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

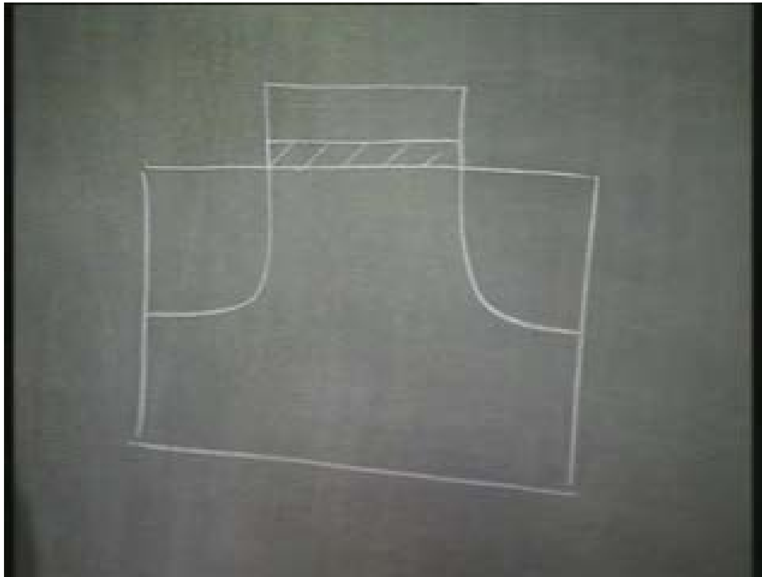
Lecture - 33
Problems in LOCOS + Trench Isolation and Selective Epitaxy

So, we are discussing about a bipolar junction transistor in integrated circuit and the first problem that we are discussing is the problem of isolation. You have already seen that in older day bipolar junction transistor technology, the isolation was provided by a deep pn junction. So, this was called the junction isolation technique, where the reverse biased pn junction would act as the electrical isolation between adjacent devices. But, this older technology proved to be not quite up to the mark, as the device dimensions were reduced and as we wanted faster and faster devices. So, on both counts junction isolation is not a viable technology, because number one, the diffusion will always have some lateral spread; therefore, as the device dimensions become smaller, it becomes more difficult to control the tub width, isolation tub width and secondly the capacitance associated with these junctions, the collector isolation capacitance which was going to hinder the speed of the transistors.

So, from junction isolation technique, we came to dielectric isolation technique and the most common, most useful dielectric, silicon dioxide was used for isolation and therefore we had the LOCOS technique, which is nothing but local oxidation. Local oxidation that is oxidation is carried out at some regions of the semiconductor; the other regions are prevented from getting oxidized by using a silicon nitride mask. So, it is actually the interesting property of silicon nitride that it does not allow oxidation to proceed underneath that has been utilized and LOCOS technique therefore, is a very interesting and very widely used isolation technique today. But, as the device dimensions went on becoming smaller, the LOCOS technique also started to show its limitation and what are these limitations?

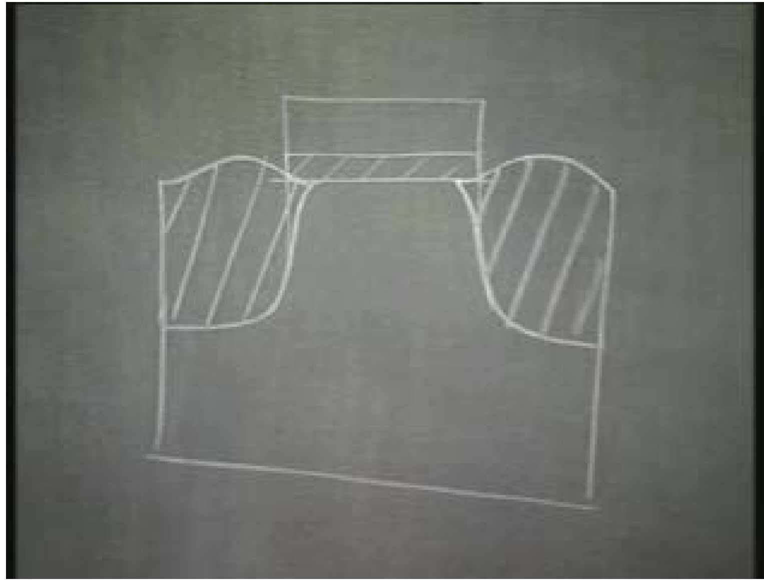
First of all you see, in LOCOS what do you do?

(Refer Slide Time: 3:50)



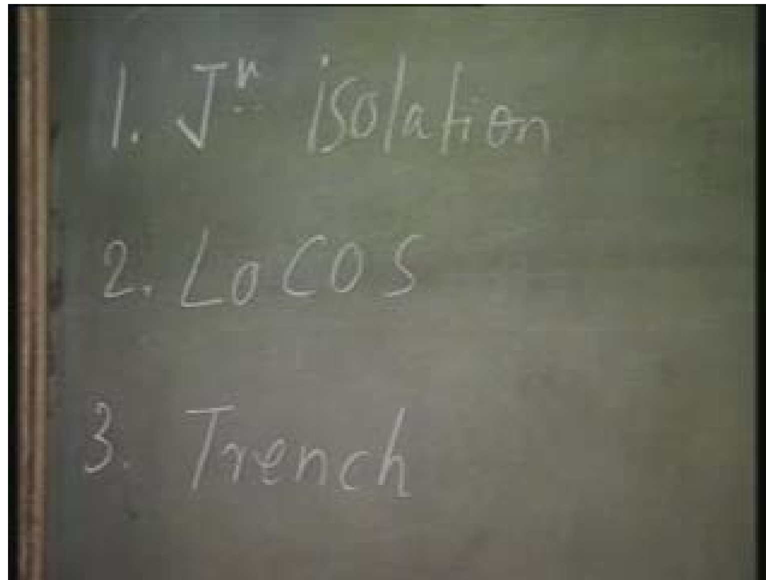
You have the cross section of the semiconductor and you have your active region protected by a layer of silicon nitride. Underneath, there is a thin oxide. This thin oxide, the pad oxide which is called, it is needed, because otherwise silicon nitride will not adhere very well to the silicon. There will be a lot of stress induced. So, if you have a thin pad oxide and on top of that you deposit silicon nitride, then the stress will be less. Therefore, it is common practice whenever you want to deposit silicon nitride, you first have a thin layer of oxide and only then, after that, on top of that you have the silicon nitride. So, now let us say I want to etch these portions and then I want to grow the oxide that is the recessed local oxide isolation.

(Refer Slide Time: 5:01)



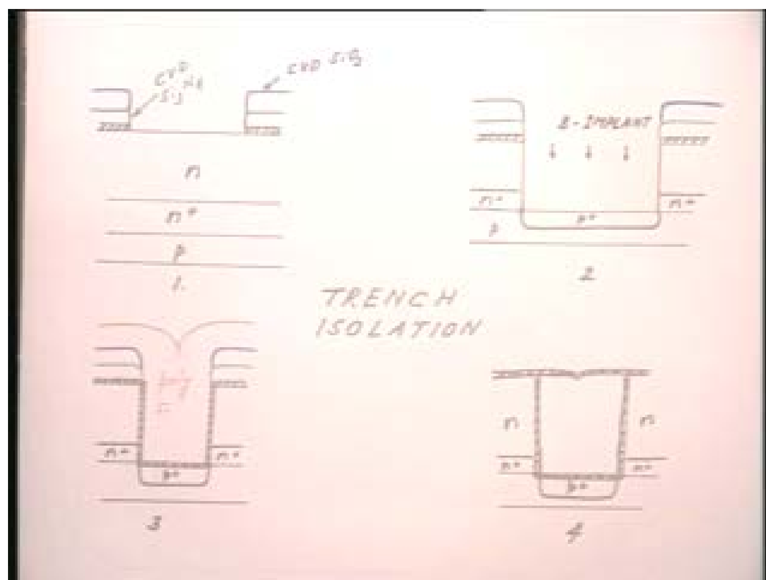
When I am doing that you see, there will be a little bit of encroachment of oxide here. Underneath this silicon nitride there will be a little bit of encroachment of oxide here and because it is shaped like a bird's beak, this is called the bird's beak problem and this is called the birds crest problem. This is because oxidation is proceeding not merely on the horizontal surface, but also at the vertical side walls. That is why there is a slight encroachment under the silicon nitride. So, now it is easy for you to appreciate that as we reduce the device dimensions, even this small encroachment may start to pose a problem. If this entire active transistor area itself is very small, then even a small encroachment will be a considerable percentage of the total active area. So, even the isolation by local oxidation will start to suffer from these problems, because of the encroachment of the oxide. So, then it was thought that instead of trying to grow the oxide, if we can etch and then fill up this etched regions with deposited some substance which will be non-conducting, instead of growing the oxide you etch required portions in silicon and then fill up this etched regions with some deposited material and that is in a sense what trench isolation is all about. So, from LOCOS we now go to trench isolation.

(Refer Slide Time: 7:20)



So, you see, we first started with junction isolation. Then, we have seen, instead of junction isolation we have seen LOCOS and now what we are going to discuss is called a trench isolation. So, in a trench isolation, just as the name suggests this is exactly what is done. You cut a trench in the semiconductor and then you fill the trench with a non-conducting material. So, let us see, what are the steps in this trench isolation process?

(Refer Slide Time: 7:56)



Look at the figure 1. What do I have? I have the p substrate. On top of that I have the n plus buried layer and on top of that I have the n epitaxial layer. So far everything is identical to the older, the essential basic bipolar junction transistor technology, right. You have started with the p-type substrate, you have the n plus buried layer and on top of that you have grown the n epitaxial layer. After that you have a pad oxide and then a double layer, one of CVD silicon nitride and then on top of that a CVD silicon dioxide. You deposit this all over the substrate and then remove it from certain regions where you want the isolation region to exist, where you want the trench to be etched. So, we had a pad oxide, a silicon nitride and then a silicon dioxide and then we have patterned this layer, so as to expose certain region of semiconductor, while other regions are protected by this layer of silicon nitride and silicon dioxide.

Now, in step 2, we have etched down vertically, vertical etching is done, you see and that is possible only when you use dry etching. So, you see, with the advance of technology, to have such vertical side walls, chemical etching, wet chemical etching will not do. So, we have to have dry etching with a lot of directionality, so that we can obtain an etch profile with perfectly vertical side wall and that is a very, very crucial step in trench isolation, to obtain an etch profile with perfectly vertical side walls and then once you have obtained this perfectly vertical side walls, you do a boron implant, so that you realize a p plus region here. Why we do it, will tell you a little later, right. For the time being, let us just accept that once this vertical etching is done, we do a boron implant, so that in this p-type substrate, we create a p plus region separating the two n plus regions. In between these two n plus regions, we have a p plus region by boron implantation. So, again you see, it is the advance of technology; only by implantation process, we can afford to have a thin p plus region so exactly defined. So, you have utilized two more modern technologies. One is dry etching, in order to obtain the vertical side walls and the other one is iron implantation to have a tight control over the p plus region.

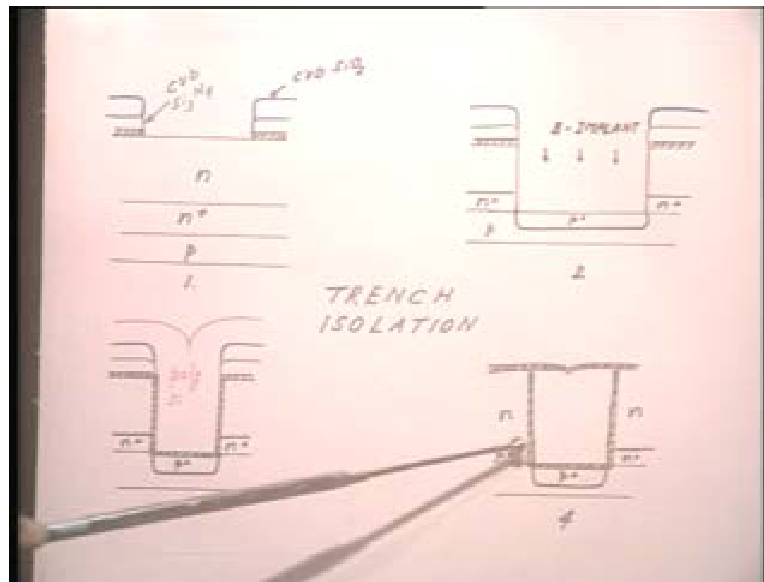
Let us come to step number 3. Now, you have your trench. So, in trench isolation this was essentially the first step, to etch the trench. What should we do next? We should fill the trench using a non-conducting material. So, we fill the trench by undoped polysilicon;

we simply deposit undoped polysilicon. But prior to this polysilicon deposition, before we do this polysilicon deposition, notice that the thin oxidation step is carried out. After the boron implantation, you carry out a thin oxidation. So, where will oxidation progress actually? Oxidation will progress only on the side walls and on the bottom surface. It cannot progress here and here, because these portions are protected by silicon nitride, right. So, I have just a thin oxide layer on the side walls as well as on the bottom surface, just on top of the p plus region and now we deposit undoped polysilicon. So, undoped polysilicon is deposited like this. This is another very crucial process, because you want this polysilicon to be deposited uniformly in the trench. Unless you adjust the parameters of deposition carefully, it might not fill the entire trench, there might be gaps. So, you have cut your trench. The trench walls are lined with oxide and you have deposited polysilicon on top of that. So, you have filled up your trench.

Next point is planarization, because topography is very important in integrated circuit. We must have a planar surface. This surface definitely does not look planar, right. So, what we have to do? We have to etch it back. Using silicon nitride as the etch stop layer, we etch it back, so that you have the poly coming only up to here, up to the top of the trench. Then, you remove the silicon nitride and silicon dioxide and simply carry out another oxidation, thin oxidation, so that the top of the polysilicon layer is also oxidized. So, now look at the trench. What do you have? You have the bottom p-type substrate and then you have this p plus region and you have two buried n plus regions on both sides and on top of that the two regions of the n epitaxial layer. So, one transistor will be housed here, the other transistor will be housed here. In between we have a trench which is lined in all directions by silicon dioxide and filled up by undoped polysilicon; therefore, this trench region is non-conducting. So, it provides the electrical isolation.

Now we come to question, why did we need the p plus implantation? What is the importance of this p plus boron implantation? Look at the bottom of the trench. What do we have?

(Refer Slide Time: 15:12)



We have an n plus region here, another n plus region here, separated by a thin oxide and then on top of that the undoped poly and we have the p region here. So, does it remind you of something? If properly biased it can actually act like the channel region of a MOSFET. I have two n plus regions separated by a p substrate, on top of which I have a thin gate oxide. So, there is a possibility that this channel could start conducting, in which case these two n plus regions will no longer be isolated. So, how do you prevent that? By putting a very heavily doped p plus region here. If the substrate of the MOSFET is very heavily doped, it is more difficult to invert the surface, right. The conduction can only take place when the substrate has got inverted. So, if the substrate is heavily doped, it is that much more difficult to invert the surface. So, by putting a p plus region here, we effectively stop the channel from conducting. So, this p plus implantation is therefore called a channel stop implant.

It is a channel stop implant which you carry out, in order to ensure that these two n plus regions are electrically isolated. Even inadvertently, because of the parasitic MOSFET action it should not get turned on. In fact, you know it is a very common practice in MOS technologies - between two adjacent MOS devices, you always put a channel stop implant, so that the drain of one MOSFET and the source of the other, next MOSFET do

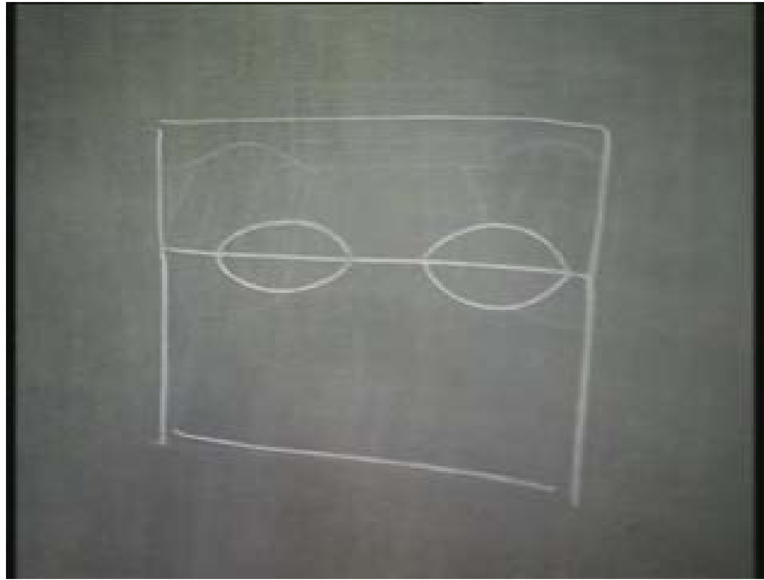
not get electrically connected inadvertently and you do that by simply putting a channel stop implant in between the two. We will discuss more about it when we discuss MOS technology. But, even in bipolar junction transistor it may be a problem. Along with the bipolar junction transistor, you may get a parasitic MOSFET and you must ensure that this parasitic MOSFET does not start conducting and you ensure that by giving a channel stop implant, so that the substrate is very heavily doped and it is very difficult to invert the surface. So, this is what trench isolation is all about. You cut a trench, you fill the trench and then you do planarization. So, this trench isolation is very useful particularly when we are talking about submicron devices. As we go towards smaller and smaller device geometry, it is very, this technology is becoming a very important point.

Another way of course, we can think of providing isolation and that is if we have what is known as selective epitaxy.

(Refer Slide Time: 18:41)

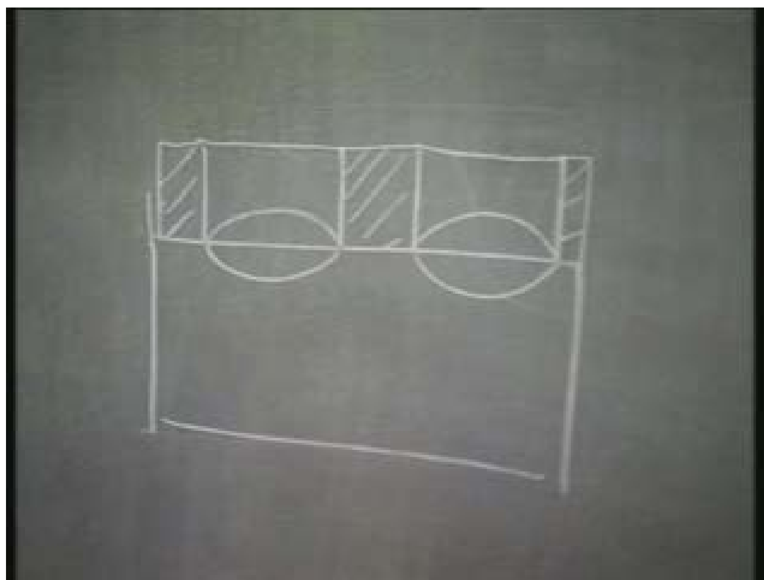
What exactly is selective epitaxy? Let us see.

(Refer Slide Time: 19:11)



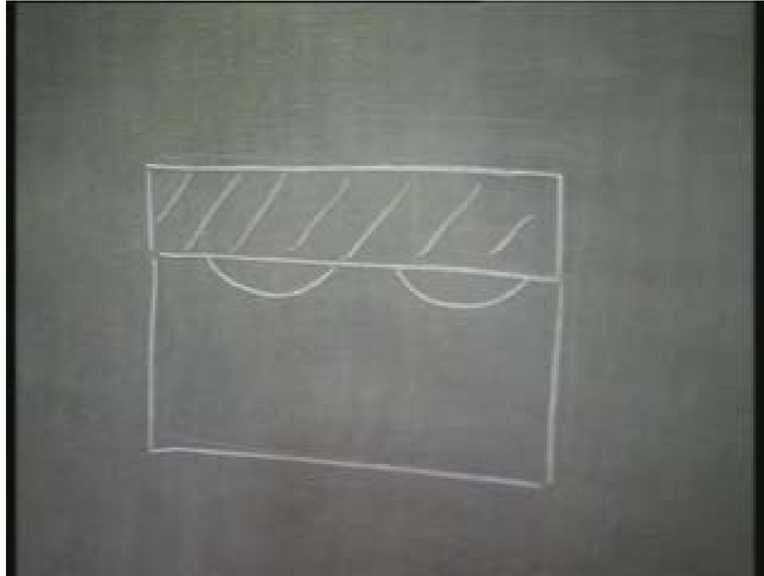
Suppose I started with the p-type substrate, I had n plus buried layer here and here and then, normally we have the n epitaxial region all over the place, right. So, a common n epitaxial region was grown for all the transistors. Now, instead of doing this, why not grow epitaxial regions only on top of the buried n plus layer?

(Refer Slide Time: 19:47)



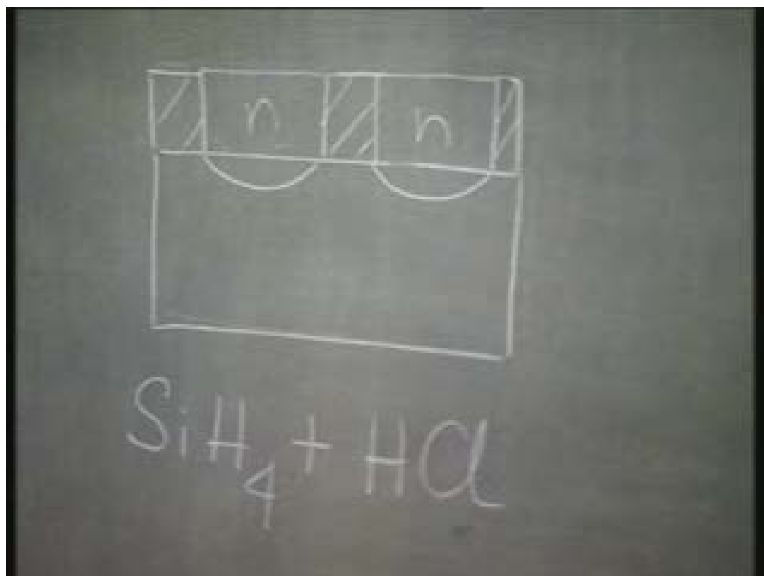
Why cannot I have …. If I can have that, then my problem of isolation does not even exist and as for as planarity is concerned, I can have these regions oxidized. So, how do we go about it?

(Refer Slide Time: 20:27)



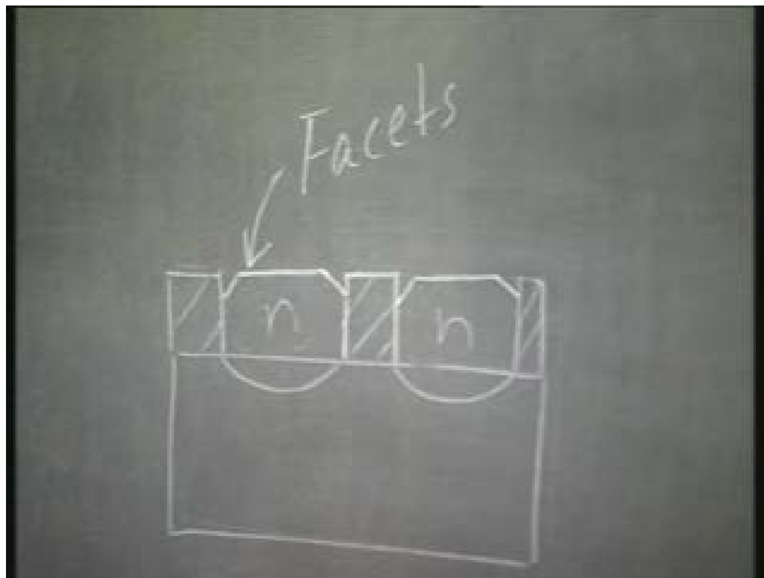
We start off, by having the p-type substrate and the buried layers. Then, we grow a deposit, a thick layer of oxide.

(Refer Slide Time: 20:52)



We etch the oxide in certain regions aligned with the buried layer, right and then, we do a selective epitaxy. You know epitaxy actually means arranged upon. So, an epitaxial layer will be formed only on top of the regions, where it can come in contact with silicon. So, it will grow here and here. So, effectively I have n epi islands sitting aligned with the n plus buried layer and this can be done by using HCl with silane. When you are doing the epitaxy, you can do it by adding HCl to silane. If you add HCl to silane, then the formation of silicon on silicon dioxide will be suppressed. There will be no silicon deposition on these regions. There will be only silicon deposition on the available substrate. So, you add ….. You do epitaxy using silane. In addition to silane, you also put HCl, so that there is suppression of silicon formation on oxide and you prefer a low pressure process. So, basically you have a low pressure epitaxy. But, selective epitaxy is still not a very perfected technique.

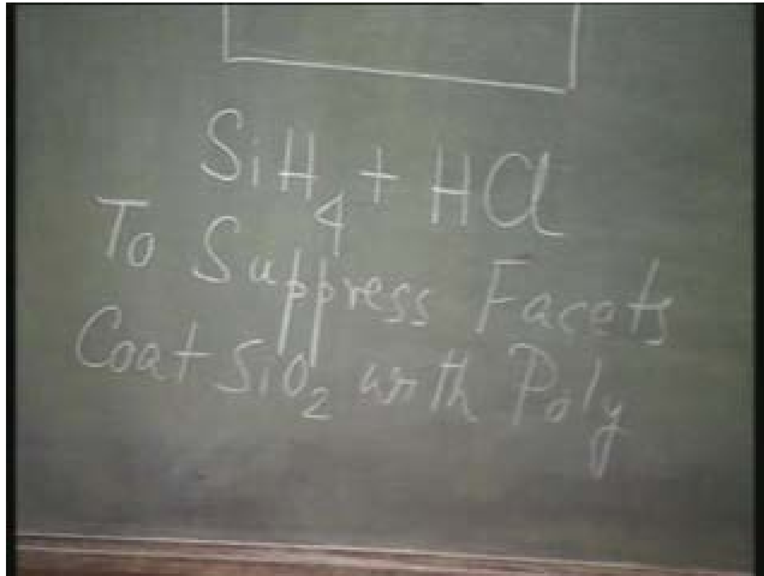
(Refer Slide Time: 23:05)



Selected epitaxy has a lot of problems and one of the most important problems is called the faceting and this is what faceting is. This is what is known as faceting. That is on the sides where it is coming in contact with the oxide, you do not have epilayer growing in a planar manner, going all the way up. You have facets forming there and in order to

suppress this facet formation you have to take more pains that is you have to coat the silicon dioxide with polysilicon.

(Refer Slide Time: 24:08)



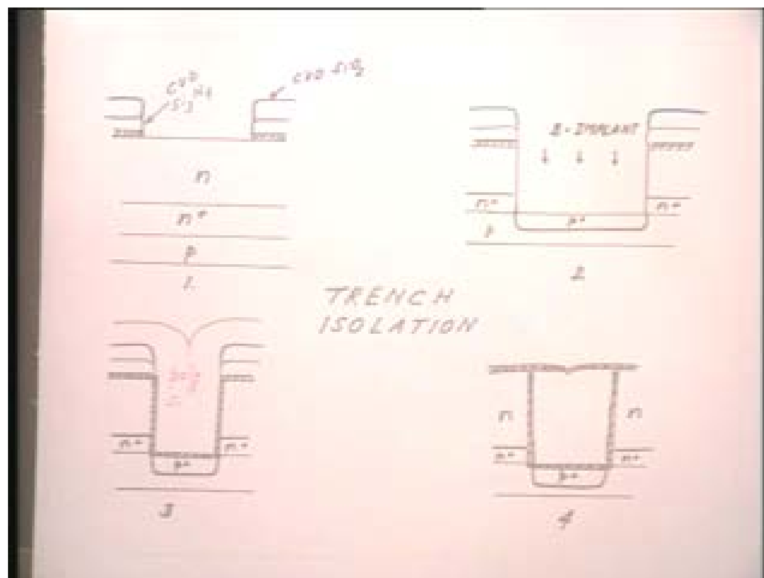
So, to suppress facets, see the whole thing is happening, because you are using this silane plus HCl, right, because you want to suppress silicon layer formation on oxide. This is the side wall of the oxide. So, even there actually you are going to suppress silicon layer formation. That is why these facets are formed, because the layer cannot grow all the way up. In order to suppress that again, you have to coat silicon dioxide with poly; coat this side walls with poly, so that the facets will grow all the way up and will give you a planar structure. But still, selective epitaxy has a lot of defects. So, these are the four possible ways of isolation between adjacent transistors in integrated circuit technology. You can either have junction isolation; in fact, junction isolation is still quite a lot in use, particularly for analog devices where the speed may not be of that much importance.

As you go to higher speed devices, most modern transistors they will use either LOCOS or more advanced transistors will use trench isolation. Both of them, incidentally they use the same sort of philosophy I would say, that you etch certain regions in silicon and then you fill that etched region. In LOCOS technique, you fill that region by growing oxide.

Because the oxide is being grown, there will be little bit of encroachment on either side, which is the bird's beak problem. In case of trench isolation however, since you are filling it by polysilicon deposition, there is no question of any encroachment.

But obviously, in trench isolation, your control has to be very high; you must have perfectly high vertical side walls and understand that we are doing all this not only for performance enhancement, but also because we want to waste less and less area for isolation. Essentially, the area that is required by isolation is a waste, isn't it? It is not housing your active transistor, so you want to keep only very small regions for isolation. At the same time, the isolation must be effective. So, in trench isolation, actually the dimensions of the trench that is very small. The trench width is less than 1 micron. So, it involves a high degree of control, to etch vertical submicron trenches.

(Refer Slide Time: 27:39)



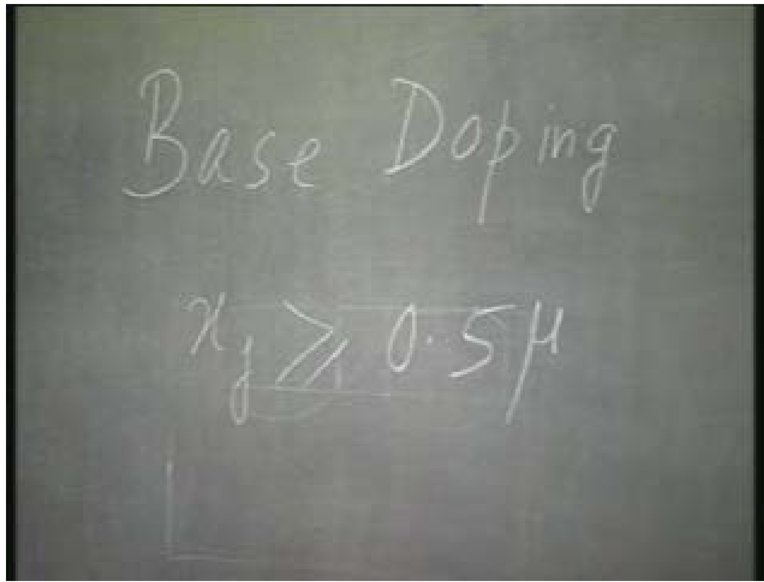
That is what I am trying to say is, you see, the height of the trench will be much more than its width, can be much more than its width. So, you have to have very high control on the etch profile. It should be perfectly vertical and then of course, the deposition of polysilicon also is another very crucial step and finally, this thin oxidation. You see, you need this thin oxide, because you want to seal the polysilicon on all sides. Polysilicon,

undoped polysilicon is not going to conduct, but unless it is sealed on all sides by silicon dioxide, you cannot be perfectly sure of electrical isolation. Silicon dioxide is a much better insulator, right; so, I need to have silicon dioxide on all sides.

At the same time if the thickness of this oxide is too high then remember, there is a seam in the polysilicon; it is getting deposited from both sides, so there is a seam here. So, if the thickness of this oxide is too much, then you know oxidation can progress down the seam and it can push the two polysilicon regions apart. So, all the steps have to be very carefully controlled and standardized; both and all of them, all three of them. That is etching of the trench, then filling of the trench by polysilicon and then the final oxidation step. So, this is a very important step in bipolar junction transistor fabrication, namely the problem of isolation. So, depending on your requirements, you can choose from among the list. If junction isolation will do, you can use junction isolation or else you can use LOCOS or use trench isolation or even selective epitaxy. A lot of work is being done in selective epitaxy, but we cannot say that the problems are completely sorted out.

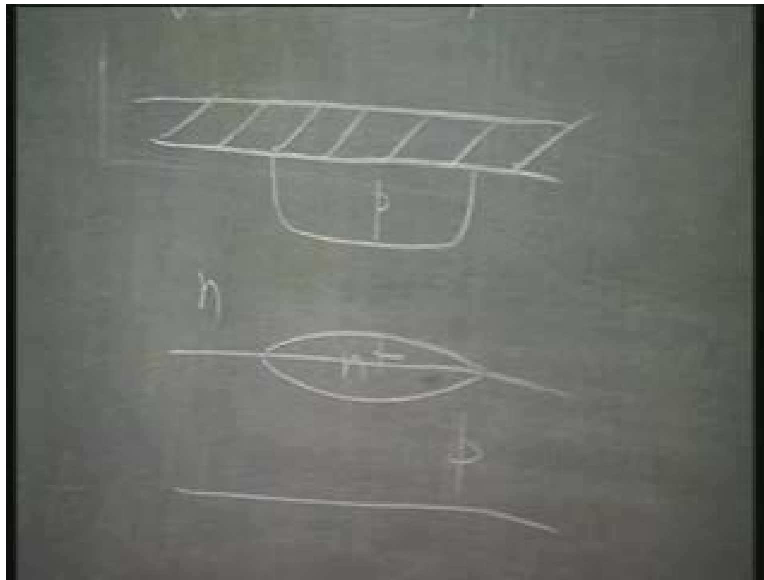
After the isolation step, the next step in bipolar junction transistor is the base. We must have the base region. See, we already have the buried layer for collector, we already have the epitaxial region and then we have the isolation and now is the time to have the base region defined in the epitaxial area. This is the next processing step that we are going to discuss, the question of base doping. You know already that for doping I have two techniques, diffusion as well as ion implantation. Now, what, which technique are you going to use in order to dope the base, will depend on the base junction depth.

(Refer Slide Time: 31:07)



If you have a junction which is, if you want the junction which is say, greater than equal to 0.5 micron, if x j is greater than equal to 0.5 micron, then you can use diffusion or you can use ion implantation. All the older day bipolar junction transistor technology, they used diffusion based technology. At that time, the junction depth was greater than or minimum junction depth was about 0.5 micron, so you can use diffusion. You use the two step diffusion which is the most common thing. That is you first have the infinite source diffusion and then follow it up with the drive-in, in order to realize the requisite junction depth and you see, drive-in was usually done in an oxidizing ambient, so that it serves two purposes. One is you drive the dopants deeper and at the same time, you also grow an oxide layer on top of the base, because you know the next step will be emitter, so for that we have to protect the base region and then open a small window in the base, right. So, the drive-in in oxidizing ambient actually served both purposes.

(Refer Slide Time: 32:37)



You have the base doping that is this is your n epitaxial region, this is your p and while doing this itself, you grow a layer of oxide on top. So, after the two step diffusion in base, this is what you are going to achieve. Next step, you will open a window in this base region and then diffuse the emitter. But then, as ion implantation became a viable technology, there were a lot of points, there were lot of points in favor of ion implantation. Why people wanted to use ion implantation for the base doping? You see base doping is a very crucial parameter.

First of all the beta of a transistor, it depends very much on the base doping, on the total amount of dopants, total amount of charge that is placed in the active base region. So, that is called the Gummel number. The total charge present in the active base that is called the Gummel number and that is very important. That must be low for realizing very high beta. So, if you want a high gain transistor, you must control this Gummel number; you must control the total amount of charge that has been put in the base. But, you see, you cannot afford to have this Gummel number too low. If you put too low a charge in the base, then there is always the possibility of the base getting punched through. Doping is very small, therefore the depletion region can extend all the way and can just punch through the base. So, this is the problem of doping. I must have a low

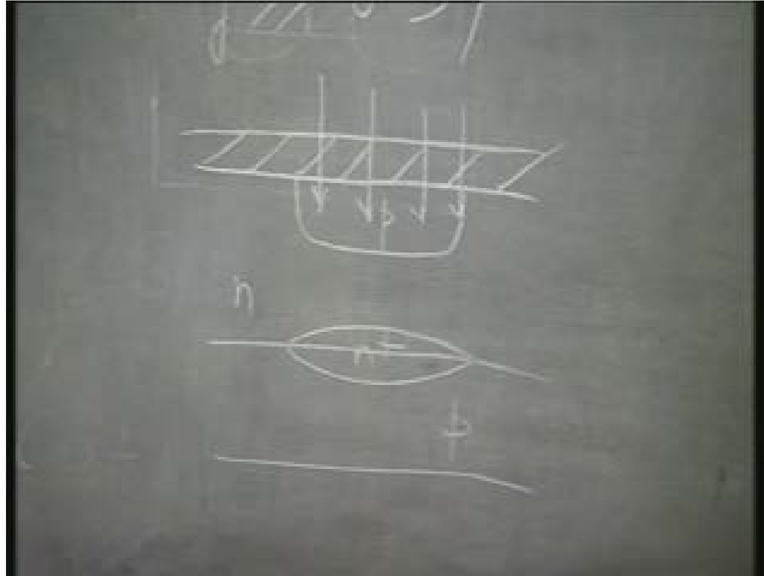
Gummel number, but not too low a Gummel number. At the same time, the base junction depth must not be high; I must have a narrow base, right. Otherwise there will be all recombination in the base region. So, one must have a narrow base also to have high beta and you know that it is much easier to control these parameters using ion implantation.

Ion implantation offers you a tighter control over the heat resistance. That is it offers you tighter control over the dose. How are you controlling the dose in ion implantation? Simply by adjusting the beam current; absolutely technology dependent, you can control it precisely. While in case of diffusion, you know, it is not so easy to control; a slight temperature fluctuation, diffusion coefficient changes, everything changes, the total dose changes. So, it is much easier to control the total dose in an ion implantation process. Therefore, it is much easier to control the heat resistance. In fact, heat resistance is widely used as the parameter during transistor fabrication. As soon as the base region is doped, you measure the heat resistance. Does it conform to the specified value? If yes, fine, you proceed for further processing. If it does not conform, well, you have to take some remedial measure. So, the heat resistance after the base doping is a very crucial point and this can be very well controlled if you use ion implantation.

However, there is a problem. What is the problem? I do implantation. There is going to be damages; I have to anneal out these damages. If I try to anneal out these damages in oxidizing ambient, I have a problem because there will be dislocations, right. So, I have to anneal out these damages in an inert ambient and also, remember, it is not just the annealing step; any subsequent oxidation step will give rise to dislocations. So, how do I then realize this oxide on top of the base? Simple; all I have to do is a chemical vapour deposition. So, after the base implantation is over, instead of oxidizing we do a chemical vapour deposition of silicon dioxide and then we carry out the annealing in an inert ambient. So, that will serve two purposes. What are these two purposes? First of all the damage due to implantation will be annealed out and secondly the CVD oxide will get densified by this heat treatment.

The other thing that is done is if you want to decrease the amount of damage in the active transistor, the other thing that can be done is implant boron through a silicon dioxide layer.

(Refer Slide Time: 38:52)



That is before the base implantation itself, you grow an oxide on top of the base, right and then you do the implantation through this. It is a thin oxide; it is not going to completely mask the implantation. The only thing that you are going to do is you must adjust the energy of the implantation, such that the peak of implantation is at the oxide semiconductor interface, right.

(Refer Slide Time: 39:19)

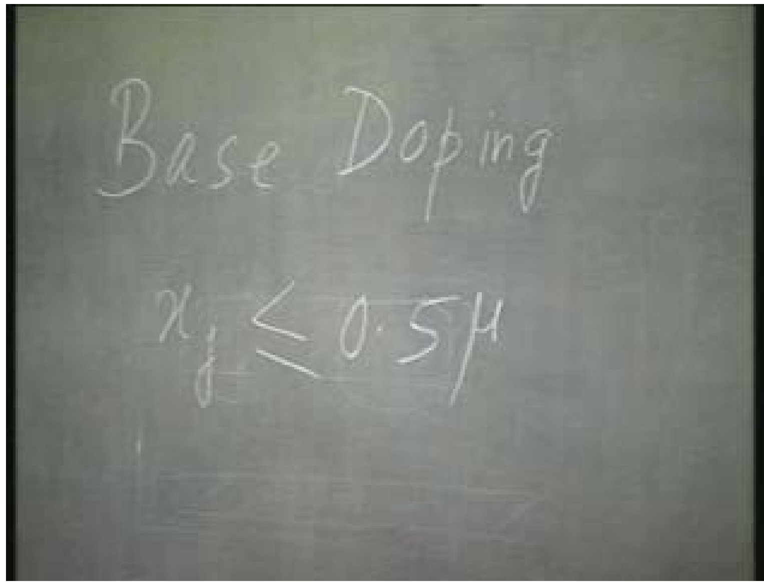
That is what I want you to have is you have oxide and then you have silicon. You adjust the implantation energy, so that the peak is here. What is the advantage? You know that the peak of the damage is slightly in front, because the maximum damage is caused in the region when the ions are still in motion. Peak of the implantation actually signifies that region where most of the ions are coming to rest, right; that is the peak, in very crude terms. In a Gaussian distribution, the peak is where most of the ions are coming to rest. That is the average, right. So, the damage peak is actually just before that. When most of the ions are still in motion that is where it is creating maximum damage. So, if you have the implantation through an oxide layer, through a thin oxide layer and if you can adjust the energy such that the peak of the implantation is occurring right at oxide semiconductor interface, you are ensuring that the damage peak is in the oxide. That is most of the damage is contained in the silicon oxide, in the silicon dioxide layer and your active transistor region is not getting damaged, right.

If you do base doping by implantation, these are the factors you must keep in mind. First of all, how to reduce the damage? You can do that by having the implantation done through an oxide layer and secondly, remember that in order to anneal out the damage, you must do the annealing in an inert ambient. Otherwise you will have dislocation,

decrease in beta, your main purpose will be defeated. If there are dislocations, beta is anyway going to be down and there will be an increase in the leakage current; increase in the leakage current, breakdown voltage will come down, all kind of undesirable effects if there are dislocations, right. So, you must ensure that after the implantation the annealing is not done in an oxidizing ambient and in order to keep the annealing time and temperature at minimum, you should have the implantation done through an oxide layer. So, these are the normal techniques, so long as the junction depth, the base junction depth is greater than or equal to 0.5 micron.

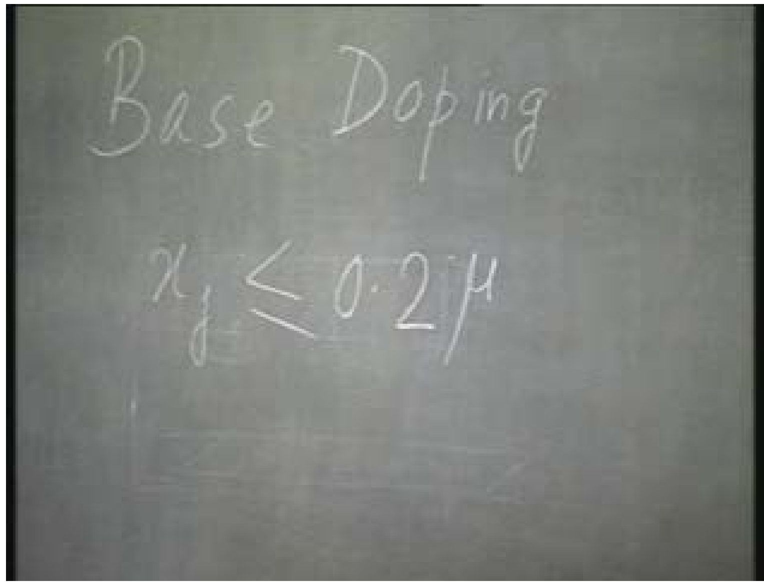
If however the base junction is less than 0.5 micron, you have obviously no choice, but to go for ion implantation. It is going to be very difficult to control it using diffusion. You know why, because in diffusion, always the surface concentration is going to be very high, right. You can bring down the surface concentration only by prolonged drive-in. But, if you have a limit, a restriction on the junction, then you cannot carry out a prolonged drive-in, right. So, if you want to keep the surface concentration down and at the same time you want to realize a shallow junction, then there is no choice but to go for ion implantation. Diffusion cannot answer to the problem. That is why as soon as the junction depth falls below 0.5 micron for the base, we have no other alternative but to go for ion implantation.

(Refer Slide Time: 43:24)



But, we are still not talking about really very shallow junction; we are just talking about x j less than equal to 0.5 micron, slightly less than 0.5 micron. It is not much, much less than 0.5 micron, but you know around there; 0.4 micron, 0.35 micron, something like that. So, you have again, the only thing you have to make sure is that the problem of dislocation does not arise and you can do that by you know, like I said, by implanting and then annealing in an inert ambient. You have deposited CVD, deposition of silicon dioxide and then you anneal it in an inert ambient, so that you have the damages annealed out, at the same time you have a dense oxide layer on top of the base which you can use in a subsequent step when you are doing the emitter doping, to protect the base.

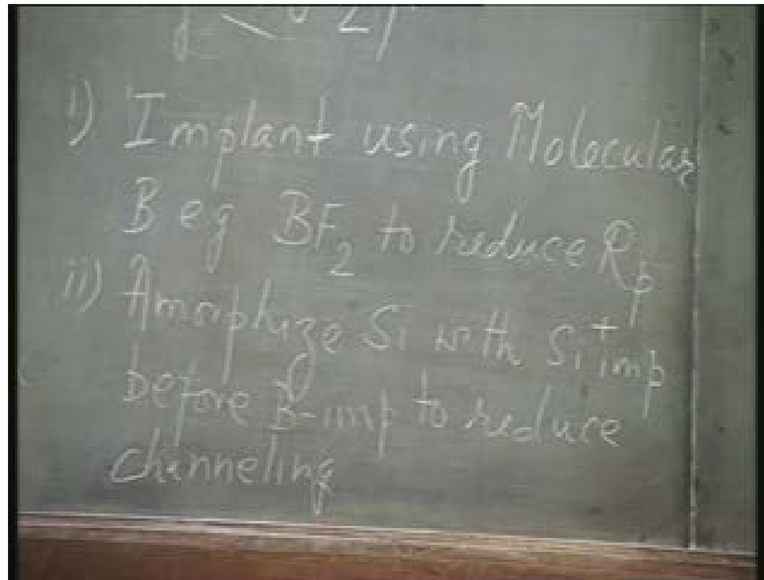
(Refer Slide Time: 44:41)



However the problems become more severe when we really go to a very shallow junction that of say, junction depth much less than 0.5 micron that is say x j less than or equal to 0.2 micron, for example. Even by ion implantation it is going to be difficult and you know why because, boron is a light atom. Boron is a light atom; boron is therefore very prone to channeling. You have to ensure that sufficient amount of boron goes in, right and it is going to be very difficult, because boron is very light. It will go through a considerable distance before coming to rest, so that is going to be very difficult.

How can I reduce the problem of both? Well, I cannot do much about boron being a light atom. Can I do anything? I cannot do anything. It is the property of the material; boron is a light atom. I cannot wish to make it heavy. However, I can play a small trick. I can use molecular boron. Instead of using atomic boron, I can use molecular boron, so that the weight problem is to some extent reduced. So, one thing is you implant from a molecular boron source, for example BF 2, so that your projected range is reduced; R p is reduced.

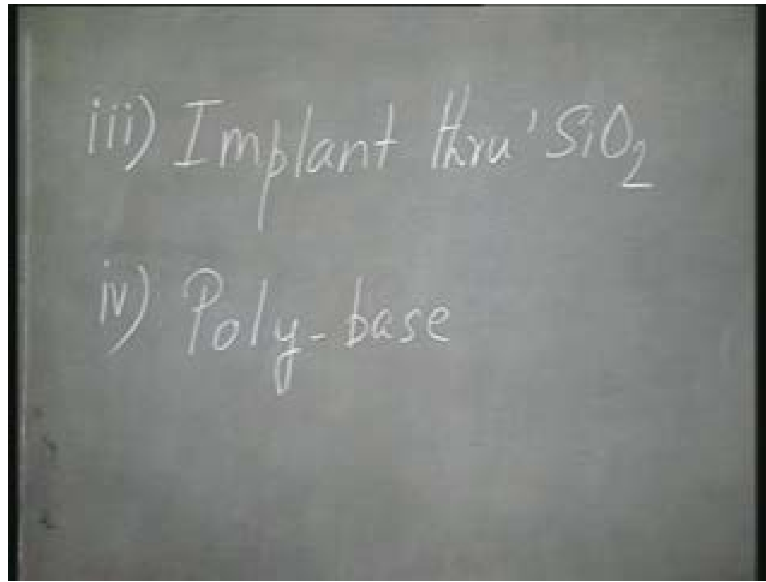
(Refer Slide Time: 46:14)



So, first step, first choice is … and then how can I reduce channeling? The one way to reduce channeling is amorphize silicon by a self-implantation prior to boron implantation. That is you do first a silicon implantation. That is the self-implantation, right; silicon implanted with silicon, silicon implanted with itself that is called the self-implantation. So, you use sufficient high dose energy of silicon implantation, so that you have a thin amorphous layer of silicon. So, first do this amorphizing and then you implant boron. Then your channeling problem is going to be minimized. So, that is the second possibility, amorphize silicon with silicon implantation before boron implantation in order to reduce channeling, right.

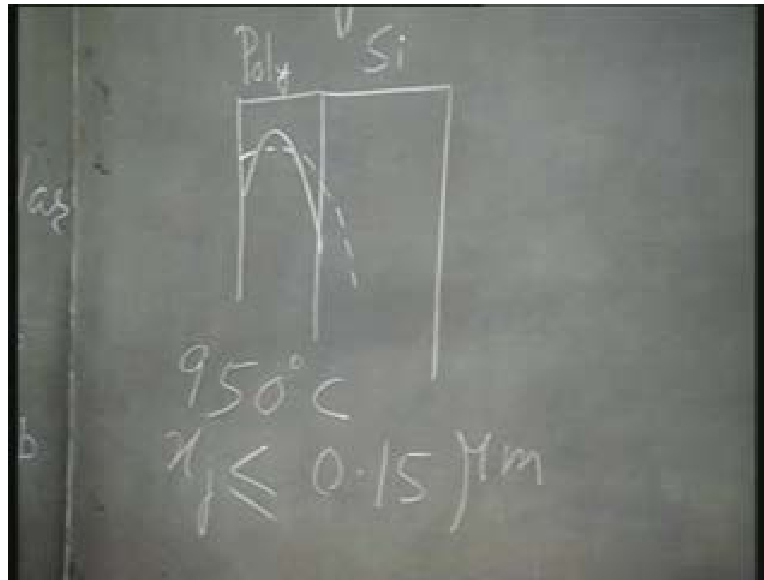
The third possibility is of course, you have a thin silicon dioxide layer on top of silicon, so that the range is cut down, right. Instead of implanting straight into silicon, you are implanting through a thin silicon dioxide layer, so that some of the implantation is getting masked by this silicon dioxide layer.

(Refer Slide Time: 48:55)



So again it is possible to realize a shallow junction and the fourth possibility is what is known as a poly-base technology; diffuse through a polysilicon layer. That is we can just briefly outline this; we can expound on this in the next class. What you want to do is you first deposit a polysilicon layer, undoped polysilicon layer over the entire base. Then, you do the implantation. See, you have no restriction on the polysilicon layer thickness. You can have a thick polysilicon layer and then you implant in this polysilicon layer making sure that this implantation does not reach the actual crystalline silicon.

(Refer Slide Time: 50:16)



That is to say I have a, this is my poly and this is my crystalline silicon. You do the implantation, so that right here nothing goes into the crystalline silicon. Now, we use this doped poly as a source and you do, carry out annealing or drive-in, whatever you want to call it, so that actually what happens is you get a shallow junction in the silicon. So, you are not doing the implantation directly into the crystalline silicon; you are doing the implantation in the polysilicon making sure that no part of this implantation reaches the crystalline silicon region and next you are going to use this implanted poly as the source; you do a short high temperature treatment, so that this is what you get. You can realize a very shallow junction, 0.15 micron by this technique and this is widely used today for all high speed transistors. You can carry out annealing at around 950 degree centigrade and you can realize an x j as low as 0.15 micron, right.

So, you see how you are going to dope the base is going to depend on the junction depth. If the base junction is fairly deep that is greater than or equal to 0.5 micron, you can use either diffusion or implantation. But of course, people will prefer to use implantation, because it gives you a better control. However, if you are doing implantation, remember, you cannot follow it up with an oxidation step, then all subsequent, for all subsequent oxides, you must use CVD oxide and do the annealing in inert ambient. If however the

base junction depth has to be less than 0.5 micron, you have to use ion implantation. So long as it is, I mean, if the order is 0.5 micron, you know 0.4 or 0.35 micron, these same precautions will hold good if you do ion implantation.

If however you are going really to very, very shallow junctions of junction depth less than 0.2 micron, it is going to be difficult by doing it simply by ion implantation. You must now take adequate precautions that is you should reduce channeling by preamorphizing the silicon substrate, you should try to reduce the projected range of boron implantation by using a molecular boron source or you should try to implant through a silicon dioxide layer or best still, the best alternative is have a polysilicon layer deposited on the base; implant in that polysilicon layer and then use this polysilicon layer as the source, anneal at 950 degree centigrade, so that you can realize extremely shallow junctions.