VLSI Technology

Dr. Nandita Dasgupta
Department of Electrical Engineering
Indian Institute of Technology, Madras

Lecture - 37
MOSFET II Tailoring of device parameters

So, we have been discussing the basic technology of n MOS devices, right; n MOS technology in integrated circuits. You have seen how, we have started with the first metal gate MOSFET technology and then, you have switched over to the polysilicon gate technology, because it gives a self-aligned technique. Particularly, as the device dimensions become smaller, the problem of the overhang capacitance or the overlap capacitance between the gate and the source and the drain becomes more and more severe and therefore, you would prefer to have a self-aligned technique, right and for that you have been using polysilicon gate MOSFET and then, we have also told you that you use a thick field oxide; the field oxide thickness must be large, so that the parasitic MOSFET has a very high threshold voltage, but at the same time if the field oxide thickness is very large it creates a problem of surface topography, right. That is the surface has a lot of ups and downs.

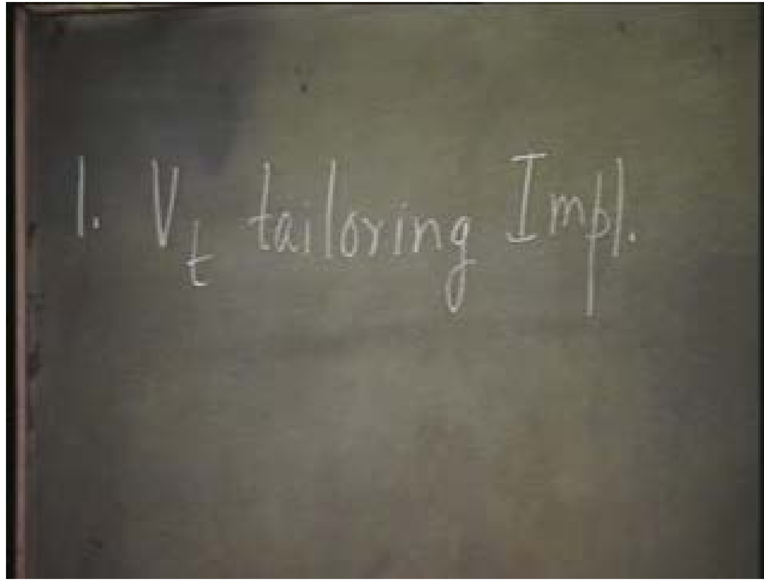
So, in order to prevent it and have a planar surface, what do you use? You use a LOCOS technique, local oxidation. That is you protect the active transistor region with a nitride oxide mask and then you have a local oxidation carried out in the rest of the region; you can either have this local oxide fully recessed or partially recessed. If it is fully recessed, then the final field oxide thickness will come only up to the surface. If it is not fully recessed, even then the topography will be considerably less ups and downs. But, in this LOCOS problem, LOCOS technique, we had some problem, the problem of bird's beak and bird's crest. This problem is nothing but oxide encroachment underneath the nitride mask, particularly severe when you have a thin pad oxide or when you have a recessed oxide structure, because the oxidation proceeds also along the side walls and that creates this bird's break and bird's crest problem and in the last class we have discussed certain modified techniques, modified LOCOS technique called the SWAMI in one case, in which case we had the masked side wall. In addition to having the active transistor region

masked by the silicon nitride, we also had the side walls; after the silicon etching, we had the side walls also masked by silicon nitride, so that the oxide encroachment problem becomes less severe or you can have a silo technique that is the sealed interface local oxidation technique, in which case you do not have the pad oxide sitting on top of silicon, instead of that you have a nitride-oxide-nitride sandwich structure, so that the oxide encroachment problem is less severe. So, essentially these are the basic features of an n MOS technology.

Let us now look at the finer points of n MOS technology, the finer techniques with all its details which are needed in today's very small device dimensions. The first important technique, I would not call it really a very modern technique, because you know, even the first n MOSFET was possible because of this, which is the threshold tailoring implant. You know that threshold tailoring implant is actually very, very important in n MOS technology, because, otherwise because of the presence of fixed oxide charges, we had always a depletion mode type n MOSFET and even today when we have more or less sorted out the problem of fixed oxide charges, it is now usual practice to have fixed oxide charges in the range of 10 power 9 to 10 power 10 for actually MOS fabrication process flow, even then, you know, if you have a substrate doping concentration of 10 power 15 per centimeter cube, the threshold voltage will be still pretty close to 0, about 0.2 volts or 0.3 volts, without any threshold tailoring implant.

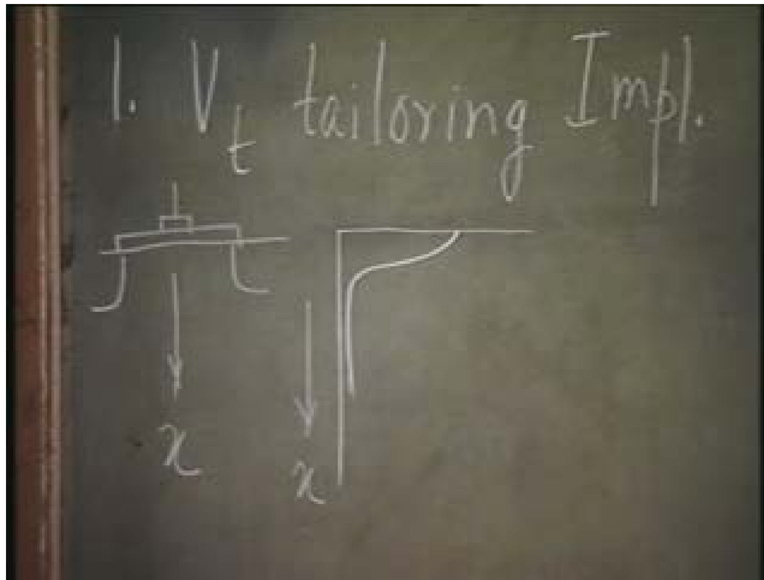
So, the problem is even if I have a positive threshold voltage, but if the threshold voltage is very small, 0.2 or 0.3 that means even when the gate voltage is 0, I have considerable amount of current flowing in the device, because you know as soon as the gate voltage is below the threshold voltage, the device does not immediately turn OFF; we have something called sub threshold, sub threshold current flows, right. So, if you have a low threshold voltage, then even for V G equal to zero there will be considerable sub threshold current in the device. So, there will be a problem of power dissipation, right.

(Refer Slide Time: 7:30)



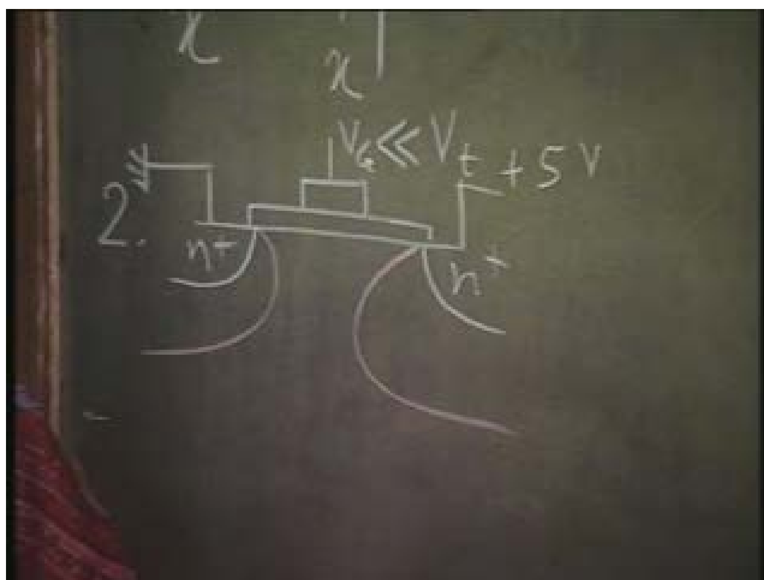
Therefore, threshold tailoring implant is a very important step; very important step in n MOS technology, V t tailoring implant and you know how that is done. You have a very shallow boron implantation carried out, so that it resides very close to the channel surface; it is very close to the channel surface, very shallow boron implantation, therefore it is going to increase the total amount of depletion layer charge. Therefore, the threshold voltage is going to be more positive, but at the same time, since it is not going to be anywhere near the depletion layer edge, it is not going to affect the body effect or it is not going to affect the junction capacitance. So, you have a very shallow boron implantation carried out in order to tailor the threshold voltage. So, this is one implantation with a profile something like this.

(Refer Slide Time: 8:49)



I am going inside the device, right. This is my device, source and drain and gate. This is the direction of x. This is the direction of x. For the threshold tailoring implant it should be something like this, the profile should be something like this. I have the implantation carried out very close to the surface. Deep inside the surface, deep inside the device, nothing is there, it is flat. This may not be the only tailoring I have to do.

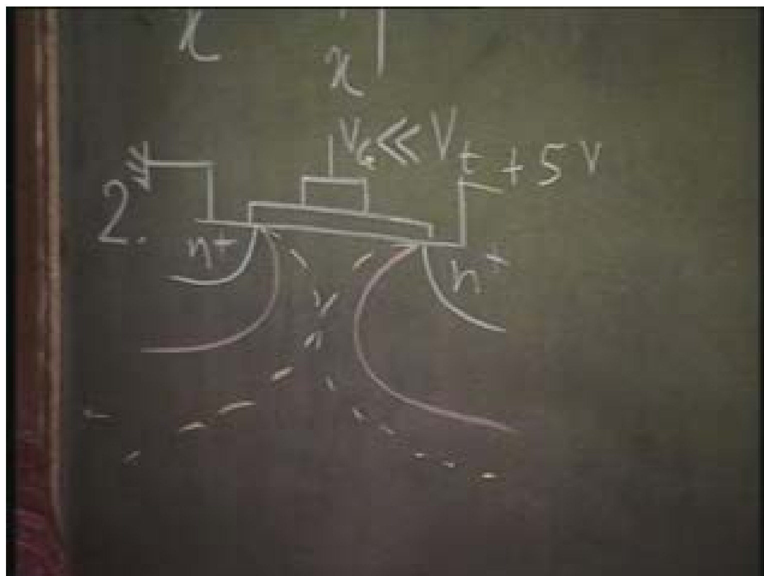
(Refer Slide Time: 9:54)



I have to take other factors into consideration and what is that other factor? I have to consider the punch through problem between the source and the substrate or the drain and the substrate. Let us see. Suppose I have a device like this; source, drain, gate. Source is grounded, drain is connected to plus 5 volts. These are n plus regions and this is my p substrate. So, you see if my gate voltage is very small, V G is much, much less than the threshold voltage, in that case what is the shape of the depletion region? I have a depletion region which is almost, the width is almost zero at the surface, because both the source and the gate, they are virtually at the same potential, something like this. Similarly, even at the drain side, obviously at the drain side, the depletion layer width will be considerably more.

Now, what is punch through? Punch through is when this source depletion region and the drain depletion region merge together. That is called punch through, when the two depletion regions are merged together. Where in this device is there maximum chance of punch through to occur? Not at the surface. At the surface, the two depletion regions are widely separated. Where? Somewhere deep inside; this is where the depletion region widths are most, right.

(Refer Slide Time: 12:10)



So, I may have a condition like this and that will be my punch through. Since this is not taking place at the surface, but somewhere deep inside, it is called subsurface punch through. That is the punch through is taking place below the surface, subsurface punch through. Now, you must prevent this subsurface punch through. You cannot allow the two depletion regions to merge together. How do you prevent the subsurface punch through? By making sure that the doping concentration of the substrate where the punch through is most likely to occur is high. See, I can, I can altogether reduce the punch through problem if my substrate doping concentration is made high. If the substrate doping concentration is made high, obviously for the same applied voltage, the depletion layer widths will be much less, right. Why am I not getting any answering look?

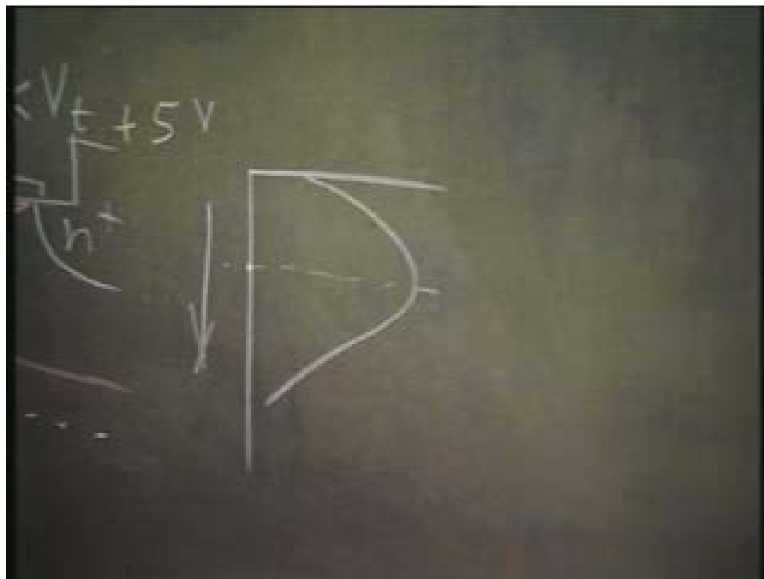
If the doping concentration of the substrate is high, what is going to happen to the depletion layer width? It is going to be reduced, right. Therefore, if you have a high doping concentration of the substrate, then you do not have a punch through problem. But, if you increase the doping concentration of the substrate, you have other problems. That is the body effect problem as well as the capacitance problem. So, you do not want to indiscriminately raise the doping concentration of the substrate. What do you want to do?

(Refer Slide Time: 14:18)

You just want to make sure that at the point where the punch through is most likely to occur, that is somewhere deep inside here, here the doping concentration of the substrate must be made high. In that case, I can separate out the two depletion regions, agreed. So, this is the subsurface punch through problem which is combated by giving a deeper implantation at the substrate.

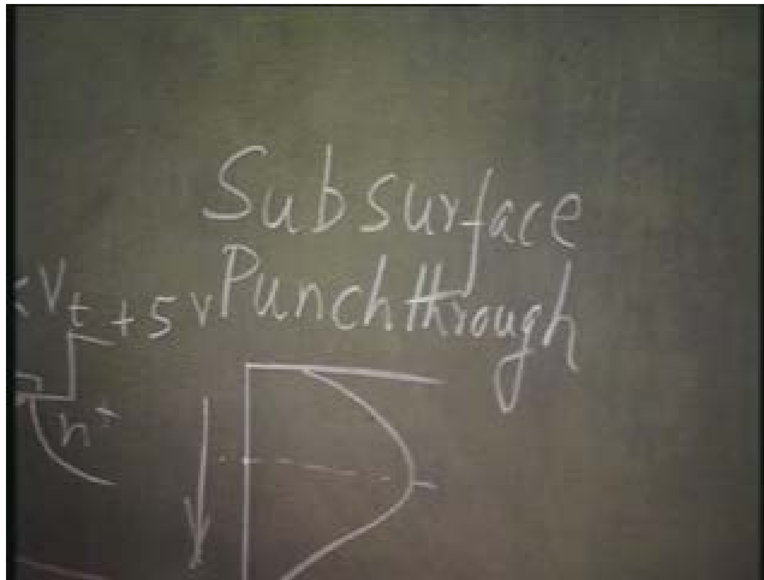
So, notice, I have first of all given a very shallow boron implantation to the substrate. That was in order to tailor the threshold voltage. Now, I am talking about a deeper boron implantation, in order to prevent subsurface punch through and this boron implantation peak should be approximately at the source and drain junctions, approximately here. This is the point where it is most likely to merge.

(Refer Slide Time: 15:27)



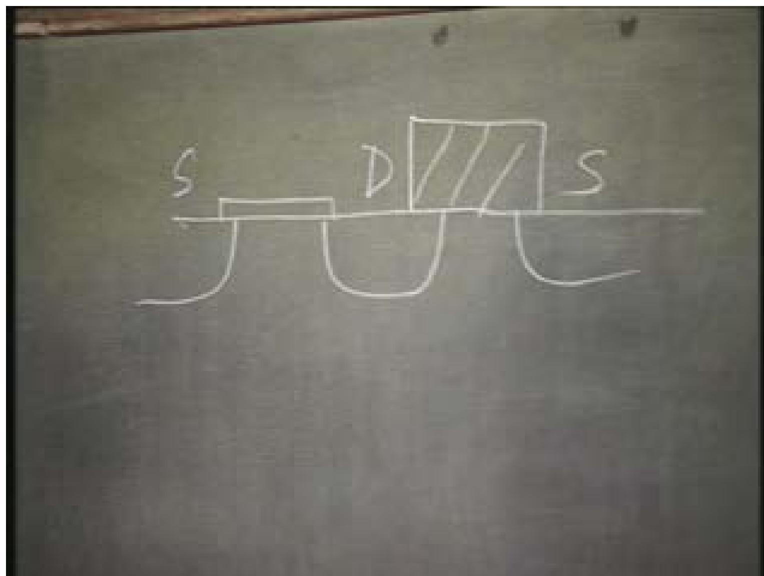
So, this implantation, the deeper boron implantation, should have a nature like this. This is approximately at the source and drain junction depth, right. This junction is approximately the source and drain junction depth. So, you see, we use not just one boron implantation at the substrate, we may need two.

(Refer Slide Time: 16:04)



One is for threshold tailoring, the other for, to prevent subsurface punch through, right and then I have to make sure that the devices are electrically isolated. That is the parasitic MOSFET does not start conducting.

(Refer Slide Time: 16:47)

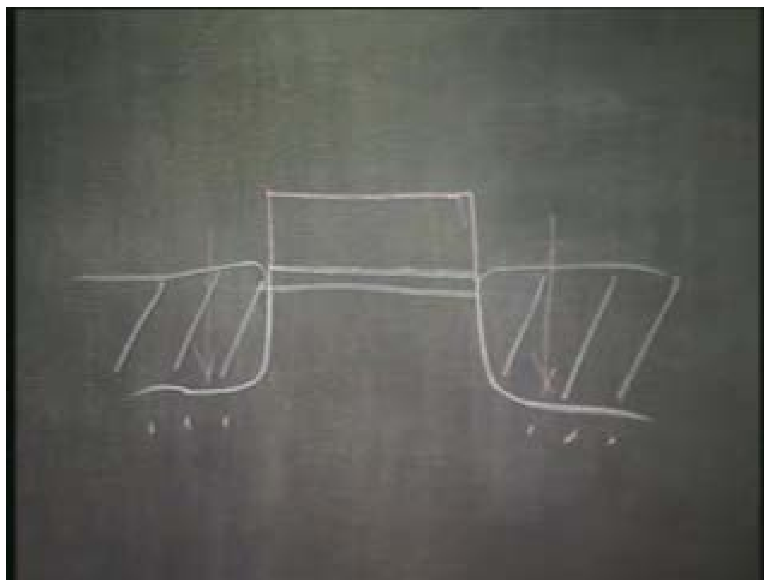


In a MOSFET, my only problem is this that I have the source and drain of one transistor and then, I have the source of another transistor. This is my gate oxide, this is my thick

field oxide and I can have the metal line and the polysilicon line running all over the place, so that I may have a conducting line running on top of the field oxide. In that case, you see, I have another MOSFET here with the thick field oxide acting as its gate. Now, I have told you that because of this very reason, I have kept the field oxide thickness sufficiently high, so that inadvertently, this parasitic transistor does not get turned ON. I do not have a conducting path here. By keeping this, the thickness of the field oxide very large, I am going to ensure that.

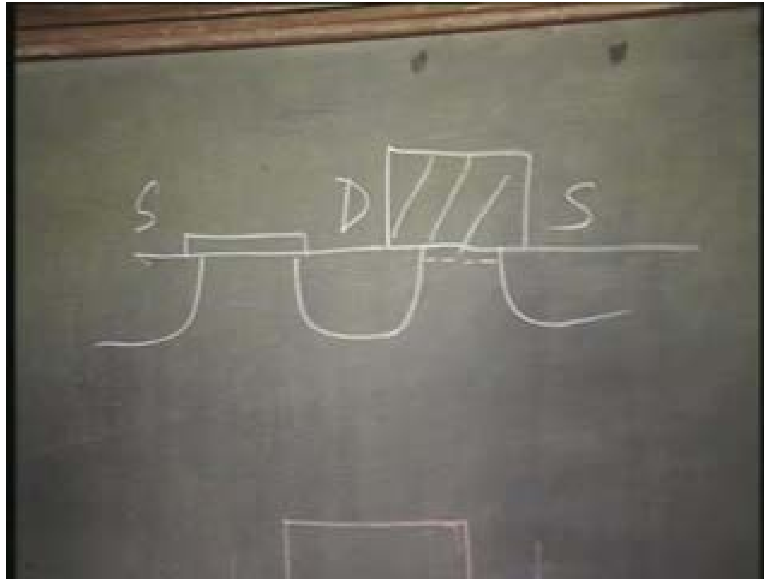
To add to this, I do a channel stop implantation. That is you know, now you know that the field oxidation is done by LOCOS. So, before doing LOCOS, what did you do?

(Refer Slide Time: 18:12)



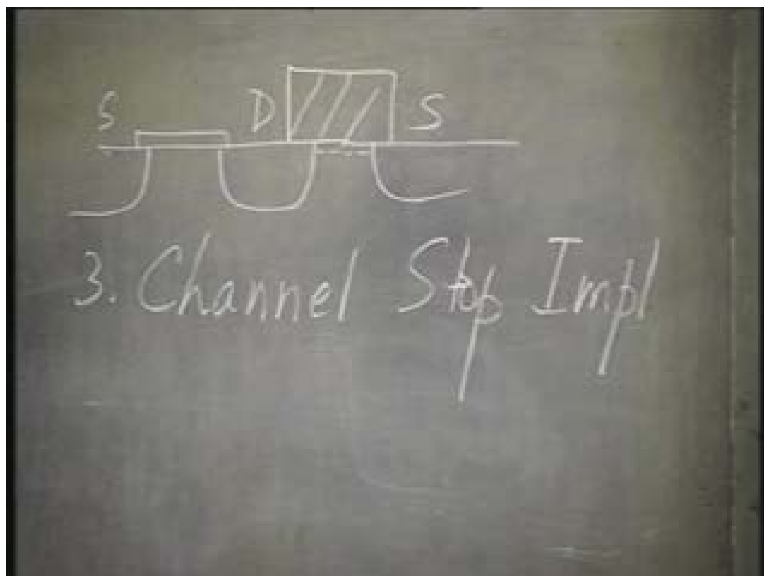
Remember, this was your active transistor region which is protected by this thin pad oxide and on top of that silicon nitride, right and then, if you want to have a recessed structure, you have etched it out, right and then you are going to grow the thick field oxide. Now, before growing this thick field oxide, you do an implantation here and then you carry out the oxidation. So, what is going to happen?

(Refer Slide Time: 19:06)



Underneath this thick field oxide, you have also done a channel stop implantation; see, corresponding to this implantation, the channel stop implant underneath the thick field oxide. What is the objective? I am raising the threshold voltage of this parasitic MOSFET even more, right. I am providing the precautions on two accounts. One is by keeping the field oxide thickness much larger than the gate oxide thickness and the other is by raising the substrate concentration under the field oxide.

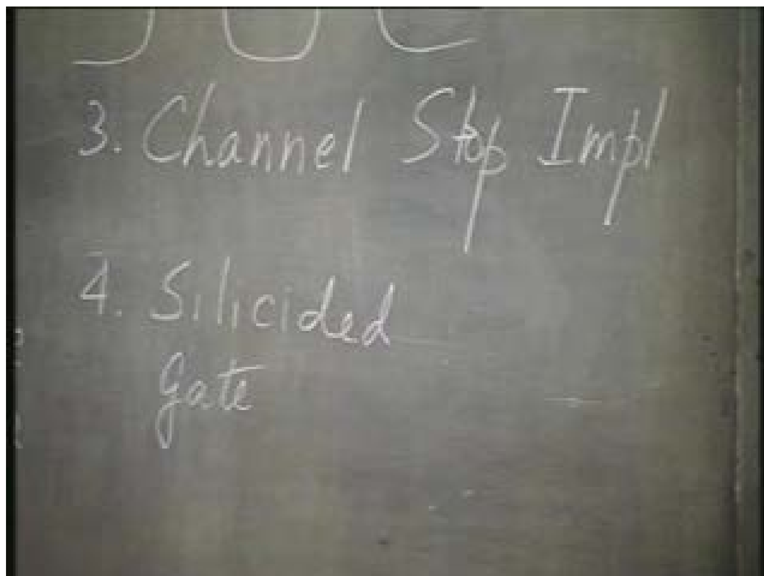
(Refer Slide Time: 19:59)



This is called the channel stop implant, channel stop implantation. So, three implantations so far, right; we have done a threshold tailoring implant, we have done a deeper boron implant in order to prevent subsurface punch through and now we are doing a third implantation that is the channel stop implantation, right.

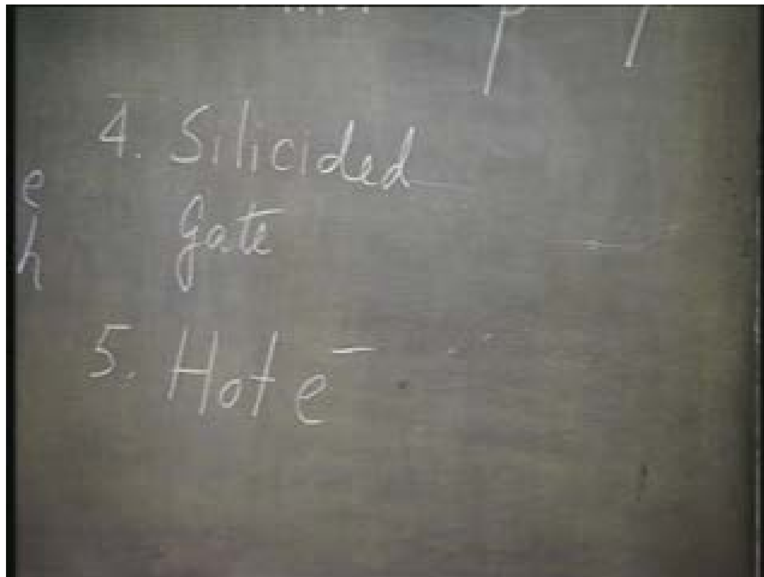
Now, what do we have to do? In order to fine tune the performance of the n MOSFET, we must look at the device once again. What do we have? We have a polysilicon gate. Usually for the n MOSFET, it is common practice to have an undoped poly deposited first and then at the time of source and drain doping, which is also done by ion implantation in most cases, when you are doing the source and drain doping you also dope the poly, so that your poly is also n plus doped, right. So, this heavily n plus doped poly is going to act as your gate contact. But you know, by now you know, that this may not be good enough, particularly now when we have long interconnection lines running, the R C time constant, R C delay may become considerable, because even though the polysilicon is very highly doped, its resistance is still much larger compared to a metal line, right. So, what do you do? Instead of just having a polysilicon, we have a polysilicon plus silicide, right.

(Refer Slide Time: 22:06)



So, most of the modern n MOS devices, they use silicided gate technology. We cannot have metal; we have sacrificed metal for the self-aligned technology. On the other hand, even heavily doped polysilicon is found to have a fairly large resistance. Therefore, we try to reduce the resistance of the interconnection by having the silicided gate and you do not want to destroy the well understood polysilicon dioxide interface. Therefore, you have the polysilicon, you have the gate silicon dioxide, you have on top of that the polysilicon and then, on top of that you have your tungsten or platinum silicide, in order to lower the resistance. So, silicided gate and finally, we have a problem called the hot electron problem.

(Refer Slide Time: 23:14)



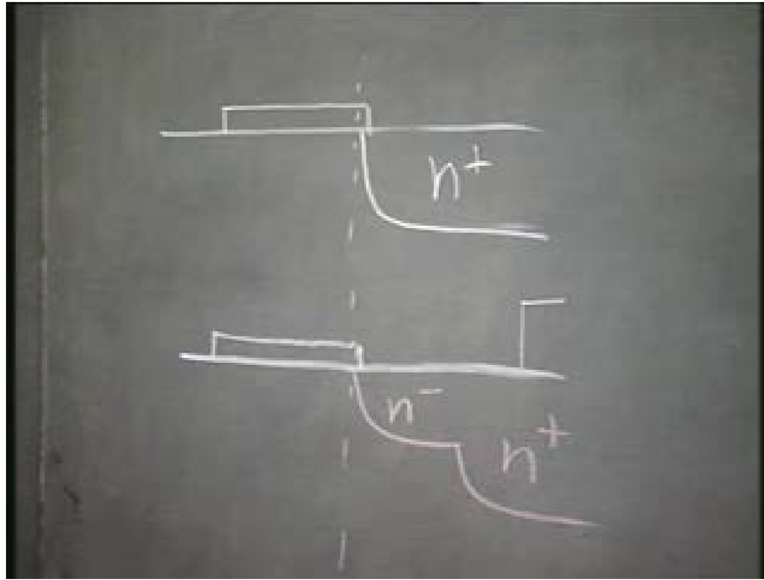
What is this hot electron problem? I have a MOSFET. I have reduced the channel length as well as all the other device dimensions, so that the channel is now very short. But, I have not really scaled the voltage levels that much, right. I still have 5 volts gate voltage. So, what is happening? Effectively, I am increasing the electric field in the channel, right; electric field is essentially, you know potential, means it is dV dx, right. So, it is actually, if it is uniform electric field, you could say it is potential divided by channel length; that is the average electric field in the channel, very crudely speaking. So, if the channel length is decreased that means the average electric field in the channel is also increasing.

So, the electrons, as they are moving from the source to drain, they are under the influence of a very high electric field and under the influence of this very high electric field, by the time the electrons reach the drain end they might acquire sufficient energy, so as to surmount the silicon-silicon dioxide potential barrier and get injected into the gate that is at the drain end, here. Electrons are moving from source to drain. As they reach the drain end, they might have sufficient energy to surmount the silicon-silicon dioxide barrier and get injected into the gate. If it gets injected to the gate, then it is going to have oxide trapped charges; it is going to create oxide trapped charges there. So, the threshold voltage is going to be shifted. So, they can either create oxide trapped charges or they can create interface states, in which case, in general, the threshold voltage will change and this is called the hot electron injection.

So, hot electron injection is finally going to set a limit on how long your device is going to perform adequately. It is characterized by MTF, mean time to failure. See, this is not taking place at the time of device fabrication. After the device is fabricated, when you are using the device, when you are connecting it to the gate voltage, to the source, etc., at that time this hot electron effect is coming into picture. So, this is a cumulative effect. The more you use the device, more and more hot electrons are getting injected into the gate and finally it is going to irrevocably change the performance, irrevocably change the threshold voltage of the device beyond the tolerance level, right. So, this is the hot electron effect and during the device fabrication itself, therefore we must have some remedial measures, so that the hot electron effects are not very severe.

So, we have what are called hot electron resistant structure. The principle behind this hot electron resistant structure is quite simple. I must reduce the electric field at the drain end. As the electrons are moving from source and drain, they are acquiring maximum energy when they reach the drain end, right. So, I must reduce the electric field. How do I reduce the electric field? By inserting a lightly doped region between the drain and the channel.

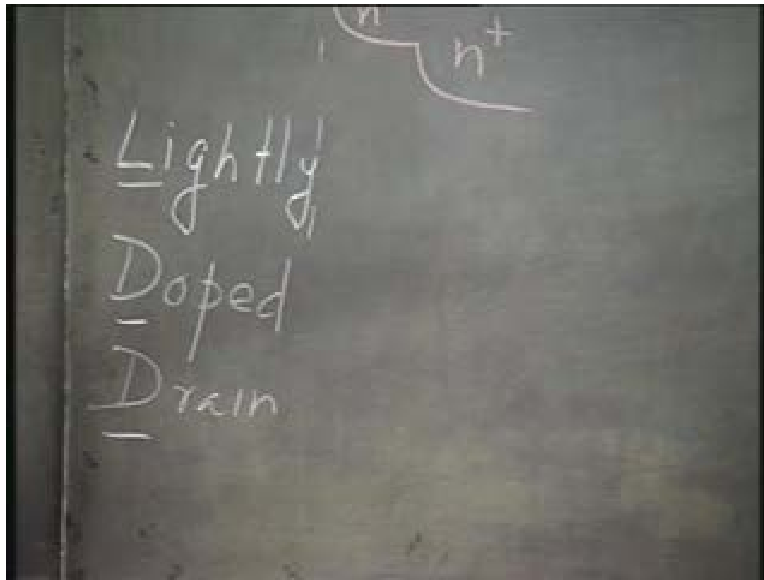
(Refer Slide Time: 28:03)



That is instead of having a drain like this let me have a drain region which will be a, which will be a, ok, this is the first structure. Instead of having this, let us have something like …. This is lightly doped. This is heavily doped. In this case, the whole thing is heavily doped. I have not done anything to the channel length. That is this point where the drain ends is still the same as this point. But, what is the difference now? Here, in this structure, the entire drain is heavily doped. So, the resistance drop in the drain region is negligible, so that this point is at the potential of V d. Now, here what is happening?

I have a voltage drop here. In the lightly doped drain region, I have a voltage drop here, right. So, I am effectively reducing the potential at the drain end of the channel, right. If I am reducing the potential here at the drain end of the channel that means effectively I am reducing the electric field, when the electrons reach the drain end of the channel. This is the basic principle.

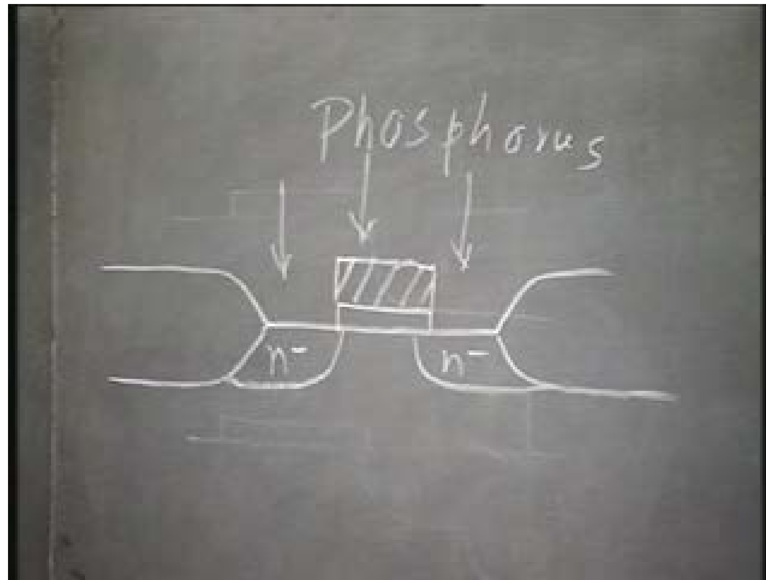
(Refer Slide Time: 30:14)



You have what is called a lightly doped drain, LDD or sometimes called the double doped drain, DDD; double doped drain, DDD or lightly doped drain, LDD. Now, you see, the point is this that the hot electron phenomena, hot electron effect is actually a, it can be reduced considerably, even if you reduce the peak electric field here by a small percentage. That is if the peak electric field here is reduced to 80% of its previous value, the hot electron effect will be reduced drastically, because it is related exponentially to the value of the electric field. So, even if you reduce it by a small percentage, the effect will be phenomenal.

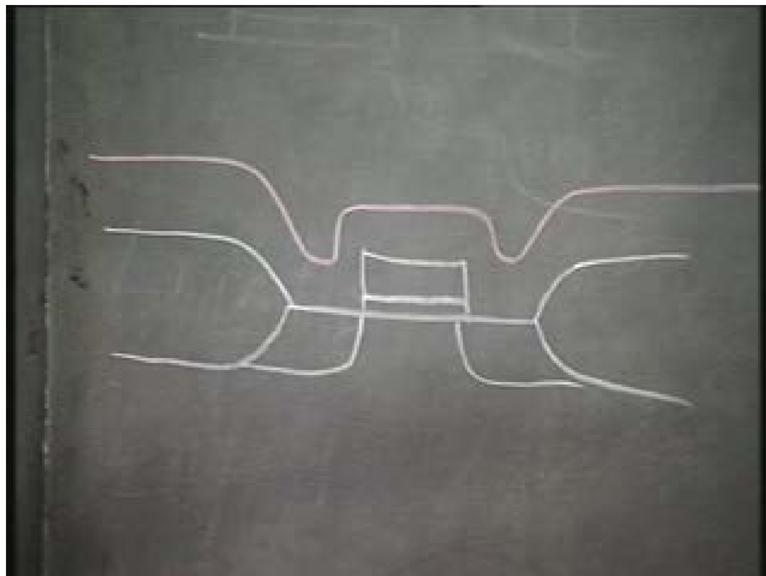
Now the question is, therefore, how do you realize this lightly doped drain structure? So, you understand that particularly when the channel length is being reduced, the hot electron problem is a very real problem. As the channel length is becoming reduced, electric field in the channel is increasing. Therefore, as the electrons reach the drain end, they may have sufficient energy to get injected into the gate. In order to prevent this, we must therefore have a lightly doped drain structure. The question is how do we realize this lightly doped drain structure? This can be done in various ways, by at least two different ways. Let me discuss at least one way.

(Refer Slide Time: 32:17)



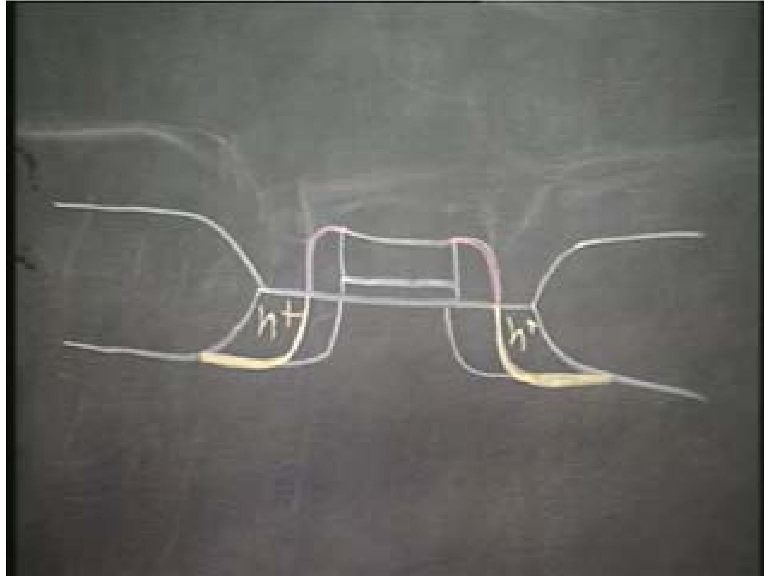
See, I have proceeded up to so far. This is my thick field oxide, this is my gate and this is my polysilicon on the gate, right. I have carried out the field oxidation; I have carried out the gate oxidation and patterned the gate with poly. Now is the time for me to do the source and drain diffusion. Now, what I do is I first do a phosphorus implantation. So, I get ….. I keep the dose deliberately low, so that I have n minus regions implanted in this and then what I do is I deposit a CVD oxide layer.

(Refer Slide Time: 33:38)



That is what I have now is …. Let me now deposit a CVD oxide, pattern the CVD oxide, so that I retain the side walls; I want to retain this side walls.

(Refer Slide Time: 34:17)

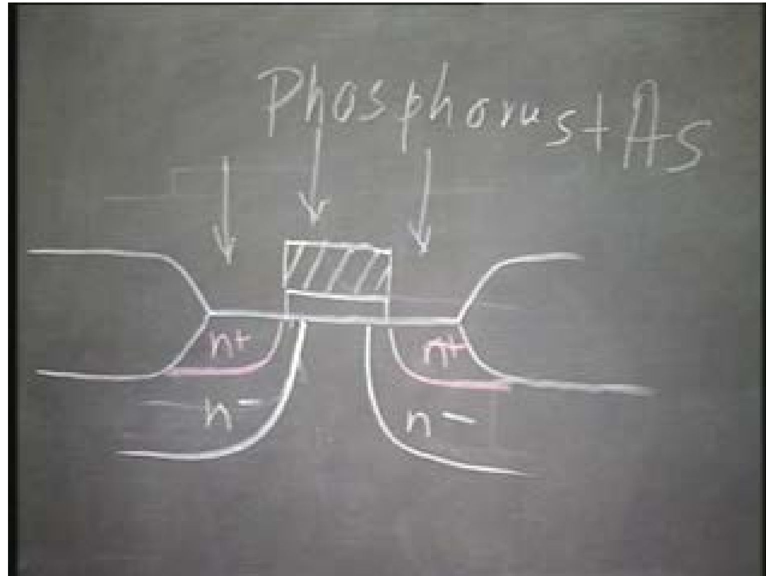


Except for this side walls, everywhere else the CVD oxide is removed. I have done an anisotropic etching, whereby the CVD oxide is removed only from the horizontal surface, but is retained on the side walls and now you see, all you have to do is to carry out a second implantation, so that these portions will now be protected by this side wall of CVD oxide and what you will have is n plus regions here and n plus regions here. It can even be deeper, anyway. So, this is literally speaking called the double doped drain, because you are using two implantation, double implantation. In order to realize the hot electron resistant structure, you have some part of the drain doped lightly, some part of the drain doped heavily, so that the peak electric field at this point is reduced.

Note that, since the MOSFET is a symmetric device, we are doing the same thing for source and drain, even though at the source end it does not really have any significance, right. The hot electron problem is only felt at the drain end. At the source end it is not felt, but because it is a symmetric structure, because in the mask design we do not distinguish between source and drain, we are doing it for both source and drain. This kind

of lightly doped drain structure can also be realized if you use double implantation through the same window, one with phosphorus and the other with arsenic.

(Refer Slide Time: 36:56)



That is in the first step itself, if through the same window you have phosphorus plus arsenic, then you see, phosphorus will diffuse more. Phosphorus has a larger R p, phosphorus will diffuse more. So, what you have is essentially this. This will be deeper and you have arsenic. So, this portion is n plus, this portion is n plus and this is n minus, n minus. Same thing is again established. Of course, the earlier technique will give you better control, because you can actually control the side wall thickness and therefore, precisely control the n minus layer width and the n plus layer width.

Here of course, you are doing a double implantation, phosphorus plus arsenic, using the arsenic lower depth to realize the n plus region and the phosphorus deeper depth, deeper and therefore more lateral spread also, you are using that region for n minus. What are the advantages? Here you are doing only one step, the process complexity is less; process complexity is less. Of course, the other technique will have more process complexity, but it will give you better control. So, these are two ways to realize the lightly doped drain

structure, but in both cases we are achieving the same thing that is we are realizing a hot electron resistant structure.

So, these five are the features of a modern day n MOS technology, apart from the polysilicon gate and the LOCOS technique. Apart from the polysilicon gate and the LOCOS technique which are now I mean understood, whenever we talk about a MOSFET technology, we are talking about a poly gate MOS and we are talking about the LOCOS technique, apart from that these are the five steps. One is the threshold voltage tailoring implantation, for which the boron implantation peak should be at the surface. Second is to prevent subsurface punch through for which the implantation peak should be approximately at the source and drain junction depth. Third is the channel stop implantation for which you have to do, carry out a boron implantation just after the silicon etching in the regions which are going to be covered by field oxide. Just underneath the field oxide you must have a heavy concentration of the substrate, in order to prevent parasitic MOSFET coming into conduction. Then, the fourth is the silicided gate technology, which we must have if we want to reduce the delays in the circuit and finally, the hot electron resistant structure, particularly important when the device dimensions are becoming smaller and that is in order to prevent the hot electron effects. So, this is so far as n MOS technology is concerned.

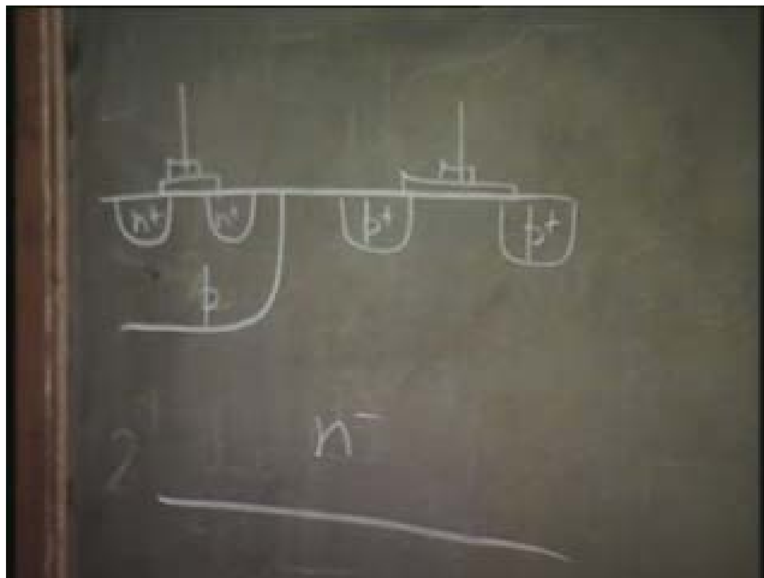
Of course, now more and more we are going towards CMOS technology. So, after n MOS technology, see, this is how the technology has advanced. First we started with the metal gate p MOS technology. Then we tried the n MOS technology, the poly gate n MOS, the silicided gate n MOS, etc, etc, right. So, these are the most detailed n MOS technology and then we have the CMOS technology. What is a CMOS technology? In a CMOS, we have complementary MOS devices. That is we have both a p channel MOSFET as well as an n channel MOSFET, right. Therefore we have to realize both a p channel device and then n channel device; various ways it can be done.

First of all, you know you can have a p-well structure, p-well structure. That is you start out with an n-type substrate. In that n-type substrate, you realize a p-well. Inside that p-

well, you will form your n MOS devices and in the n-type substrate, you will form your p channel devices, agreed. So, this p-well technology was found to be compatible with the p MOS devices, right. Except for this p-well implantation, it follows the p MOS device specification and you start with the same n-type substrate, right. So, this is the oldest CMOS. It was used in 1960's when even at that time, the p MOS technology was the most prevalent technology. So, people used p-well CMOS, so that it can be integrated in the p MOS process flow.

Now, you see, what are the advantages in this p-well technology? You are starting out with an n-type substrate. So, the n-type substrate is very lightly doped. Now, you are putting a p-well in this. So, the p-well doping has to be higher than the n-type substrate, right; it must be higher than the n-type substrate. So, that means it is easier for you to fabricate enhancement mode type device, right.

(Refer Slide Time: 43:42)

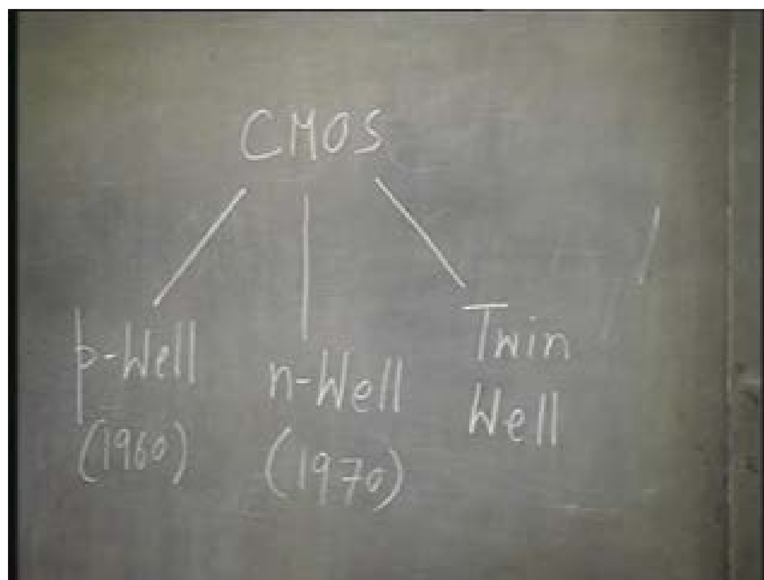


If the n-well doping is higher, sorry, I am sorry, if the p-well doping is higher, let us draw it out, so that it becomes clear; I start out with an n-type substrate, very lightly doped n-type substrate, then I form a p-well here, this is my p-well. So, obviously the p-well doping concentration is somewhat higher than this n minus substrate. Now, this p-well is

going to house my n channel device, right and this n-type substrate is going to house my p channel device. This is my CMOS, right. I have one n channel device and one p channel device. Now, you see, since this p-well concentration, doping concentration is somewhat higher than this substrate doping concentration that means my n channel MOSFET has a slightly higher substrate doping concentration. As far as the n channel is concerned, this p-well doping is its substrate doping concentration and you know, for an n channel MOSFET, if the substrate doping concentration is higher that means it is easier to form enhancement type devices, it is easier to have a positive threshold voltage. So, that is why the older technology, they preferred this; you can have an enhancement mode type n channel MOSFET, because in that time, 1960's and all, ion implantation was not, not such a commonly available technique. So, you can have enhancement type n MOSFET possible.

Then, in 1970's, when the n MOS technology became more prevalent, people started using n-well technology. n-well technology was more suitable for this n MOS compatible process, because you start with the p-type substrate and you form an n-well for CMOS, which houses your p MOS device and that is how it goes, right. So, from the p-well technique we came to the n-well technique.

(Refer Slide Time: 46:31)



In other words, we can say that the CMOS technology, it started with the p-well technique in 1960's, then we have the n-well technique in 1970's. But, in both cases what was the problem? The problem is, you see let us take up the n-well case. You started with a p-type substrate and you have doped an n region in that. So, the n region doping concentration is higher. So, it became more and more difficult if you want to match the threshold voltages of the two devices. So, as the circuit complexity increased, dimensions reduced. People started thinking that it might be better if we can control the two substrates that is the substrate for the n channel device and the substrate for the p channel device independently. So, then we came to the twin-well or twin-tub technology.

That is in twin-well technology, you start with very lightly doped substrate, almost nearly intrinsic and then you dope both an n-well and a p-well. The n-well is going to house your p MOS, the p-well is going to house your n MOS and you can tailor the two well doping concentrations independently, individually and tailor the threshold voltages of your CMOS device, right. So, these are the basic aspects of a CMOS technology. Only the well formation, we need to bother about the well formation. After that it follows exactly the same way as the n MOS technology is concerned, right. The only difference however is, you also have a p MOSFET. Therefore, you have to carry out a separate set of source and drain diffusion for that. But that apart, all the other technological considerations remain the same.

The only difference is how do I realize the two regions to house my p channel device and the n channel device? In the next class, we are going to take up this issue. First we look at a CMOS which uses an n-well technology. We are not going to discuss the p-well technology, because nobody uses the p-well CMOS anymore. So, we are first going to discuss an n-well technology and then we will go to the twin well technology and then I will show you how as the complexity of the circuit increases, how the twin well substrate starts to resemble a bipolar junction transistor substrate more and more and from there, we will go to the biCMOS technology. That is a biCMOS technology is one where we will house a CMOS transistor as well as a bipolar junction transistor using as few additional steps as possible. But in order to do that, first we will have to see what a

modern day CMOS is like, what does it look like, what are its requirements? So, we will first see a not so modern CMOS, an n-well CMOS and look at its drawbacks and then see how in a twin-well technology, these drawbacks can be ironed out and what are the requirements in this twin well technology and then we will look at the biCMOS technology.