Introduction to Semiconductor Devices
Prof. Dr. Naresh Kumar Emani
Indian Institute of Technology, Hyderabad

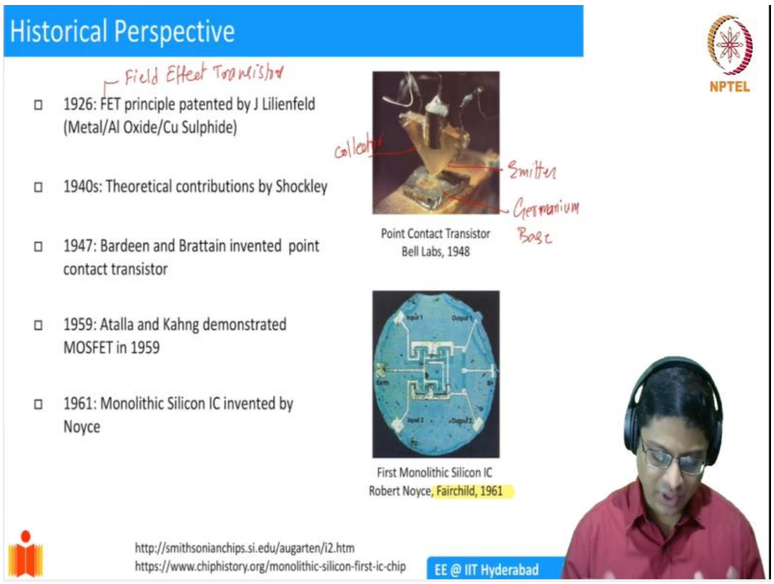
Lecture – 57
Introduction to MOSFET

This document is intended to accompany the lecture videos of the course "Introduction to Semiconductor Devices" offered by Dr. Naresh Emani on the NPTEL platform. It has been our effort to remove ambiguities and make the document readable. However, there may be some inadvertent errors. The reader is advised to refer to the original lecture video if he/she needs any clarification.

Hello everyone, welcome back to Introduction to semiconductor devices. So, far in this course, we have studied some fundamentals, fundamental aspects of semiconductor devices, we understood how a P-N junction works and then we understood how a mask capacitor works. All of that lays foundation to understanding one of the most important semiconductor devices which is a MOSFET.

It is not possible for us to really imagine a world without a MOSFET. A MOSFET has revolutionized semiconductors and it has paved the way for the modern technological revolution. It is not an exaggeration to say that, if there was no MOSFET possibly there is no semiconductor device right now. But there is no the information technology revolution would not have happened. It is that crucial.

(Refer Slide Time: 01:05)



So, today we will begin our journey into understanding how a MOSFET works. So, before we do that, let me start with a brief historical perspective. So, back in the early 19th century, telephone was invented and then the telephone systems were being laid all across America and Europe mainly. And they were using vacuum tubes in the telephone systems to connect calls and electrotype automate.

Some of that and those vacuum tubes were not very efficient and people realize that there needs to be a solid-state alternative to the vacuum tube. It was quite well known, and in the year 1926 a scientist by named Julius Lilienfeld had proposed the concept of a FET, which essentially stands for Field Effect Transistor and this was in 1926, and he proposed that you could use a combination of metal aluminium oxide and then cadmium sulphide.

Cadmium sulphide is a semiconductor. So, if you use this combination, you can demonstrate field effect is what he said. But it turns out that this is 1926 and it was not possible to exploratory demonstrated he patented the idea but exploratory did not work. And almost 20 years went by without much progress. In the 1940's AT and T Bell Labs was one of the first foremost research organizations in the world.

And Shockley was part of Bell Labs and he had developed a lot of theoretical concepts necessary for understanding semiconductor devices. And he was also independently I think, proposed the concept of an equity interest trying to demonstrate that. So, the imperative to actually discover a solid-state transistor was actually much stronger by 1940s. So, there was searching for this and then there were many failed attempts.

And Bardeen in Brattain also were contributing to this effort, and Bradeen I believe in 1945 proposed that there are the surface states in the semiconductor which are very important. And that helped them in their theoretical understanding and finally in 1947, Bardeen and Brattain invented what is known as the point contact transistor. I showed this image I believe in the introduction, but it was a replica.

This is a 1 more replica of the point contact transistor that was invented in Bell Labs and if you look at the system very good device. So, it had a piece of plastic, a triangular shaped wedged which was coated with gold. And then at the tip of the you know, the stretch there is a cut, the tip there was a cut here, and then essentially one side of it was like working like a emitter of carriers the other side was working as a collector.

And this wedge of plastic coated with gold was pushed on a piece of germanium. This has germanium, and then literally it was a base of the device. So, it was pushed against this base consisting of germanium by a spring and they managed to show transistor action in such a device. It does a very, very good device, it was not at all reproducible, because you shake it, the spring moves and then the other structure is lost and so on.

It is not a very reliable thing. But this was the first device that showed transistor action, that is, if you can control the current which is flowing into the base, you could, which is usually a very small amount of current. You could control a much larger conductor or emitter current that was a principal. And they could show amplification. So, this is one of the most significant inventions in the history of transistor.

This is in 1947 December, and in early 1948 or so and Shockley, Bardeen and Brattain came up with the concept of junction transistors, bipolar junction transistors which are known as BJT is now. And after about 10 years you know they were actually searching or they were actually trying to make a field effect transistor, but they could not make it work. And in the process, they happen to discover a bipolar junction transistor.

And the challenge as I said was actually the interface between silicon and silicon dioxide. Whenever you have this interface between semiconductor and insulator that interface has to be extremely good. So, back in Lilienfeld concept the interface between cadmium sulphide and oxide was not at all but it does have a lot of defects and it did not work. And even whatever Bardeen and Brattain tried also did not work.

The most, the breakthrough came when in 1957, also a scientist by name Atalla, he was working on diffusion of dopants through a mask, silicon dioxide masks and they noticed that if you have an interface between silicon and silicon dioxide it had a very good quality interface. And that led you know immediately to the idea of a MOSFET which is essentially metal oxide and semiconductor.

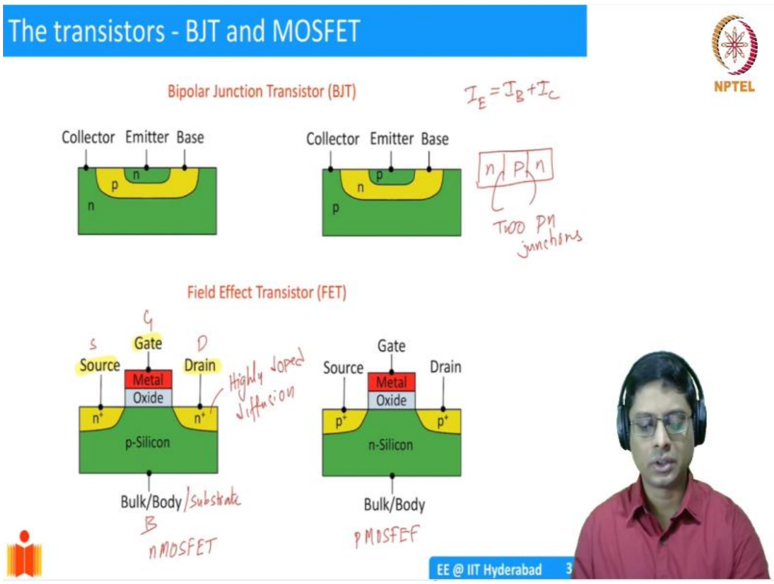
So, the MOSFET was demonstrated by at Atalla and Kahng in 1959. So, about 10 years after the invention of a point contact transistor. And immediately in a few years, Robert Noyce who was actually one of the founding members of Intel, he proposed and demonstrated a concept known as an integrated circuit. So, he integrated various MOSFETs on a single piece of silicon and this is one of the first IC's that was demonstrated.

It was demonstrated at by Fairchild semiconductor where Robert Noyce was working at that point of time later, he split up in along with Gordon Moore and others, they founded Intel which makes all our processor even today. So, you see here is you know, this idea of using a single point contact transistor having these elements is a very crude one but quickly this idea of integrating these things on a single chip.

So, you see that you know, there are multiple inputs that are these transistors you will understand slowly in the next couple of weeks. Why these you know what these things mean? We will talk about that later. But we had fabricated one of the first integrated circuits and this was a very significant breakthrough in terms of science. And thereafter it is essentially you know, increasing the scale of these devices or making them smaller and smaller.

And by the way, the bipolar junction transistors were not predominant in 1960s but by 1970s MOSFETs became the dominant technology world over. And today you see that MOSFETs are the most commonly used device conductors.

(Refer Slide Time: 08:02)



So, what are these transistors? When we can even think of junction FETs and things like that, but I will just broadly talk about two classes which are bipolar junction transistors a BJT and a MOSFET metal oxide semiconductor field effect transistor. So, what is a bipolar junction transistor? Well, essentially it is a system of 2 pn junctions you could think of it like this. So, let us imagine that you have a n - type semiconductor.

I will make a p type diffusion region in that we talked about it when we talked about pn junction, making a pn junction you will diffuse it. And after that, I will again diffuse another intent. So, I will have 3 areas. Corresponding to n p and n, that is one option or you could also have pn and p, these 2 of combinations are possible. And you know schematically you could show this as simply a piece of silicon consisting of three regions n p and n, or you could have p and n p.

So, three blocks. So, essentially you have two junctions pn junctions now. We understood how single pn junction works and there is a BJT is essentially simply to pn injunctions and you have again depletion and whatever we talked about. Forward bias live advice all that works in this. And it turns out that one of them we call us emitter one of them we call us base and one of the call as collector.

So, we can study this in greater detail, but in this course, I have decided not to really focus on BJTs. The reason for that is the BJTs have been dominant in 1960s. After that they are not that much used MOSFETs are what are more commonly used now. And I felt that you know, I would spend time discussing MOSFET in greater detail than actually talk about BJTs. I believe BJTs are part of the gate syllabus in AC or you know, here also I believe to some extent.

But you know, having understood the concepts of pn junctions I am sure you can follow the concepts of BJTs by looking at some other material. I can also post some references for that but even you know theory has it and even has it everybody has some chapters on BJTs you could read up a little bit about that. Basic ideas are that you know in BJT you could basically have three currents and just giving you a glimpse of it later I will not talk about it.

So, you have emitter current and that will be equal to base current less collector current. This is how it is emitter base collector. So, usually the base current is a very small current and by controlling the base current we could control the emitter current that is how you get the amplification. There is something called as beta and alpha you can look up all that it does not take too much time, but I will not cover that in this course.

So, what we will talk about in depth is a MOSFET, which is basically metal oxide semiconductor which we already discussed the MOSFET system. In you know when discussing the MOS cap and what we will do is, a MOSFET consists of a MOS capacitor plus 2 additional regions. This we have already seen metal oxide semiconductor, but now we are adding n type diffusion region and another attempt to enter diffusion regions.

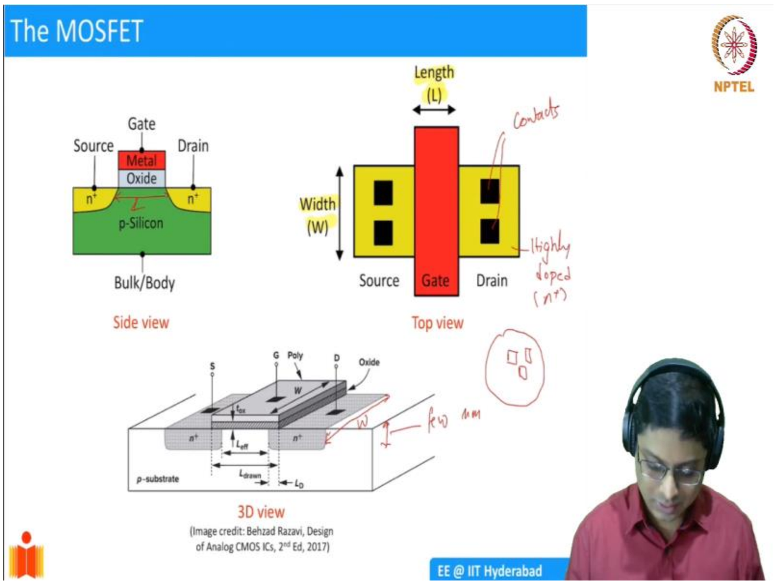
These are highly doped high end implicit diffusion region. We call them a source and every one of them is called source one of them is called a drain. There is no real, physically there is no difference between the source and the drain; it is just a matter of what voltage we are applying to it. That is where the connection comes about. But you know in terms of doping, typically they are symmetric.

You can interchange the brain and the source and nothing much happens in the transistor will still work. And then the gate anyway we have already seen and we will also introduce another term. In the MOS capacitor discussion, we did not mention it specifically we just call it substrate now; you could call it bulk or body or even substrate. So, typically people do not use substrate because you know, we tend to use source to S you know.

If you put substrate again that becomes less and it is confusing. So, we use bulk or body typical. So, this we call as you know, this D terminal drain source gate and bulk. And we apply various voltages to these four terminals and that leads to some effects. So, this configuration is called as n MOSFET. You know, by now that you know n MOS is when you have a p type substrate and you have an inversion layer in.

That is why it is the n MOSFET and you can have a complimentary device which is a P MOSFET which is consisting of n type substrate and p + drain n source, this diffusion means that p + so that is it. So, basically, we have the MOS capacitor plus two additional terminals. So, how does this help us? Well, to understand that let us you know look at the structure in a couple of ways.

(Refer Slide Time: 13:21)



We have already seen the side view and you know, if you take a cross section of a mosfet, this is how it looks. We could also have a top view, if you look at it from the top how does it look like you know. So, it will be you know, you have a gate which is having a certain length. So, this is important terminology. So, the length of the gate is essentially the distance between the source and the drain. So, this is my length of the transistor.

And this was only you know, in two dimensions, but if you extend it into the volume of the semiconductor into the paper, it is a 3-dimensional structure, and that 3-dimensional structure is showing you that you have this region again here, this is the width of the transistor. From the top view, it looked like this. So, what you have is a wafer of silicon, you have source and the drain if you diffusion regions here which are shown in yellow here.

These are highly doped diffusions. If it is a n MOS it will be n +, if it is a p MOS it will be p +, the diffusion region and then they are separated I mean it is not really separated in the same plane. You know it is out of the metal is actually sorry, the gate is out of the plane. But when you look at it from the top, it looks like they are separate. The gate is separating the source and the drain but it is only from the top, physically we see that there is this oxide in between.

So, this is a top view and you also have a 3d view which essentially shows you the same things. So, you can have a length of the transistor, the width of the transistor which is again another important parameter. So, on a piece of semiconductor I can make many of these you know devices for example, if I take a wafer, I will take you know one MOSFET second MOSFET and third MOSFET.

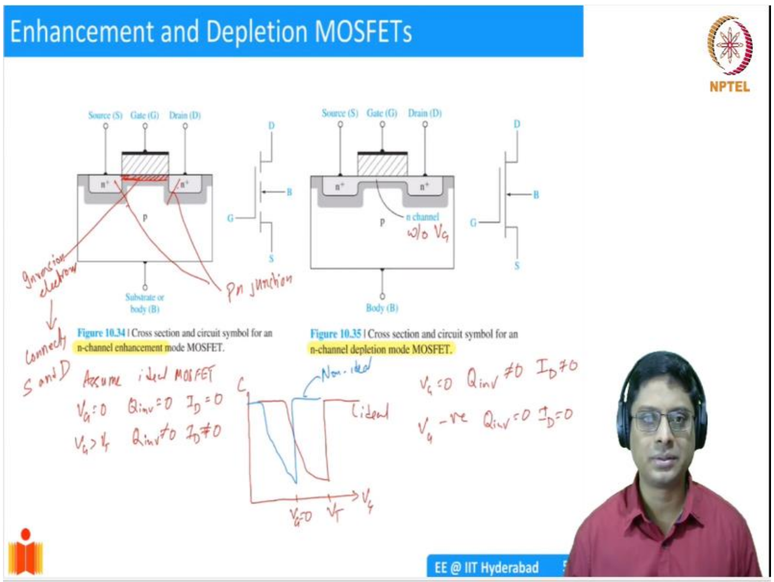
And we can connect them and that is an integrated circuitry that we talked about. So, we are already like if you take the most modern silicon chip which is let us say an Intel processor, if you take a Xeon processor, it will maybe it will be of size of you know one or one point by square inches. And it will have a few billion transistors. So, many, many of these devices are integrated and the challenge comes in how to integrate all of that.

So, and we already told you that the essential physics happens only a few micrometres here, this is few micrometres. Below that, it is all just silicon bulk silicon. So, this is how a transistor structure looks and this black thing here you see, these are what we call us contacts. The reason is if you look at semiconductor we talked about some metal and semiconductor junctions Schottky junctions are they can be rectifying contact or Ohmic contact we say.

So, it is not enough to have a piece of semiconductor but we have to connect it to the external world. We have to connect it relates it to another transistor how do we connect it if you directly put a piece of metal, it will have a contact resistance, we mentioned this already. So, now to avoid that we can engineer these contacts in such a way that I mean one contact would suffice but we just typically make more contacts to minimize the contact resistance.

So, and these things I can connect so this will be connected to metal another metal piece and then so on. So, that is how the transistors work. So, this is a physical structure but, how does it operate?

(Refer Slide Time: 17:01)



So, this is again another view of it. So, the same story that we have been talking about so you have a substrate, you have a bulk, you have a drain, we have a source and we have a gate. So, essentially what you have is in between these 2 regions here, there is a junction. This is a pn junction, this is another pn junction, these 2 are pn junctions. In between them you have a MOS capacitor you could think of it like that.

So, if it is an ideal MOS capacitor, let us say there are no means there is no band bending at all. There are no carriers, nothing is ideal MOS capacity. The moment you apply gate voltage what happens? If you apply gate voltage, let us say positive voltage in this case, you can create speed at substrate if you have no positive voltage, you will create an inversion layer. So, let us say we think we apply gate voltage greater than threshold then we have in an inversion layer.

So, that essentially forms here below this, just below the gate oxide, you will have an inversion layer. And what is the inversion layer consisting of? Well, this is electrons inversion electrons. So, what does it do when you have an integration layer like this just under the oxide? Well, essentially you are connecting source and the drain. So, current can flow from source to drain but the moment I let us say make the voltage 0 this inversion layer disappears.

And there is no current between source and drain. This is a fundamental you know, in a very simple way, this is how switching happens. When let us say I would assume ideal MOSFET. Basically, MOS cap ideal MOS cap FET structure. When you have such a scenario, if your V G = 0, I will say Q inversion = 0. That means, we could talk of drain current or source current both are will be equal here.

Because oxide is actually perfect oxide, there is no leaking leakage. So, typically source and the drain currents are equal. So, I will simply talk of drain current = 0. But when VG is greater than VT, Q inversion is not = 0, there is some substantial amount of inversion charge. So, ID basically is not 0 so current flows. This is what we call as an enhancement type MOSFET. Why do we say enhancement? Well, there is no inversion charge at 0 voltage.

But the moment you apply positive voltage, we are actually creating an inversion charge. So, in the low voltage state like VG = 0, it does not conduct current. But the moment you apply a voltage conducts current, so this is called as enhancement mode MOSFET. But if you remember, you do not need to have this as a scenario all the time because we already studied the C V.

We said, if you have a V G like this, and let us say this is C. If you take an ideal MOS capacitor, let us say the C V is something like this, I am sure the let us say and this was your V T. And let us say this is your V G = 0, 0 volt, so now V G = 0, the devices may be mildly depleted. But it does not conduct current, there is no inversion charge to conduct current. This is how a C V could look like.

But we also saw that there can be many things like metal work function difference, and then they can be also some fixed offset charges. Because of wage, you may not have this condition, but you might have a situation where your CV looks like this. You get the drift current and this is depletion, this is inversion. So, this is basically with some work function differences. Let us say it is not ideal and this was ideal scenario.

So, if you have some sort of shift in the V T now, because V T now has become close to 0. So, even if you do not apply any voltage, there will be an inversion layer. So, here, you have n channel without V G, or at V G = 0, you have a connection. So, you do not have to apply any voltage it simply conducts. But then, if you apply negative voltage to this, then you will push electrons away and then you will have the channel removed.

So, in this case, let us say if V G = 0, there is Q inversion is not = 0. So, I D is not = 0. It is conducting when there is no voltage, but then when V G is less there is let us say, a negative. I will just say that it has to be sufficiently negative then Q inversion is = 0, because it pushes away the electrons that depletes the substrate I D = 0. So, we have to deplete the environment to push away the innovation charge and depleted.

So, this is called as a depletion mode MOSFET. I would say this is mostly academic, you know, it is not really now we will typically deal with enhancement mode. We do not really worry too much about depletion mode. So, this is how the basic structure of a MOSFET looks. So, in the next video, I will talk about the various current conduction mechanisms or current conduction mode MOSFET. I will see in the next video. Thank you.