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

Lecture - 38
CMOS Technology

So, we are talking about CMOS technology, CMOS technology which is one of the most important technologies in today's VLSI. Now, from the point of view of technology, in a CMOS, we can realize both a p channel and n channel MOSFET in three different ways. That is one is you start out with an n-type substrate and form a p-well there, which is the p-well technology and because you are starting out with the n-type substrate, essentially it is compatible with the p MOS fabrication process flow; so, n-well technology which is, sorry, p-well technology which is compatible to the p MOS process flow and then, similarly we can have, we can start with a p-type substrate analogous to an n MOS process flow and in that p-type substrate, we can create an n-well.

(Refer Side Time: 2:41)

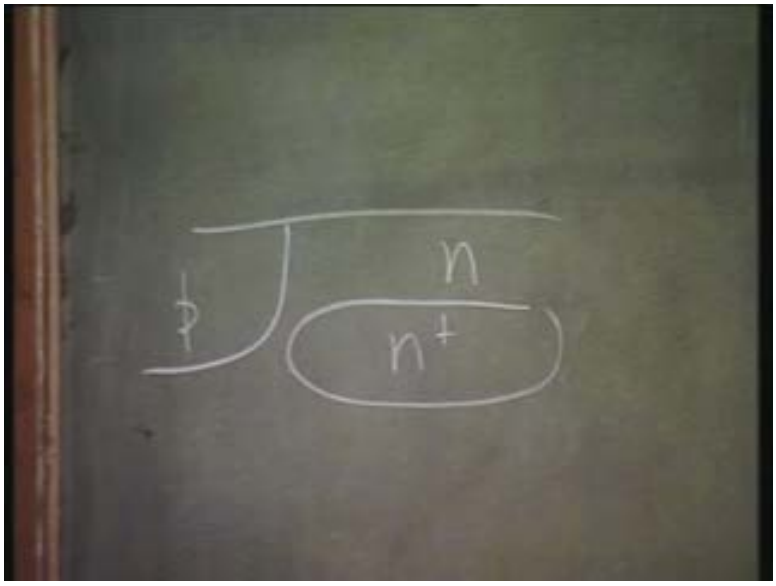
So, we can have in CMOS technology, either a p-well or an n-well or better still, we can start with a very low doped substrate material, nearly intrinsic and we can have both the

wells, which is called the twin well or twin tub technology; twin well or twin tub. These are the three different ways in which a CMOS can be realized.

(Refer Side Time: 3:32)

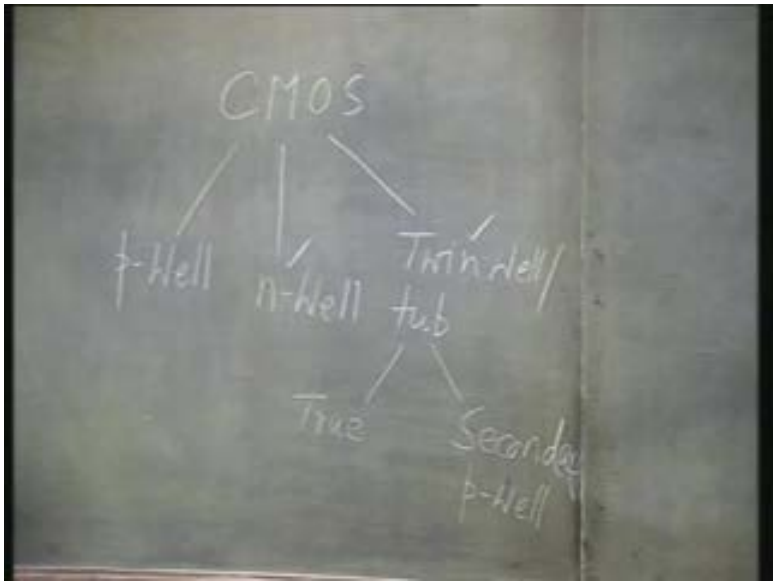
Now, in this twin well technology again, we can have a either a true twin well, that is, in which case you really start out with a nearly intrinsic substrate and you dope both a p-well and then n-well; the other possibility is to have a secondary p or n-well. The secondary p or n-well, you know, in this twin well is essentially an extension of either this p-well or this n-well technology. In this secondary case, again essentially you start either with an n-type substrate and then essentially you create, in that n-type substrate you essentially create a p-well only. But, you do a deep substrate implantation to prevent the punch through in the n-type substrate itself. So, what is happening?

(Refer Side Time: 4:42)



In the secondary, let us say if it is secondary p-well, you are starting with an n-type substrate; you are actually creating a p-well here. So see, basically it is an extension of this p-well only. Only thing what you are doing is in this n-type substrate, you are having a deep, somewhere deeper inside, you have an n to prevent punch through. So, this is the secondary p-well. In the same manner, you can also have a secondary n-well, in which case you basically have an n-well, but in the p-type substrate also, you carry out a deep implantation to prevent punch through.

(Refer Side Time: 5:42)



Now, as far as the CMOS technology is concerned, these two are the most important techniques today. Previously, because n MOS was the dominating device, the most widely used MOSFET we used n-well technology, which was compatible with the n MOS process flow and now as CMOS itself has become the dominating device in VLSI, now we use twin tub or twin well technology. So, let us first discuss the n-well technology, how exactly the well is created and then we go over to the twin well technology.

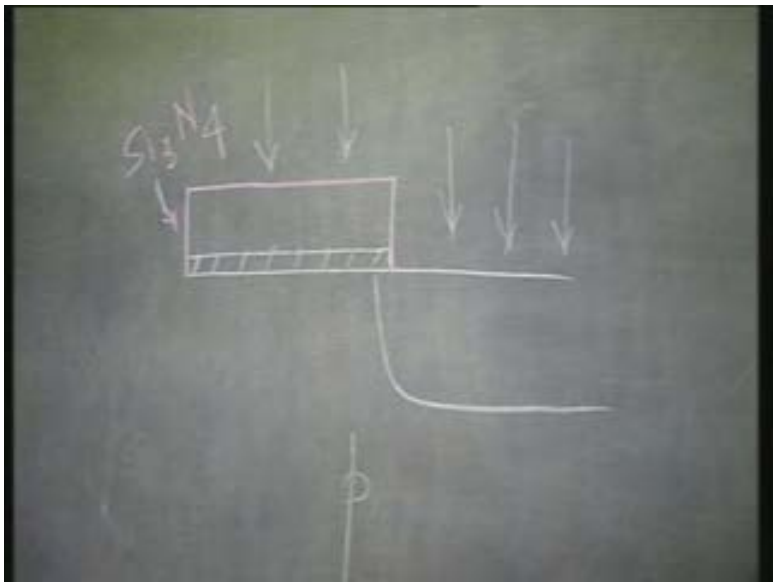
Now, in the n-well technology obviously, you start with a p-type substrate, right. Remember, with the, with better understanding of the oxide charges and the interface states, now we know that 1 0 0 substrate has at least one order of magnitude lower interface states. So, for MOS technology, universally we use 1 0 0 substrate. In contrast, the first metal gate p MOSFET which was fabricated, it was fabricated on 1 1 1 substrate. That was because, it is a very old technology; in those days, 1 1 1 was still by far the cheapest substrate, right. People did not quite understand the relation between the substrate orientation and the interface state density. But then, as it became clear, now, for any surface devices, MOSFET is a surface device; in contrast bipolar junction transistor is a bulk device, right, so for MOSFETs, you always use 1 0 0 substrate, right.

So, in the conventional n-well architecture, you start out with a p-type substrate, 1 0 0 p-type substrate, right.

(Refer Side Time: 7:55)

Now, your first mask is going to be the n-well mask. That is I must dope certain region in this p-type substrate as the n-well and how do I do that? I do that by a LOCOS technique.

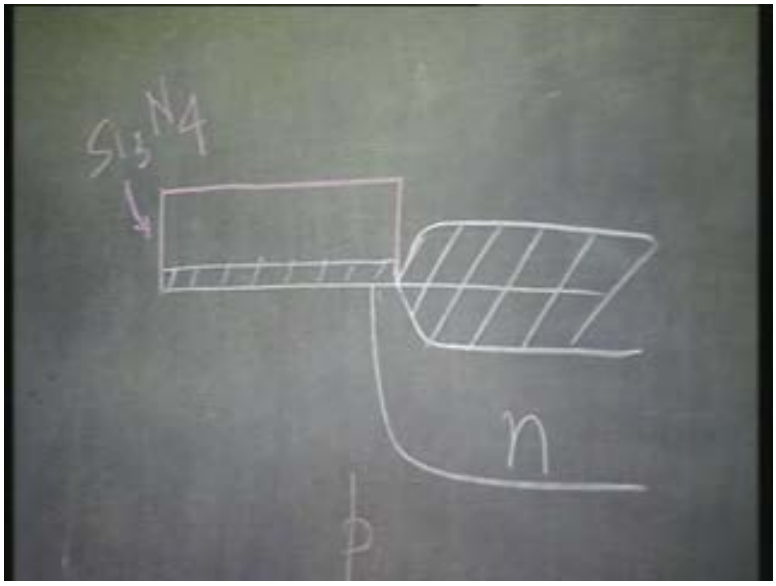
(Refer Side Time: 8:18)



That is I protect certain region of the p-type substrate using an oxide nitride mask. This is nitride sitting on top of the pad oxide. You can have finer techniques here, like side wall masking and all that; don't let us bother about that, essentially what you do is you are planning to carry out a LOCOS technique. You are going to protect certain region in this with an oxide nitride mask and the rest of the region you are going to have a phosphorus implant. So, you see, this implantation can be a blank gate implantation; certain regions are protected by this silicon nitride-silicon dioxide mask. So, you can just, with this mask on you can just put it on for a phosphorus implantation, right and once you have the phosphorous implantation, you have certain regions doped n-type.

Now, in the next step, you carry out the drive-in in an oxidizing ambient. You see, you want to have an …, you want to have a deep np junction, right. So, you carry out the drive-in in an oxidizing ambient. So, what will happen? Oxide will form over the regions which are not protected by the nitride and oxide mask.

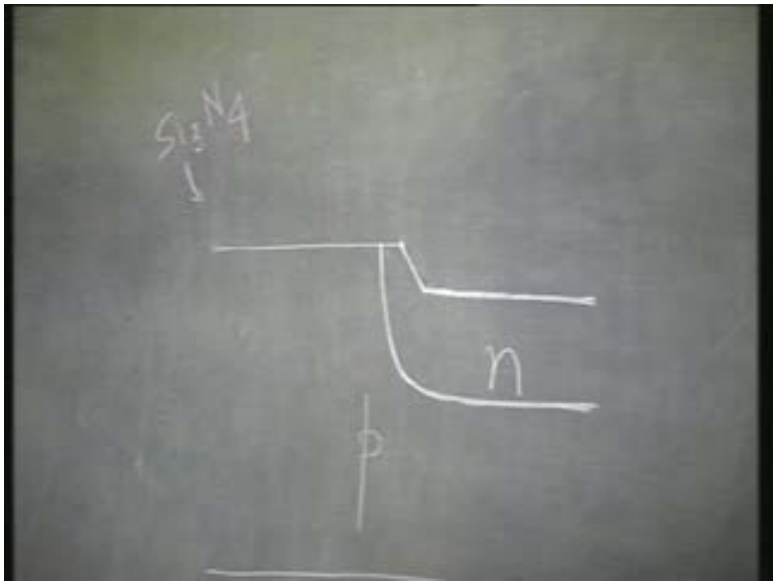
(Refer Side Time: 10:25)



So, what I will have is I will have the oxide grown like this and in the process this whole thing will be driven deeper like this. So, this is your n region, this is the nitride and this is the oxide. Now, I want to strip off the nitride as well as the thick oxide that is grown on

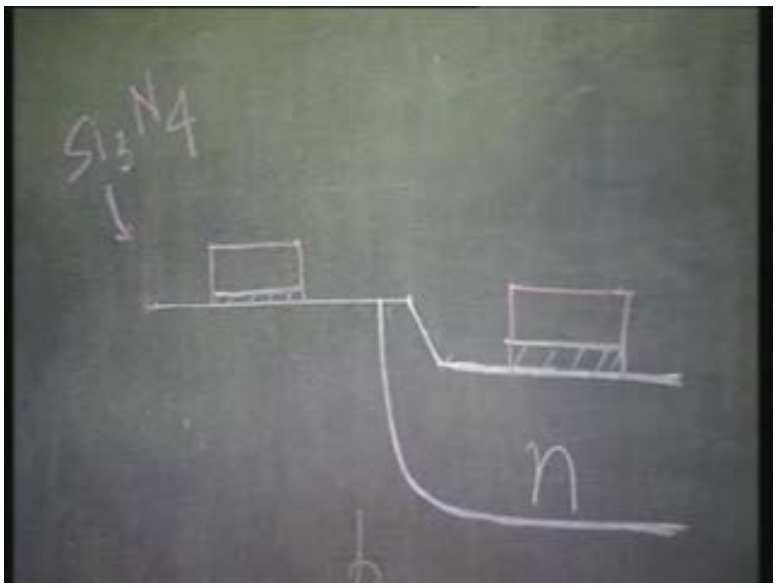
top of the well. I want to strip of both this oxide and as well as this nitride and oxide. So, what do I have if I strip it off? Right?

(Refer Side Time: 11:33)



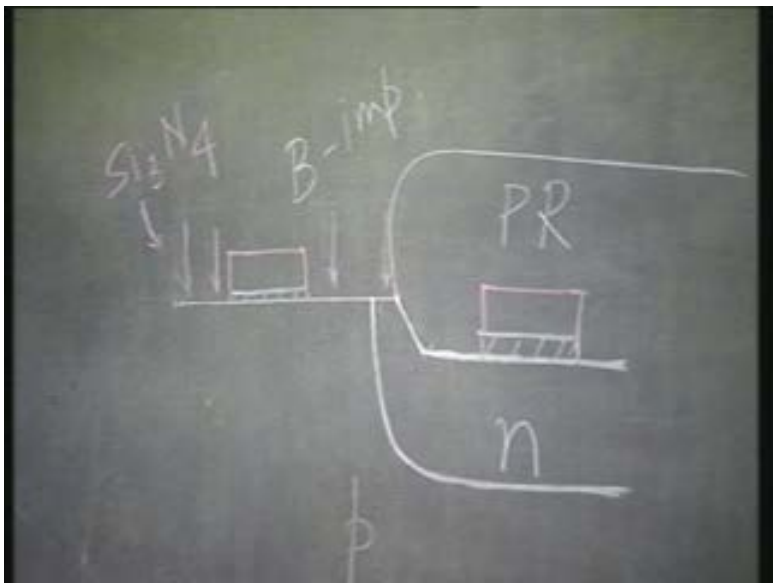
I create a step in the semiconductor. What is the purpose of this step? This step actually signifies the position of the n-well. So, now you have marked the two places. This is your, this is the region where you are going to form your n MOSFET, this is the region where you are going to form your p MOSFET and this step demarcates the two regions, right. Next is of course very simple. After you have achieved this, in these two regions, now you have to have the field oxide and the area for the active transistor.

(Refer Side Time: 12:31)



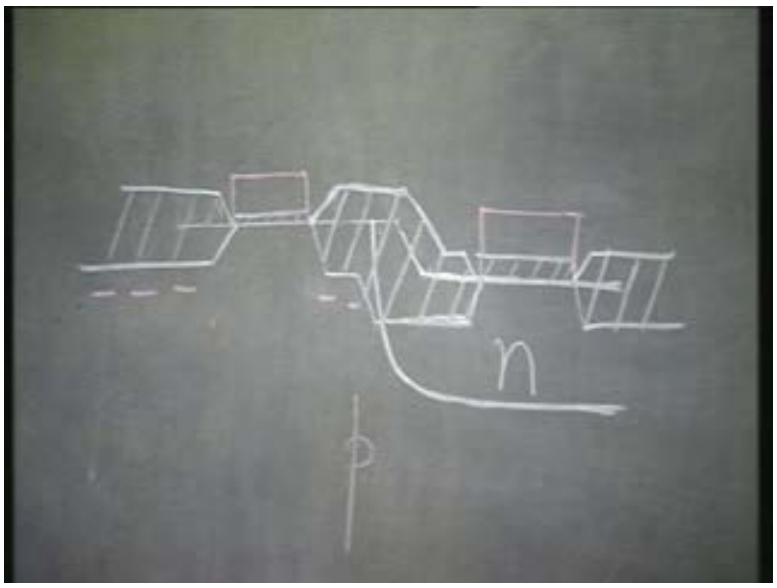
So, again you see, you have to carry out oxide nitride combination and pattern it and then carry out LOCOS. But, before carrying out LOCOS, what you can do is you can do the boron implantation for the channel stop, right. Remember, we mentioned this yesterday that underneath the field oxide the substrate doping must be raised, so that the parasitic MOSFET does not get inadvertently turned ON. So, after you have protected certain regions you may or may not want to etch the silicon. If you want a recessed oxide, you will etch silicon to the required depth. If you do not want a recessed oxide, you do not have to etch.

(Refer Side Time: 13:48)



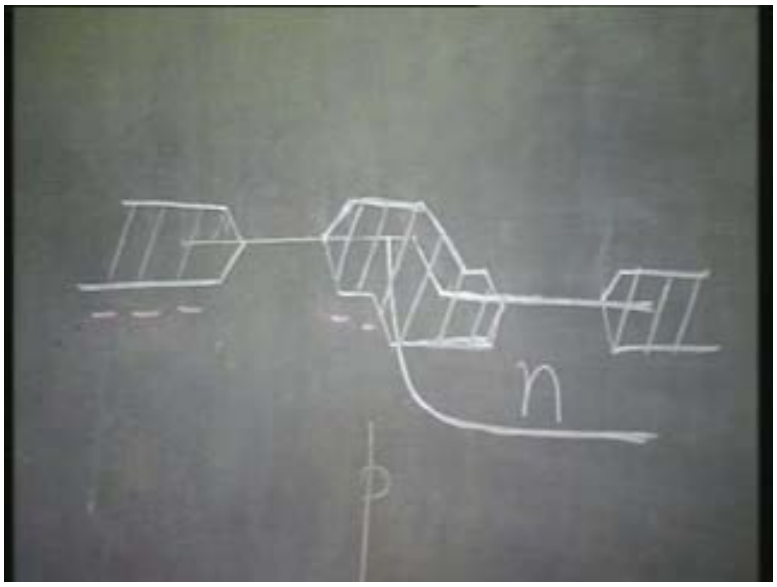
But, you do a boron implantation and of course when you are doing this boron implantation, you have to prevent this implantation from going into the n-well. So, this region must be masked with photoresist. You are using ion implantation, so no problem, photoresist can be used as a mask and there also you see, this indentation will come in very useful to align the photoresist mask on to the n-well. So, now you have done the boron implantation, which is the channel stop implantation for the n MOS and in the next step, you grow the thick oxide.

(Refer Side Time: 15:04)



So, now underneath this you have boron implantation; may be I should just show it with dots and now I can remove this silicon nitride-oxide mask.

(Refer Side Time: 16:16)



So, you see, you have your two active transistor areas defined, right; rest is all protected by thick field oxide. You have your n-well. This is the region where your p MOS is going to be spaced, this is the region where your n MOS is going to be spaced; you have done

the channel stop implantation to stop the parasitic MOSFET from getting inadvertently turned ON. Next, you can just follow it up with gate oxidation, gate polysilicon deposition, patterning and source and drain diffusion, right.

So what were the steps? The first step was to have a nitride oxide mask for the n-well. First you have a nitride oxide mask to demarcate the regions where your n-well is going to be placed. Next step, you do a phosphorus implantation to form this n-well. Then, you carry out the drive-in in an oxidizing ambient, so that the n-well is driven deeper and at the same time you have a thick oxide forming on the n-well. After this, you do the stripping of nitride as well as oxide. Because there was a thick oxide formed on top of the well, you have a step in the semiconductor surface, which helps you to identify the n-well regions.

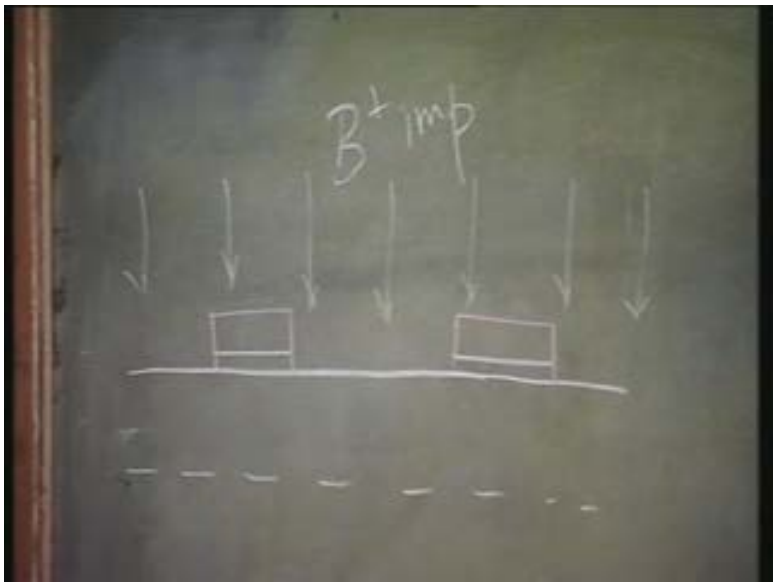
Now, in the next step, you again do a nitride oxide mask, mask step, but this time, this is in order to define the active transistor regions and with these masks in place, you carry out the field oxidation. But prior to that, you do a channel stop implantation. Now, you do the field oxidation. So, you see, apart from the active transistor region, the rest of the surface is protected by thick field oxide. Underneath the thick field oxide in the n MOS region, you have the channel stop implantation and you have both your p-well and n-well, I mean both your p MOS regions and n MOS regions clearly defined. So, now you can follow it up with gate oxidation, patterning, source and drain diffusion. So, this is the conventional n-well technology, which is compatible with the n MOS technique.

Now, in the twin tub technology of course, we will form actually two wells. In this case, we have formed only one well that is the n-well. In the twin tub technology, we are going to actually form two wells. So, you start with a very low doped p-type substrate, very low doped p-type substrate.

(Refer Side Time: 19:33)

So, let me start with a very low doped p-type substrate. At the beginning again you have a LOCOS.

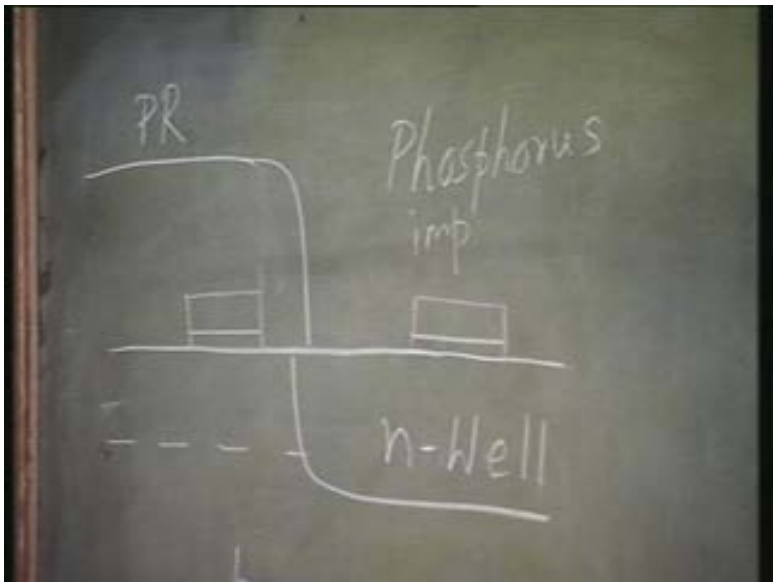
(Refer Side Time: 19:58)



That is you first of all protect certain regions using the pad oxide and the silicon nitride. Remember, in this case I only had one mask. I only had to protect the n MOSFET region.

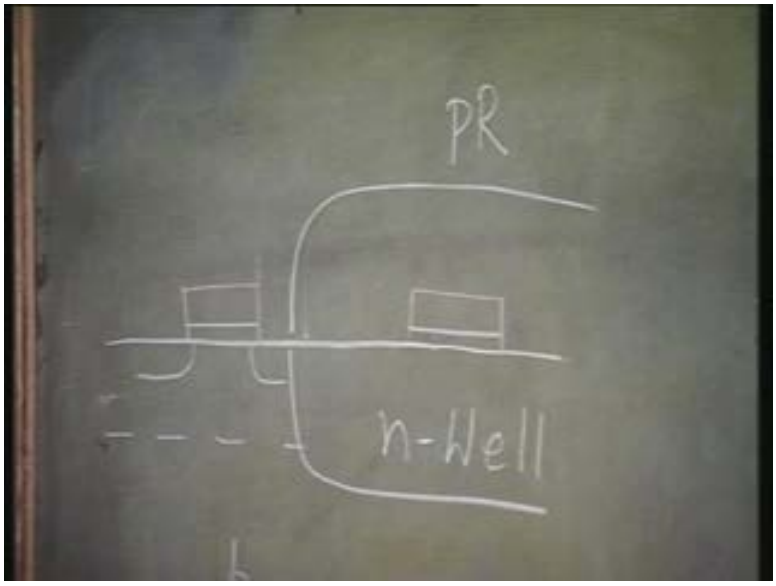
Now, I have two masks, because I am going to have two wells; I must protect both the regions. After this, you can do a boron implantation. Usually this is a blanket implantation that is I do not use a mask. With this silicon nitride-silicon oxide in place, I carry out a blanket boron implantation; it is diffused p, implanted p all over. Now, you have the n-well mask and you implant it. So, what do we do?

(Refer Side Time: 21:55)



We just use photoresist as the mask and the rest of the region we do phosphorus implantation; may be it is not so critical, so, I can actually come up to this. If you do a phosphorus implantation, this implantation dose is much higher than the previous boron implantation. So, what you have essentially is you have the n-well. After this n-well is formed, you do a boron low energy implantation, low energy boron implantation.

(Refer Side Time: 23:24)

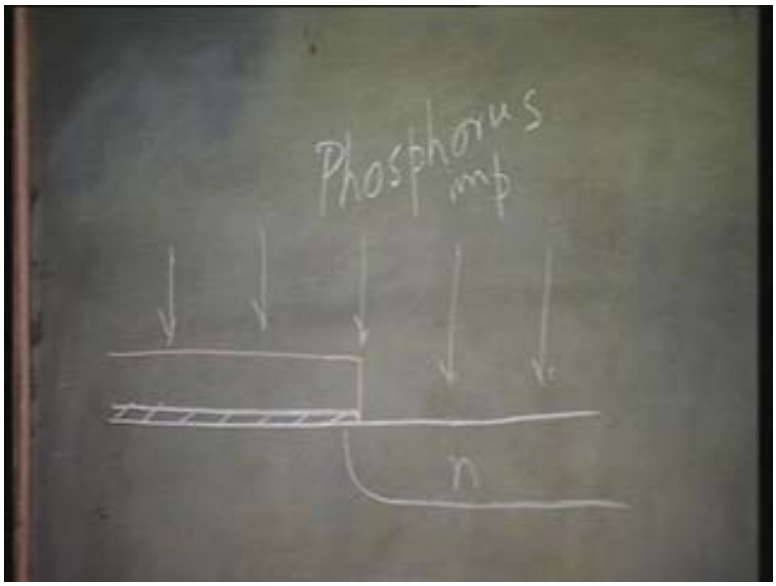


So, what you have is you just protect this region now, it is a photoresist here; do a low energy boron implantation. Because you are doing it at low energy, this mask will mask the boron implantation; you will have just boron here and boron here. So, essentially what you are doing now is the channel stop implantation for the n channel MOSFET. So, you see, this is actually the example of the secondary twin well, secondary twin well. I told you that in twin well, I can either have a true twin well or I can have a secondary well. This is essentially a secondary well in which case, you see, I have actually formed the n-well. I have started out with a p-type substrate. I have actually formed the n-well, analogous to in this particular case. But, there is one additional step involved in this.

What is this one additional implant? That additional implant is the first boron implant which I have done, so that I will have a buried, deep B implant. This is done in order to prevent the subsurface punch through. So, essentially this technology is analogous to the n-well technology. The only difference is in this additional boron implantation, which is helping to prevent subsurface punch through. So, this is an example of the secondary n-well twin tub. This is not a true twin well technology. This is a secondary n-well technology.

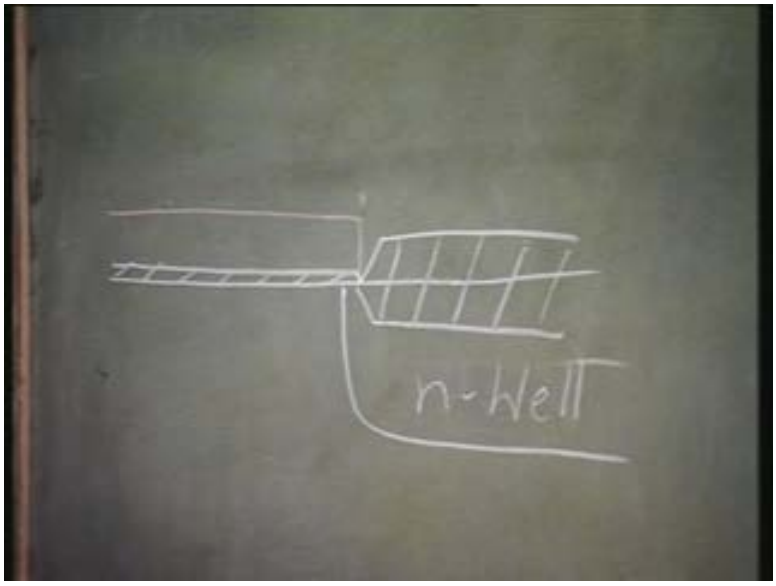
Now, let us come to a true n-well technology, sorry, a true twin well technology. In the true twin well technology, you actually can start with either p-type or n-type substrate. The only requirement is that your substrate must be very low doped, it should be near intrinsic. So, I do not really put a restriction on what should be the type of the substrate.

(Refer Side Time: 26:16)



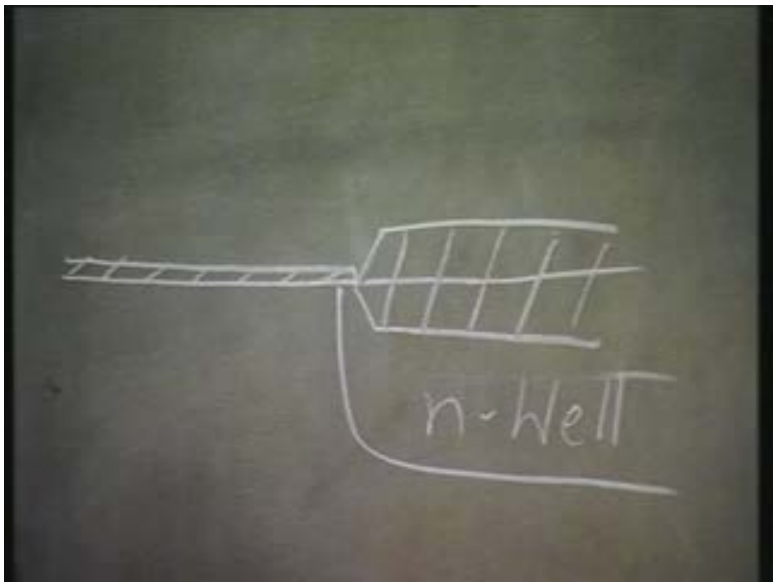
I can have either very lightly p-type substrate or very lightly n-type substrate. Only thing I do is I protect certain region in this by using an oxide nitride mask and then I do, first in this particular case I have taken up a phosphorus implantation, so that you have your n-well. Now, you see, now you carry out an oxidation over the n-well; I mean you subject the entire substrate to an oxidation process. The oxide is going to be formed only over the n-well, because the rest of the portions are protected by the oxide nitride mask, right and while you are doing this oxidation, this n-well will also be driven deeper.

(Refer Side Time: 27:49)



So, what you would have at the end of that oxidation process is something like this. You will have an oxide and underneath will you have the n-well. So, you have already formed your n-well. Next step, what do you have to do? You have to form your p-well. In order to do that, first step you strip off the nitride.

(Refer Side Time: 28:13)

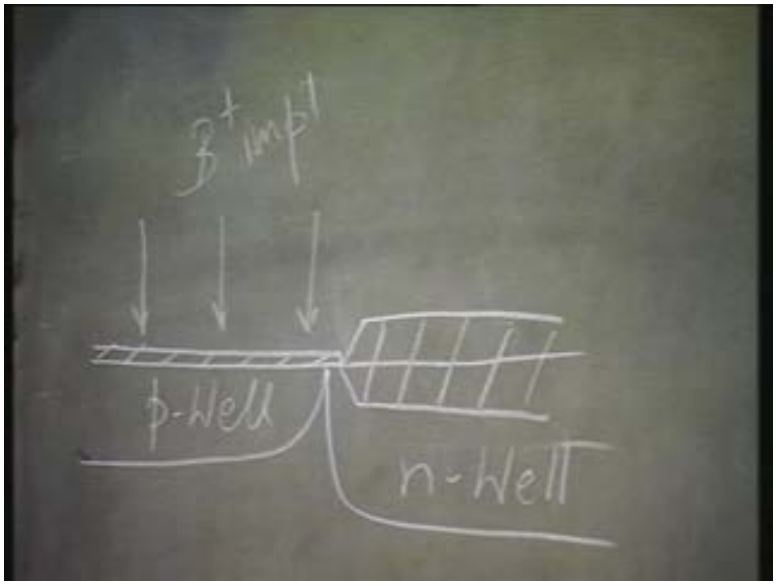


Now I carry out a boron implantation I do a p-well implantation. Now, the interesting thing is usually in these kind of twin well technologies, you know, we use mask for one well. Already I have used that mask for the n-well. I protected certain regions using the oxide nitride mask and then I have done the phosphorus implantation. Now, you see, I have grown a thick oxide on top of the n-well, while on top of the rest of the portion, there is only a very thin oxide. So, the implantation energy is such that it can penetrate this thin oxide, but it cannot penetrate this thick oxide, right.

So, now when I want to carry out the p-well implantation, I do not have to bother about any mask; I can do a blanket implantation. The whole point is like this. As we are reducing the device dimensions, it becomes more and more critical if we have a lot of alignment, lot of mask alignment to do. The smaller the device dimensions are, the more difficult it is to align. So, we always want to go for self-aligned technology, right, where one region is automatically aligned to the other. This is what we are achieving here. You see, we first defined a mask pattern. Using that mask we have our n-well implanted and driven in in an oxidizing ambient, so that there is a thick oxide formed on top of the n-well, to protect that n-well region, right. Once this protection is there, now, I do not have to bother about another mask, in order to have my p-well. I can do a blanket implantation. The p-well and the n-well are automatically aligned to each other. This is the interesting point of this.

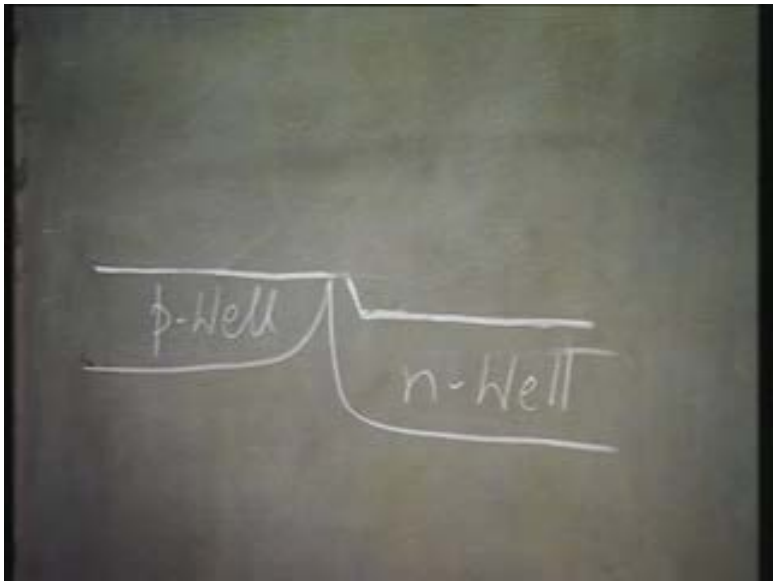
So, then next what do you do? You do a boron implantation.

(Refer Side Time: 30:38)



Then you have your p-well. Then, I mean you can do a common drive-in step together, so as to obtain your necessary thicknesses of p-wells and n-wells, after removing all this oxide.

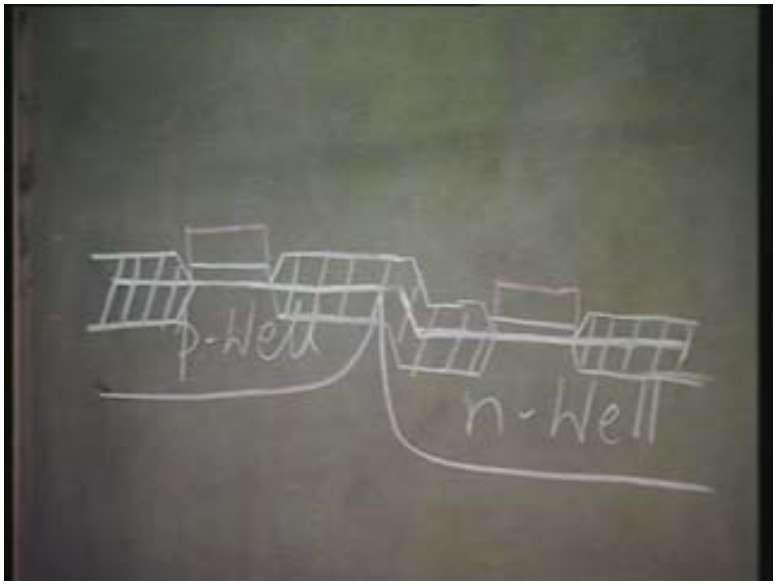
(Refer Side Time: 31:16)



You have your p-well and your n-well and demarcating the two regions, you also have a step, right. So, you know the region where you are going to form your n MOSFET and

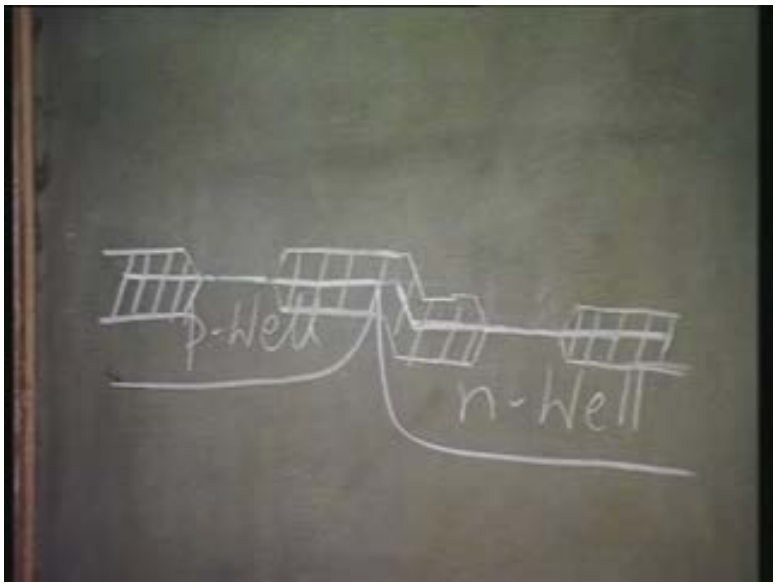
the region where you are going to form your p MOSFET. So, what is going to be the next step? Next step is going to be your active area definition.

(Refer Side Time: 31:56)



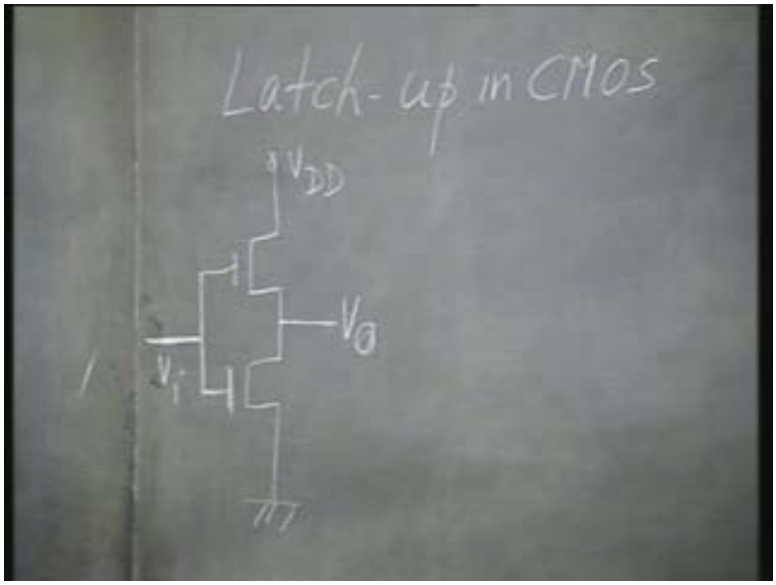
So, again you form a LOCOS mask and then carry out the thick field oxidation and then remove the LOCOS mask, strip off the silicon oxide and nitride.

(Refer Side Time: 32:47)



So, you see, you have your p-well, you have your n-well, you have certain regions protected by the thick field oxide and this rest of the portions are for your active transistors. So, this is a true twin well technology, a true twin well technology. You have formed really both a p-well and an n-well. As the CMOS becomes more and more complex, we have to add other process steps in addition to this basic step. So far whatever CMOS technology we have discussed, you know, we are only using a bulk substrate layer, right. We do not have to use, so far we did not have to use a buried layer or an epitaxial layer; so far we did not use that. So far we only have p-well and or n-well, right. If you are having a twin well technology, you have both an n-well and a p-well; otherwise you have one or the other.

(Refer Side Time: 34:49)

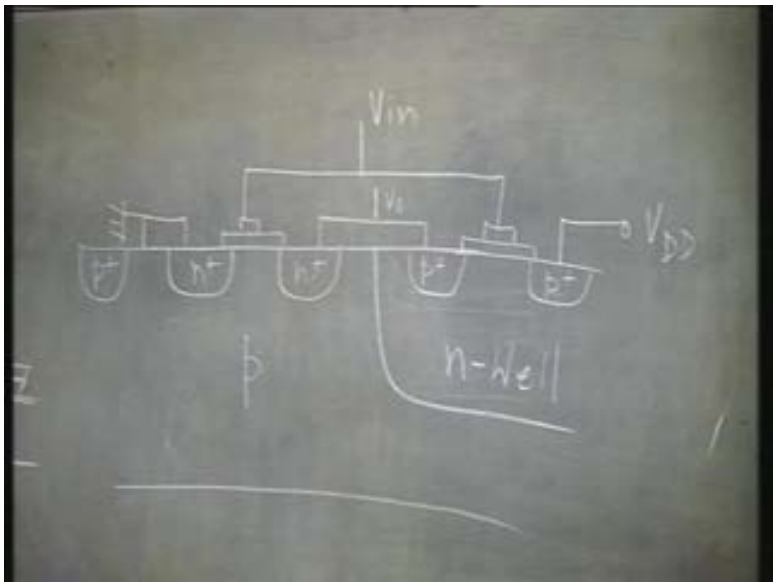


Now, CMOS devices however, can have a very serious problem called the latch up, latch up in CMOS devices. Now, you see, a CMOS inverter, a basic CMOS inverter is something like this. You have, this is your input, this is the output. This is grounded, this is V DD. This is the basic CMOS inverter. The top one is the p channel device, the bottom one is the n channel device. You use the p channel device as the load, the n channel device as the driver, right. So, this is your basic CMOS inverter. Very simple; if your input is high, this gate has a high voltage. It is an n MOS, so it is going to conduct.

We have a path for the current, therefore the output is going to the loop; inverter. If the input is low this gate is not going to conduct, the output is going to be high, right.

Now, if a CMOS inverter is connected like this in a transistor, I mean in a circuit, if I have a CMOS inverter like this, what do I have?

(Refer Side Time: 36:51)

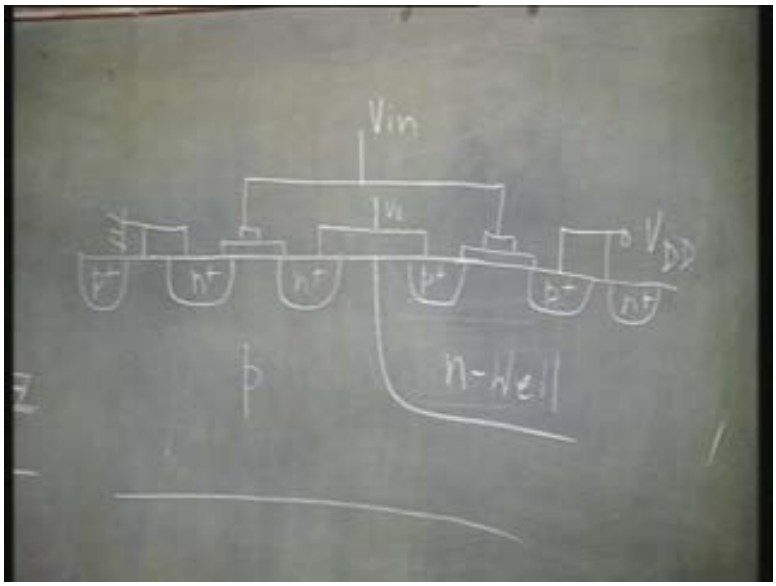


I have a, let us say, this is my n-well, so this is my p MOS, p plus, p plus. This is n plus, n plus. Two gates are tied together, right. These two are also tied together. This n plus is grounded. We will have a p plus here. This is a substrate contact which is also tied up this. Now, you see, if because of the presence of a noise signal this v output goes below the ground potential, suppose because of the presence of a noise at the output terminal you have a negative voltage spike appearing, so that this pn junction gets forward biased, you see, this p is tied to this, it is grounded. So, this p is at ground potential. At this n, if you have a momentarily negative voltage appearing, then this pn junction is going to get forward biased. So, what will happen? This n plus region will inject electrons here. Now you see in this particular structure I have two transistors as well, npn, right. This n plus, this p-type substrate and this n-well; there is an npn transistor. Similarly this p plus, this

n-well and this p-type substrate, I have a pnp transistor also. I have both an npn transistor and a pnp transistor.

Now what will happen? These electrons can enter this npn transistor. You are injecting electrons here; it can go out in this. So, what is basically happening? It is an npn transistor; you have forward biased the base emitter junction, right.

(Refer Side Time: 41:19)



So, it starts conducting and so, these electrons are finally collected out here. This is the substrate contact for this n-well. So, they are finally collected out of this, go out in through V DD. If there is sufficient resistance in this n-well, remember, your well doping concentration is pretty low; if there is sufficient resistance here, there could be a potential drop as the current is flowing, right. There could be a potential drop and it could so happen that this potential drop is sufficient to turn ON this pn junction as well. So, the process will be regenerative. Essentially what you have is called a npnp structure, right. You know that thyristor is as npnp or pnpn structure, right.

Now the problem in this is that once the noise spike has occurred, which is sufficient to turn ON this junction and there is flow of electrons and in that process there is a

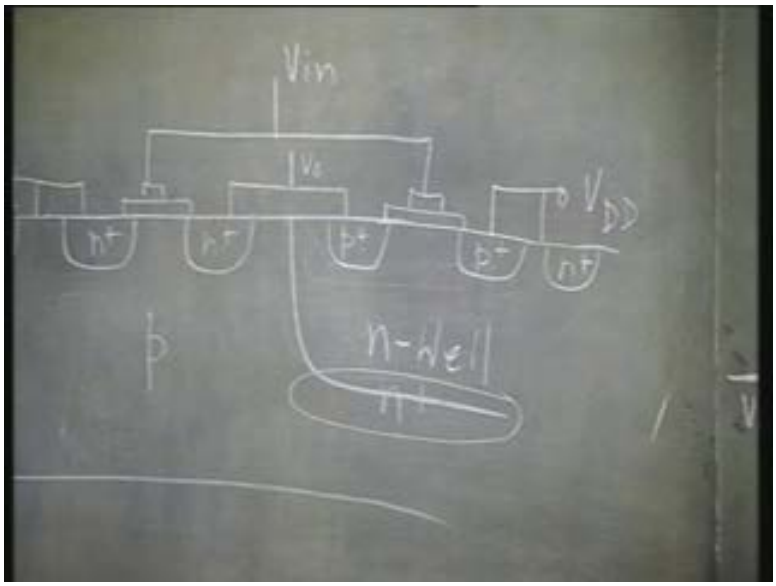
sufficient potential drop, so that this junction again gets forward bias and keeps injecting holes here, the process just goes on and on. It becomes a regenerative process. So, even after this noise signal, even if this noise signal is momentary and even if it is withdrawn afterwards, right, it does not stop; the process just continuous, right. So, this is called the latch up; this is called latch up. This is because, every CMOS, in every CMOS device, if you have a CMOS inverter, because of the way it is connected, you always have a npnp or pnpn combination, which is a potentially dangerous combination. Once it gets fired, you know, once the thyristor gets fired, you do not have a control on it, except by switching OFF, right. So, once this npnp structure fires itself, even after the noise signal is removed, you cannot really control it. So, the latch up is actually a major problem in CMOS.

One way you can prevent this latch up is, you see, essentially we have two transistors, npn and pnp, by keeping the beta of these transistors low and the one way, this is the older technique of keeping the beta of the transistor low, that was by doing gold doping or neutron irradiation, so that the life time is killed. But, these techniques are not really used very much these days, because you cannot accurately control. So, that technique of controlling beta is not very much in use. The other option is of course you understand that all this is happening because of the substrate resistance, right; the whole point of the regenerative mechanism. Suppose electrons are injected from here, fine, they are collected out of this V DD. If the matter would have stopped there, fine, but what is happening? In the process, we have a sufficient resistance drop here, so that you know this is actually connected to V DD. This point between this and this you have a sufficient voltage drop, so that this n-well is actually at a lower potential. That is what is making this pn forward biased. That is what is causing the regeneration problem, regenerative problem.

So, the other possibility is you try to keep the well resistance low. You do not allow a sufficient resistance to be there, in order that this regenerative mechanism can take place. But that has its drawback. How can I control the well doping concentration, how can I allow the well doping concentration to go high? If the well doping concentration is high,

then the device is more prone to the body effect as well as to the capacitance effects. So, we cannot just like that allow the well doping concentration to go high. What then can we do? We can have buried layer, we can have an epi material; you can have this n-well as a … you can have the substrate material itself as, you know, p on p plus type or what you can do is you can use what is called the retrograde structure and what is a retrograde structure? In a retrograde structure, you have buried layers.

(Refer Side Time: 47:01)



That is if I can have an n plus region here, then the current will always flow through this, right. But, as far as the depletion layer is concerned, it will still see, still see the low doping concentration of the n-well. I am cutting down the shunt resistance path without actually affecting the doping concentration of the well. So, this is called the retrograde structure and the retrograde structure has now become quite popular in CMOS in which case, what we do is aligned with each well, p-well and n-well, we have a p plus or an n plus buried layers; buried n plus layer with the n-well, buried p plus layer with the p-well, which actually brings the CMOS technology quite close to the bipolar junction transistors, because now you need to have the buried layers, right.

So, you see, one possibility will be to use steps analogous to the bipolar junction transistor. You just have a buried n plus as well as a buried p plus and then, on top of that you grow an epitaxial region and then you just create your wells in that epitaxial region. Make sure that the epitaxial region is aligned to each other; I mean the well is aligned to the buried layers. So, this is one way in which CMOS technology is coming close to bipolar junction transistor technology and which is in fact utilized, when one is making the biCMOS device. So, we will talk about the retrograde wells with epitaxy for CMOS in the next class and then we will see how from there we can realize a bipolar junction transistor in the same process flow.