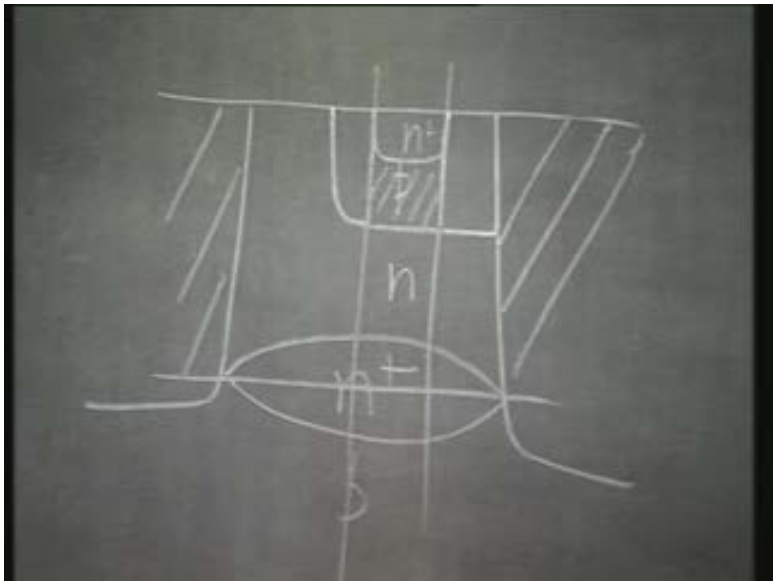
VLSI Technology
Dr. Nandita Dasgupta
Department of Electrical Engineering
Indian Institute of Technology, Madras

Lecture - 34
More about BJT fabrication and realization of p-n-p transistor

So, we have been discussing about the different processing steps used in fabricating a bipolar junction transistor in integrated circuit. We have started with an npn transistor. We have seen how the isolation between the transistors has to be obtained, either by junction isolation or by dielectric isolation. In dielectric isolation, you have also seen LOCOS as well as trench isolation and then, we have said that it can even be done by using selective epitaxy. After the isolation, comes the base doping step and there also we have outlined the basic strategy. That is depending on the junction depth, we use either diffusion or ion implantation and in the extreme case of very shallow base junction, we have to use a poly base. That is in that case, we first deposit an undoped poly, implant in that poly, taking care that no implantation goes into the crystalline silicon and afterwards, using this implanted poly as the dopant source, we do a short thermal anneal and we create a very shallow base.

Now, please understand that this strategy is as far as the active base is concerned. Now, what do I mean by the active base? What I mean is simply this.

(Refer Slide Time: 3:05)

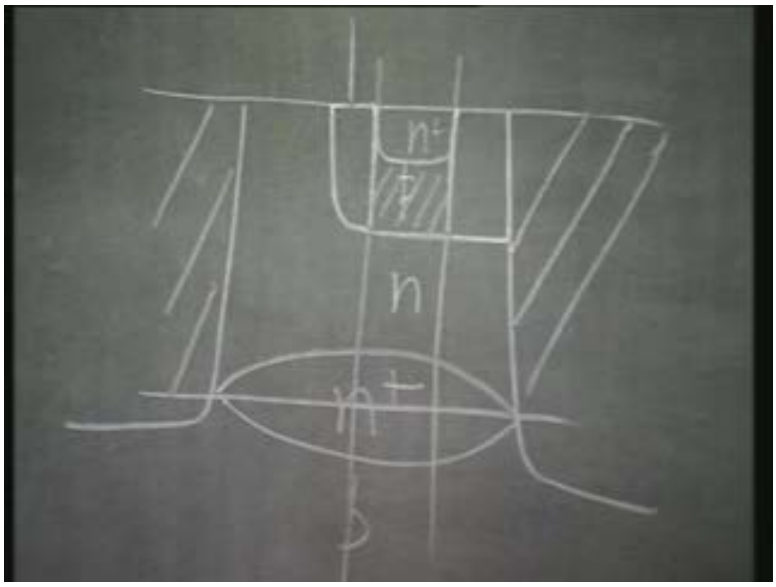


I have this transistor, you know, started from a p-type substrate, had the n plus buried layer; then I have the n epitaxial layer. I can have some kind of an isolation either junction or oxide. Let us say, I have oxide isolation here and then I have the p base region and then I have the emitter. So, this n plus p n is actually my transistor, right. So, what is the active base? Active base is actually this, the p region directly underneath the emitter. This is my active base, agreed. This is the path of the current. This is the active base. So, what about this portion of the base? In the active base, in order to keep the beta of the transistor high, I must have a low Gummel number. That is the base doping must be low, but not too low, so that there is no punch through. Keeping that in mind, I must keep the Gummel number as small as possible, so as to have the high beta of the transistor. But, what is the requirement in the rest of the portion, which is called the extrinsic base, right? So, we have active base and extrinsic base.

(Refer Slide Time: 5:18)

The extrinsic base is needed only to provide the base contact.

(Refer Slide Time: 5:50)

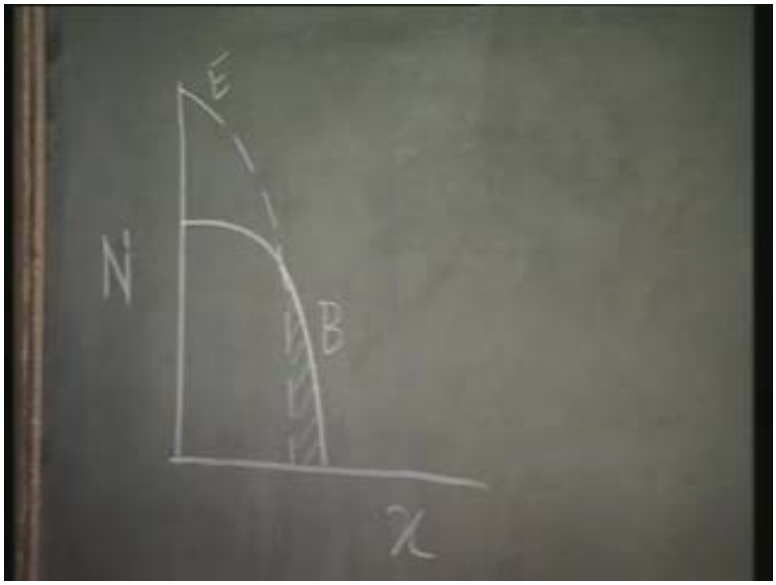


So, the p region surrounding the emitter is the extrinsic base and that must be doped high to provide a low resistance path to the active base, because you see, your base contact is taken from here. So, when you are injecting a base current, the current must flow like this and go into the active base, right. So, you do not want unnecessary voltage drop in the

extrinsic base. That means you want the extrinsic base to be as highly doped as possible, in order to cut down the voltage drop in this region, yes. So, what we do is we use our strategy which we have discussed already for the active base and in order to have a high doping in the extrinsic base, we have to do something else.

How do we keep the extrinsic base resistance low? By keeping it, making it highly doped and this, that is the dual requirement that of high doping in the extrinsic base and low doping in the active base, can be achieved either by using a double implantation or by using a single implantation. How? Let us see.

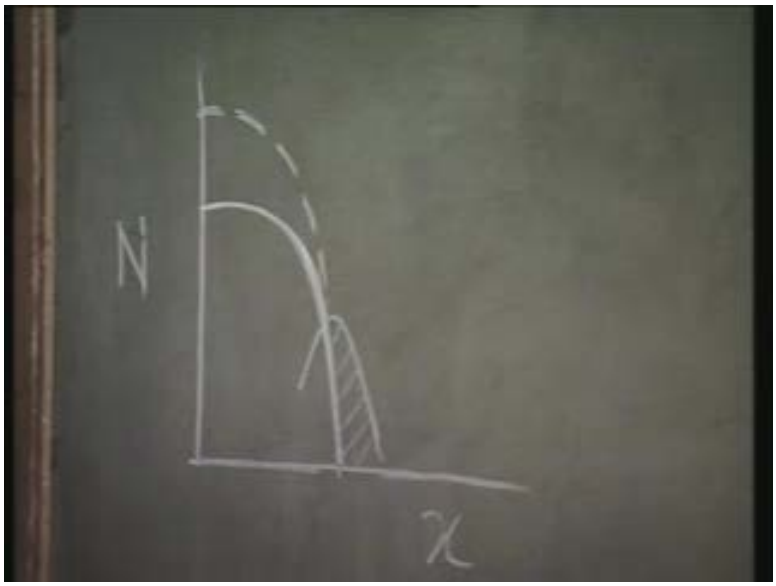
(Refer Slide Time: 7:11)



Suppose, what I am trying to plot is the doping concentration with the distance. Now, let us say, I use an implantation like this for the base region, an implantation profile like this for the base region and let us say, my emitter doping is going to be …. The dotted line is for the emitter, the solid line is for the base, right. Then, as far as the active base is concerned, active base is directly beneath the emitter, directly below the emitter. As far as the active base is concerned, this is the profile in the active base, right. I have only a small total charge in the active base region.

However as far as the extrinsic base is concerned, I see a much higher total charge with higher surface concentration. That is the region surrounding the emitter has the full benefit of this implantation profile and it has a much higher doping concentration and therefore, much lower resistivity, clear. I can of course, get the same thing by doing a double implantation also and the advantage in the double implantation will be that it will give me a better control. In that case, what do we do?

(Refer Slide Time: 9:21)



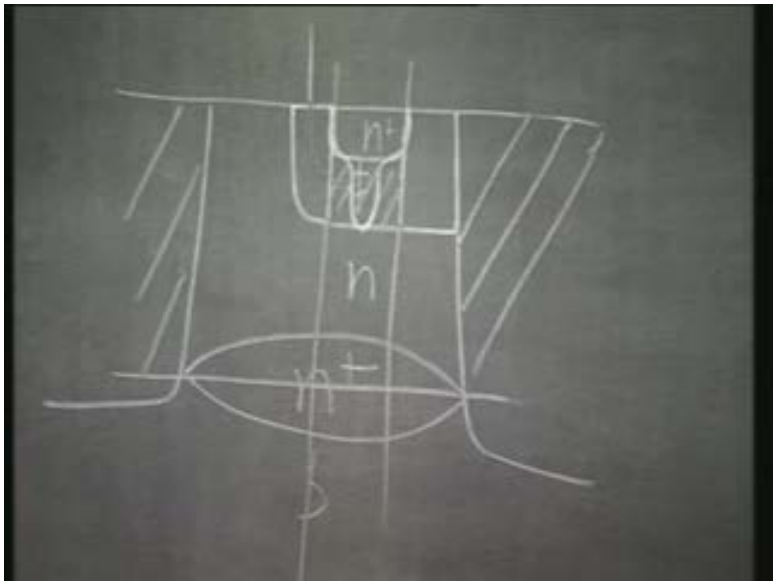
That is to say, I use two implantations for the base. This is for the extrinsic base and this is for the active base and this is my emitter implantation. I adjust the base implantation depth such that the extrinsic base implantation depth coincides with the emitter and I have a second base implantation, in order to obtain my active base. So, this is actually the active base immediately below the emitter, whereas as far as the extrinsic base is concerned, you are getting the full benefit of the doping concentration. This of course, will give you a better control over the base depth as well as over the base doping concentration. So, please understand that in integrated circuit, along with the active base, it is very important to have the doping concentration of the extrinsic base also properly designed.

The extrinsic base must be highly doped, in order to cut down the resistive drop in that region, parasitic drop in that region and therefore, the base implantation has to take into account both the requirements for the active base as well as for the extrinsic base. This can be achieved by suitably tailoring the doping profile either by using a single implantation or by using a double implantation - one implantation for active region, the other implantation for the extrinsic region. So, this is as far as the base doping is concerned.

Next step is of course, the emitter doping and even in emitter doping we simply go by the junction depth requirements, so as to decide whether we use diffusion or ion implantation to realize the emitter and even in that case you know, the same thumb rule holds good. That is if the junction depth is greater than equal to 0.5 micron, you can do it either by diffusion or by ion implantation. For diffusion, we mostly use phosphorous because, you could use arsenic, but arsenic has out diffusion problems, so the maximum surface concentration achievable is not as high as in case of phosphorous. So, mostly we use phosphorous which also diffuse faster. If your junction depth is less than 0.5 micron, mostly we use implantation and for implantation, arsenic is preferred; implant arsenic.

If your junction depth requirement is much, much less than 0.5 micron, then just like in case of base, even here, you use a doped poly as the source. That is on emitter you deposit poly, you implant n plus arsenic in it and then from that you do a short thermal anneal to realize a very shallow emitter. For shallow junctions, obviously we prefer arsenic, because arsenic has a lower diffusion coefficient; it will give you a shallower junction. This is basically the npn transistor, with no special requirements. This is an npn transistor with no specific requirements or specifications. The problems we may have in this case are of course, you know, if you use phosphorous as the emitter, you can have an emitter push effect, which will widen the base and therefore, reduce the beta of the transistor. If you use aluminium as the metal contact, you can have aluminium spiking the junction, particularly when your junction depths are narrow, when we are talking about shallow base, you can have junction spiking. You can also have electromigration problem and if there is localized enhanced diffusion, then you can have an emitter collector pipe.

(Refer Slide Time: 14:32)

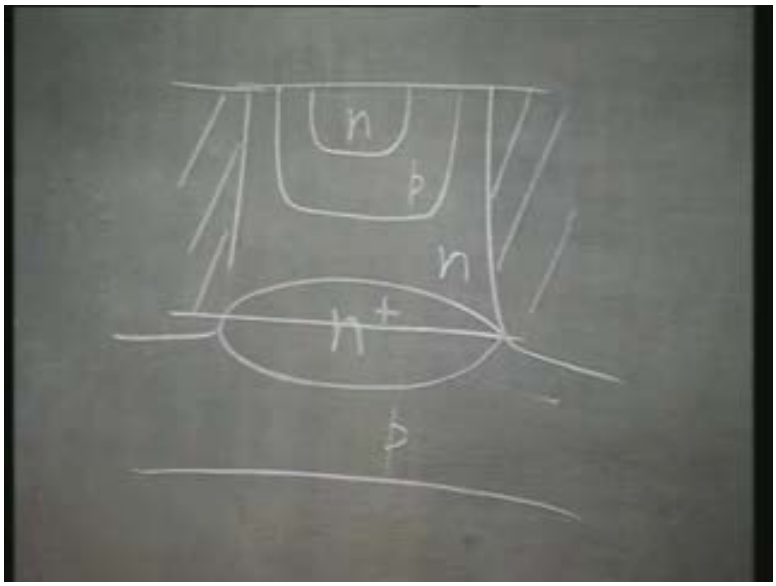


That is if there is an enhanced diffusion like this, emitter and collector are getting shorted. So, these are the three basic yield problems. As far as an npn bipolar junction transistor is concerned, you have an emitter collector pipe due to enhanced localized diffusion; you can have a base widening effect, because of emitter push if you use phosphorus as emitter and you can also have junction spiking. Particularly when we are talking about shallow base, aluminium could spike the junction, unless you have taken proper precaution by using an alloy which contains silicon along with aluminium.

So, you see, so far we are discussing only about npn transistor. npn transistors have some specific merits as far as integrated circuit technology is concerned. Some of them are like, you know, you can have a higher emitter doping concentration; if you use n emitter rather than if you use p, then you can have better facility of making ohmic contacts. It is more difficult to make ohmic contact to moderately doped n region. So, if you have a pnp transistor, where your base is moderately doped, you will have difficulty in forming an ohmic contact. In npn transistor, there are no such problems. But still, for some specific applications, particularly in case of analog devices, we use pnp transistors as well as npn transistors.

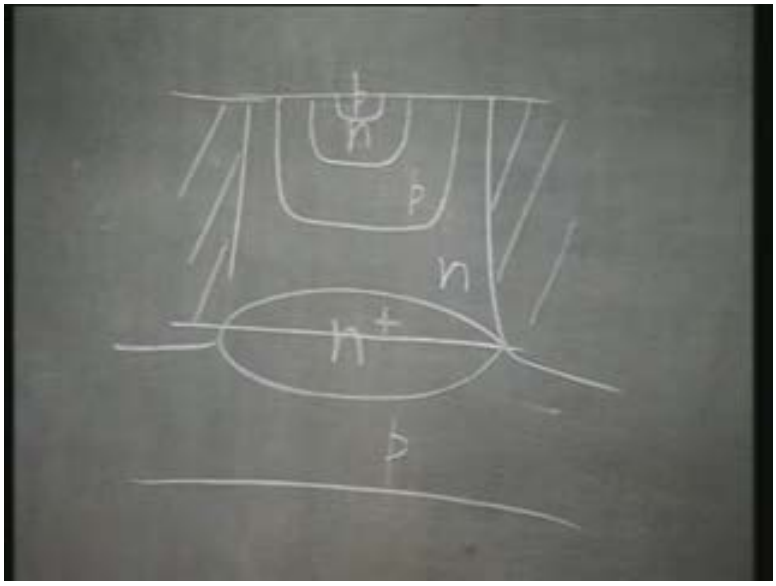
So, how does one fabricate a pnp transistor or more importantly, I have an already established npn transistor process flow, so how do I use this same process flow, add may be one or two more steps and realize a pnp transistor in the same npn transistor process flow? So, let us try to realize a pnp transistor in an npn process flow.

(Refer Slide Time: 16:57)



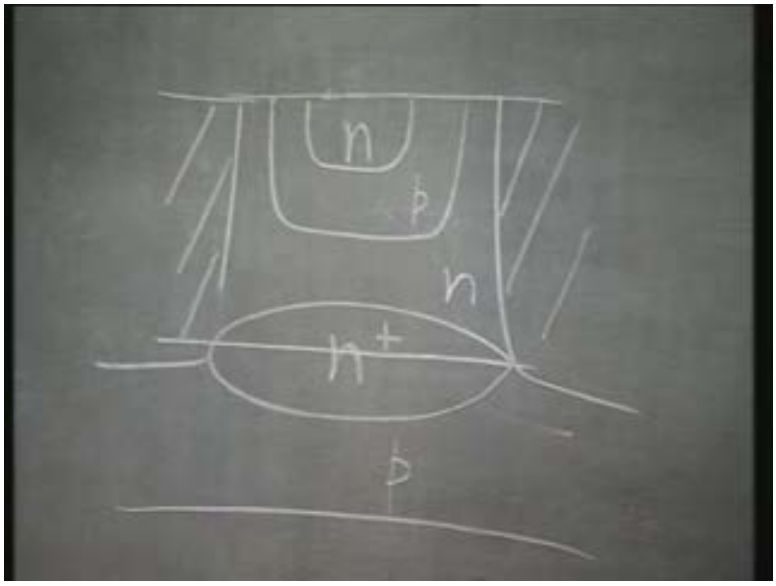
Essentially, this is my npn transistor, npn, right. Keeping this basic process flow intact, if I want to realize a pnp transistor, I have two possibilities, two choices.

(Refer Slide Time: 17:52)



That is I keep the npn transistor flow as it is and use one extra p diffusion, so that I get this pnp transistor. This is easy to say, but very difficult to achieve. You see, realizing an npn transistor by double diffusion process itself is quite tricky, because you are going to use two subsequent diffusions one after another; one for the base of the npn transistor, the other for the emitter of the npn transistor. Now, the process we are talking about is a three diffusion process. You have another p diffusion and notice, remember that diffusion means this is the highest surface concentration, then this, then this, then this. So, it is going to be extremely difficult to realize a pnp transistor by this three successive diffusion process. So, even though this is a possible solution, this is not really feasible.

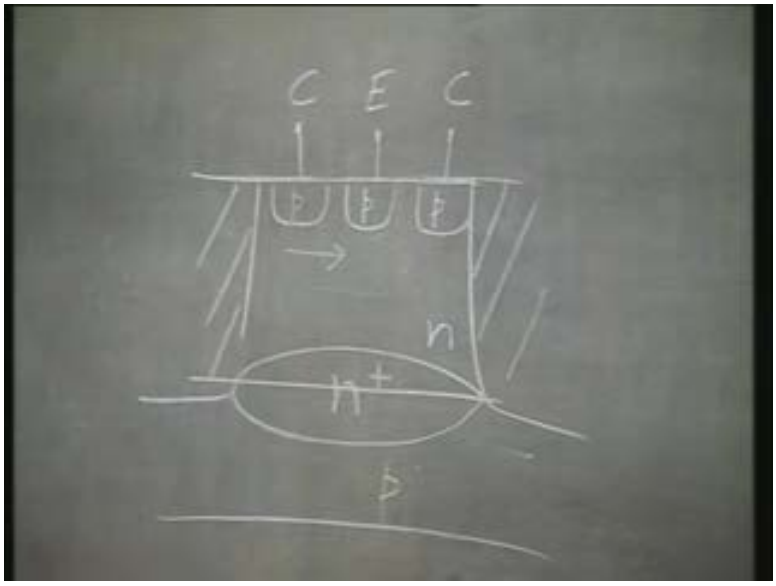
(Refer Slide Time: 19:07)



What is the other possibility? I could use this pnp transistor. That is I could use the base collector substrate pnp transistor; base collector substrate pnp transistor, right. I have already one pnp transistor integrated in this npn process flow; I could try to use this pnp transistor. That is I am going to use the substrate as the collector of the pnp transistor. But, this will create a major problem, because you see, your substrate is lowest doped. You will have a very high collector resistance and on top of that, you have the n plus buried layer which is going to be in the base of this pnp transistor. So, essentially this is going to cut down the gain of the pnp transistor, because you are increasing the total charge in the base, you are increasing the Gummel number, right. So, both these avenues seem to be not quite feasible.

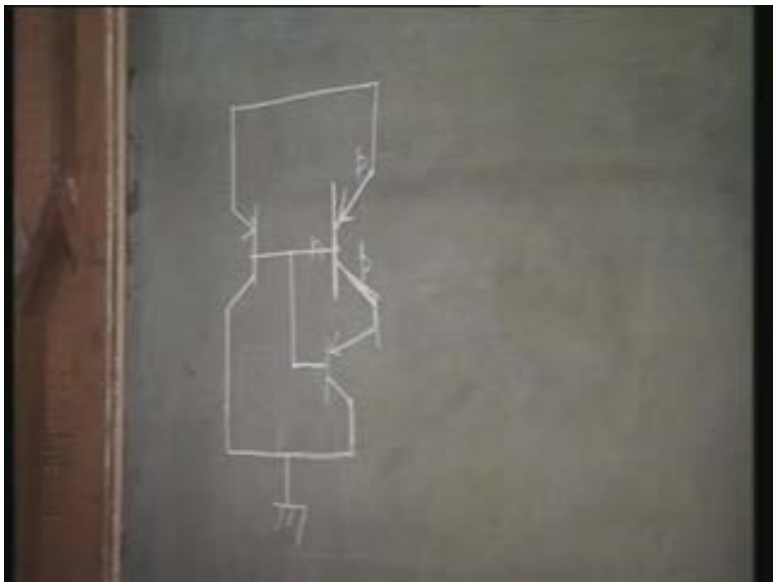
In other words, if I want to realize a vertical pnp transistor, a vertical pnp transistor that is emitter on top, base under that and collector under that, current flow is in the vertical direction; if I want to realize a vertical pnp transistor, I am going to have problems. So, the remedy is, let us try to use a lateral pnp transistor.

(Refer Slide Time: 21:17)



In a lateral pnp transistor, I have the emitter surrounded by the collectors and the current flow is ….. This is p, this is n, this is p; pnp transistor. So, the current flow is no longer in the vertical direction, it is in the lateral direction and that is why this transistor is called a lateral pnp transistor. Now, look at this transistor very carefully. What do I have?

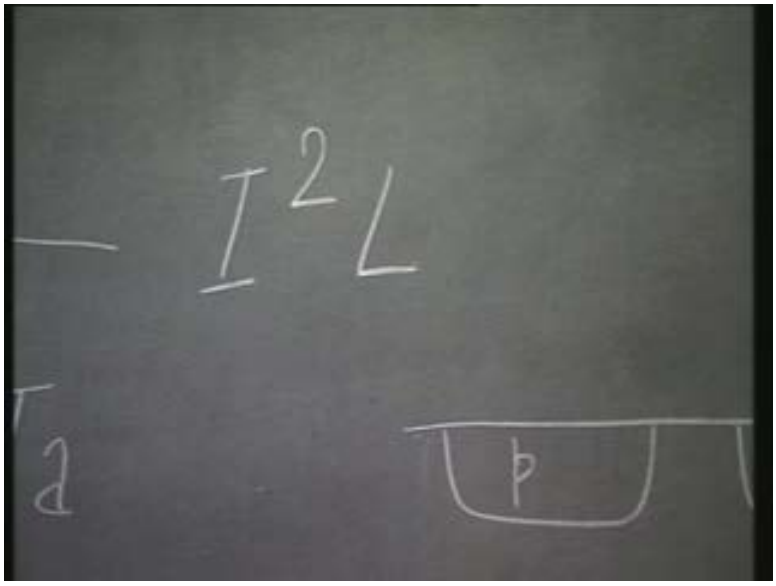
(Refer Slide Time: 22:32)



I have this collector p; then, I have this base n. Let us say this is the emitter p and this is the collector p. Base is n, emitter is p, collector is p. Now, look that there is another pnp transistor also available in the same diagram. What is that pnp transistor? One is collector, base, substrate; emitter, base, substrate. In both cases, the base is our common. So, I have this base; I can think that the substrate is grounded. At the same time I have …., right. I have this lateral pnp transistor as my main transistor. This is my main transistor, pnp, but using the collector of this pnp transistor, the base of the pnp transistor and the substrate as the p, I can have another pnp transistor. Similarly using the emitter, base and the substrate, I have another pnp transistor. All of them share the same base, agreed. So, you see, every lateral pnp transistor is going to have two parasitic pnp transistors associated with this.

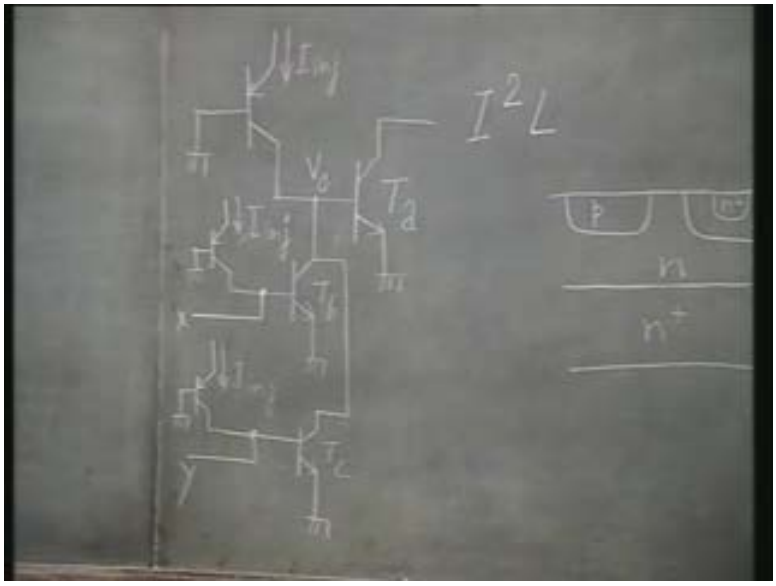
However if you look at this parasitic pnp transistor, its gain is very low, because of the n plus buried layer. Because of the presence of the n plus buried layer which is the base of this pnp transistor, the gain of these parasitic transistors are going to be very small and therefore it is not going to affect your actual pnp transistor performance to a large extent. So, this lateral pnp transistor is a feasible design and in fact, in almost all integrated circuit bipolar junction transistor, pnp transistor is always a lateral transistor, while npn transistor is a vertical transistor, because you want to incorporate a pnp transistor in the same process flow as the npn transistor and this can only be done as you have just seen. If we use a lateral design, then we do not have to change the substrate, we do not have to change the process flow; simply by having another mask for the pnp transistor, we can realize the lateral pnp transistor.

(Refer Slide Time: 26:37)



Now, even though pnp transistor is mostly used for analog circuits, we use pnp or rather a combination of pnp and npn transistor in digital logic design also for a particular type of logic circuits called the I-square L circuits, I-square L, integrated injection logic, IIL; integrated injection logic, IIL, commonly referred to as I-square L circuits and what I have here is the basic I-square L circuit. Let us see how this circuit is going to perform. What do I have? I have a pnp transistor and an npn transistor.

(Refer Slide Time: 27:03)

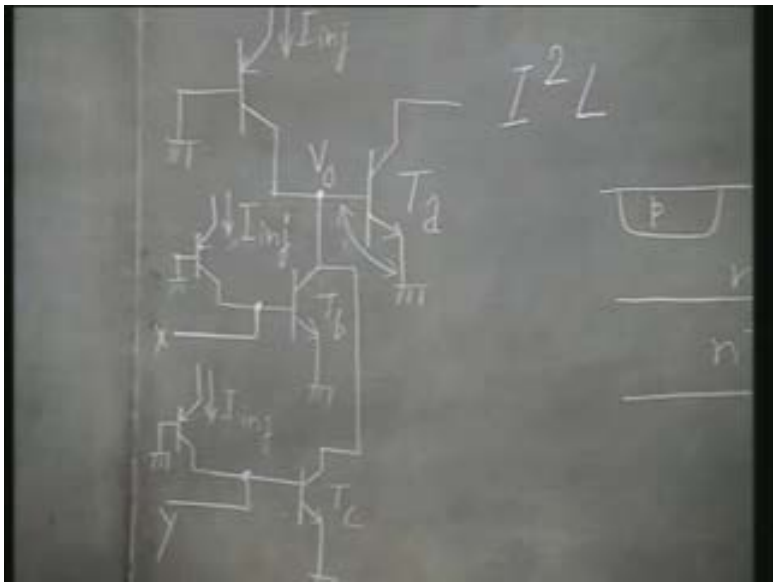


This pnp transistor is being used as a current source; it is injecting a current. So, from there comes the term injection logic. The same injection current is being given to all the blocks. I have one block here, this is one block, this is another block, this is another block, right. That is T a and this pnp transistor is one block, T b and this injection logic is another block, T c and this injection current is another block. All the three blocks are identical. Do you see? Now, let us see how this circuit is going to perform. I have two inputs x and y. Let us say both x and y are zero or at logic level, zero.

(Refer Slide Time: 28:11)

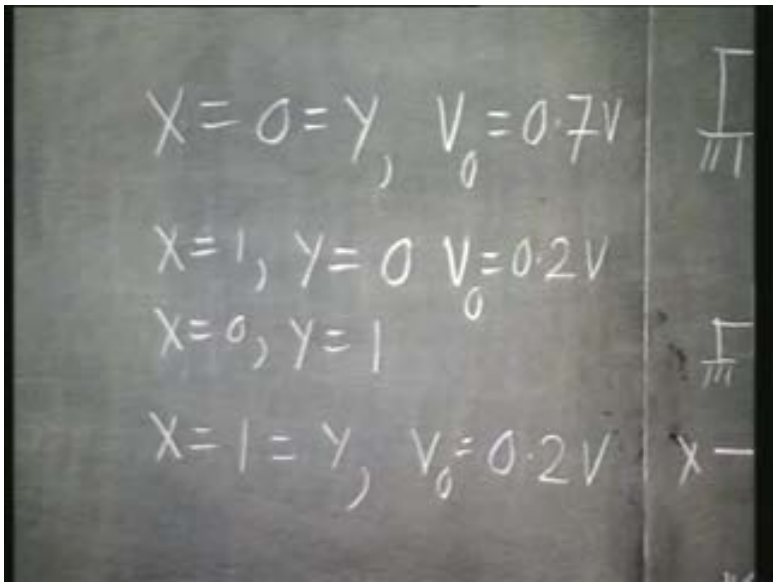
In that case, both T b and T c are cutoff, right. x and y are given at the base of T b and T c. If they are both at logic level zero, then both T b and T c are cutoff. In that case, the current here has only one path to flow that is in the transistor T a, right and then, what is going to happen to the output voltage? What is the output voltage in that case?

(Refer Slide Time: 28:46)



The output voltage is simply the base emitter drop of this forward biased diode. T a is an npn transistor, right. So, this is only the base emitter drop of the forward biased diode. How much is that? 0.7 volts.

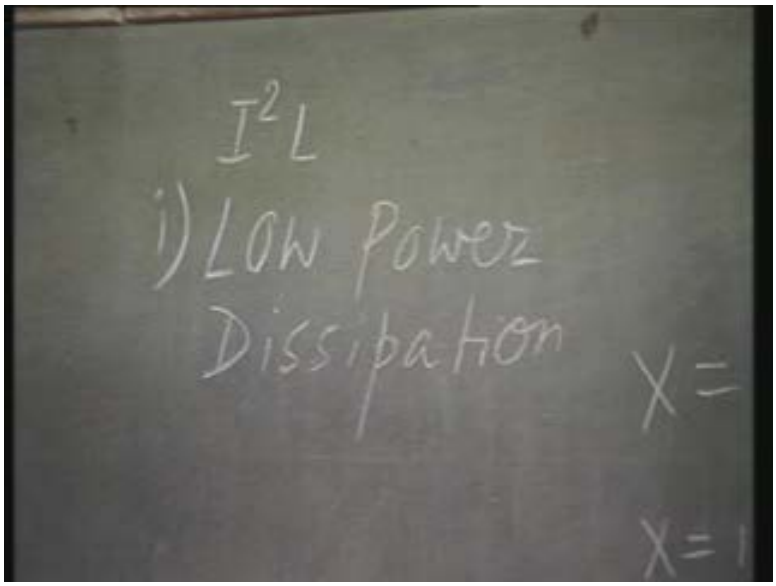
(Refer Slide Time: 29:07)



So, when x is zero and y is zero, V output is 0.7, right. Now, suppose x is 1; x is 1, y is zero; x is 1 that means this transistor is in saturation, T b is in saturation. So, then what is the output voltage? This output is connected to the collector of T b and therefore, if you ….., look that emitter of T b is grounded. So, actually V 0 is nothing but the collector emitter voltage of a saturated transistor, of a transistor in saturation or V naught is nothing but, V CE SAT. What is V CE SAT? 0.2. Identical is the case if x is 0 but y is 1. In that case, T c will be in saturation and V 0 will still be the collector emitter voltage of a transistor in saturation, V CE SAT and even when both x and y are 1, so you see, I could say, therefore that the output voltage V 0 is going to be high, only when both the inputs are low and the output is going to be low, when any one or both the inputs are in the high state. In other words, the basic I-square L circuit is a NOR gate; 0 0 it is 1, in all other cases it is going to be 0. So, this is basically a NOR gate.

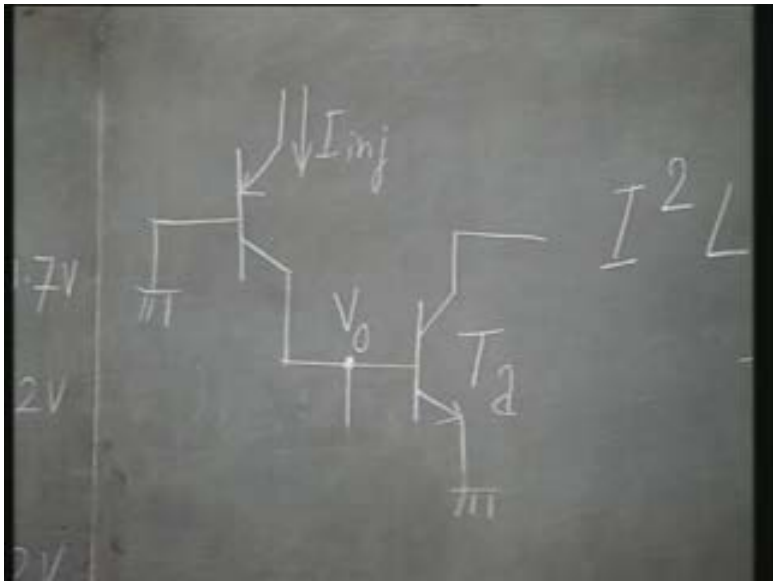
Now, there are several advantages of this I-square L circuit. What are these advantages? The advantages are that first of all it has very low power dissipation; it has low power dissipation.

(Refer Slide Time: 31:51)



That is the first major advantage of I-square L, it has low power dissipation and there is another advantage which will be evident to you if you look at, just concentrate on one building block. Let me rub off all the other extraneous matter and just concentrate on this one building block.

(Refer Slide Time: 32:32)



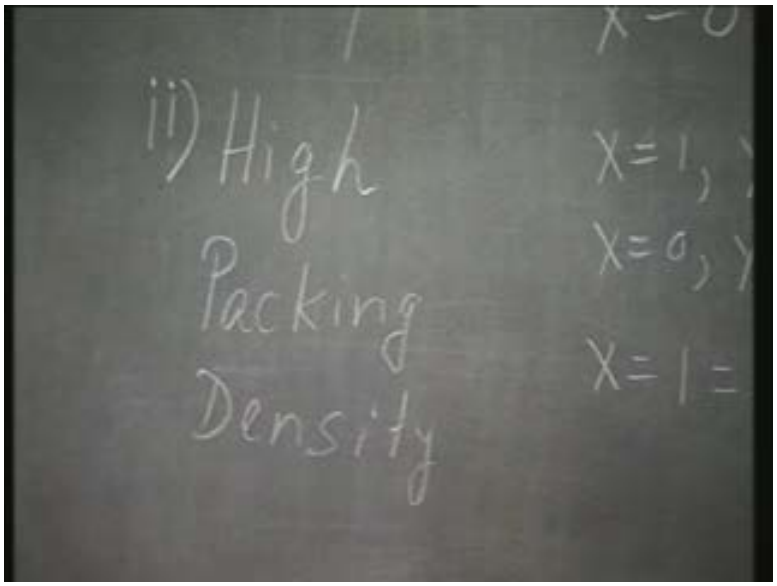
This building block is replicated any number of times in the I-square L, right. So, let me just concentrate on this one building block. What do you see? I have a pnp transistor and an npn transistor, right. Now, let us look at this pnp transistor. Its base is n-type and it is grounded and I have an npn transistor whose emitter is n-type and is also grounded. So, I could use the same n region both for the base of the pnp transistor as well as for the emitter of the npn transistor, agreed. See, another very interesting thing. The collector of the pnp transistor is fed to the base of the npn transistor. That is the collector of the pnp transistor is shorted to the base of the npn transistor. So, again I could use the same p region for the collector of the pnp transistor as well as for the base of the npn transistor.

In other words, these two transistors are going to share certain regions. These two transistors are not going to be two distinct entities, they are going to share certain regions or in other words, these two transistors will be merged.

(Refer Slide Time: 34:25)

I will have a composite structure in which two transistors will be merged together which gives I-square L its other name called MTL, merged transistor logic, merged transistor logic. So, obviously since these two transistors need not be unique entities, but they can share certain regions, packing density of I-square L circuit is going to be very high.

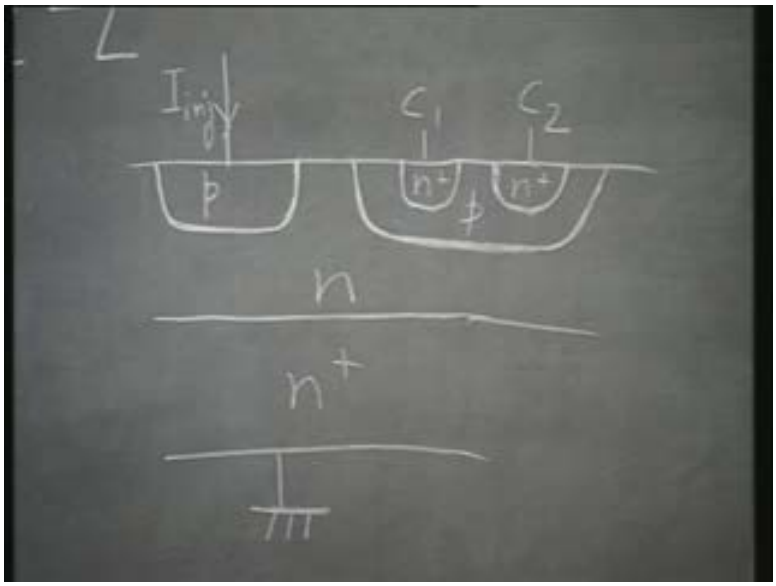
(Refer Slide Time: 35:00)



So, I have the second advantage, it has high packing density. These are the two great advantages of the I-square L circuit. One is low power dissipation; the other is high packing density.

Let us now look at the basic I-square L circuit from the technological point of view.

(Refer Slide Time: 35:31)



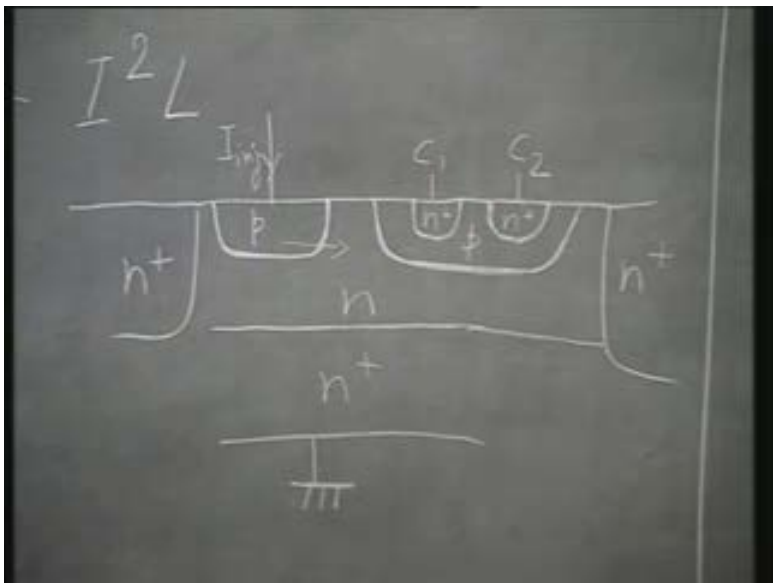
I have started with a substrate called n on n plus, right. That is I have n plus as the bottom layer on which an n epitaxial layer is grown. I have used this n on n plus epitaxial substrate. Now, in this, I have diffused two p regions, right. So, you see, I have a lateral pnp transistor, pnp. This is going to be my current source. In the other p region, I have diffused multiple n plus regions. I have shown only two here, it could be multiple regions. Now, what is it? Let us see. This is a pnp transistor and you see, this pnp transistor's base is shared with the emitter of the npn transistor. So, this must be the emitter of the npn transistor, which should be permanently grounded. Base of the pnp transistor as well as the emitter of the npn transistor is permanently grounded and you see, the collector of this pnp transistor is also the base of the npn transistor and both purposes are being catered to by this p region. So, this is my pnp transistor and this is my npn transistor.

The interesting point here is that you find that these are the collectors of the npn transistor and this is the emitter of the npn transistor, which is actually the inverse of the npn transistor designs we have done so far. In other words, the npn transistor in this I-square L is operating in the inverted mode and that gives rise to the disadvantage in the I-square L circuit. That is since the npn transistor is operating in inverted mode, you have low speed.

(Refer Slide Time: 38:11)

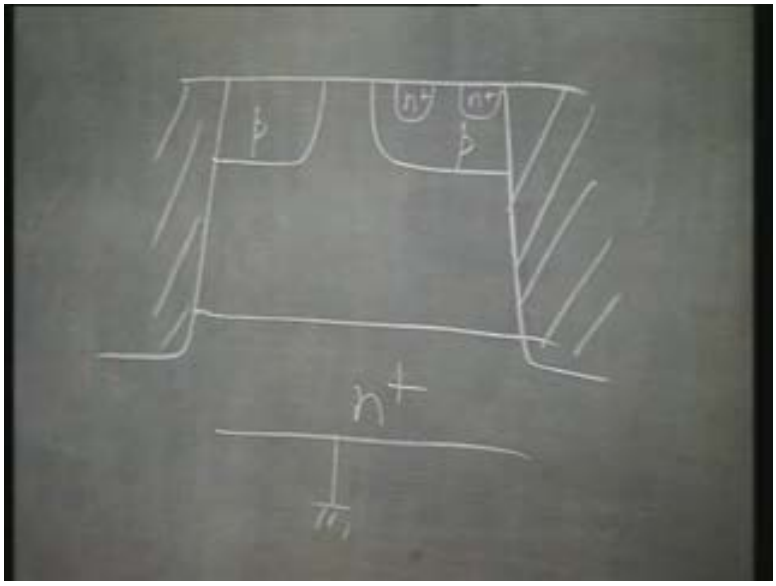
So, while the first two are the positive features, this is the negative feature of the circuit. Now, from the technological point of view, I-square L circuits have another advantage. That is in these circuits basically you do not need any isolation. Why? Because remember, all the blocks whatever I drew here, all the blocks they had grounded base for pnp transistor and grounded emitter for npn transistor. Therefore, I do not have to isolate them. They can all share the same substrate. The substrate can be grounded and all the transistors can share that same substrate. So, strictly speaking, in I-square L, an isolation is not necessary, right; you do not really have to have an isolation.

(Refer Slide Time: 39:38)



But of course, in order to make the circuit perform in a more efficient manner, you can use two n plus regions called the n plus collars on both sides of the p diffusion. This is simply done in order to facilitate hole injection. You see, you are injecting current, right; you want the hole injection in this direction, pnp. You want the hole injection in this direction. However, as far as holes are concerned, it does not see any difference between this direction or this direction. However, if you add this n plus collar, then because the doping concentration in this regions are going to be far more than in this region, the hole injection will be predominantly in the active direction, in the direction you want it. So, adding n plus collars will actually facilitate hole injection and that is why usually n plus collars are added in the I-square L circuit, though for the basic circuit performance you do not really need to have any isolation and of course, as you have seen in case of junction isolation compared to dielectric isolation, if instead of junction isolation we have dielectric isolation, it is always better. So, a more advanced I-square L design will be to replace the n plus collars by an oxide. So, in that case, we will have a circuit like this.

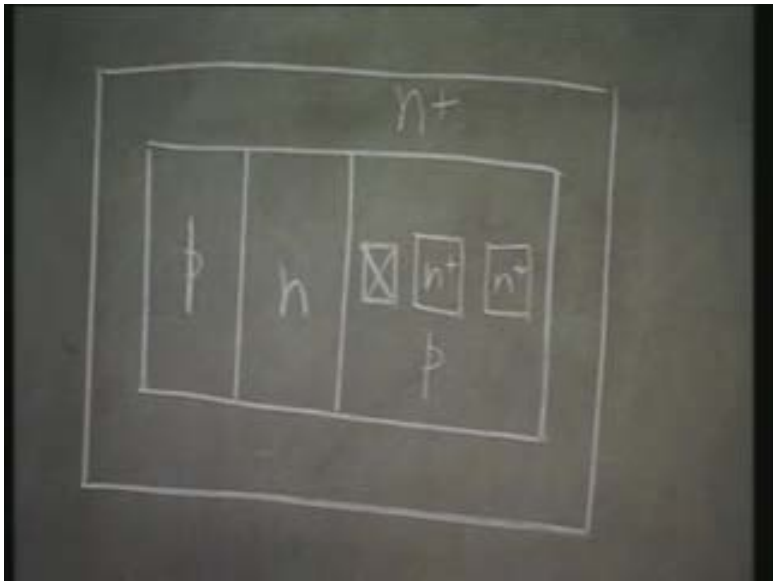
(Refer Slide Time: 41:37)



Notice the advantage. I can have my p regions stacked against the oxide as well as my n region stacked against the oxide. In this particular case, these n plus regions must be separated from the n plus collar. But in this case, I do not have any such requirement, because I am using an insulator for the isolation. If you look at the top few of these two I-square L designs that is one using n plus collar and the other using the oxide isolation, the difference will be clearer to you.

Let us first talk about the top view of the mask, where I use the n plus collar.

(Refer Slide Time: 43:08)

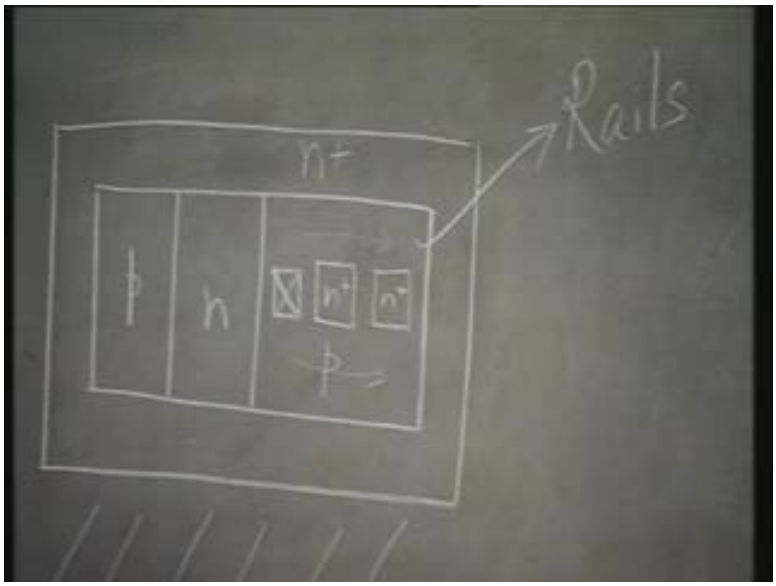


I have, this is the active region and I have n plus collars surrounding this active region. This is my n plus collar, right and in this active region, I have the consecutive pnp regions. So, this is my p, then this is my n; this is my p, in which I have this n plus regions, right. Remember, I must have a contact here. This is my output point. I must have a contact here, so I should have a contact for base in this region. So, this is n plus collar, this is my pnp transistor and this is my npn transistor.

(Refer Slide Time: 44:39)

If I compare it with the oxide isolation case, it is surrounded by the oxide and I have this p region here, this n region here. Now, my n region can extend all the way up to the oxide where as in this case, they must be separated from the n plus collar by the intervening p region. Do you understand the basic difference? Here, I can stack the n plus layer directly against the oxide, but in this case, I must have a separation from the n plus collar by this p region.

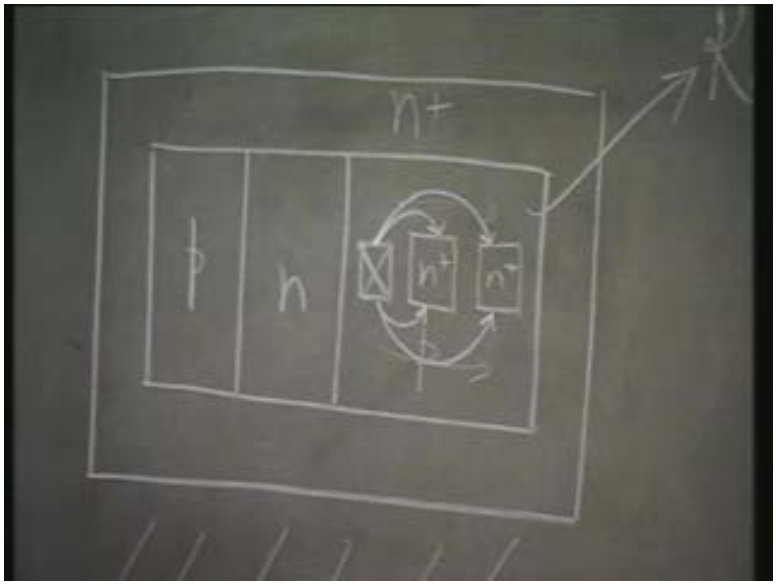
(Refer Slide Time: 45:37)



So, this is actually called the rails. On both sides, the region that we have is called the rails. So, as is obvious, if you compare these two mask designs, I have a lot of space saving if you use oxide isolation. Just compare this with this. I can achieve the same using a much smaller region. See, finally everything is determined by your lithographic limitations. So, it depends on how small an n plus region you can have. Here you must have a separation; here you do not need to have a separation. Therefore, the dimensions of this can be much smaller than in this case.

But, even though space saving is a great advantage, the oxide isolation case does have a small problem and what is that small problem?

(Refer Slide Time: 46:46)

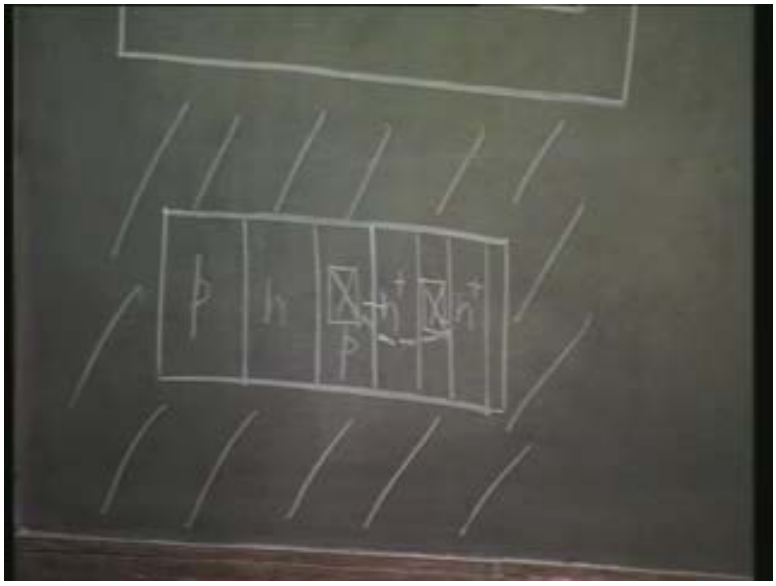


Look at the way the current flows from the base to this. I have to have multiple collectors in the I-square L; multiple collector means I have so many fan outs. So, the path of the current in this case is like this. Current can flow through the rails. I have p regions on both sides which are the rails and the current can flow; go to the second collector, to the third collector, through the rails. In this case however, how will the current flow? From this base to this first collector, it is not a problem. How about the second collector? It must flow underneath the first collector in order to reach the second collector, because I do not have any space on both sides for the current to flow, right.

(Refer Slide Time: 47:48)

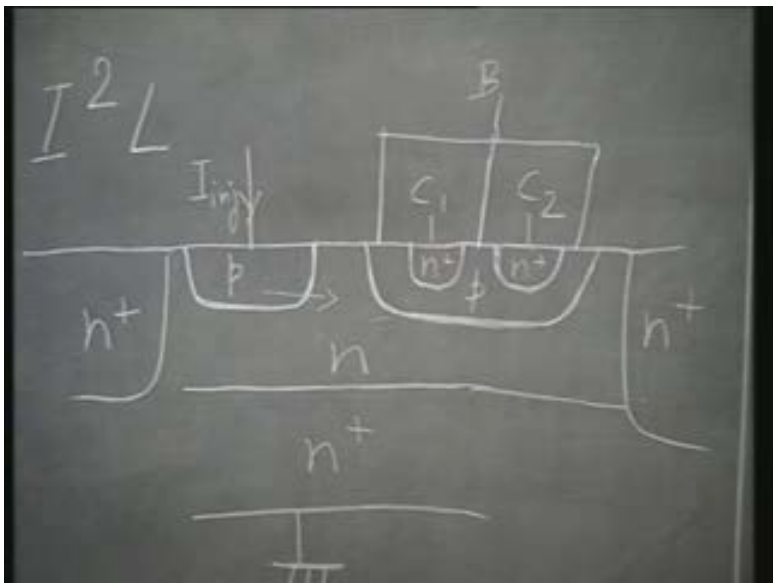
So, here it is like this. In the next case, it is going to be like this. The dotted line signifies that the current is flowing under the first collector, in order to reach the second collector and if I have a third collector, current has to flow underneath the first two collectors to reach the third collector and whenever I have the current flowing underneath the collector I am making it flow through a constricted region and therefore, the resistance in that path is going to be larger. So, I will have unequal voltage drops and which will lead to unequal drives. Remember, the multiple collectors actually represent the multiple fan outs. So, different fan outs will receive different drives. So, this is the problem in the oxide isolation case. By removing this rails which we thought was the unwanted region, by removing this additional p regions, we have indeed saved a lot of space, but in the process, we are having unequal drives. So, what is the solution?

(Refer Slide Time: 49:19)



The solution is to have multiple base contacts. One - this contact here, another, this contact here, another, this contact here, so that the current, in all cases they flow through equal resistance paths.

(Refer Slide Time: 49:46)



Instead of having one base contact, you have one contact here, one contact here, one contact here; all of them are tied together to have the common base. This requirement of

having multiple base contacts actually leads to a self-aligned I-square L structure, which we will discuss in tomorrow’s class.