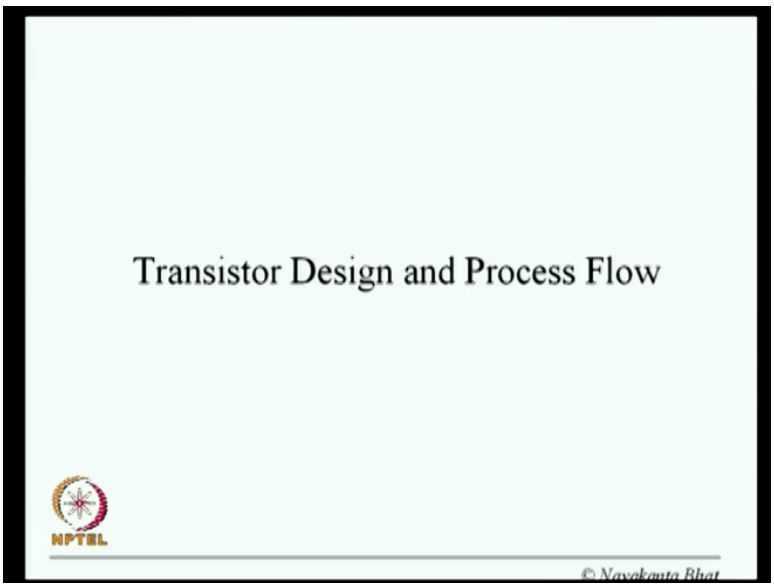
Nanoelectronics: Devices and Materials
Prof. Navakanta Bhat
Centre for Nano Science and Engineering
Indian Institute of Science, Bangalore

Lecture - 07
CMOS Process Flow

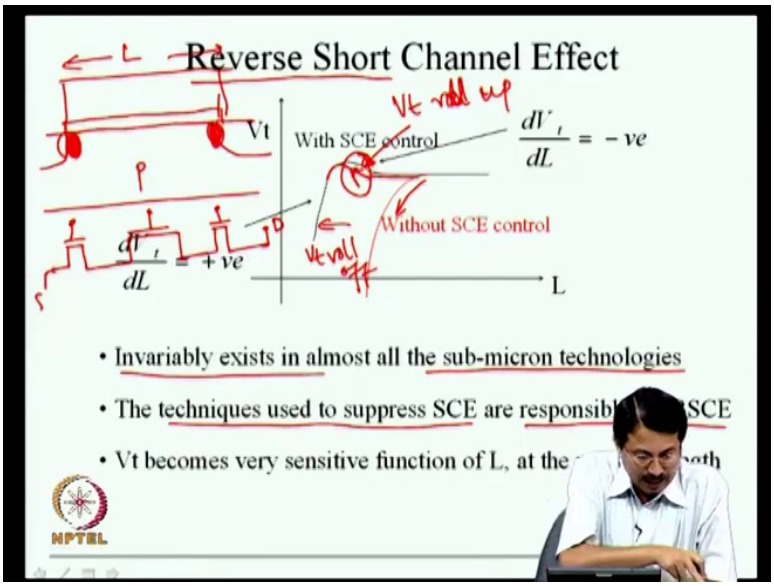
In the last lecture we have seen how do we use channel engineering and source drain engineering, in order to design a transistor at nanoscale. In this lecture we will look at two aspects; the transistor design metrics, how do we go about designing a transistor, what is the metrics that we want to achieve, and then subsequently go through a conventional CMOS process flow, right.

(Refer Slide Time: 00:40)



So, that is the target for today's lecture. Before that I just wanted to highlight one very important aspect that you see in today's CMOS transistor technology, and that is something called reverse short channel effect, which we have not yet discussed. We are discussing this right now, because we have discussed the channel engineering process, wherein we introduced pocket halos. It turns out, because of the introduction of pocket halos, we end up with a new phenomenon, which is called reverse short channel effect right, as the name suggests.

(Refer Slide Time: 01:25)



As you know, if you have a short channel effect the threshold voltage at long length is constant, and as you start you know coming down in terms of channel length, your threshold voltage starts coming down very drastically. And we have now discovered various techniques; such as channel engineering and source drain engineering which we discussed in the last lecture. If you follow all that, we can really improve this roll off, rather than this starting to roll off here. We can extend this even to much shorter channel length.

But in the process often times what we see is that, before the Vt starts coming down for ultra short channels, you may see a small region, where your Vt is actually increasing with respect to decreasing channel length, and this is what is called reverse short channel effect. Again this is relevant when the channel length is very small, but very small window of channel length, over which the Vt increases as a function of decreasing channel length right. I mean this is something interesting for you to understand. It invariably exists in almost all submicron technologies, which have utilized channel engineering and source drain engineering. The reason for that is, it turns out the techniques that are used to suppress short channel effect, more particularly pocket halo implant, itself is responsible for reverse short channel effects right, because that is a very interesting lesson to understand.

And you know Vt of course, becomes sensitive, it can either go up or come down, depending on which region you are operating; however, it has been extended, the flatness has been extended too much shorter channel length. Eventually it is going to roll down anyway you see.

Now, it is not difficult to understand this effect. Again if you were to look at this transistor that we have, let me just sketch that right. This is my gate length. And you see when we talk of pocket halos at an angle, we have increased the doping concentration in these pockets you see right. If this is an N channel transistor with P type doping, and eventually you will make N plus junctions here, and the concentration here has gone up, this P concentration. The P concentration here has gone up. A very simple way to understand river short channel effect is to really model this transistor as three transistors in series.

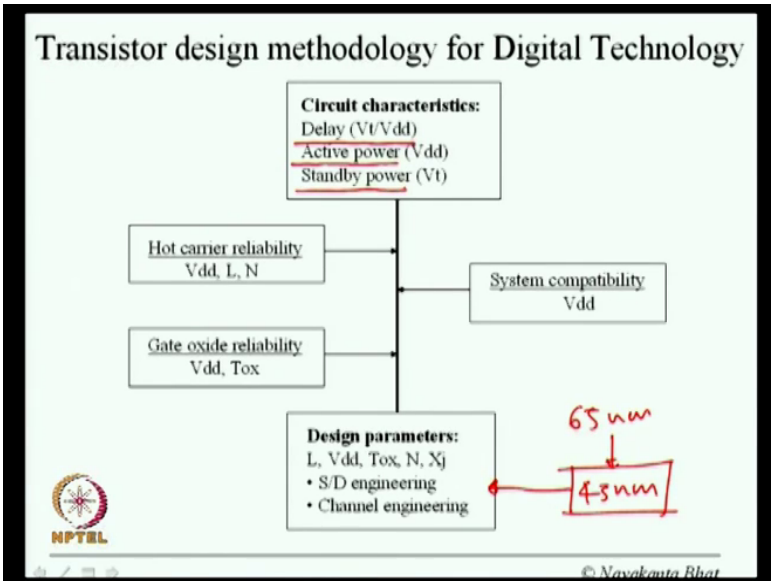
There is a very tiny transistor here, which is due to this increased pocket halo implant that I have, and then there is another transistor that I have, which is essentially most of the channel length that I have. And then again there is another tiny region here, and you can sort of think about it as if from source to the drain. This is my source and this is my drain terminal right, because there is a non uniform doping concentration, this is a transistor with high doping concentration, effectively larger threshold voltage. Very simple zeroth order Vt model that we have looked at in the very beginning. You know that increased doping concentration means increased Vt, right

So; obviously, in this region, this small channel length transistor has much higher Vt compared to the Vt in this region. And again this is exactly identical to this. Now you see again when the channel length is very large. This region is so small that it is very insignificant, you do not see the effect of these regions at all right, but once you start decreasing the channel length, before you actually begin to see the short channel effect, there could be a region where this length is now comparable to. This is the region, where these lengths are comparable to this middle length, which mean the total conduction property of this channel is governed by these two high Vt transistors as well, in addition to the lightly doped channel here, and because of that you could expect over a very short region, the threshold voltage starts increasing.

This is where, these small regions of pocket halos are actually becoming important in determining the transistor conduction, but once you start decreasing the length further, you know then, you know you our conventional short channel effects starts coming in, because the length is. So, small that you know you can essentially have the huge depletion width here, which will essentially govern the conduction of the transistor, and then you have a conventional Vt roll off, right

So, this is your conventional Vt roll off. Whereas, this is Vt roll up, if you will right, Vt increases here. It rather than coming down it actually goes up right. This is what we mean by reverse short channel effect, and as I said this is present in almost all technologies. And again depending on which technologies now you look at, if some foundry has put in lot more halo, you may see lot more reverse short channel effect. Whereas, if another foundry be as little less halo, you may not see as reverse strong, we reverse short strong reverse short channel effect as you would expect fine.

(Refer Slide Time: 06:57)



Now, given all that background, when we are designing a transistor for a new technology, that is for example, let us say you know, I am going from 65 nanometer technology scaling down to 45 nanometer technology. It does not matter it could be any technology node right. So, how do you design a transistor, to meet the requirement of this new technology node?

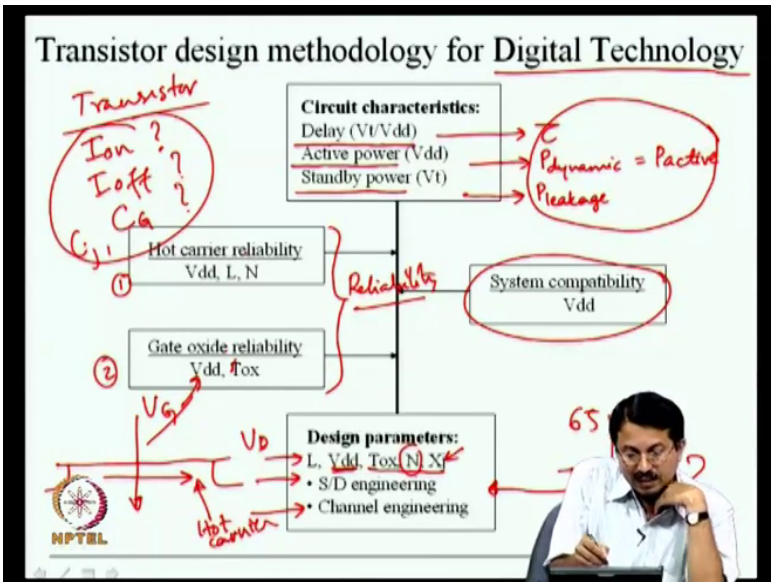
It turns out eventually the reason why we are building this transistor, is eventually make some chips right, and hence to make some circuits. So, these circuits in turn will have certain metrics. They need to have a certain delay, as we will see it turns out this is a very important function of the ratio of threshold voltage to supply voltage, they need to meet certain active power specification, they should not really consume too much of power. And they also need to meet certain standby power or leakage power consideration.

Again we are talking of digital technology. We are not talking of analog technology. In other words we are not talking of building amplifiers, building A to D converters, if that is your metric, then you are different kinds of metrics. For example, then gain linearity all those things become very important, whereas here when we are talking of digital logic circuit. The most important thing is speed of the circuit, and the next important thing is power dissipation of the circle ok.

So, when I go to a new technology node, I have a requirement that my microprocessor should operate at 10 gigahertz, rather than at 8 gigahertz, that should in turn translate to what should be your transistor speed, that information comes from experience ok

So, if you have to build transistors with certain architecture. If the transistor has a propagation delay of let us say 5 picosecond that may likely give you a power microprocessor at 10 gigahertz right. So, that kind of benchmarking is available typically right, and you know those who are making these chips routinely would have all that information. So, make use of that information, and for this new technology node you know what market demands, in terms of where do you want to position your chip in terms of speed, power and so on and so forth.

(Refer Slide Time: 09:28)



And then translate it back to the fundamental building block, which is transistor right. Then you come up with the specs what should be the delay or propagation delay, which we typically call tau for a transistor. What should be the active power dissipation, which we call sometimes P dynamic, or simply Pd, or we also call this as P active sometimes just P act. And accordingly you will also come up with leakage currents back right, it is essentially P leakage. This will essentially tell you eventually, all this numbers, in turn will tell you for a transistor eventually, what you are interested, is to translate from here to ask the question, what should be I on, what should be I off, what should be capacitance CG and also C junction and so on and so forth, all transistor capacitance.

So, this will eventually translate to the numbers that we can deal with, in terms of transistor design. Now I know, I need to meet this on current specification, this off current specification, this is the on current to off current ratio, I need to build for transistor at this technology node. Similarly I need to make sure that my capacitances are within the limit. So, that your tau which is Cv over I if you recall, is also met, and you reach the required speed specification for that technology you see. So, these are the things that we identify.

Given this, we go down and we know what length that we want to design the transistor. If it is 45 nanometer logic technology, the transistor length will be even smaller than 45 meter as we have already discussed earlier, and you would also know what would be

your Vdd, because you have the historic scaling trend, you do not want to let the electric fields go up very significantly right, you come up with the Vdd. These are all you know external specification, and that also sometimes get governed by you know where will this chip go, what kind of system is it going to interface with, and what supply voltage is required there, and you determine that, and then you go down further, what kind of oxide thickness I need to have for this transistor to get certain capacitance to get certain noncurrent.

What kind of doping concentration that I need to have Na or Nd, and what kind of junction depths that I need to create for this transistor. And in turn as we have already discussed, we do not want to do a very blind transistor process. We make use of source drain engineering shallow extension and deep source drain, channel engineering, which is pocket halos and super steep retrograde channel.

Using all this, we essentially iterate right, start with the best guess, may be iterate once or twice, to achieve all these metrics that we have; that is how we satisfy the circuit metric, and in turn if you make a more complicated chip such as pentium or whatever you have you know that will also meet the speed specification power specification. One other important point that I need to highlight depending on what product that you are making, what technology that you are going to use. You also have certain reliability consideration for transistor.

Transistor when fabricated in a chip should not only give this performance today, next week, next month, but may be for a year or 10 years right, and that is essentially governed by your reliability, is transistor going to fail in 1 year or 5 years. If it is a chip which is going in mobiles for example, you know, people change keep changing their mobiles may be once in couple of years, then there is no point in doing a reliable transistor for 10 year lifetime. You know it is an overkill right.

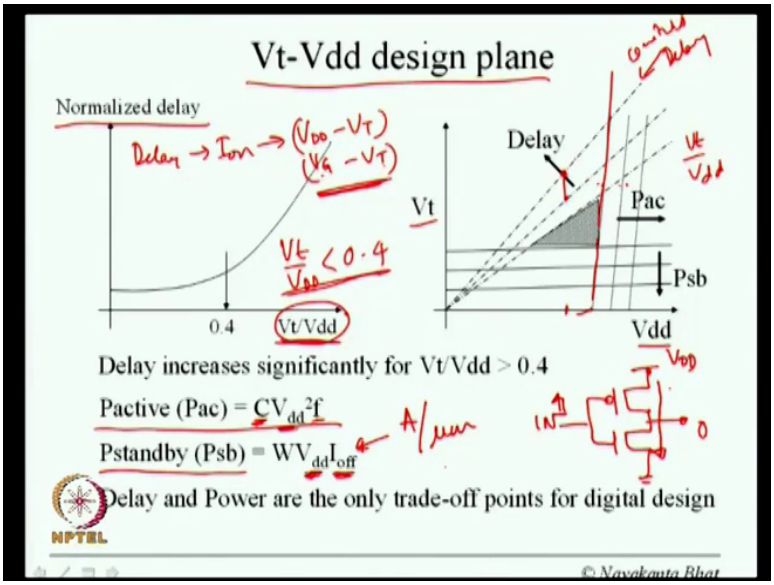
You will spend all your energy is trying to make it so reliable that you know they do not even use that chip for so long, whereas if the chip is going in satellites, you know or some other very mission critical applications. Then maybe you know you need to focus more on reliability. So, all I am saying is that reliability also becomes very important, and any particular application that you have, that in turn will have a reliability specs, may be 1 year 10 year 5 year reliability, and the reliability is essentially two kinds of

reliability. Number 1 is called hot carrier reliability, and number 2 is called gate oxide reliability.

We will talk quite a bit on gate oxide reliability in 1 of the future lectures, but we may not talk about hot carrier reliability, but suffice it to say at this point, that these 2 reliability come about, because of 2 different electric fields. There is 1 electric field in this direction; there is 1 electric field in this direction you see. You apply a gate voltage, which sets up a vertical electric field. You apply a drain voltage which sets up a lateral electric field. Hot carrier reliability is due to the lateral electric field, and gate oxide reliability is due to the vertical electric field. If you have too higher vertical electric field you suffer in terms of gate oxide reliability. If you have too high a lateral electric field from source to the drain, especially at the drainage, you suffer in terms of hot carrier liability ok.

In other words the message is that you need to also engineer this transistor and make sure that the fields are within certain limits. So, that you meet the hot carrier reliability, and gate oxide reliability spec, whether it is for 1 year, 5 year ,10 year whatever. So, this was how you essentially do the transistor designing. These are the metrics, and these are the knobs that are available for you, to reach these metrics ok.

(Refer Slide Time: 15:36)



Now, let us try to sort of revisit these 3 points in the context of digital technology again. This is what is called a Vt Vdd design plane. Vt is a threshold voltage of a transistor, Vdd

is a supply voltage of a transistor. And first thing that we recognized is, no delay we just normalize this scale. So, we do not want to put numbers here. The delay very interestingly, is a very strong function of Vt Vdd which is not very hard to understand you see, because delay depends on current, on current depends on your Vdd minus Vt right, your Vdd and vg are same right, it is the same supply voltage, because your gate on current, is if it is in saturation it is vg minus Vt whole square.

And similarly if it is, not necessarily in such I mean its transistor which is velocity saturated, and then we may not have a quadratic dependence, it may be little less than 2, but nonetheless it is this difference vg minus Vt . In other words we really are interested in Vt over Vdd ratio right, and that in turn determines how what is this difference. When Vt over Vdd ratio is small the delay is also small. When Vt starts approaching vdd. For example, in the limit as Vt tends to Vdd you have infinite delay, because you do not have any current to switch on your capacitors right. Remember CMOS circuits, you need to switch capacitors from Vdd to ground ground to vdd, you need to have current to do that. So, if your Vt starts approaching closer and closer to vdd, it is not a good thing.

So, a typical Vt to Vdd ratio that we try to target for especially high speed high performance application is, your Vt to Vdd should be preferably less than 0.4 that is if you have a 1 volt supply, then your Vt should be 250 milli volt max 0.25 lower the better of course, it is not good for leakage current, but it is good for on current. Now we are talking of on current, and how does it impact the delay you see. So, Vt over Vdd is important and hence in this Vt versus Vdd design plane, these lines which have constant slope, you see anywhere along this line. My Vt to Vdd ratio is constant, is not it, and that is obvious here Vt versus Vdd is of what we are looking at here right. So, tan theta is, theta is same here Vt to Vdd ratio is same.

In other words these are called constant delay contours along this line, anywhere I go along this line I have the same delay more or less. There may be some minor variation. We are not talking about those minor variation, and when I go from this line to this line, the delay increases right. and that is not hard to understand right. Going from this line to this line, meaning for the same Vdd I am going from this Vt to larger Vt . Now for the same Vdd correct; that is our transition from this lower line to the upper line. So; obviously, Vt by Vdd is increasing, your delay is increasing.

So, all these are constant delay contours, and the delay is increasing in this direction, very important to understand this. Let us come to two other matrix we talked about for digital technology right, by active power, what is it dependent on, I have already mentioned it sometime in some time ago c Vdd f in the context of Vt to Vdd design plane P active is a quadratic function of Vdd, which means P active is a very strong function of Vdd. So, along this line, these lines are called constant active power contour for all practical purposes. Again barring minor variation as long as I keep my Vdd constant, my power is same along this line, when I transition from this line to this line, I am going from lower Vdds transistor which operates at 1 volt versus transistor, which operates at 1.2 volt. So, active power would increase then. So, the P active as the arrow shows along this direction, your active power is increasing.

Let us come to the last part, which is standby power. What is standby power transistor is not switching right, but still there is a leakage current, a very simple illustration is that I have a CMOS inverter PMOS and N channel connected in series, my input is steady state 1, your output is steady state 0 digital logic right, Vdd ideally when your input is 1, this PMOS should be completely shut off, and hence no current, but we have already seen, there is sub threshold current.

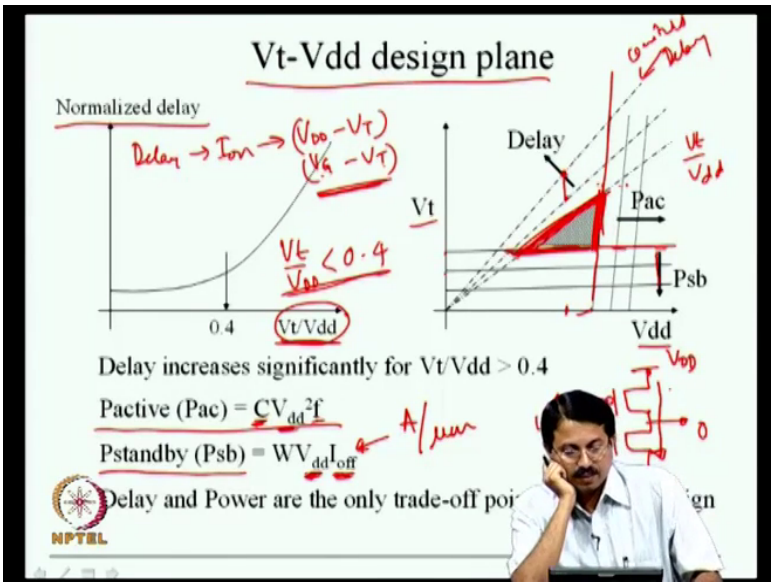
So, there is a nonzero current which flows here. So, that is off current times vdd. Typically when we talk of off current, we talk of all off current normalized to the width. In other words, the unit of off current here is ampere per micrometer, and that is why you multiply with the total width of the transistors that gives you ampere times volt, which is essentially your leakage power, then notice that standby power is a linear function of Vdd and linear function of off current; however, off current as you may remember from our sub threshold discussion of current is an exponential function of threshold voltage, a very small change in threshold voltage. Remember the sub threshold slope discussion results in exponential increase in off state current.

In other words what you see in terms of Vt Vdd design plane is standby power is an exponential function of Vt standby power is a linear function of vdd. So, for all practical purposes, we can say that once we fix Vt the standby power gets fixed essentially because of the Vt value. Although there is a minor, because of Vdd right, we are not really worried about the minor variation; that is why these lines will not be perfectly horizontal. You see these lines will be you know they will have some slope which

illustrate that these are constant power, you know if it is, we call them here as a constant standby power contour, ok.

But in reality even though Vt is same, when Vdd increases power increases a little bit. I know that from this equation; however, when I go from this Vt to this Vt there is a huge change in standby power. So, that is why and also standby power is increasing as the arrow suggests here in this direction lower Vt higher standby power standby power has increased exponentially. So, you see what has happened now, this is a Vt Vdd design plane, it gives you; now what is the allowed operating regime for your new technology. Your new technology may have certain delay number.

(Refer Slide Time: 23:55)

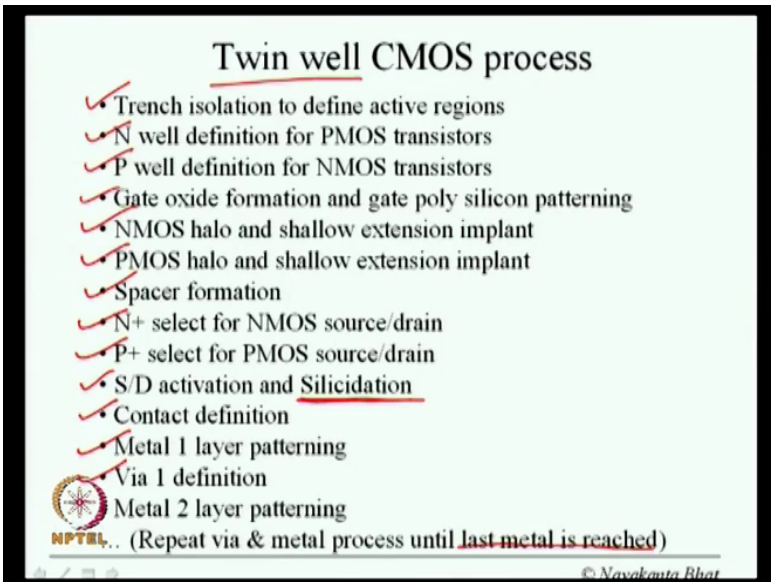


Let us say this is my delay requirement, it has certain active power requirement let us say this is my active power requirement, it has certain standby off power requirement let us say this is what it is. What it means is that, you can be anywhere in this triangle and still meet all the specifications that are required.

In other words there is no unique solution right, you can have different combination of Vt and Vdd you see, and meet this specification that you had to begin with, the delay specification, active specification, standby power specification, and that is why if you look at technology is, let us say 45 nanometer technology from different foundries, let us say Intel's technology, Tis technology, and Infineon, St Microelectronics, TSMC. They may all be meeting same specification, but the transistors could be quite different.

The way their transistors are made, because there is a window of operation, and each share of foundry may have some preference in designing a transistor, still meeting all those specification that you have right. So, that is also very important point to recognize, although nominally all of them are 4 5 nanometer technologies, they are not identical, there could be huge difference between each of these technologies, but all of them will meet certain specifications.

(Refer Slide Time: 25:27)



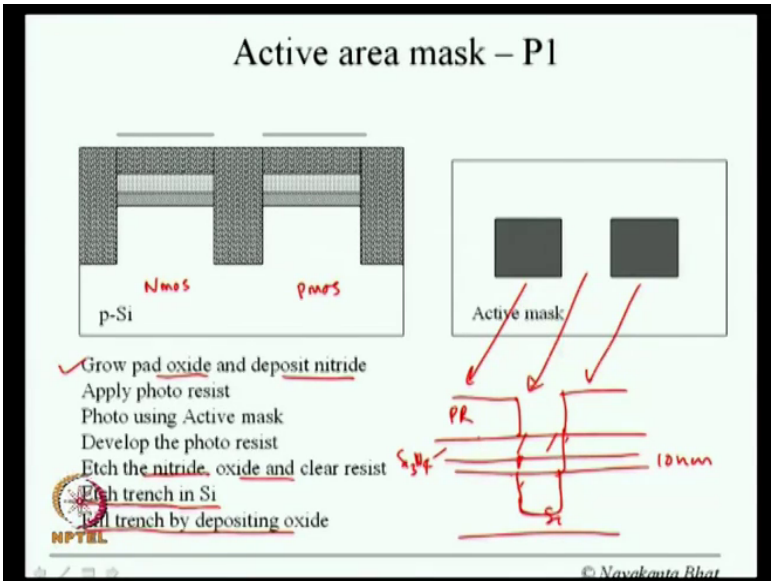
So, with that let us look at a typical CMOS process flow. Today's technology, most of the modern CMOS technology uses what is called a twin well CMOS process view. What it means is that, the starting wafer type is inconsequential, you can start with N type or P type, because either for NMOS or for PMOS, you have to do the channel engineering source drain engineering, you have to define all the doping wells the way you want right, which means you have to define both the wells anyway, the way you want it, and that is why it is called a twin well CMOS technology. It starts with defining isolations which in today's technology as I mentioned trench isolation technology. Then you define wells n wells for PMOS and p wells for NMOS transistors.

After that you form what is called gate stack, which includes gate oxide formation and polysilicon deposition, and polysilicon patterning, then channel engineering. Do the halo implants for N channel transistors. For N channel transistors at the gate edge you see you have to increase the P type dopand, whereas, for P channel transistors the substrate being

N type at the gators you need to increase N type dopand; that is why you need to have different halo for NMOS and different halo for PMOS. Then you need to form spacer as we discussed last time.

Then do a deep source drain implant, shallow extension, before the spacer you know along with the halo you also do the extensions right, becomes clear as we move on, and then you know do the N plus P plus deep source implant, after implantation as I have mentioned earlier you have to do a heat treatment and that is what is called a annealing, you do the annealing. And there is another very important step we are not yet discussed I will talk about it later, something called solicitation, define your contacts for metal, deposit metal 1 layer, then via 1 on top of metal 1. If your technology has 10 metals keep doing that metal 2 via 2 metal 3 via 3. All the way up to the last metal is reached, and that completes your process flow, ok.

(Refer Slide Time: 28:17)



So, let us look at it a little more carefully. Now let us say that I want to build this twin well CMOS process, wherein I want to build let us say an inverter, an NMOS transistor here a PMOS transistor here, eventually connected, the 2 gates are connected 2 drains are connected. So, let us say that is the structure I want to build which is a simple representative circuit. So, the first thing that I need to do, is the whole silicon wafer is semiconductor right. I need to define isolations, we have already talked about isolation

right; 1 transistor should be isolated compared to the neighboring transistor, how do you do that put in oxide in silicon.

So, the way the process sequence goes as this you know, you start with bare silicon wafer and you grow what is called pad oxide. So, what is shown out here you know is really the pad oxide, grown on top of silica? And then you know you apply photo resist, then you do a photolithography use this mask, and this mask will tell you that you know these region should be blocked, only these regions should be filled with the put oxide, and those regions should be opened, you do that and you know then you etch the nitride and oxide stack that you have, you see you have oxide and nitride that you have deposited initially you etch. Then etch the trench in silicon and fill that trench using depositing the oxide ok

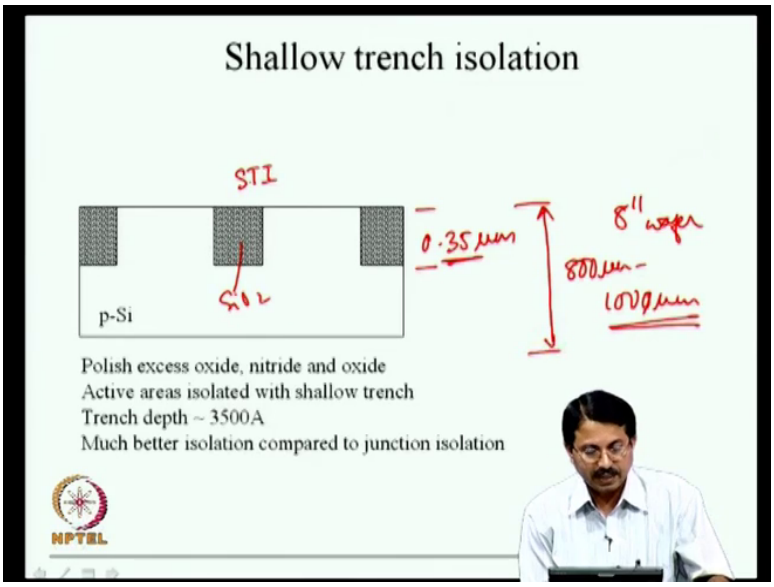
In other words you start with bare silicon wafer, grow oxide everywhere. This is what is called pad oxide, typical you know pad oxides that we talked about you know are of the order of may be 10 nanometer. As the name suggests it is its role is sort of cushioning silicon nitride has, if you put it directly on silicon it can actually impact the silicon substrate, it can create defects in silicon right; that is why you put oxide first and then put silicon nitride. Silicon nitride, because silicon nitride is a very good barrier for oxidation right, that is why we are putting silicon nitride. And using photolithography right we say that in this region you know I want to build a transistor, in this region I want to build transistor, but in between I want to fill oxide.

So, what I do is that I put photo resist and if I want to define this right. So, I open the photo resist here, correct using lithography there is this mask, in this this is the region that I am talking about, this region corresponds to a transistor here this region corresponds to another transistor here, this region opens the photo resist PR and then you etch this nitride that is what is mentioned here, etch the nitride oxide etch the oxide. So, this whole thing is etched. Let us go back to this right.

So, this whole thing is etched here, and then you have a silicon opening right, in this area silicon is opened to the external world. You continue etching silicon, silicon is etched further here, and you know that results in a trench inside the silicon; that is what is meant by etch trench in silicon in this region, and then you fill the strange by depositing oxide ok.

This is how I define some regions to be oxide; some other regions which are protected by this stack continue to have silicon underneath. And once this job is done, I strip off this silicon nitride and I strip off this oxide ok.

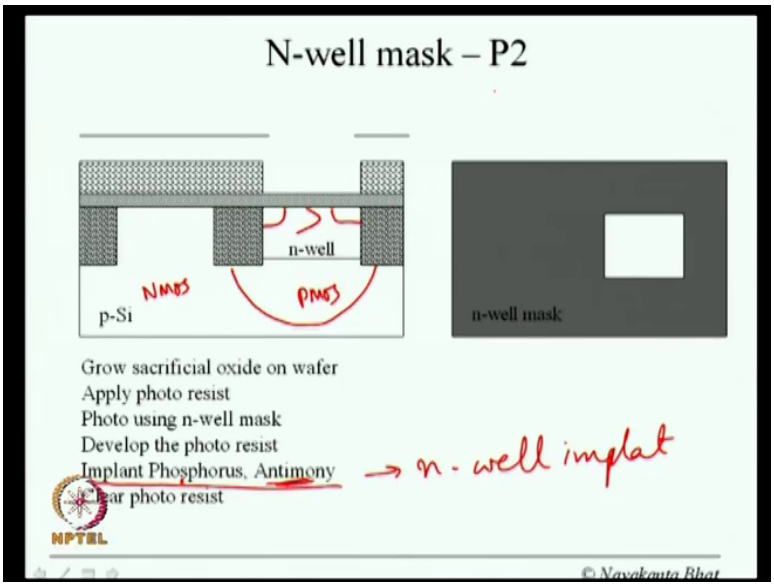
(Refer Slide Time: 32:32)



Having done this you will have a substrate now which looks like this right, part of the region is now Sio 2, this is the isolation and the rest of the region is very fine silica. The shallow trench isolation is done. And the way it is done is that you polish the excess oxide remove nitride, and you know when you do all that. You come back with a very flat surface of the silicon wafer.

And typical depths that we talked about here are of the order of you know 350 nanometer or 0.35 micrometer, compared to the wafer thickness this is not to scale remember that. If you talk of 8 inch wafer for example, that may be close to 1000 micron thick, anywhere from 800 micrometers to 1000 micrometer. This is an 8 inch wafer, compared to this dimension 0.35 micron is insignificant right, and that is why we call shallow trench isolation or STI. This trench is very shallow compared to you know the dimensions that we are looking at.

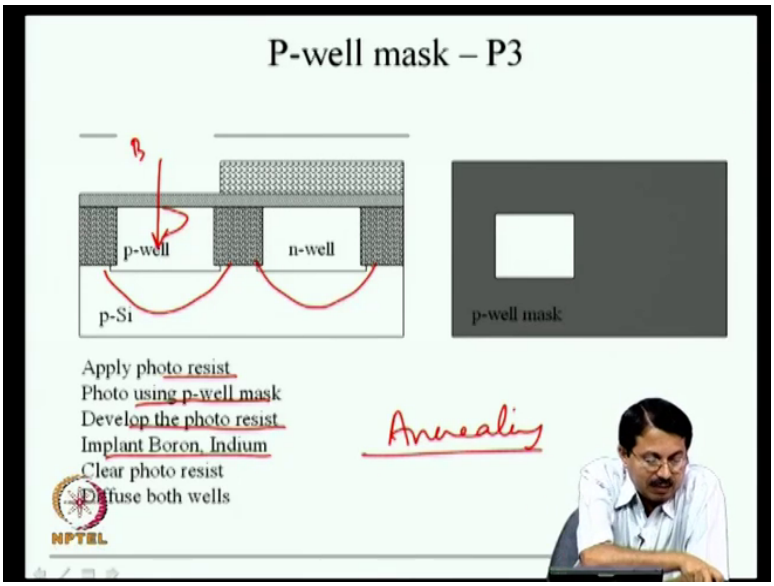
(Refer Slide Time: 33:49)



Well then. So, we have isolated, what is the next step well definition. In some regions remember I want to make PMOS transistors, and in some region I want to make NMOS transistors. Wherever I want to make PMOS transistor, I want to put down photo resist first, open the regions the mask will look something like this right. This is the region which is open.

So, using this I can etch the photo resist here in this region, and you do n well implantation right, this is n well implantation phosphorus and antimony. Phosphorus decides the deep junction depth of the well right, because we want well to be very deep, and antimony as you remember will determine a very steep retrograde profile here, which will help in subsequently when I make the junctions. I have not made the junctions yet right remember that. So, that you know this will help us engineer the transistor in a better way as we have discussed right. So, when you are doing this in n well region, the whole p well region is completely protected, because you do not want these implants to go there; that are why you need a photolithography step.

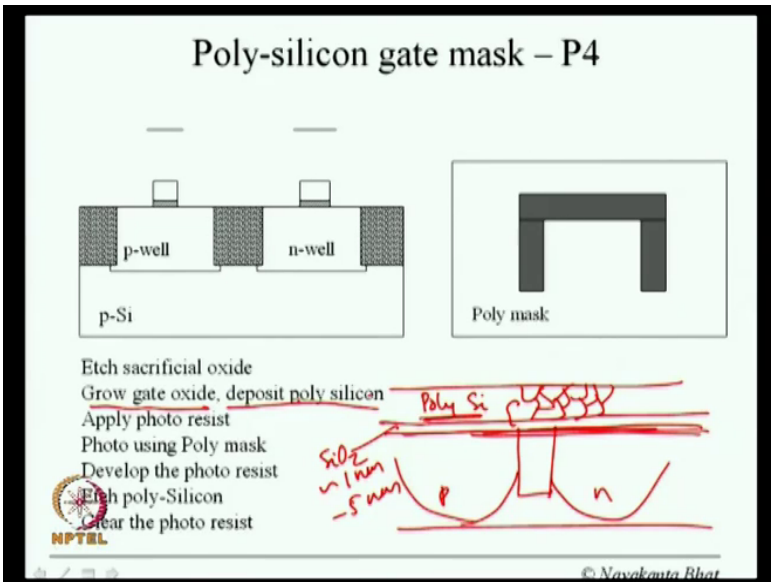
(Refer Slide Time: 35:09)



After this compliment you need to define p well. The mask is complement of that. In the active regions, where you do not have n well, you will make p wells, open those regions again using this mask, which again apply photo resist photo lithography using mask, develop the photo resist you know, you open these areas then implant boron and indium here right.

Again boron determines the depth indium sits here. And you know when you do that you clear the photo resist, and then you do after implantation invariably you have to do annealing, because implantation if you remember creates damage in silicon, and during that annealing there is also going to be some diffusion. The wells will diffuse and you will eventually get some n well depth and P value. So, we have defined wells we have defined isolation and active region, we are ready for gate stack.

(Refer Slide Time: 36:10)



How do you do gate stack; you again start with you know when we are doing the previous implants you would have had something called sacrificial oxide. There is some oxide you need to strip that, again get a bare silicon wafer, flat silicon wafer that silicon wafer will look like this, there are these oxide isolations, there are these wells formed p well formed n well formed with a flat silicon surface at the top. I need to grow gate oxide, ultrathin gate oxide right that is what we are talking about today, 1 nanometer kind of gate oxide.

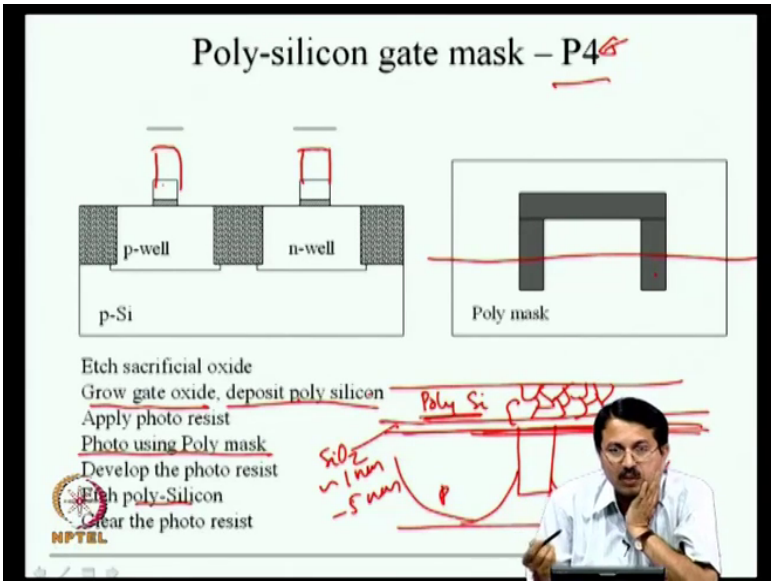
So, grow gate oxide deposit polysilicon because polysilicon is our electrode, polysilicon gate technology. We will later talk about metal gate technology, but you know for long years CMOS has been silicon oxide based polysilicon gate technology. You deposit polysilicon what happens here Sio 2, very thin. Let us say of the order of 1 nanometer to 5 nanometer depending on what technology you are looking at then polysilicon. It is polysilicon, because we deposit silicon because silicon is being deposited on oxide, which is an amorphous I cannot get single crystalline silicon. I end up either with amorphous silicon or polysilicon, but in this case I want poly crystalline silicon, because I want as low resistivity as possible.

So, I intentionally go to high temperatures of processing, and that will give me polysilicon. So, I get polysilicon which essentially means that there are multiple crystallites of different orientation, crystal orientation is not the same, lots of crystallites

polycrystalline, but you see now polycrystalline is everywhere on your wafer, meaning all gates are shorted correct I need to do the lithography to isolate the gates let us say this is where I am building inverter, there will be a gate here remember, this is the length of the transistor and width of the transistor is perpendicular to this screen, and when I see the top view this is the length and this is the width, you know it is going down here, and that is what is shown here, and this gate and this gate needs to be connected in an inverter, and you will apply an input to this right connected gates.

Let us say I have this mask, I use that mask and photolithography using polymath, this is called poly mass, which is P 4, which means fourth layer of the mask, develop the photo resist etch polysilicon, right

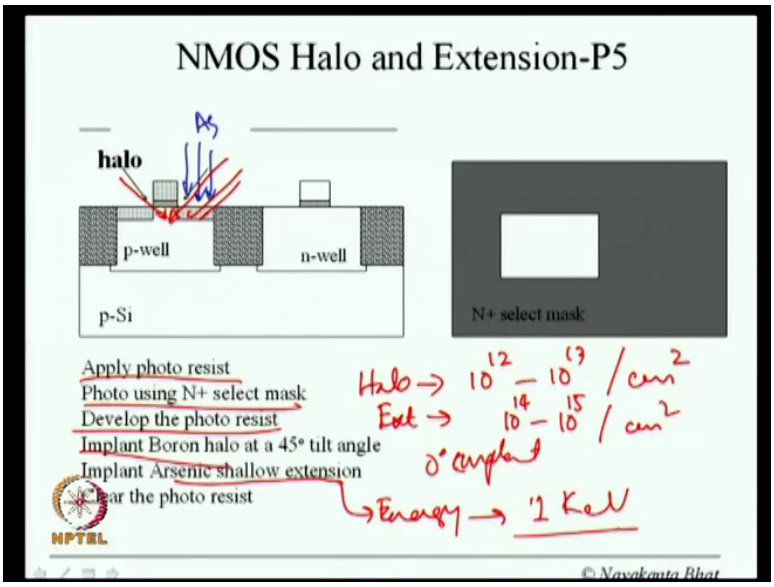
(Refer Slide Time: 38:54)



So, when I develop the photo resist photo resist will be remaining only here, and here rest of the region photo resist is gone. I expose the polysilicon to etch end all polysilicon and oxide will be gone in this region. In this region photo resist was protecting that, strip off your photo resist, you will be left with a polysilicon oxide stack. This is what we call a gate stack .And polysilicon oxide stack if you take a cross section here, this is what you see right the cross section is not taken here. If you take a cross section here you will essentially see 1 strip connecting the 2, correct the core cross section is essentially taken here.

So, isolation done, n well p well done, oxide silicon oxide extremely important gate oxide and polysilicon is done, polysilicon, at this time is still undoped, it is not doped yet, when I do the source drain polysilicon also gets doped, ok.

(Refer Slide Time: 39:57)



So, then I am ready for channel engineering. What do I need to do channel engineering, halos implants and also extension implants? So, I open the NMOS regions, and block all PMOS feel; that is why this mask is called N plus select mask. It selects only N channel region and blocks as you see here a line here, indicates blockage blocks P channel region. So, whatever you are doing is only for N channel transistors. So, apply photo resist photo using N plus select mask, you open it develop photo resist means open set, you have photo resist protecting everywhere else implant boron at an angle; that is the halo. I am increasing the P concentration selectively in this region, at the same time I also do extensions.

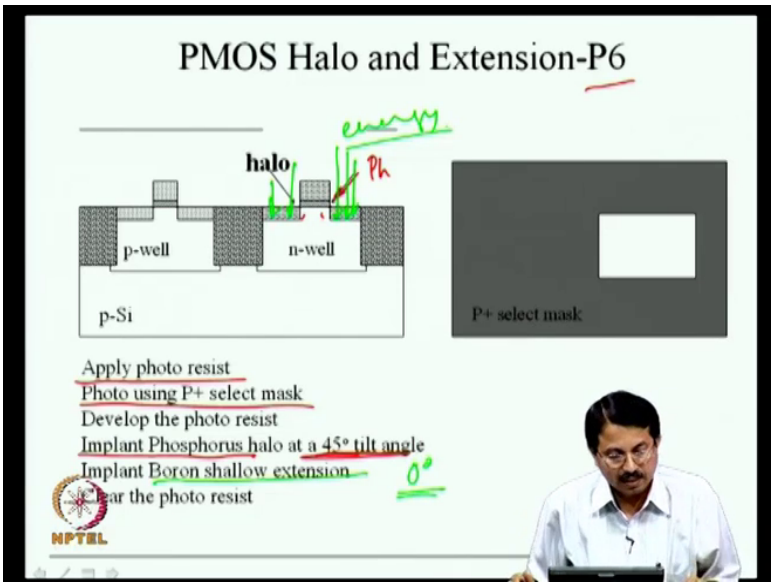
So, I follow this boron is done at 45 degree, follow this with arsenic implant at 0 degree, 0 degree implant. What it means is that, you are, let me just see here, let me get a different color here you have to sort of illustrate that. So, this is your arsenic implant, it comes at a 0 degree to the normal meaning, it is coming perpendicular to the wafer surface, I can set that in my equipment. Whereas, when you are doing your halo, halo was coming at an angle.

So, what have you done? You see boron got implanted everywhere, but in this region that boron was overcompensated by a very significant implant. Just to give you an idea the typical halo implant dose, it is the dose which determines concentration. The halo implant dose could be anywhere in the range of 10 to the 12 to 10, 10 to the 13 per centimeter square, that is how we define the dose. You see dose is different from the doping concentration in silicon. Dose is what is coming down per unit area, how many ions are coming down, how many ions have put in per unit area that is the dose. When that goes inside you look at the volume and get the volume concentration.

On the other hand if you look at the extension implant, extension implant easily is in the range of 10 to the 14 to 10 to the 15 per centimeter square. You see two orders of magnitude higher and that is why even if there is some P here boron that is overcompensated by arsenic. So, this region becomes N plus. Whereas, this region becomes p, enhanced P type doping, and I have been able to introduce pocket halos, I have been able to introduce shallow arsenic extension. Why shallow, because implantation energy is very low here.

Typical arsenic implantation energy, because you see it is energy that determines the depth, is of the order of 1 kilo electron volt. Whereas, when we talk of deep implant for arsenic; that is more like 5 kilo electron volt 7 kilo electron volt. And this energy determines the depth, and the dose determines the concentration. I choose the dose such that in this region P becomes n, and I choose the energy such that I get a shallow implant. This is all in the N channel region, by blocking the P channel region using the mask that I have.

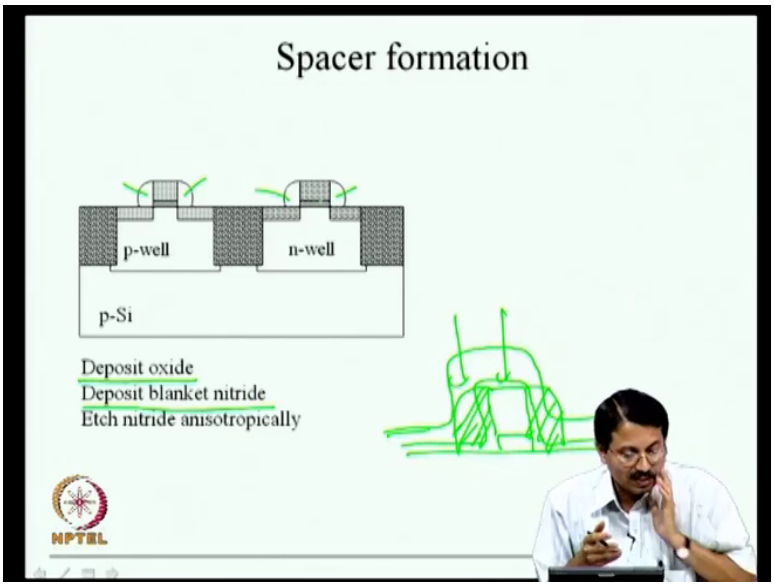
(Refer Slide Time: 43:48)



Now, I do the compliment for a P channel transistor. Do another lithography this is a sixth lithography step, I block all my N channel regions, which means photo resist photolithography using P plus select mask, which opens only PMOS region. Then phosphorus is what we use for halo, phosphorus, because I want to increase this N type concentration selectively, phosphorus is N type dopant, and that is at 45 degree tilt angle you see. Then follow it up with the other implant you know, which is a boron shallow extension at 0 degree, boron is a P type dopant as you know and that is coming down here at very low energy.

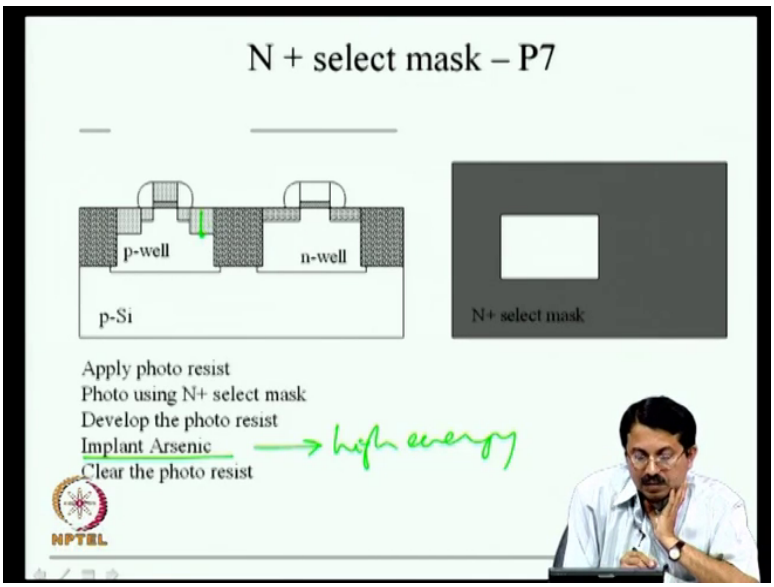
And again the doses are 2 orders of magnitude different, boron extension know this at least 2 orders higher than phosphorus halo dose, you see, and that will make sure that this become P plus region here, and also very shallow, because energy is chosen properly. Now, I have done most of the transistor design here, you know whatever design I had and I have implemented that in the process.

(Refer Slide Time: 45:00)



Except last aspect which is deep source drain. So, spacer needs to be defined. How do you define? We have already discussed deposit oxide deposit nitride. So, then you have oxide and nitride everywhere, except that because you had this gate stack, your oxide covered like this, and your nitride also came like this. And hence when you did the anisotropic thing chain you left with this region correct, and that defines my spacer here, itself align meaning no lithography here. Wherever there is a gator it has identified by itself automatically, and created a spacer, using this kind of a process sequence.

(Refer Slide Time: 45:50)



Ready for final implant N plus needs deep arsenic implant. Again use a same N plus select mask block all P regions, open this N region arsenic implant, now much higher energy. So, that it gives a deep junction; a compliment for a P channel transistor.

(Refer Slide Time: 46:15)

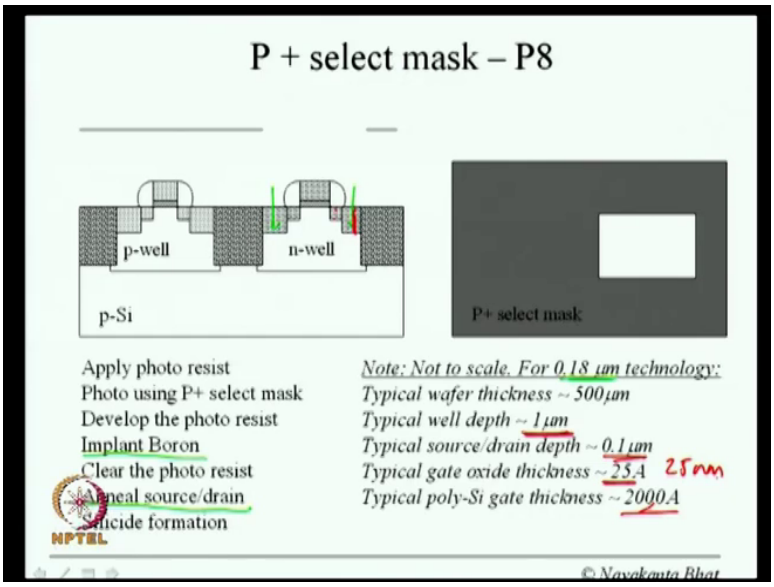
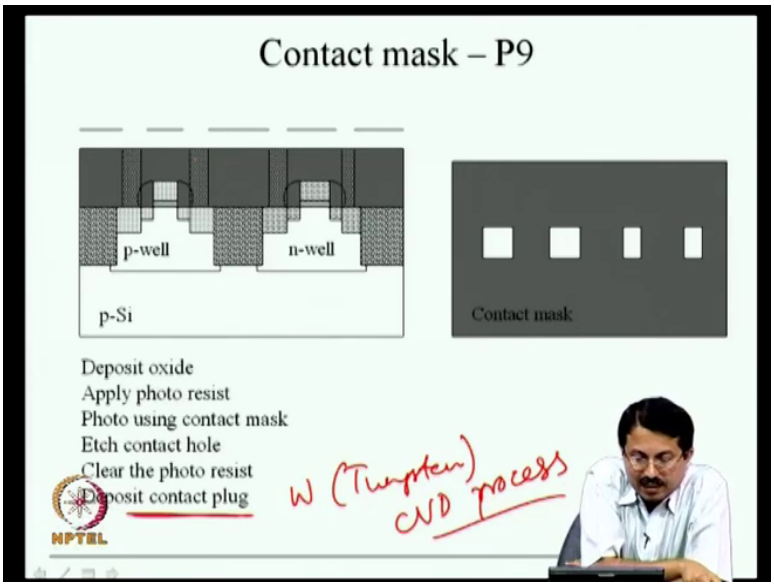


Photo resist you know implant boron here. So, boron comes here. So, after this implantation again you have to anneal the source and drain, because there is a damage you need to anneal the damage, and there is a silicide formation, will come to that in a while. Let us just look at some of these dimensions here.

I have just take 0.18 micron as an example for an illustration. So, the typical wafer thickness depending on what is the wafer diameter. If it is a 6 inch wafer it may be 500 micron, 8 inch wafer is much thicker right. Well, depths are of the order of 1 micrometer, wells are deeper, whereas trenches are shallower than this as we have already seen which is 0.35 micron trench, source drain junction depths, these are deep junctions are of the order of 0.1 micrometer.

And if you would look at the extensions, extensions are even shallower. They may be just 50 nanometer or so, and polysilicon gate thickness about 2 thousand angstrom or 200 nanometer. Gate oxide thickness, depends on your technology again this 25 angstrom is 0.18 micron technology 25 angstrom, as you know is 2.5 nanometer correct; that is 25 angstrom, but this can be anywhere from 1 nanometer 2 nanometer 5 nanometer, depending on which technology you are looking at.

(Refer Slide Time: 47:59)

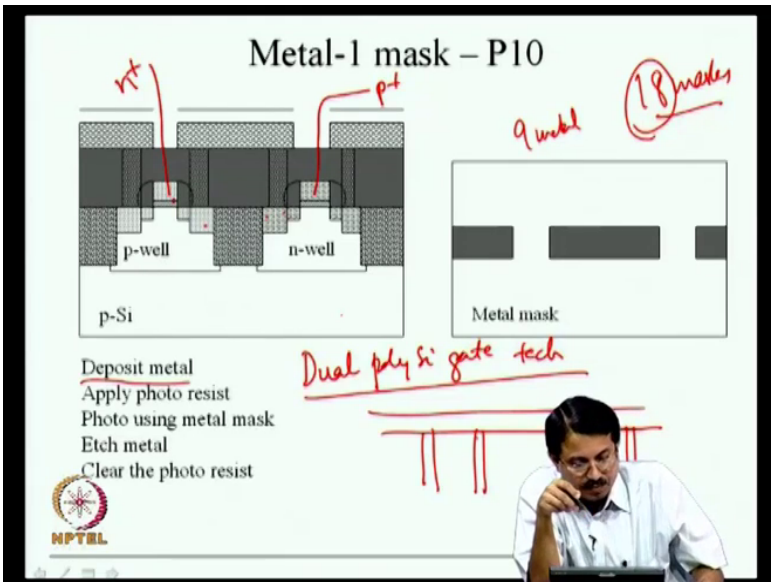


So, almost done with the transistor; now we are ready for contacts, I need to make contacts to the transistor you see. So, the contact is to be done again using a mask, I need to make a gate contact sourced in contact and so on and so forth. So, here what will happen? Again first of all you deposit insulator which is oxide, and in oxide you open the holes, wherever you want to make contacts. These are plugs, which will be conductive plugs, and you fill these plugs.

In fact, it is called contact plug, typically it is filled using tungsten, using a CVD process chemical vapor deposition process. So, if I want to define contacts to source and drain of NMOS, source then drain of PMOS you need 4 openings at least right. If you also want to have an opening for the gate, you need to define the gate opening as well, but that may not come here, that may come at a different location, because whenever you want to make a gate contact you see gate would have been connected like this, eventually you make a gate contact here.

So, again when I take a cross section here you will not see that, and that is what I have shown only 4 contacts. The source and drain, source and drain here. And now this is how I would eventually take this you know this has been filled with tungsten, this has been filled with tungsten. All I have done is that taken these connections to the next level. I need to do appropriate interconnection using metal 1, now that requires metallization ok.

(Refer Slide Time: 49:50)



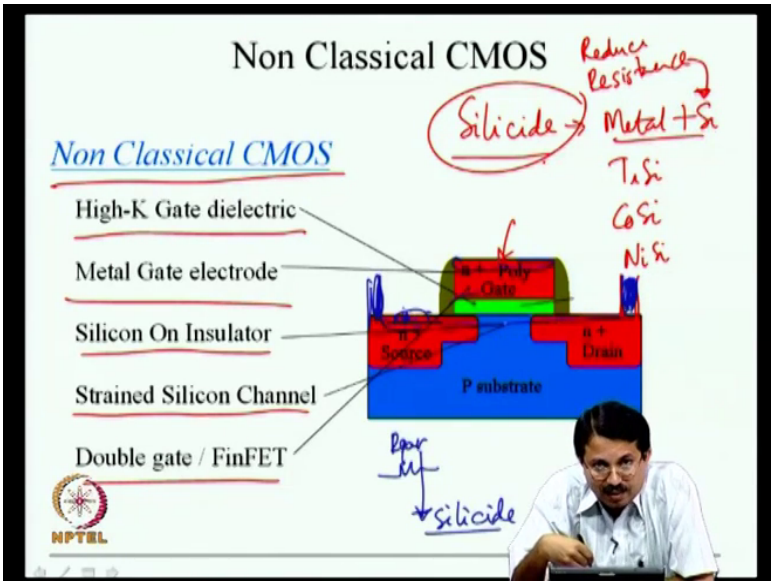
If it is aluminum metallization technology you deposit aluminum everywhere; that is you had these contacts, and you deposited this aluminum. Aluminum is shorting all contacts, this is not what we want. We want to now etch aluminum selectively. If I want to connect only this contact and this contact, because this is the drain of N channel transistor, and this is the drain of a P channel transistor, in an inverter only these two should be connected, but this should not be connected here.

So, I will use a lithography mask, I will protect this region, open this region, open this region, etch out aluminum here, etch out aluminum here, and I have connected the drain of N channel and drain of P channel together, to form the output of the inverter, and this will be the ground potential or Vdd, depending on whether it is N channel or P channel in this particular case it is N channel transistor. So, this you can connect to ground later, and this is P channel transistor which you can connect to supply voltage. Now you have an inverter; a very simple circuit fabricated through a thin will CMOS process.

But when your circuit is more complex there are so many transistors; 1 layer of metal is not sufficient, and that is why you need to start forming multilayer metal into connects. You again deposit insulator, put a via wherever you want to take the connection to the next level, metal to pattern metal to and so on and so forth. You go up to metal 10 or whatever it is right. So, each metal layer will require a vin metal, which is 2 masks.

Now you can see right, I mean if you have to add another 8 metal 9 metal layers, you need another 18 masks, because 1 via and 1 metal another 9 metal layers will be 9 metal plus 9 vias total 10 metal layer into connect let us say. So, 18 mask, you already have 10 mask 28 mask process, very complex process. All these masks have to be perfectly aligned to the patterns that are already printed; otherwise you know you will not have a working chip right, and this is really a very complicated you know technology that we have been able to routinely practice today.

(Refer Slide Time: 52:14)



So, this is what we call a conventional CMOS, and in the rest of this course we will look at lot of nonconventional things in the transistor; high K gate dielectric. In fact, from the next lecture we will start discussing the issues with silicon oxide today, and why do you have to replace silicon oxide, with high K dielectrics. Metal gate electrode, polysilicon technology it is called a dual polysilicon technology I forgot to mention that to you, when I make N channel transistor, the polysilicon in N channel transistors region is doped N plus.

Whereas, this is dope P plus, because whenever you are doing P implants that implant will also go into this polysilicon whenever you are doing N plus implant that will also go into the gate. It will not come in the channel, but it will go in the gate electrode, and hence it sometimes also called dual polysilicon gate technology. Dual meaning, there are 2 kinds of polysilicon here, in terms of doping concentration; one is heavily N type dope,

the other one is heavily P type dope, and hence the name dual polysilicon gate technology.

Now, we are talking of replacing that polysilicon with metal gates, and we will talk about some interesting aspects related to that. And subsequently in the course you will also look at what is called silicon on insulator technology. I briefly mentioned this in the context of better sub threshold slope. You will also look at strained silicon channel technology, intentionally creating strain in the channel. And not only silicon after this discussion you will also be looking at germanium based technology, or you know compound semiconductor based channel and things like that, and also you will also see things related to double gate and finFET technology.

All these we call as non classical CMOS right. Meaning classical CMOS was very simple silicon oxide as the gate, insulator, polysilicon as the gate electrode material right, and bulk silicon as the substrate, not soi, and single gate transistor the gate is only on the top, there is no gate on the bottom. When we talk of double gate, there is a top bottom gate or two side gates right, or you know all around gate, you know gate is surrounding the entire channel. So, we will have more discussion on this in the next lecture, and may be in the next lecture I will also tell you about or may be one of the subsequent lecture, what we mean by solicitation, you know that is something that we skipped during the discussion of CMOS process flow

Suffice it to say at this time that silicide. What we mean by silicide, is essentially, it is a complex of metal plus silicon. A combination of metal plus silicon is called silicide. So, then you can have titanium silicide, it is a compound right titanium silicide, or you can have cobalt silicide, or you can have nickel silicide and so on and so forth. The thing about silicide is that, silicide has much lower resistance compared to silicon, because you are putting some metal in it ok.

But certainly it is not as lower resistivity as a metal is. So, it is somewhere in between silicon and metal. The purpose of using silicide is essentially to reduce resistance, compared to what it was for silicon. Silicon has higher resistance, especially in the parasitic transistor region, you see you need to reduce the resistance. In other words let us say the opening, contact opening that we talked about let us say there is one contact here, and let me just say that there is a contact here and there is a contact here, but

remember how the current flows right, current is flowing like this. This is a region which can be controlled by the transistor.

This region cannot be controlled, and I have talked about this earlier, and this in fact, is called parasitic resistance. Now if this region which is silicon can be converted into a metal silicon complex, this parasitic resistance that you have effectively are parasitic, can come down for silicide. So, then what all we try to do is, convert this region of silicon into a silicide, which is a combination of metal and silicon, and that is what we do today; otherwise you would have had much higher resistance.

That is why it is process that is used wherein you put the metal on top of it, you anneal that metal at high temperature, that metal will react with the silicon and form the silicide here, and it will also form the silicide here as well, which is also good because that will also reduce the gate resistivity. It is polysilicon becoming a silicide there. In this region you have insulator. So, there will not be any silicide; that is good, because otherwise you would have a gate and drain shorting to each other that does not happen. So, that is the purpose of silicide. So, good, we could cover that today.

So, we will stop the lecture today. And in the next lecture we will start talking about silicon oxide scaling, and why do we need high K dielectrics.