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

## Lecture - 36 MOSFET I –Metal gate vs self-aligned poly gate

So far, we have discussed about bipolar junction transistor technology in integrated circuit. Today, we are going to start the discussion about the other branch of the most important device in integrated circuit that is the MOSFET. Let us try to trace the history of development of MOSFET technology. You know, the name MOS, MOS actually means metal oxide semiconductor. So, the MOSFET as it started originally, it did have a metal gate structure; so, these were the older generation metal gate MOSFET. From metal gate MOSFET we came to self-aligned polysilicon gate MOSFET technology. We will first see what was the important processing steps in the metal gate MOSFET technology is going to, you know, going to give an enhanced performance of these devices and then finally we will see how many more modifications were added later to the MOSFET technology, so that we find the modern day submicron MOSFET.

To begin with, the MOSFET that was first made for integrated circuit, it was a PMOS. That is the source and drain are boron doped, p-type; substrate is n-type, source and drain are p doped. So, the steps for this metal gate MOSFET was something like this. (Refer Slide Time: 3:30)



First of all you start with an n-type substrate. In those days people still used 1 1 1, because you know, 1 1 1 was the cheapest substrate material. It is easiest to grow silicon in 1 1 1 direction, so people still used 1 1 1 substrate and you take n-type substrate; you grow a thick field oxide. So, that was the first step; you grow a thick field oxide. Then, you have your first masking step, first lithographic step. This thick field oxide is grown all over; you do not need any mask for that.

(Refer Slide Time: 4:37)



After this thick field oxide is grown, you open windows in this oxide to diffuse the source and the drain, source and drain diffusion; boron diffusion for source and drain. In the next step, you again grow or deposit another thick oxide. Either you grow thick oxide all over again, in which case oxide will predominately grow over the source and drain where no oxide is there; oxide growth will be smaller on the regions where already a thick oxide exists, right, because you know oxidation follows a linear parabolic law. Otherwise you could of course, deposit. If you deposit, it will be deposited uniformly all over the place.

(Refer Slide Time: 5:45)



So, let us say, we have a thicker oxide. That is step number 3, thick oxide, let us say, grown all over the place or you can even use a deposition. Next step is the second lithographic step, the second masking step that is for the gate and for the contact. So, you open something here, something here and here.

## (Refer Slide Time: 6:31)



Notice that windows are opened over source, over drain and over the gate region and notice one more thing. The gate region must connect between the source and the drain, right.

(Refer Slide Time: 7:42)



When the gate oxidation is done, the gate oxide must extend all the way from source to drain, starting from the source region to the drain region, right. Otherwise the channel

will not be continuous. So, it is important in this metal gate technology that the gate actually overlaps the source and drain region. The gate must overlap the source and drain region, in order for the channel to be continuous. So, that is your next step. You grow a thin oxide over the gate. When you are doing that you also have a thin oxide grown over these contact regions, right. So, that is your step number 5, gate oxidation. After that, you have a third mask for the contact windows. So, you have contact windows opened again.





So, now you have contact metal deposited. This is for source, this is for gate and this is for drain. So, this is your metal gate p MOSFET started with an n-type substrate. These are p plus regions for source and drain. This is the gate, aluminum metal was used. If you note the interesting, the salient features of this metal gate MOSFET fabrication, you will notice that the gate oxidation and metallization are the final steps. Towards the end of the process flow, we do the gate oxidation and then follow it up with the metal deposition. This is because, once aluminum is deposited on silicon, you cannot subject it to any high temperature processing. That is why gate oxidation always has to be done towards the end of the process flow.

Number 2 is which I have already mentioned; since you are doing gate oxidation after the source and drain diffusion, you must ensure that the gate region extends all the way from source and drain. You can only ensure it by allowing for the overlap over the source and drain region. So, one feature of the metal gate MOSFET technology is that the gate always overhangs the source and drain region. Otherwise, the channel will not be continuous. The purpose of this thick oxide is to make sure that the parasitic MOSFET does not get turned ON. What do I mean by the parasitic MOSFET? Let me extend this.

(Refer Slide Time: 12:04)



I have another device here, right. That is also, let us say, that is the source of the next MOSFET. So, you see I have the source, I have the drain of two different MOSFETs and in between I have a thick field oxide. So, this is the parasitic MOSFET. This is your active MOSFET with the thin gate oxide and this is the parasitic MOSFET in which you have between the two source and drain regions, a thick oxide. Now, this oxide must be thick, so that the turn ON voltage, the threshold voltage of this parasitic MOSFET is very large. You know, V t the expression for threshold voltage it is, we could say it is almost, almost inversely proportional to the oxide capacitance C ox, the gate oxide capacitance. So, that means thicker the gate oxide, smaller is the capacitance, right; thicker the oxide,

smaller is the capacitance and since threshold voltage is inversely proportional to C oxide, thicker the oxide, larger is the threshold voltage.

So, you see, between this active transistor and the parasitic transistor, the threshold voltage of the parasitic transistor will be much larger almost by the factor, the ratio of the two thicknesses. For example, I have some data here.

(Refer Slide Time: 14:11)

If you use the gate oxide thickness, obviously this is an older transistor, gate oxide thickness is 1200 Armstrong and the field oxide thickness is about more than 10 times that. So, in this case, the threshold voltage for this p MOSFET was about minus 4 volt and in this case, the threshold voltage of the parasitic MOSFET was about minus 37 volts. So, you see if you scale the two oxide thicknesses by a factor of 10, slightly more than factor of 10, the threshold voltage also scales by almost a factor of 10. So, that is the purpose of having this thick field oxide, so that inadvertently this parasitic MOSFET does not get turned ON. So, you can see there are some basic advantages in MOSFET technology over bipolar junction transistor technology.

First of all, the steps are very simple. I do not need any epitaxy. I do not need to have an epitaxial wafer. In case of bipolar junction transistor, we have seen that you have to have an n epitaxial region. Here in MOSFET technology, the basic MOSFET technology, you do not need the epitaxial layer. You also do not need any isolation. All you have to have is this thick field oxide; you do not have to bother, so much about the isolation. As in bipolar junction transistor, you see, you do not really have to bother so much about the isolation. So, these are the basic advantages of the MOSFET technology and obviously, the packing density of a MOSFET is going to be much, much higher, right and the steps are also very simple, much simpler compared to the bipolar junction transistor; you really do not have to have double diffusion.

Like in bipolar junction transistor, you see, you have an epitaxial layer, you have to have a buried layer diffusion, you have to have a base diffusion and then an emitter diffusion. Her you do not need; you basically need one diffusion and one oxidation process, two oxidation processes - one for the field oxide, one for the gate oxide and one diffusion for the source and drain; that is it, very simple. So, these are some of the inherent advantages of MOSFET. But, this could be a little bit misleading; in the sense, it is more difficult to get a MOSFET working than a bipolar junction transistor, even though the process steps are very simple. That is simply because, a MOSFET is a surface device and everything depends on the oxide semiconductor interface and that is why even though the theory of MOSFET was discovered prior to the theory of bipolar junction transistor, for the first working MOSFET to be available, it took much more time. That is because the technology of oxidation had to be perfected. The semiconductor oxide interface had to be properly understood, in order to get the MOSFET working.

Now, for some of the problems in this early MOSFET technology; you will notice that I said that the first integrated circuit MOSFET that was built was a p MOS. But, why p MOS? One should always automatically go for an n MOS technology, isn't it, because you know, MOSFETs are unipolar devices. In n MOS, carriers are electrons. They will be much faster, right. Still, why did people fabricate p MOS? That is because they could not get an n MOS, which answer to the specification. You see, in a p MOS we have a

negative gate voltage. So, you see, you have a negative threshold voltage which works perfectly. When you apply the gate voltage, only then the MOS device is getting turned ON; fine, no problem.

By the same token, in an n MOS, I apply a positive gate voltage. So, for an enhancement mode device that is a normally OFF device, which should be OFF when no gate voltage is applied that is when zero gate voltage is applied, it should be OFF. For such a device, we must have a positive threshold voltage, right. It was very difficult to obtain an n MOSFET with positive threshold voltage, because of the presence of fixed oxide charges.



(Refer Slide Time: 19:35)

You see, if you have fixed oxide charge, the threshold voltage gets shifted by an amount of minus Q f by C ox. Unfortunately in silicon, almost always the fixed oxide charge is positive; it is always positive. So, your threshold voltage gets shifted by that amount in the negative direction. That is why all the early n MOSFETs, they had negative threshold voltage and if you have a negative threshold voltage, your device will become a depletion mode device, which is not very desirable for various circuit applications, because when you have not applied any gate voltage, you do not want the device to idle power. If it is a depletion mode device, even when the gate voltage is zero, the device is still conducting, right. So, it is actually idling power. Also, what happens is that you design the device to be an enhancement mode device and then, because of this Q f, you have actually a depletion mode device; all your circuit operation goes haywire. That is why it was very difficult to have the early n MOSFETs working in enhancement mode.

Then, people figured out that there is one way we could make the threshold voltage become positive and that is by increasing the substrate doping. Now, increasing the substrate doping that has various detrimental, various adverse effects, such as for example, if the substrate doping concentration is enhanced, is increased, then the mobility of the carriers in the channel will be reduced. So, the one major effect of substrate doping increased is mu degradation. Secondly, the higher the substrate doping, the more prone the device is to substrate bias effect, body bias effect. So, this is another problem. The other problem is of course, the capacitance of the drain to body, drain to substrate; drain to substrate capacitance will increase significantly and that is another detrimental effect. So, even though it will give you an enhancement mode device, the performance of this device will not be very good, will not be very optimum.

Then, technology came to the rescue of all integrated circuit, in the sense that ion implantation was used in order to tailor the threshold voltage. The concept was you do not change the entire substrate doping concentration.

(Refer Slide Time: 23:17)



What you do is simply you have the same p-type substrate that you wanted to have and you simply have a very thin layer, tailor this. If you have a very shallow boron implantation, if you have a very shallow boron implantation, you know that boron is an acceptor type of dopant, right. So, when these dopants are ionized, it will give you negative charge sitting at the interface, right. They are acceptors; they have accepted electrons, so they are negatively charged in the ionized state, right. So, effectively what do you have? You have a thin sheet of negative charge at the interface. What is its effect? Exactly in the direction opposite to that of Q f. In case of Q f, what did you have? You had a thin sheet of positive charge sitting at the semiconductor oxide interface.

Now, you are combating that effect by putting a thin sheet of negative charge at the interface.

(Refer Slide Time: 24:50)



So, effectively you are allowing the threshold voltage to get shifted by an amount Q B by C ox without actually affecting the substrate threshold voltage, sorry, the substrate doping concentration.

(Refer Slide Time: 25:09)



This thin sheet of charge is located very close to the surface, so that it does not really affect the depletion layer edge; the depletion layer edge is still not affected. So, it does

not really affect your body bias or even the mu degradation at the channel. None of these effects are felt. The only effect is to shift the threshold voltage back, shift it in the positive direction by an amount of the charge you put in divided by C ox. So, you see, by ion implantation, you can precisely monitor how much boron you are going to put close to the surface. So, by adjusting the dose and making the energy of implantation very low, you can have a very shallow implant very close to the surface, which is called the threshold tailoring implant, threshold tailoring implant, in order to modify the doping concentration, modify the threshold voltage. So, one major problem of n MOS technology was sorted out by ion implantation technique that is by using threshold tailoring implant.

Then, as the device dimensions became smaller and smaller, the other problem began to be dominant. What is that other problem? Remember, in the first metal gate technology, we had to have the gate region overhanging the source and drain region and that started to give a lot of problems. Number 1 problem is of course, area requirement. So, that means you have to have an allowance, right; you have to have the gate region overlapping the source and drain.



(Refer Slide Time: 27:16)

So, particularly when the device dimensions are becoming smaller, when the length of this region itself is becoming smaller, you see, in order to ensure that there is an overlap you cannot really cut down the overlap dimensions much. That depends on your photolithographic limitations. Suppose your photolithography has a resolution of 1 micron, then in order to ensure that the gate region overhangs the source and drain, the overhang must be 1 micron on each side, right, in order to ensure that the gate region is extending all the way from source to drain. I cannot go anywhere below the minimum resolution. So, now the overhang region itself becomes comparable to the channel length. If your channel is, in those days the channel was, let us say, close to 1 micron, 2 micron or something like that. So, you want the 2 micron channel length, but you have to have 2 plus 1 plus 1, 4 micron, right.

The other problem is of course more important and that is the overhang capacitance. So, this overlap will create a problem both .... As far as area is concerned, as you go to smaller and smaller dimension devices, you want to save more and more area; this is proving to be the bottle neck. The second thing is that gate to source or gate to drain overlap capacitance. So, we needed a technology where the gate will be automatically aligned to the source and drain. One way to achieve this is if we first define the gate region, so we have thick field oxide everywhere. We open an active area; then, we define the gate region. So, whatever is not protected either by the gate or by the thick field oxide is subjected to source and drain diffusion. So, automatically source and drain gets aligned to the gate. Do you understand what I am trying to say?

Suppose I have a scheme like this; you look at the top view of the masks.

(Refer Slide Time: 30:13)



First of all you have thick field oxide everywhere and then you open an active transistor region in this. Rest is all protected by oxide. The hashed region is all thick field oxide and I have opened a window in that thick field oxide. Now, let me define my gate, have a thin oxidation done, protect it with the gate metal and then pattern the gate metal, so that I have ... This is my gate. Now, look at the structure. What do I have? I have, everywhere else I have thick field oxide, only over the gate I have the gate oxide protected by the gate metal. What are the only openings here? This and this. Let me subject it to diffusion. I will have source and I will have drain and the gate will extend all the way from source to drain, right.

Let us compare it with the older, the previous technology that we were talking about. In that what did we have?

(Refer Slide Time: 32:02)



We had, the first step we had thick field oxide. Then we had opened the source and drain window in that. So, the first step was opening the source and drain window and then, after that we had to define the gate. So, what we had to do is to make sure that the gate actually overlaps the source and drain region. Compare it with this.



(Refer Slide Time: 32:35)

Here because the gate is defined first, the source and drain are automatically aligned to the gate. Then why could not we do it for this technology? For the very simple reason that we were using aluminum gate and once you put aluminum, you cannot subject the semiconductor to any high temperature processing. Whether you do diffusion or you do ion implantation followed by annealing for source and drain regions, you need high temperature processing. So, if you use aluminum gate, you cannot have source and drain done after the gate pattern. That was the reason why here we had to do the source and drain first and then follow it up with the gate.

So, the idea is, if I want to have a self-aligned gate to source and drain, if I want to have a self-aligned technology, I can no longer use a metal gate. I must use a material which can withstand high temperature and at the same time have low resistivity. Polysilicon, doped polysilicon answered to the description. So, from metal gate n MOS process we came to self-aligned polysilicon gate n MOS process. So, in self-aligned polysilicon gate n MOS process, it is very simple; I have already actually outlined the steps, let me do it once again.



(Refer Slide Time: 34:23)

You have the thick field oxide and then you open the window in that thick field oxide. Then, you grow a thin gate oxide all over the place and then you have polysilicon deposited all over and then you pattern it. That is you retain it over certain regions and remove it from the rest of the portions.



(Refer Slide Time: 35:05)

So, your gate region is defined. So, you see, you have certain regions protected by the thick field oxide and the gate region protected by the polysilicon and now you can subject it to high temperature processing and you can have source and drain which will be automatically aligned to the channel, right. So, this is the self-aligned poly gate n MOS technology.

However, there is one small problem in this self-aligned poly gate n MOS technology and what is that problem? That is the problem of topography or the problem of surface planarity. I have already told you that this thick field oxide is needed, so that the parasitic MOSFET does not get turned ON and in order to keep the threshold voltage of the parasitic MOSFET sufficiently high, the field oxide thickness must be high. But, if I have a very thick field oxide that means the surface is going to be non-planar; surface is going to be all ups and downs. So, in order to maintain the planarity of the surface layer it was,

the technology was further modified and instead of growing the thick field oxide all over, it was decided to have local oxidation technique, LOCOS.

Let me outline the LOCOS steps once again. In LOCOS, what do we do? We protect certain regions by first having a thin pad oxide and topping it up with a thick silicon nitride layer. So, what we have is something like this.



(Refer Slide Time: 37:30)

You have first a thin pad oxide and then top it up with a silicon nitride layer, about 2000 Angstrom of silicon nitride and about 200 Angstrom of pad oxide. So, the silicon nitride layer thickness is 10 times that of the pad oxide. Now, on top of this you deposit another layer of silicon dioxide. So, let us have another layer of silicon dioxide and then let us have positive photoresist. So, you have a composite layer - positive photoresist on top, underneath that CVD silicon dioxide, underneath that silicon nitride and then finally a thin pad oxide. Now, I will have to pattern it. So, I will pattern the positive photoresist first. (Refer Slide Time: 39:05)



I have removed positive photoresist from certain region. Next, I remove CVD silicon dioxide from this region. To remove silicon dioxide you use buffered HF and then in order to remove silicon nitride, you use orthophosphoric acid. So, actually that is the reason why you need to put the CVD silicon dioxide, you see. If you just have silicon nitride and then protected it with positive photoresist or try to protect it with positive photoresist, positive photoresist cannot withstand orthophosphoric acid. So, in order to protect this silicon nitride layer, I must have a silicon dioxide on top. Silicon dioxide will withstand positive photoresist, silicon dioxide will withstand orthophosphoric acid; of course, this top layer will get destroyed when you are putting it in orthophosphoric acid, fine. So, you see, after this, all you have to do is to remove the thin pad oxide from here and here. So, this is what you have.

Now, you see, suppose now I carry out oxidation. You know, oxidation cannot progress underneath silicon nitride.

(Refer Slide Time: 41:11)



So, there will be no oxidation here, but oxidation will have, will take place here and when oxidation is taking place here, you know, approximately half of it will grow above the surface and half it will grow below the surface. So, we will have something like this. So, even if you have a very thick field oxide, you are loosing surface planarity only by half the amount. That is if even I want to grow say, 10,000 Angstrom, even if I want to grow say, 1 micron thick field oxide, on the surface only 5500 will project out.

(Refer Slide Time: 42:01)

I can of course, if I want to make the surface more planar, I can of course, first carry out an etching step and then follow it up with oxidation, right. I can first carry out an etching step and then follow it up with oxidation, so that the oxide comes exactly up to the surface, right. So, this is LOCOS, fully recessed; fully recessed LOCOS. Fully recessed, what does it mean? That is after the oxidation, the top oxide surface coincides with the surface of the semiconductor. It does not protrude anywhere above the surface. You have a perfectly planar surface. But, every technology will have its drawback. LOCOS is also not entirely free of this drawback.

The problem is simply this. We said that oxidation does not proceed underneath the silicon nitride; quite right, but there is a slight protrusion underneath. There is a slight movement inside, so that actually what you have is not exactly like this.



(Refer Slide Time: 43:21)

What you have is a little bit of projection inside. You see, this problem of the oxide penetrating a little bit inside is more if you have the thin pad oxide. You see, there the silicon surface is actually in contact with the pad oxide and the oxidation can actually, at this corner actually it can proceed slightly underneath, underneath the pad oxide. But at the same time, we need to have the pad oxide, because if you put silicon nitride directly on silicon, it will give rise to a lot of stress. There will be stresses and there can be dislocations and mind you, that is going to be very serious, because this region is going to house your active transistor.

You cannot afford to have that region damaged, right. So, you must ensure that when you put silicon nitride on this region, this region, the electronic property of this region is not affected at all. That is why you are putting the pad oxide. But, if you put the pad oxide, then the problem of oxide projecting slightly inside is more and the problem is going to be much more severe when you have a recessed oxide that is when you first etch the silicon and then you want to have the oxide.



(Refer Slide Time: 45:18)

That is because at this point, you have done the etching. Let us look here. You have done etching. So, not only this surface of silicon, but this surface of silicon is also available for oxidation, right. So, obviously now it is going to project more; actually you are going to have something like this. Because oxidation is also taking place on the side wall, so there also it is going to have some oxide underneath that surface and some oxide above that surface. So, what you are going to have is a shape like this.

(Refer Slide Time: 46:12)



This portion is called the bird's beak, because it looks like bird's beak and you have a projection here. This is called the bird's head. In different books it is referred to differently, like bird's head or bird's crest or even birds nest, as if a bird is nesting under the protected region.

In order to, even you see, when I discussed the LOCOS technique in bipolar junction transistor, I said that if you have recessed oxide, you are going to have this bird's beak problem. However, in case of bipolar junction transistor, it was not very serious; because you are only using these regions for isolation, so it was not very serious. But here, it is going to be extremely serious, because you are going to have an encroachment in the active region itself. So, a lot of your active region will get eaten up by this bird's beak and bird's head problem. Particularly as the device dimensions are becoming smaller, this problem will become more and more serious.

In case of bipolar junction transistor you see, the device is actually vertical; it is a bulk device. So, nothing really gets affected; neither the base width gets affected, consequently the beta of the transistor does not get affected. Nothing really gets affected.

But here, if the channel length gets affected, that is a very serious problem. So, particularly for MOSFET technology, the bird's beak problem is a very serious problem and one tries to prevent it by various ways.

One way is called the side wall masked insulation; side wall masked isolation or side wall masked insulation. The acronym is SWAMI.



(Refer Slide Time: 48:31)

You understand that this protrusion is taking place at the side wall of the etched region of silicon. So, if somehow we can protect the side walls also by silicon nitride, then the amount of protrusion will become less. That is the principle behind the side wall masked isolation. So, in that what do we do is very simple.

(Refer Slide Time: 49:23)



You have the pad oxide and you have the silicon nitride on top. Now, you carry out the etching to the desired depth. Now, you do another, a second silicon nitride deposition. So, you deposit it all over the place like this and then remove it by doing an anisotropic etching.



(Refer Slide Time: 50:29)

You have just removed half of the silicon nitride. That is only the second layer of silicon nitride you have removed, but now you are using anisotropic etching, so that the silicon nitride is retained on the side walls and only removed from the horizontal surface.



(Refer Slide Time: 51:09)

So, now when you subject it to oxidation, even the side walls now are protected by silicon nitride. So, it cannot really penetrate; slight penetration here, but nothing much. After that you remove the silicon nitride. So, what you have is something like this.

(Refer Slide Time: 51:54)



Now, you do another CVD deposition of silicon dioxide, so that these regions are all filled up and then you can planarize.



(Refer Slide Time: 52:11)

You can just partially etch the silicon dioxide layer, so that you have ...

(Refer Slide Time: 52:26)



Nothing has penetrated in the channel region. So, this is the side wall masked isolation. The basic principle is that the protrusion of the oxide in the active region is occurring because of the oxidation of the side wall. So, we protect the side wall by putting another, a second layer of silicon nitride and anisotropically etching it, so as to retain the silicon nitride on the active region as well as on the side walls and then, you follow it up with the oxidation, so that the oxide does not penetrate in the active region. Then, you remove the silicon nitride and whatever recesses you have you fill it up with CVD silicon dioxide and then planarize it by partially etching. So, this is one possibility.

The other possibility which is called a silo technique that is actually an acronym for sealed interface local oxidation, in that case what you do is the basic concept is like this. Because of the pad oxide, you have more projection inside. At the same time, the pad oxide is needed to relieve the stress at the silicon nitride interface. So, what you have is a composite layer. You first put a silicon nitride layer, a very thin silicon nitride layer, so that the stress is not too much. Then, you put the pad oxide and then you put the silicon nitride. So, what you have essentially is something like this.

(Refer Slide Time: 54:20)



First step you have thin silicon nitride, about 100 to 150 Angstrom of silicon nitride. Then you put a slightly thicker, about 200 Angstrom of silicon dioxide and then you put your regular 1000 or 1500 Angstrom of silicon nitride. So, this and this is silicon nitride and this is silicon dioxide. So, the stress is still being relieved by the pad oxide layer. But, the pad is not in contact with silicon. At the interface, we actually have a very thin silicon nitride layer, which will not cause dislocation in the active region, but at the same time, it will prevent the oxidation from projecting underneath. So, this is called silo - sealed interface local oxidation.

(Refer Slide Time: 55:29)

Sealed Interface Local Oxidation

We will stop here today.