is essentially the meaning of work function and similarly, we can define work function in silicon as well also it is same. The distance between the vacuum level and the Fermi level of silicon is what we called work function of silicon and you know if we consider that this is a vacuum level, energy level that is and I have this band diagram which I showed you, there are electrons in conduction band and also electrons in the valence band which are closely bound electrons to the silicon.

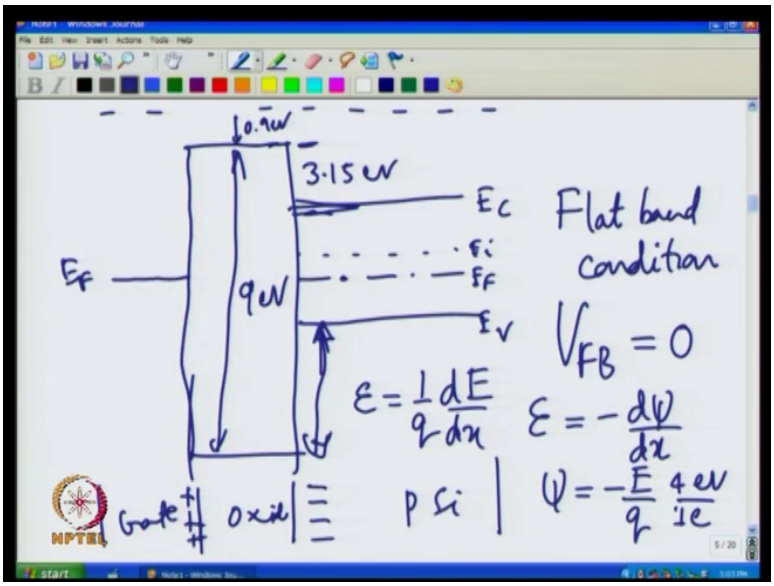
As I said this is intrinsic level and Fermi level is essentially below the mid gap for P type silicon. So, this distance here between the vacuum level and the Fermi level is what we call work function, right. You know this work function of silicon phi silicon and for work function of silicon is essentially this quantity plus half the band gap plus the bulk potential, correct and this quantity is what we call in the context of semi-conductor electron affinity that is the distance between the vacuum level and the conduction band level is what we call electron affinity.

Again it means that if you have an electron sitting in the conduction band, you need to give only this much energy to free this electron in silicon. The electron affinity happens to be 4.05 electron volt, ok and you know half the band gap as you know is 0.56 electron volt and for the doping concentration that be considered which is 0.3 electron volt, right 10 to the 15 doping concentration, so that this distance finally is 0.3 electron volt, correct. So, you know what is your work function for silicon and then, it is essentially 4.91 electron volt that is your work function of P type silicon which is doped with acceptor impurities of 10 to the 15 per centimeter cube.

Let us for the time being assume that we are using a gate metal which has a work function of exactly 4.91 electron volt meaning there is no work function difference between the gate metal and the silicon. You know metals come with a variety of work functions, right. You have aluminum for example whose work function is 4.1 electron, you have gold and platinum which have very high work function of the order of more than 5 electron volt, right. So, let us assume a fictitious metal which has a work function of 4.91 electron volt. There is no work function difference on the gate side and the substrate side.

In other words, this is a two terminal device, this is one terminal and this is another terminal. Both have the same work function, ok.

(Refer Slide Time: 16:38)



Then, if we were to look at this direction and sketch a band diagram going from the gate oxide and silicon, right the band diagram would look something like this. On the left hand side, I have gate and then, I have silicon oxide, right and then, I have this silicon which is as we already have seen is P type silicon, correct. It may not be exact due to the scale, but you know this is E c, this is E i, this is E f, this is E v valence band, right. The bands are flat. We know conduction band is flat, valence band is flat, of course Fermi level is flat and this is the gate work function I have chosen. The gate whose work function, the Fermi level on the gate is exactly same as the Fermi level on the silicon.

So, there is not going to be any movement of the carrier from gate onto the silicon or from silicon into the gate, right. That is why we have chosen identical work function. Then, we indicate the oxide that silicon oxide has a huge band gap that is why we call it as an insulator. The band gap of the silicon oxide we treat it as 9 electron volt that is huge and the electron affinity in silicon oxide is about 0.9 electron volt, that is this is your vacuum level this is 0.9 electron volt and this as you know is 4.05 electron volt and there is this band offset here which is what we call a band offset that is an electron which is sitting in a conduction band of the silicon see is a huge barrier of about 3.15 electron volt to go in this direction. That is why oxide is a very good insulator.

Similarly, the holes which are setting in here also see band gap. In fact, the band offsets for holes is even more than the band offset for electrons. You know you can go back and

compute this as an exercise for yourself. You can do it very easily because you know this is 9 electron volt, you know this is 4.05 here and you know what is the band gap and you can compute the barrier for the holes, right and you see that will be more than 3.15 electron volt, right. This is what you see in terms of energy bands, right. This is what we call as a flat band condition.

In other words, we have chosen a condition, we have chosen a metal such that V FB is equal to 0 meaning without doing anything, without applying any gate voltage, already the bands are aligned, they are all flat aligned. Meaning they are flat, there is no band bending. This is the condition that we have and now what we need to do only at the surface? Remember that only at the surface we need to convert this P region into N region and not in the bulk. In the bulk, nothing will happen in the bulk. It will continue to remain as you know your P type silicon. In other words, remember it is essentially the picture of that we had. This is P type silicon, this is oxide and this is gate, right. That picture is sort of flipped around, right. I mean we are looking in this direction going from gate oxide into P type silicon, ok.

Now, let us see what happens? I start applying positive voltage to the gate. When I apply positive voltage, a non zero voltage, then the two Fermi levels will not be aligned anymore because the electron energy on the gate side and the silicon side is not the same because you are applying a potential difference in fact because we are looking at electron energy in the energy band diagram. When I apply a positive voltage on the gate, the electron energy will go down here.

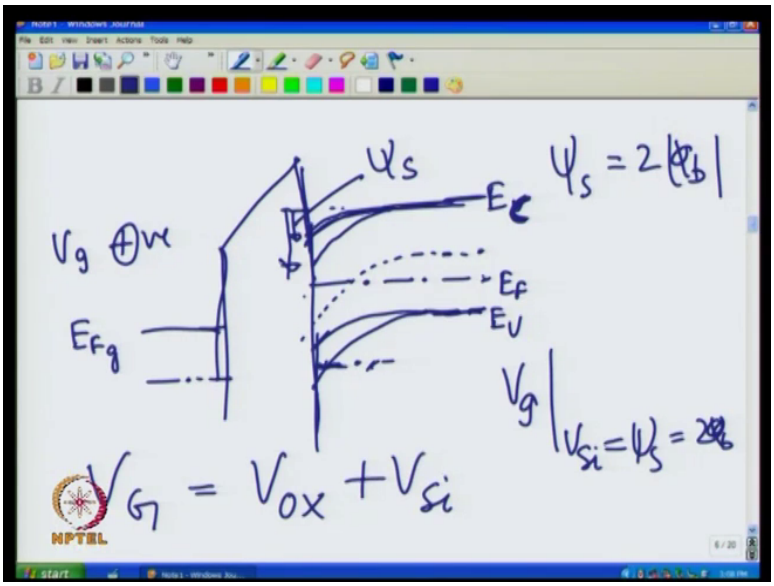
The Fermi level on the gate side will go down compared to the Fermi level on the silicon side. That will set up an electric field in the system. It is not very hard to imagine, right. I mean you apply voltage, there is an oxide insulator, right. It is a capacitor structure applying positive voltage essentially means that I am putting positive charge on the gate. This positive charge has to be balanced by equal and opposite negative charge on the substrate, right. If you have a positive charge on the gate and negative charge in the silicon, there is going to be an electric field that is going to be set up, right and this electric field results in what is called the band bending; the bands in the oxide starts bending; the bands in the silicon starts bending.

In fact, whenever there is a gradient with respect to position in energy band diagram, there is an electric field. So, electric field and the energy band diagram gradient are related, right and that is very easy to understand. Remember electric field is a negative derivative of potential, right and the energies here remember potential is defined with respect to positive test charge whereas, the energies that we are drawing here are for electron. These are electron energy and also these are not expressed in joules, but they are expressed in electron volt. In other words, potential is related to this energy or electron volt, right. How is it related?

Essentially potential is given as minus E over q, where E is an electron volt, q is charge, but charges is not in coulombs, right. Charge is one electron, right. So, when you know 4 electron volt divided by 1 electron, you get 4 volt as the potential you see, right. So, that is how potential and energy in the energy band diagram are related to each other. This q is only for dimensional consistency you see because E is expressed in an electron volt whereas, potential is expressed in volt. So, you are converting electron volt into volt by dividing by 1 electron. That is about it. So, obviously now you can see if you replace I with respect to minus E by q, then you can immediately see that your electric fields are essentially related as d e by d x, where x is the position, ok.

In the flat band condition, E does not change with respect to x. E is flat you see, right and there is no electric field that is obvious, right. There are no charges, no electric field, but whenever I have charges setup by applying the voltage, now there has to be electric field and hence, there has to be a gradient in band diagram. So, with this background now we can very easily sketch this band diagram under the condition that I have a non zero op positive voltage applied.

(Refer Slide Time: 23:30)



So, what happens you have non-zero positive voltage applied and because of that you will have a condition which would essentially look like this. Remember applying a positive voltage, right. So, the Fermi level will be right up here because the energy of electrons on the silicon side is more than the energy of electrons. This is the Fermi level.

On the gate side and this is Fermi level on the silicon side. So, what would then happen away from this region, nothing will change, right. So, this will be exactly like what you had. This is P type, E f is below E i by certain amount, but as I start approaching you know there is this band bending, correct. As you can see now the Fermi level which was closer to the valence band here has gone little further away from the valence band meaning the hole concentration here has decrease and electron concentration here has increased, right.

In fact, the conduction band, this is conduction band, right has come closer to the Fermi level, right. And hence, electron concentration is slowly increasing. As I start increasing the gate voltage, electric field will be set up higher and higher; more electric field will be set up and more band bending will take place, right and this band bending how much this band have bent at surface with respect to bulk is really the potential drop in silicon. You see it is essentially a series circuit of oxide and silicon. When you apply a voltage, part of that voltage will drop across oxide and part of that voltage will drop across silicon and this is what we call as surface potential psi S.

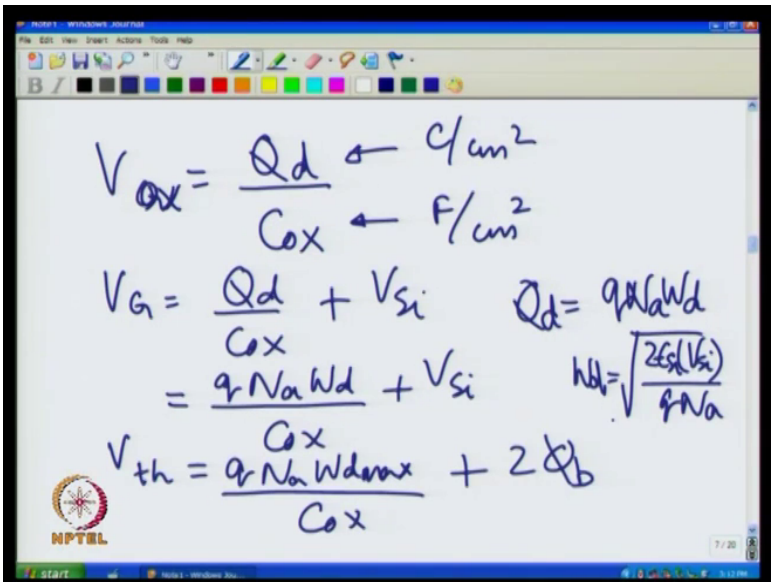
So, as I start increasing the voltage, I am applying positive voltage. As you know I start increasing the voltage, more positive charges are put here, more negative charges have to come in here. Initially the negative charges are coming entirely by acceptor ionized impurities because electron concentration is very small although it is increasing remember compared to 10 to the 15. Acceptor impurities, this is very small, but as you start applying higher and higher voltage, let us consider another case where V g is much higher, then the bands will bend even further you see.

Now, you see the conduction band has come very close to Fermi level. You will have a situation where electron concentration at the surface could reach as high as the hole concentration in the bulk and that is the condition we define as inversion condition. Now, let us understand how much these band should bend for inversion condition, right. We say that this surface potential psi s, psi s should be equal to two times the bulk potential that we had. In other words, remember bulk potential, right. What is the bulk potential distance between Fermi level and intrinsic level? Here it was p type and see here what has happened, intrinsic level is essentially between E c and E v. Intrinsic level has come below the Fermi level by the same amount phi b, right.

Here your electron concentration is as much as you know bulk concentration that you have whole concentration. In other words, your bands have bend by phi b plus phi b is that correct, right. Your band bending has to be 2 phi b in order to reach inversion condition. In other words, inversion is a condition; right inversion is a gate voltage V g when your V silicon which is also what we call surface potential is two times phi b that is your inversion condition. Now, this is the definition of inversion, right that our electron concentration at the surface is as much as whole concentration in the bulk.

Now, we need to find out then what is the expression for threshold voltage, right. So, remember I said it is a series circuit V g drop partly across oxide and partly across silicon, correct.

(Refer Slide Time: 28:24)



So, V g for the oxide drop what we say is that it is MOS capacitor, right. You have an oxide in between; there is some positive charge on the gate and negative charge on the substrate. You take this charge which we called a depletion charge, right Q d divided by C o x. So, you know charge by capacitance gives you the voltage across the oxide. I am sorry this is oxide voltage, right because oxide is a capacitor that you have Q d by C o x, ok.

So, in other words your V g is Q d by C o x plus V silicon and an important point I want to highlight that dimensions here is coulomb per centimeter square and this C o x is farad per centimeter square, right. We are looking at per centimeter square capacitance and per centimeter square depletion charge unit area depletion charge. Now, also it turns out Q d. Remember Q times doping concentration N a times depletion width W d, correct. You see dimensionally N a is number per centimeter cube multiply with charge, you get coulomb per centimeter cube. Multiply with depletion width, you get coulomb per centimeter square, correct and that is your Q d.

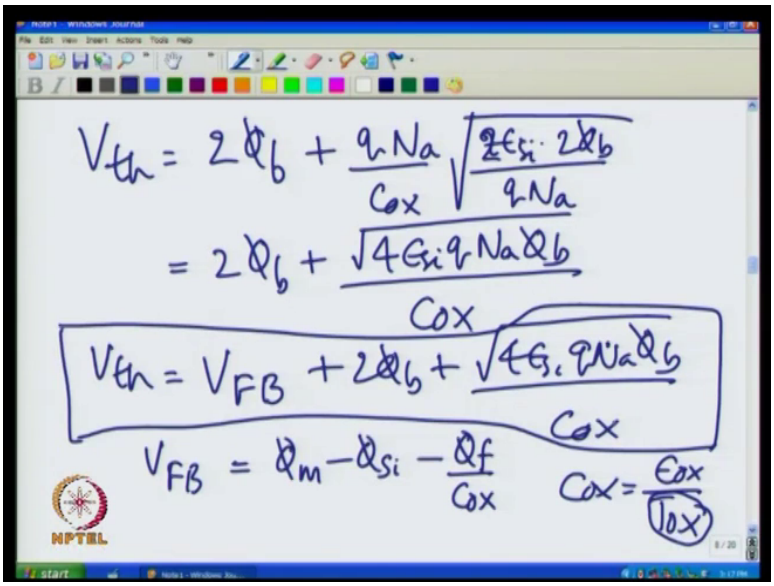
Now, as a result of that at any given applied voltage, your gate voltage is equal to q N a correct W d divided by C o x plus V silicon as I am increasing the gate voltage. You remember initially V silicon was zero flat band condition. I just start increasing the gate voltage V silicon increased slowly and accordingly more and more acceptor impurities, where uncovered and depletion width also started increasing accordingly, right. W d also

started increasing, your V o x is increasing, V silicon is increasing because you are increasing a gate voltage. Part of it drops across oxide, part of it drops across silicon. It is as simple as that, right. So, what is a inversion condition? Now, inversion happens when V silicon is equal to 2 phi b, correct. So, in other words V-th is equal to this part.

Q N a W d i reached a maximum depletion width. When I reach inversion, it is slowly increasing and it reaches a maximum. I denote it at W d max divided by C o x plus 2 phi b, where phi b is a bulk potential. So, this is essentially a threshold voltage, right because this is the gate voltage that I need to apply to create electron concentration at the surface which is same as the whole concentration in the bulk because the band bending now is 2 phi b and I ventured back to electron concentration condition that I set forth to begin with.

Now, what is W d max? You can use your simple one sided junction approximation. Remember that your W d is always given by 2 epsilon silicon V across the depletion region. What is the voltage across depletion region? It is silicon voltage because depletion region is in silicon and V silicon is the drop across that silicon voltage, right. So, V silicon is the voltage across the depletion region divided by q N a, correct. Under inversion I reach W d is equal to W d max that is when V silicon is equal to 2 phi b, you replace V silicon by 2 phi b here and you get W d max, correct. That is the maximum depletion width, right.

(Refer Slide Time: 32:35)



So, now you can you know substitute that here and then, you will get an expression for your threshold voltage which is 2 phi b and then, you know what you had out here q N a by C o x. We will replace W d max with the equation that we just wrote, right.

So, what is that 2 epsilon silicon? That is V silicon times 2 phi b by q N a and this intern becomes equal to 2 phi b plus this is q N a and this root q N a and this is q N a. So, what you really get is that 4 epsilon silicon q N a phi b divided by C o x. So, this is the classical threshold voltage equation that you see in any text book, right. So, you know how it came about now, right. It is very easy to understand and derive. This is the condition when I have voltage drop across silicon which has ensured 2 phi b band bending. And hence, at the surface my electron concentration jumped up to 10 to the 15 which is exactly equal to the whole concentration in the bulk silicon which is 10 to the 15 N a and accordingly at that junction, this is the depletion charge that I have and hence, this is the voltage drop across the oxide, correct.

Of course, in reality I started with the fictitious metal which has a work function exactly equal to silicon work function, but in reality you will not have that situation. In other words, you will have a non-zero flat band voltage. It is analogues to a built in voltage in your diodes that is even when you do not have any external voltage applied, there is a built in field due to the fact that the Fermi level on the gate and the Fermi level on the silicon are not aligned to begin with and hence, you have that correction term in equation that you see which goes as V FB plus 2 phi b.

Then, this equation, this part 4 epsilon silicon q N a phib divided C o x, this V FB in ideal condition is 0, but in non-ideal condition you know you just look at what E v f B v f in general is given as phi m minus phi s which is the work function on the metal minus work function on the silicon. They are not aligned to begin with if they are same. It is 0, but in addition oxide may not be ideal as well oxide may have some charges inside to begin with and that can also modulate your flat band condition if there is some charge in the oxide which we called fixed oxide charge.

Then also, even without applying external voltage, you have a built in electric field that results in band bending. So, this is your V FB, this is a correction term that you will put in and you have the classical long channel V t equation. Now, look at this equation. This equation says that V t depends apart from V FB. Let us not worry about it on doping

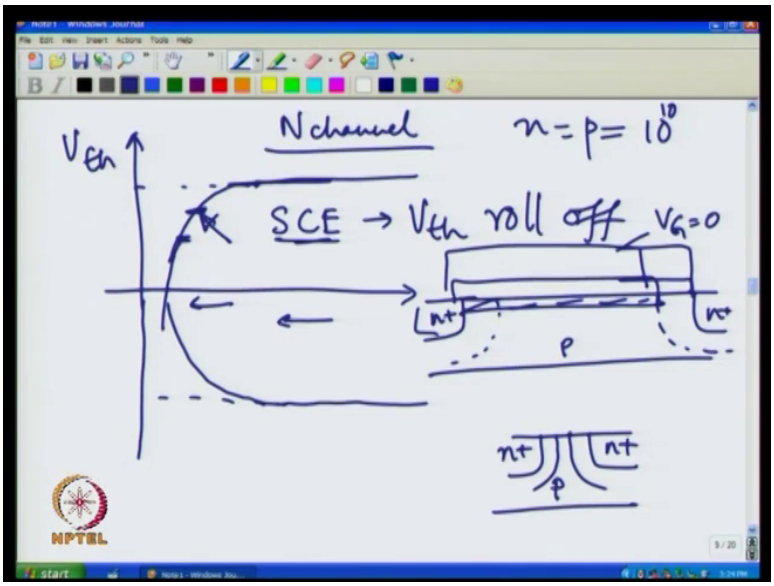
concentration because phi b also depends on doping concentration. Higher the doping concentration, higher is the V t.

Not difficult to understand because instead of 10 to the 15 per centimeter cube, you started with the P type which is 10 to the 18 per centimeter cube. You need to put in more effort on the gate to convert that minority carrier concentration which is 10 to the 2 to begin with to bring it all the way to 10 to the 18 you see and hence, you need to apply more voltage on the gate.

Obviously, it is directly proportional to the doping concentration. In addition, C o x remember I said- C o x is per unit area capacitance which is epsilon o x divided by T o x epsilon. O x is relative permittivity of the oxide and times free space permittivity divided by oxide thickness, right. That gives you farad per centimeter square. I have not multiplied that with the area, right that is unit capacitance which also says threshold voltage is a strong function of oxide thickness because T o x eventually will be going to the numerator epsilon o x over T o x. Larger the oxide thickness, more is the threshold voltage.

Again very not to you know understand, right. Larger the oxide for the same, voltage field is lower. So, I need to apply more voltage to cut the right field in the silicon to create that band bending, but the point is that the classic V t equation tells you that the threshold voltage of the transistor is only a function of oxide thickness that you used to build V t and substrate doping concentration. It does not depend on width of the transistor, it does not depend on length of the transistor, right because length does not appear anywhere in the equation, width does not appear anywhere in the equation.

(Refer Slide Time: 38:00)



In other words, if I were to build transistors of different length and take them in the lab and do I v measurement and extract threshold voltage, I should get a relation between V t versus length which looks flat because in V t equation you do not have l term appearing explicitly, right. This is what we had seen when we were building bigger transistor. 100 micron, 50 micron, 10 micron, 5 micron, V t was exactly flat, but as we started decreasing the length especially you know submicron region 0.5 micron, 0.2 micron, when we took these transistors in the lab and started measuring the threshold voltage, it no longer stayed flat as per this expectation, but instead V t starts doing this and this is what we call short channel effect. SCE for short or also referred to as threshold voltage role off, so that if you really do a very small transistor, this is all remember n channel transistor which should have a positive threshold voltage. It could be so much, so that your n channel transistor is always on it has such a lower threshold voltage and the same thing will happen even if it is P channel transistor.

Again you will have a negative threshold voltage which should have been flat all the way, but look no longer is the case starts rolling off V-th starts decreasing per P channel transistors also, right. Whether N channel or P channel does not matter, V t decreases with decreasing length and this problem is very severe when we talk of 100 nanometers of 100 nanometer transistors, correct. So, let us first understand you know why is this V t roll off and in order to understand that we just need to look at the picture that we have, ok.

We started with this N plus region and this P region in a long channel transistor and now, we are looking at a very short channel transistor N plus and P region. Remember whenever you have N plus P junction, there is so-called built in potential. In other words, there is going to be a depletion region already even without applying any voltage external world, right. You know you have V g is equal to 0 because there is PN junction. There is already a built in potential and some of the region here, right just at the edge of source and drain region is already depleted

Remember what do I need to do invert this transistor. It is heavily P doped. I need to decrease the hole concentration. I first need to create depletion region. Remember our discussion earlier that is when the whole concentration is decreasing, electron concentration is slowly increasing, but both electron and whole concentrations are much lower than the impurity concentration. That is why we call it depletion region, right because your N can be 10 power, P can be 10 power 10. So, your NP is equal to N i square, but you see both N and P are much smaller compared to 10 to the 15 which is my impurity concentration, right.

That is what we call it depletion region, but none the less you see in the depletion region source and drain have already helped you to do some work because they have helped you to bring the electron concentration from 10 to the 5 which was the concentration in P type region to 10 to the 10. You need to do some extra work to bring it all the way to 10 to the 15, right because when I am talking of inversion I am talking inverting this entire channel,. Part of the channel entire channel has to first go through deletion and then, go to inversion, but the part of the channel is already depleted because of the built-in potential from source and drain except that when the transistor is very long, you see this depletion region is so small. It is so miniscule compared to the total length of the transistor, right.

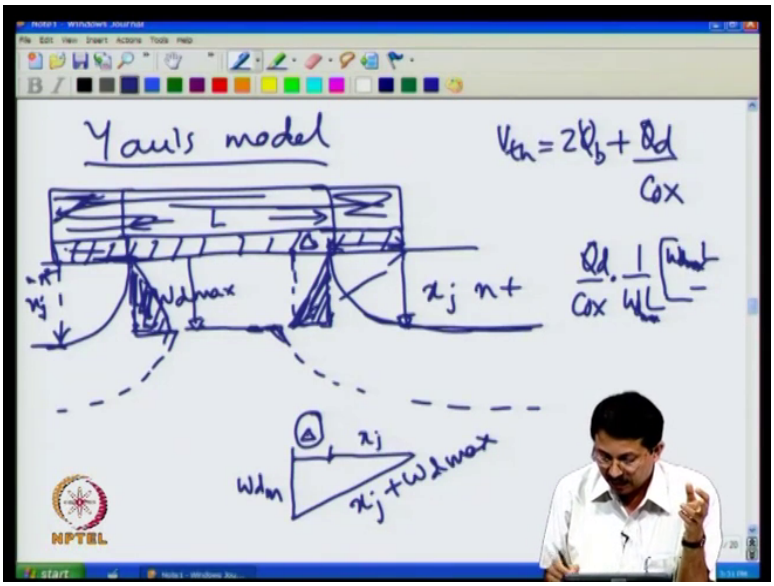
So, it really does not matter whether a very small you know 0.1 percent or even smaller fraction is already depleted by source, then built-in potential has no consequence because you still have such a large volume to be depleted and inverted, but when I scale the transistor you see now the depletion region can be a significant fraction of your channel length you see, right. So, now I can no longer ignore this you know.

In other words, your gate need not put in so much effort to invert the channel because source and drain have already helped you to invert the channel. You see they have partly helped you, right and when you miniature the transistor further, deletion region becomes even significant fraction of your channel length and hence, you just need to apply much lower voltage. Not this voltage only, this voltage to invert because 10 percent of the charge you know that is required here. You supported by source entry and here maybe 50 percent of the charge is supported by source and drain region, right because the transistor length is going down further, depletion width is not changing.

Depletion width is same for all length transistor because it is only dependent on the doping concentration, correct and this is the reason for short channel effect, the fact that it is 2d device; not just 1d device that we saw. Although it is a 2d device, when the lengths are very large, I can make 1d approximation, but when the lengths are very small, I cannot make that approximation. Hence, I do not need to apply so much gate voltage to invert the transistor and hence, your threshold voltage is lower. Go to even smaller transistor, your threshold voltage even lower and hence, lower the channel length, lower is the threshold voltage.

So, this is essentially what is the reason for the short channel effect. The source and drain built in potential already help you to invert the transistor. When that region is significant fraction of your transistor which happens when the transistor length is small, you have a lower threshold voltage because I do not need to you know put in so much effort to invert the transistor.

(Refer Slide Time: 44:35)



In fact, this is really explained with a very simple model that is Yau's model his coworkers gave, right. You know this is what is called a charge sharing model. Let us suppose I have a transistor which has a source and drain region and we typically you know this is my junction depth x j. This is n plus region and similarly here this is x j and this is also n plus region. Now, typically when we talk of inversion, right this is the transistor that we have.

We have apply voltage, right and this is the region where you are creating depletion, correct. This is your depletion width. This let us say w d wd max and inversion condition w d max is not necessarily equal to x j. They are not to scale here. Do not worry about that too much, right. So, this is depletion width which extends something like this. Wherever you have this depletion width extends like this and this is your length of the channel, right typically your gate needs to support this entire charge, this depletion charge.

Now, what we say is that part of the charge here in this triangle is supported by drain junction and part of this charge here is supported by the source junction and this is why we call a charge sharing model, right. There is this depletion charge that needs to be created first. Not entire depletion charge is supported by the gate. Part of that is supported by the built-in voltage that you have here between the n plus p junction and other part is a essentially supported you know by this region, right.

So, in other words this when we remember when we did the V-th. What will be V-th? It is phi b plus Q d by C o x that Q d was essentially the charge in this rectangle. That is what we had assumed earlier and that is indeed true when this is really large you see because these triangles are insignificant compared to the large region that we are looking at, but these triangles are no longer negligible compared to this region that we are looking at, right. So, the Q d is not entirely supported by the gate and only part of that it can be said that the charge in this trapezoid supported by the gate and the charge in these two triangles are supported by the source and the drain. So, that is what we need to do, right.

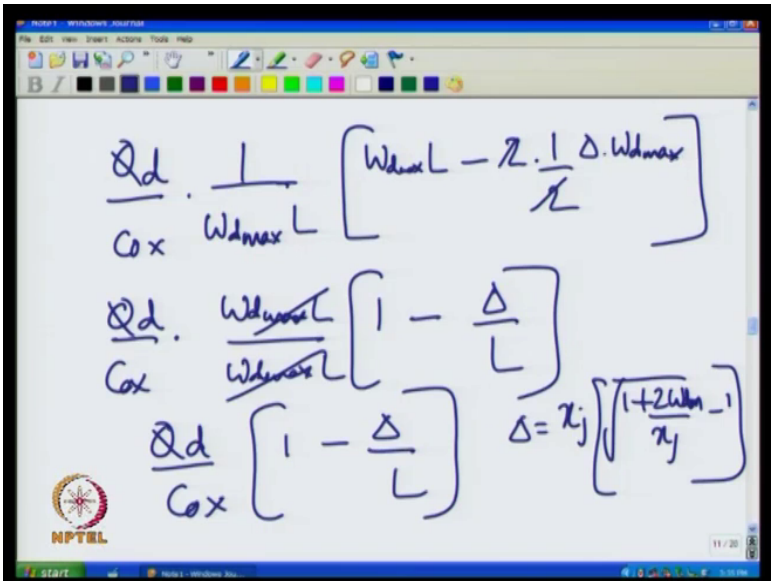
So, in other words, there is this delta, right. This is the delta, this base of the triangle is what we are calling delta here and this of course you know the junctions are shown like this just to sort of you know drive home the point that when you actually build a transistor, the junctions are made using diffusion and when you do the diffusion junctions actually spread laterally, right. In fact, the exact picture of the transistors should look like this, right. This whole thing here is gate and this is your oxide here Si O2. You do the implantation of the impurities and they diffuse. When they diffuse, they also diffuse laterally, right. A simple approximation is that if you have x j as a junction depth, they essentially go as a radius which is x j laterally, right. There is more lateral diffusion here.

As you start going down, there is less lateral diffusion and hence, junctions have this approximation. That is why we have that and you know what we are looking at now is that this is if you look at this triangle that we have, this triangle here is this is delta and this is x j and this is w d max and this one here is really x j plus w d max. So, now because this is x j and this is the depletion width that you have that is the w d max that you have, this is the triangle that I am looking at this triangle, this is the hypotenuse of the triangle which is x j plus w d max and this is w d max and this is delta plus x j. What I am trying to get at is really what delta is, ok.

If I find out what delta is, then I can tell you exactly what is the change in threshold voltage, how much did the threshold voltage decrease because of the charge sharing phenomenon that we are talking about. In other words, what we can say here is that you know your Q d by C o x w term needs to be modified. This second term will only take that second term Q d by C o x 1 over w l where w d max l where l is the length of the transistor and w d max is a depletion width that I have right here, right the w d max times

l minus the area of the 2 triangle, right. The fraction of the area that which is the trapezoid area really that needs to be supported by the gate, ok.

(Refer Slide Time: 51:11)



So, let me just re-write that here just to avoid that mess, right. I have this Q d by C o x here, then w d max times l, right l minus. There are two triangles. Area of the each triangle, what is the area of each triangle, right delta times the height which is w d max delta times w d max, correct. Now, this is really the area. What I am saying is that I am waiting Q d by C o x by this fraction which is essentially because it is not you know the entire rectangle that is supported in the case when delta is equal to 0. As you can see this w d max l cancels with wd max l and that reduces to your classical V t equation which we have derived already, right. Only when this delta is becoming comparable to the l, right then it becomes important to consider this.

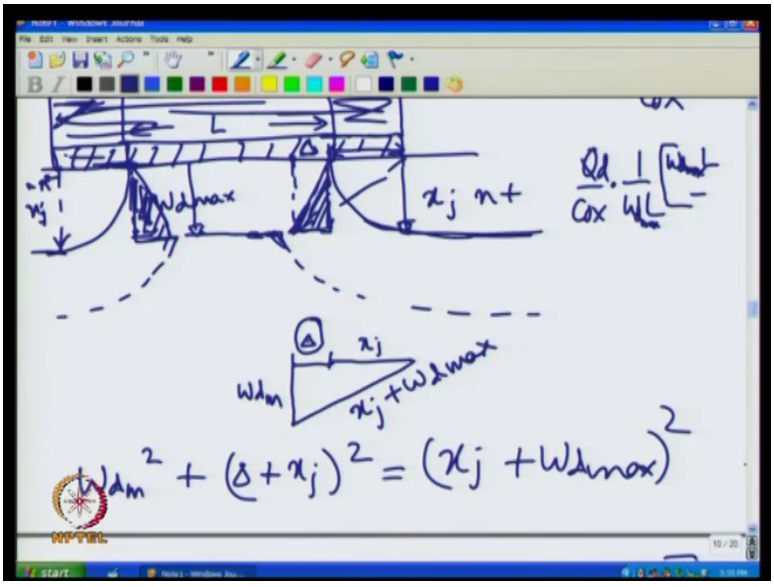
In other words, you can re-write this C o x times w d max l w d max l into 1 minus. So, what you get here is delta over l because this is w d max, this is w d max remember that, right. So, what you have? The second term in your V t is really Q d over C o x into 1 minus delta divided by l. As you know when delta is very small compared to l, this term can be completely ignored and that reduces to your classical equation, but when delta starts becoming comparable to l, right. In other words, we are decreasing l.

You see when we are looking at different length transistor, l is coming down very drastically and then, at some point l starts becoming comparable to delta, then I need to

take this one minus delta over l term and as l start decreasing you know, the change in V t becomes more and more and the fact that your V t is decreasing with respect to l as we saw in the equate in the plot here is very well captured.

Now, by this simple model that says when l is decreasing further, there is much more degradation in threshold voltage. Threshold voltage also starts decreasing. Now, what you can do, I will leave that as an exercise. You can go back and do this exercise yourself.

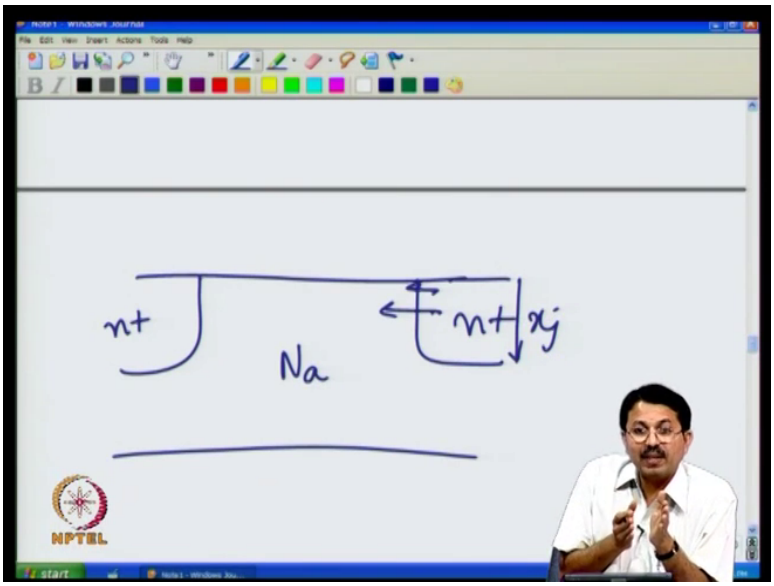
(Refer Slide Time: 53:58)



You take this, you write the equation essentially as w d max square plus delta plus x j whole square is equal to x j plus w d mx whole square and now, simplify this equation further and further and eventually you will arrive at an expression which will give you the expression for delta. So, the expression for delta will essentially come. I will not really derive this.

You can actually using that expression that I had, you can actually derive that. It goes as 1 plus 2 w d max by x j minus 1. This is how the expression goes for delta, right. So, delta if your x j is small, delta is also small, right and it turns out if your doping concentration is large, your delta is also small. So, that is very important to recognize, ok.

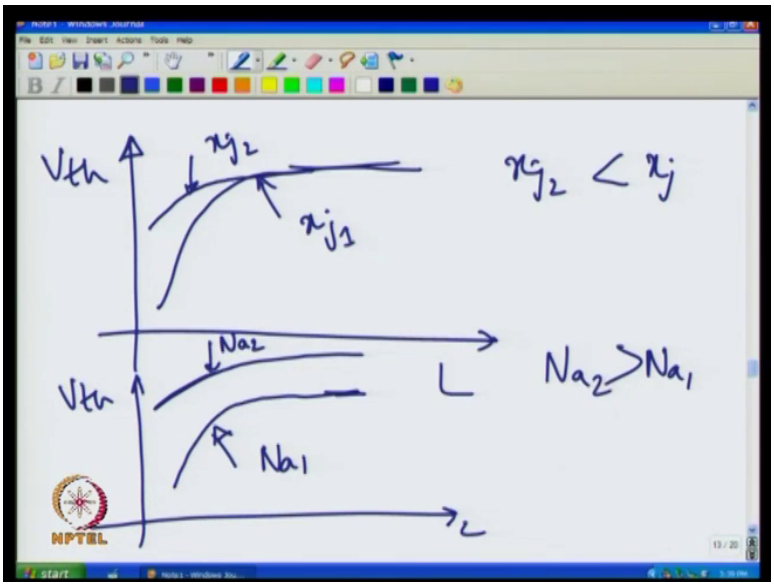
(Refer Slide Time: 55:17)



Qualitatively what we are saying essentially is that if I have this transistor, this is junction depth x j and this is doping concentration. If your junction depth is small that is good for short channel effect meaning your V t does not decrease as much, another way to think about is that you know short channel can come effect, come essentially because of two-dimensional effect. Two-dimensional effect meaning the drain starts influencing the channel. Deeper the drain, more region of the channel is influenced by the drain. Shallower the drain, it will not influence that bigger region of the channel, right.

This is another qualitative intuition that you can have. Similarly, if you have very high doping concentration, m delta itself is small because high doping concentration will not let this deletion widespread so easily into this region. If delta itself is small, then you start seeing short channel effect at much lower channel lengths.

(Refer Slide Time: 56:28)



In other words, what I am trying to tell you is the following. If you look at transistors with different designs, let us say there is one transistor which was the role off which look like this. This is threshold voltage as a functional channel length. Let us say this has a junction depth of x j 1 i create new transistor again different length transistors. I measure experimentally in the lab, if the junction depth is smaller x j 2 such that x j 2 is less than x j 1. X j 2 will not role of as fast as x j 1. In fact, this is also giving you intuition to design transistors. If you want a better short channel behavior, better role off, you do not want the V t to decrease as much, you better design very shallow junction. You figure out how to minimize diffusion of impurities in your source and drain junction, then you can build a better transistor.

Similarly, if you have two different transistors and one transistor which has a role off with N a 1 and I create another transistor which has higher doping concentration which is N a 2, that will probably have a role off which looks like this N a 2. Remember there is little difference that have shown compared to this graph, whereas the two transistors had the same V t for long channel transistors, but here if you increase doping concentration remember the classical V t equation which says for the long channel transistor if you have larger N a, you should get larger threshold voltage, right and that is why even when the transistor length is long like 10 micron, you have much higher V t and this is a condition N a 2 greater than N a 1. So, this has really given some insight into really engineer the transistor.

So, let me then summarize, right. So, threshold voltage as you know is very important consideration in the transistor MOSFETs. Classical threshold voltage equation based on 1d approximations says it only depends on doping concentration and the oxide thickness. These are the only governing factors not on length of the transistor, but all of sudden when we started building shorter transistor, we actually measure the V t. And we saw that V t is no longer as predicted and that is because of the 2d effect and a very simple model is a charge sharing model given by Yaw and that also gives some insert in designing the transistor.

Let us stop the lecture here and continue in the next lecture.

Thank you.