

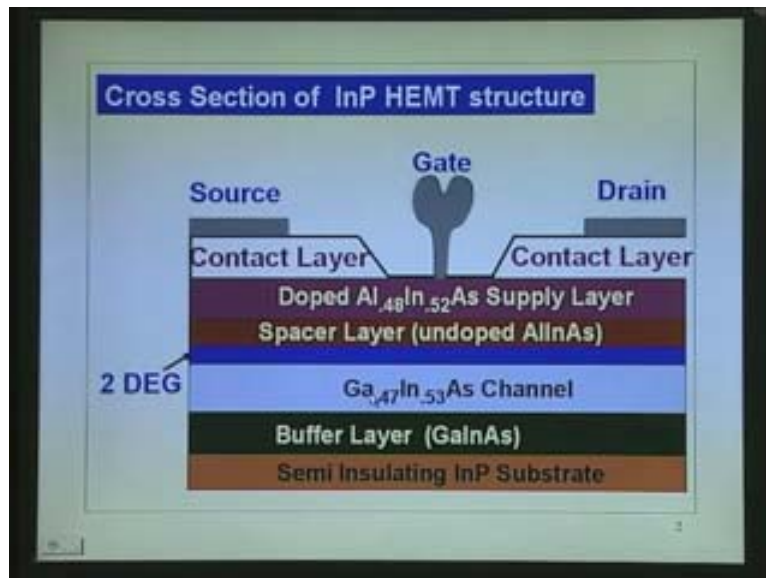
**High Speed Devices and Circuits**  
**Prof. K. N. Bhat**  
**Department of Electrical Engineering**  
**Indian Institute of Technology, Madras**

**Lecture 37**

**Pseudomorphic HEMT & Heterojunction Bipolar Transistors**

We have been discussing the high electron mobility transistor. We have discussed most of it – how it works, what are the structures that are there for gallium arsenide based devices, and also on indium phosphide based devices. Today, I will continue on that a little bit, on a structure known as pseudomorphic HEMT and then, we will go on to the topic on bipolar transistors.

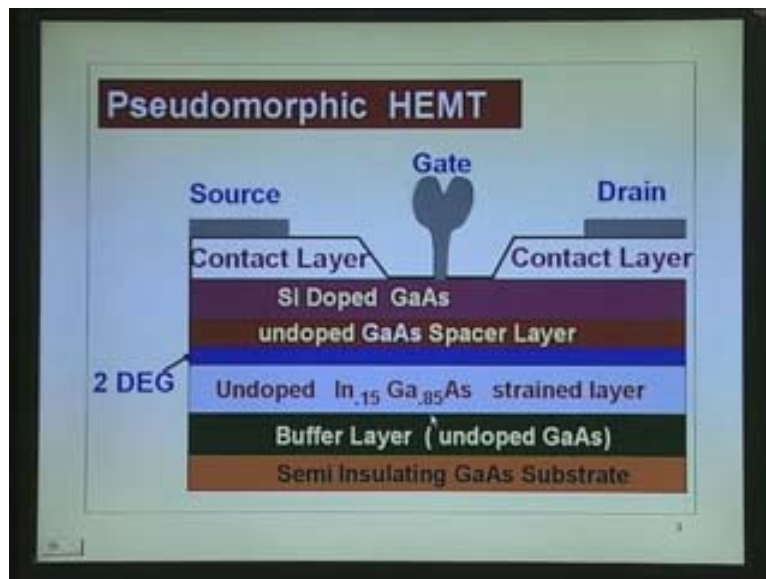
(Refer Slide Time: 01:40)



In fact, just to remind you, I am showing this structure again, that is, the indium phosphide based HEMT, where the structure is **common, similar**. On the top, you have got a heavily doped layer. In this case, it is aluminum indium arsenide, because it has lattice match with gallium indium arsenide, which has lattice match with indium phosphide. You have got a buffer layer, undoped layer, which is a channel layer, a spacer layer to improve the mobility so that **there is (Refer Slide Time: 02:18)** two-dimensional electron gas does not at all see the dopants, dopant layer. So, there is a spacer layer and then, aluminum indium arsenide.

We saw this; there are various versions because this type of structure (Refer Slide Time: 02:35) breakdown voltage. What they did was change this top layer with indium phosphide which has a problem with Schottky barrier contact, then another structure that was put... keep the top layer like this (Refer Slide Time: 02:46). Instead of gallium indium arsenide, put indium phosphide. Now, the problem there is lattice match is there, but the mobility of indium phosphide is low. So, you get high ON resistance. All these problems are there. In this structure, a slight problem is because of high aluminum content. The moment you have high aluminum content, you get of course good  $\Delta E_C$  here even compared to gallium arsenide based structures. What was thought of was modify this entire structure, go back to gallium arsenide based devices, but you would like to have gallium indium arsenide, because it gives better mobility than gallium arsenide.

(Refer Slide Time: 03:30)



This is another structure called the pseudomorphic HEMT. This is not indium phosphide based; this is gallium arsenide based. We have a semi-insulating gallium arsenide substrate – you can get it with very high purity. Still, you can put a buffer layer so that (Refer Slide Time: 03:48) defects do not get transported on to the undoped layer and this layer is indium gallium arsenide. You can see that you do not use that indium 0.47, but use indium 0.15, mostly gallium. It is gallium-rich indium gallium arsenide, but still, it has mobility better than that of gallium arsenide, because indium arsenide has got very high mobility compared to gallium arsenide.

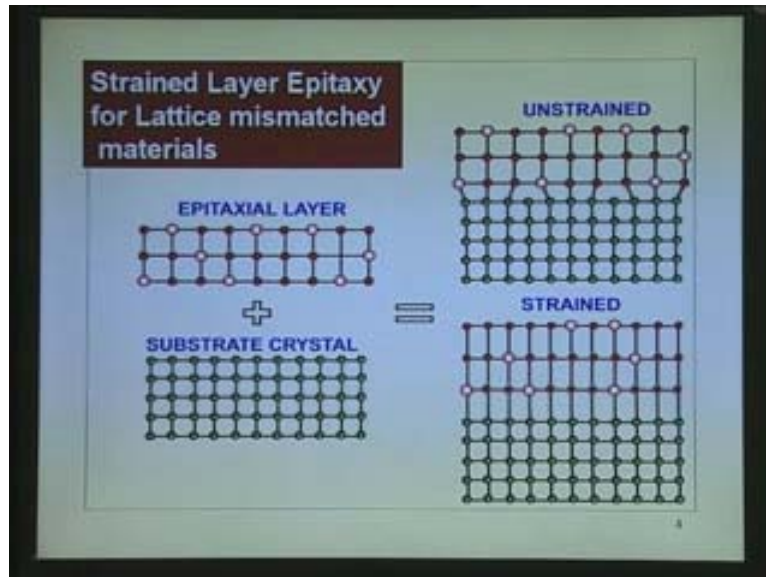
Indium gallium arsenide is somewhere in between, depending upon how much indium you put. You do not want to put fully indium, because the bandgap is low. Here, the bandgap will be close to 1 or so and gallium arsenide is 1.43. You have got now yourself a structure that has gallium arsenide, indium gallium arsenide with undoped gallium arsenide, silicon doped gallium arsenide.

In fact, you can replace this top layer with aluminum gallium arsenide also. You get better  $E_C$  if you like because you can get higher bandgap here. This top structure (Refer Slide Time: 04:48). Instead of gallium arsenide, put aluminum arsenide, AlGaAs. What is the problem with this? Right at the beginning, we were telling that you must have layers that are lattice matched to each other. If I put gallium arsenide layer as the channel layer and the doped layer AlGaAs has a very good match.

If I use indium phosphide substrate, gallium indium arsenide has a good match provided gallium is 0.47. Now, what we are talking of is indium gallium arsenide does not have lattice match with gallium arsenide, grown on gallium arsenide. This layer, if it is very thick, it will have lattice constant of this indium gallium arsenide. If you recall, indium has a tetrahedral radius of 1.44 angstroms compared to 1.26 of gallium. The lattice constant of indium arsenide is much larger than that of gallium arsenide. When we mix indium arsenide and gallium arsenide, the lattice constant is in between. How much closer is it to gallium arsenide depends upon how much less indium is present. Here, it is close to gallium arsenide, but not as much as gallium arsenide. This layer could be defective, but it so turns out that if you make this layer very very thin, that does not have those dislocations or the defects.

In fact, this is a defect-free layer but it is a strained layer. That is why it is called pseudomorphic layer. It is not truly a defectless layer. There are no defects (Refer Slide Time: 06:48) dislocations, but it is a strained layer. Its properties will not be exactly like indium gallium arsenide, but slightly different. Now, let us take a look at what happens. This is a layer that gives us high mobility. It also gives the ability for gallium arsenide to grow on top of that. It (Refer Slide Time: 07:04). When you grow a layer like this, defect-free with physically no defects, but it is a strained layer, you call it as pseudomorphic layer. It is not genuinely defect-free. When you say it is not defect-free, it does not have dislocations, etc.

(Refer Slide Time: 07:32)



Let us take a look at the structure here. We will go back to the device afterwards. We are talking of a layer that does not have lattice match, grown on a substrate layer. This is the substrate layer, which has a certain lattice constant. It could be gallium arsenide or some lattice constant. On the top of that, you grow an epitaxy layer whose lattice constant is more than that of this. If you count 1,2,3,4,5,6,7,8,9,10,11... here, there are less number of lattices, that is, the spacing between the atoms is more. If this entire layer is grown on that, the tendency of this layer, for example if it is gallium indium arsenide, is to take this structure.

If we are to grow up this, there are two ways. One way is this structure. The bottom layer has the lattice structure of gallium arsenide or material 1. On the top surface, it has lattice structure of this epitaxial layer. But in between, you can see, finally it has to go join hands with the thing to complete the structure. You can see these bonds are stretched out there. There are some of the bonds that are not formed at all. There are more number of atoms here, less number of atoms there. This atom does not have a bond there. It is a defective layer.

Similarly, this atom does not have a bond there; it does have a neighbor, because all the other neighbors are occupied there; neighboring atoms form bonds with these atoms. Two different colors or sizes are shown. It can be gallium indium arsenide, a compound; that is why it is shown like that, but the top layer now, when you keep on growing, it will have a layer that is defective,

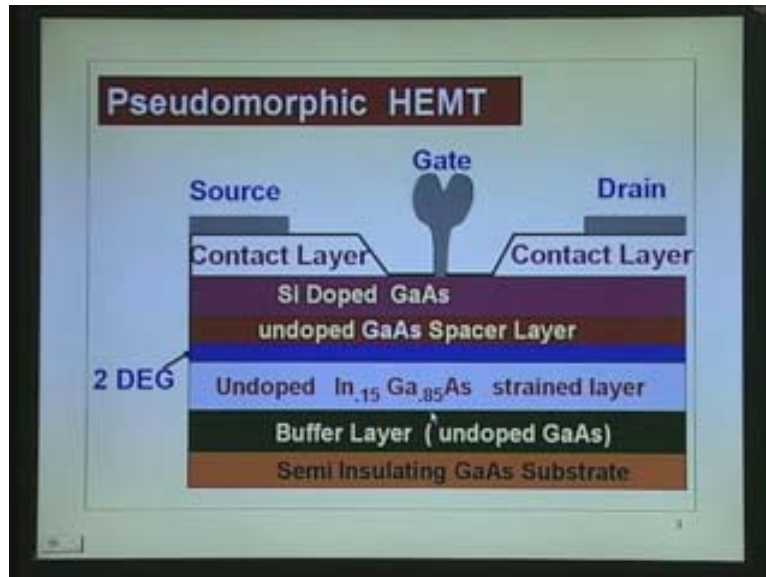
or alternately, if the top layer is very thin, what happens is this bottom layer exerts stress on this top layer, pulls the atoms together to collapse it down.

We can see here. All the atoms are aligned with respect to that. The number of atoms here are same as this. What has happened is the lattice constant here or the spacing between the atoms compared to the original atom here is reduced. This is because as you start growing the layer, the initial layer will not have defects. The initial layer will take the same lattice structure as the substrate, because the stress that is brought in from the substrate will affect the thin layer. This layer will not get affected because it is very thick. The layer that is rolled gets affected. It experiences the force of these atoms and these atoms comprise the lattice. This layer is the strained layer and it is this layer that we call as the pseudomorphic layer.

As we keep on growing, this layer becomes thicker and thicker. Then, it has its own ability to come back and pull back into its original lattice structure. What we are telling is you grow a layer whose lattice constant is different from the substrate till you reach some critical thickness – 20 angstroms, 30 angstroms or about 40 angstroms, depending upon the relative difference in the lattice constants. It will still have the single crystal structure without defects, without dislocations and that layer is the strained layer and that is the layer that we call as pseudomorphic layer. It has the high mobility.

In fact, this has (Refer Slide Time: 11:35) properties that sometimes give even higher mobilities than the unstrained layer, but it will keep on growing thicker and thicker layers. Then, it is pushed back to its original size, its minimum energy position and the lattice breaks here. There is a defective layer here. Once this was realized, people said why not make devices like this. I do not need a very thick layer to make the channel; all that I need is a thin layer.

(Refer Slide Time: 12:13)



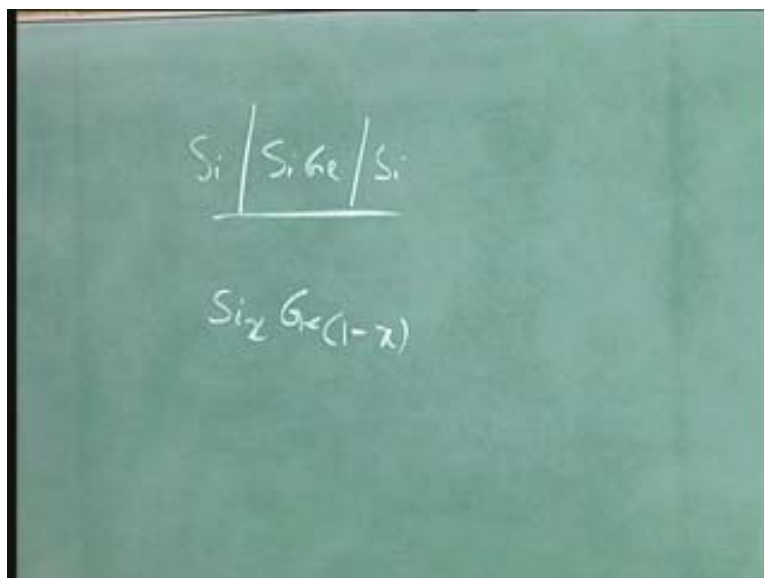
That is how they made this type of layers. They grew indium gallium arsenide on gallium arsenide, though there is no lattice match, this is a strained layer. Then, of course, on top of that, you can put gallium arsenide. This actually gives better performance in terms of the mobility. It also gives better performance in terms of bandgap of this layer compared to indium, which is 0.47, because this bandgap is slightly higher. We may have a slight problem with the electron confinement, because the gallium arsenide bandgap is 1.43. This may be about... you can calculate by substituting the formula that I gave, about 1 or so, or 1.1. So,  $\Delta E_C$  here may not be much, two-thirds of the difference; if there is 0.4, two-thirds of that will be 0.8 by 3, so, it is slightly less than 0.3.

If you want better electron confinement, what you can do is switch over from gallium arsenide to aluminum gallium arsenide. There is no difference between gallium arsenide and aluminum gallium arsenide as far as the lattice constant is concerned, but as far as bandgap is concerned, it depends upon how much aluminum content is present. The advantage of changing over to aluminum gallium arsenide is you can get a higher bandgap, you do not have to go all the way to 0.3, 0.4 aluminum content, because this bandgap is smaller. You will get  $\Delta E_C$  or  $\Delta E_g$  much higher than (Refer Slide Time: 13:49) gallium arsenide and indium gallium arsenide, like changing over to aluminum gallium arsenide.

We may remain within the safe region by putting aluminum content that is not (Refer Slide Time: 13:59), 0.1, 0.2, you do not even have to go to 0.3. Still, you will get better  $\Delta E_g$  and because of that, better  $\Delta E_C$ , higher electron concentration, higher maximum current that you can get for the given device. I just thought I will point out this to you because in many places, we will come across people talking of strained layers, strained layer epitaxy and pseudomorphic layers or devices. All that I am summing up here is that it is a thin layer. So long as this layer is thin, we can grow epitaxial layers without defect. We cannot grow thick layers of epitaxy. We cannot think of growing 10 micron epitaxy, we cannot think of growing 1 micron epitaxy, but we can grow 20 angstroms, 30 angstroms, 50 angstroms – of that order. In fact, this is the principle that is used for making heterojunction devices with silicon germanium.

Now, you can talk of not merely compound semiconductors with gallium arsenide indium phosphide, that is, III-V semiconductors. We can talk of compound semiconductor devices with silicon itself. You may not talk of high electron mobility transistor, but we can talk of bipolar type of transistors, in the case of silicon germanium. The problem with silicon germanium is germanium bandgap is smaller, but if you mix silicon with germanium, you can get a bandgap that is in between germanium and silicon. Both are indirect bandgap semiconductors. So, you can make a heterojunction device with silicon and germanium, provided you can grow silicon germanium. We will come back to this silicon germanium little later.

(Refer Slide Time: 16:00)





Silicon on silicon germanium on silicon. You can grow this type of structures provided this layer is very thin (Refer Slide Time: 16:14). If this layer is thin, even though this lattice is not matched with the silicon lattice, this will be a strained layer; strained layer without defect. So, we get mobilities as good as silicon germanium or even sometimes better, the strained layer. So, this is  $\text{silicon}_x \text{ germanium}_{1-x}$ , where  $x = 0$  is germanium and  $x = 1$  is silicon. This sort of structures are possible today. It is a very popular device or material, silicon germanium and silicon are high speed devices.

I think we will have occasions to talk about that little later in the course of the coming few lectures, but very popular for not for high electron mobility transistors. Neither of them have high electron mobility. Silicon is 1500, germanium is 3600, silicon germanium will be in between, but we will see that this is suitable for making bipolar transistors. Let us just go back and see. You can see there is an opening for silicon germanium, silicon-based devices. The silicon guy is very happy, because silicon is cheaper than gallium arsenide, much cheaper than indium phosphide. (Refer Slide Time: 17:46) can make devices with silicon and heterojunctions with silicon, then that opens up another area that has been very popular for high speed. In fact, I will come back to this particular thing. Many of the industries look into these types of devices and have been working on that.

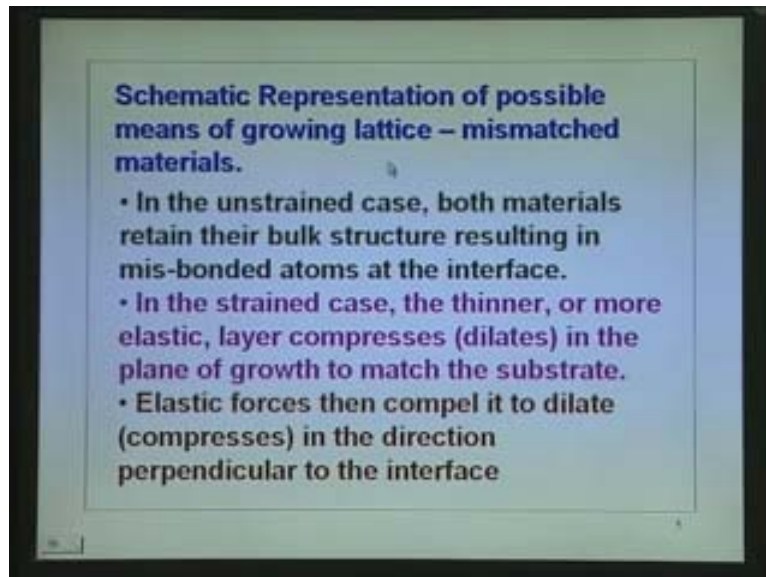
Let us just take a look at... Right now, we close the chapter on field effect transistors, with which we have discussed MESFET or JFET or both – same performance. Instead of Schottky barrier, we put junction, it becomes JFET and then high electron mobility transistors with gallium arsenide based, indium phosphide based. You can classify all these as FET. We are not talking of MOSFET in the case of gallium arsenide, indium phosphide, etc., because we have discussed and said (Refer Slide Time: 18:47) oxides are not good in those devices. They lead to lot of interface state densities. Silicon we are fortunate that (Refer Slide Time: 18:57) oxide actually gives very good interface state, very low interface state density, but instead, we would talk of MIS type of devices (I will come back to that; right now, let us leave that because that is one area that we have been working) – MIS is counterpart of MOS.

Instead of putting metal oxide semiconductor, put some insulator, deposit some insulator. I will come back to that, particularly gallium arsenide based devices, MIS type of devices, MIS field



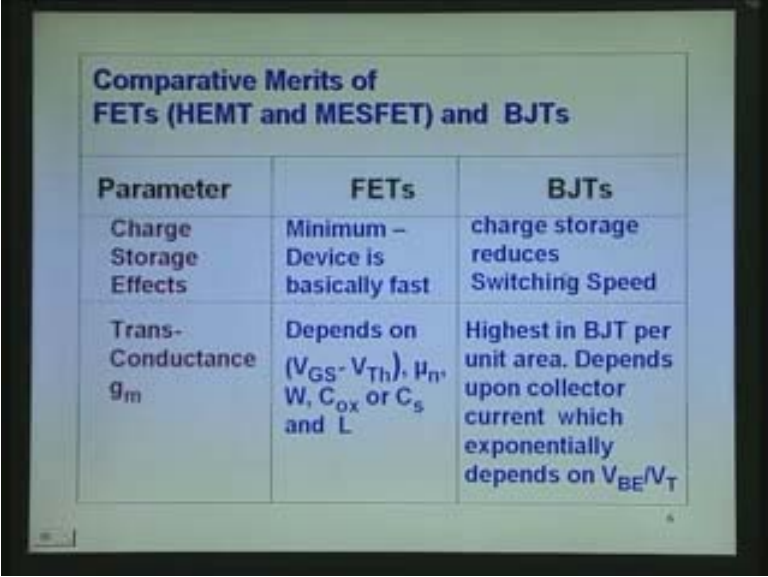
effect transistor with silicon nitride. In fact, that work has been done in IIT, Madras in great detail in PhD by one of my students on MIS field effect transistor using gallium arsenide. We will come back to that afterwards.

(Refer Slide Time: 19:49)



But in the meanwhile, let us see whether we have to dwell on FETs at all. This is just a summary of what we have said, that is, this particular aspect that I have just mentioned orally. Schematic representation of possible means of growing lattice-mismatched materials. That is the footnote for this. In the unstrained case, both materials retain their bulk structure resulting in misbonded atoms at the interface. This is unstrained (Refer Slide Time: 20:12). In the strained cases, thinner or **more elastic...** the thinner layer gets flexible, compresses in the plane of growth, it compresses in that direction. That is what we have shown there. This gets compressed in that direction, elongation in that direction, that gets compressed so that it matches with that. That is what we have shown here.

(Refer Slide Time: 20:36)



Parameter	FETs	BJTs
Charge Storage Effects	Minimum – Device is basically fast	charge storage reduces Switching Speed
Trans-Conductance $g_m$	Depends on $(V_{GS} - V_{Th})$ , $\mu_n$ , $W$ , $C_{OX}$ or $C_s$ and $L$	Highest in BJT per unit area. Depends upon collector current which exponentially depends on $V_{BE}/V_T$

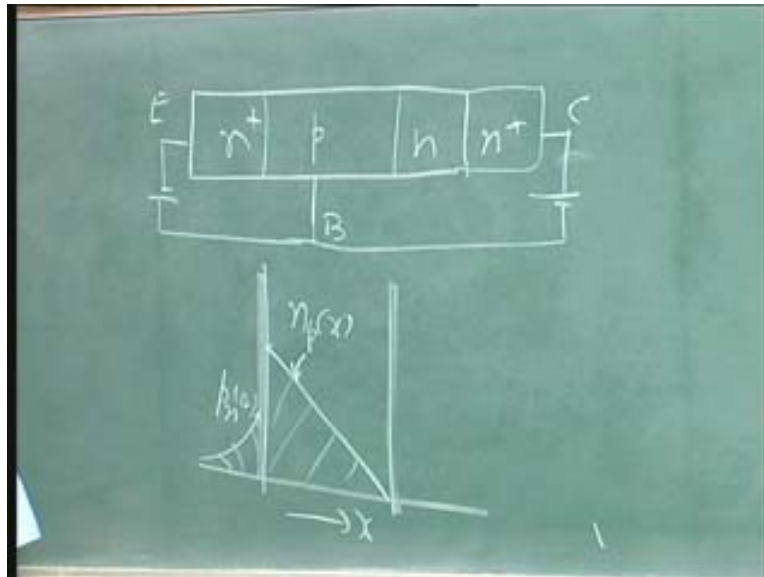
Let us leave this now and go back to this. We have to totally ignore other type of devices like bipolar junction transistor. Is there any merit in bipolar? Why we are taking a look at this is one of the things is silicon germanium has come up in a strong (Refer Slide Time: 20:58) silicon with the possibility of making a bipolar junction transistor. Number two: if you take a look at the fastest logic circuit that is the emitter coupled logic, bipolar transistor.... This gives us a feeling that maybe we should take a second look at the bipolar transistor and what are the problems in bipolar transistor.

What are the merits of bipolar transistor compared to FET? If it has no merits, we do not have to look into it at all. It so turns out that the bipolar transistor is the one that was invented first before the FET was invented, but the search was for FET. (Refer Slide Time: 21:54) were working on searching for field effect transistor. They were working on getting a working FET, but then the problem of which landed up led them to bipolar transistor. For several years, it ruled the market for about two decades. In 1970s, the FET came, but all the time, we have to say that the bipolar transistor has been coming back with a vengeance. You can see here that if you have high electron mobility transistor, etc. here, what do we have here?

Let us compare the FETs and BJTs. In the case of field effect transistor, it is a majority carrier device, unipolar device. Between the source and drain, you have the channel. All the charge that

is present is in the (Refer Slide Time: 22:50) region that is (Refer Slide Time: 22:51). C is equal to C into V. That is a small thing. So, turning on and turning off of the device is easy compared to the bipolar transistor.

(Refer Slide Time: 23:11)

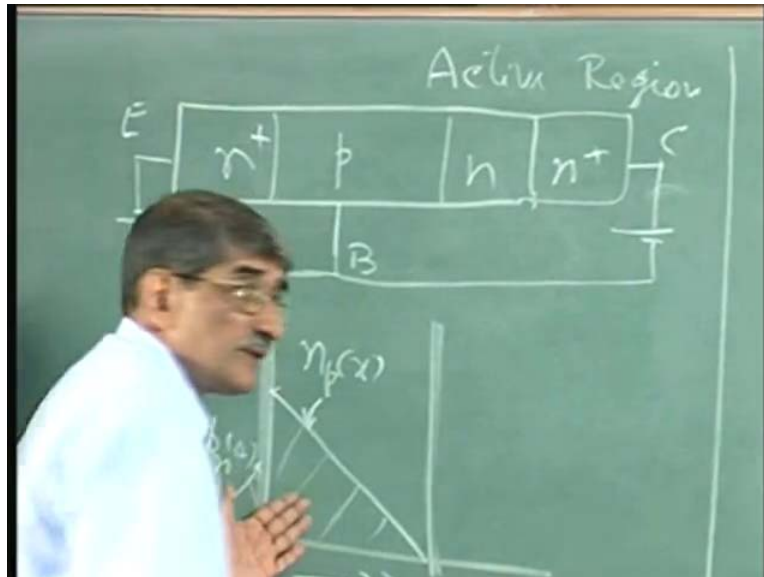


We take a look at the bipolar transistor n, p, n. We can put n plus because emitter and then the base and then, we will have always an n plus layer. It is always present so as to reduce the series resistance. That is the forward bias. For the sake of illustrating, I just put this diagram. Not that is (Refer Slide Time: ) common base, ultimately in the active region... this is reverse bias, this is forward bias. Here you can see in the active region, you have got the carrier concentration like this, that is  $x$  and on the emitter side, of course, there is a small transition region. Here, of course, there is a transition region (Refer Slide Time: 24:25) and from here, we have got  $p_n(0)$ .

These are all stored charges, this is stored charge. (Refer Slide Time: 24:42) the device, this charge from here must be removed and if the base is open during turn off, the charge will decay by recombination only and that slows down the process. So, if we turn off, (Refer Slide Time: 24:56) turnoff, we must have the base driven in the reverse direction to draw this charge. This is one of the problems. How do we reduce this charge? Reduce the base width. This is the active region.

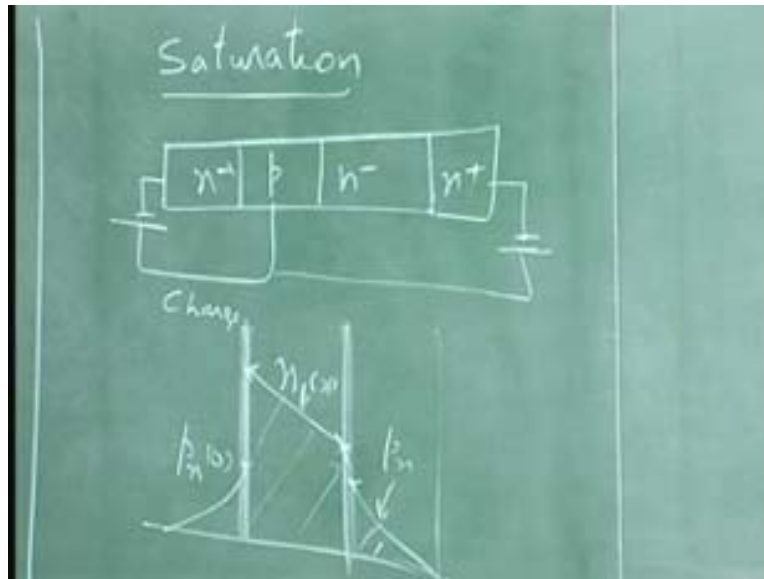
In fact, in the emitter coupled logic circuit, it never goes into saturation. In a TTL, you prevent the transistor going into deep saturation to better speed. If the transistor is in saturation, how does it happen?

(Refer Slide Time: 25:32)



This is the active region. That is the stored charge that we are talking of, stored charge here, stored charge here. We must reduce this width, we must reduce this width; all the widths must be reduced and then, saturation.

(Refer Slide Time: 25:52)

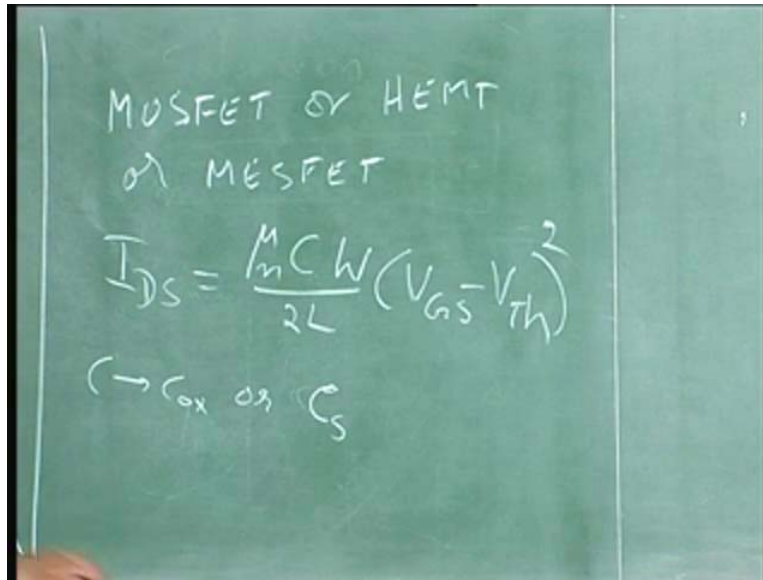


If we allow the device to go to saturation, both emitter-based junction and collector-based junction are forward biased. Then, what do you think? We must put it like this. I will put the same diagram, but it is like that, forward bias. What is the structure? What is the charge distribution? You remember when we go into discussions later. This is the carrier distribution charge or electrons or holes. Here, that will be like that (Refer Slide Time: 26:52). What happened to the base? The collector is forward biased, **carrier concentration here is up, and carrier concentration here is up**. It is almost like that. In fact, if base width is small, that is linear, but forward charge here, tremendous amount of charge here.

If it goes into saturation, remove the charge. We must pump out the charge from that base region by reverse draw of the current in the base. What about here? When we forward bias, the **killer** is here also. I have a small transition region in all the cases of course, which I have not shown here. From there, this is  $n_p$  of  $x$ ,  $p_n$  of  $x$  and this is  $p_n$ , injected holes, forward bias. The p-n junction forward bias, holes are injected there. That will be another thing. Now, we can see this problem is compounded when it goes in saturation, but if you use emitter-coupled logic, you can cut down at least that stored charge here. We can cut down this charge but still we have this charge (Refer Slide Time: 28:10), stored charge. If you take the MOS transistor, it is much smaller, (Refer Slide Time: 28:18) in the form of majority carrier. Now, let us see the other problems. This is

one of the dreaded problems in bipolar transconductance. Let us see (Refer Slide Time: 28:29) merit of the bipolar transistor. What does transconductance depend upon?

(Refer Slide Time: 28:35)



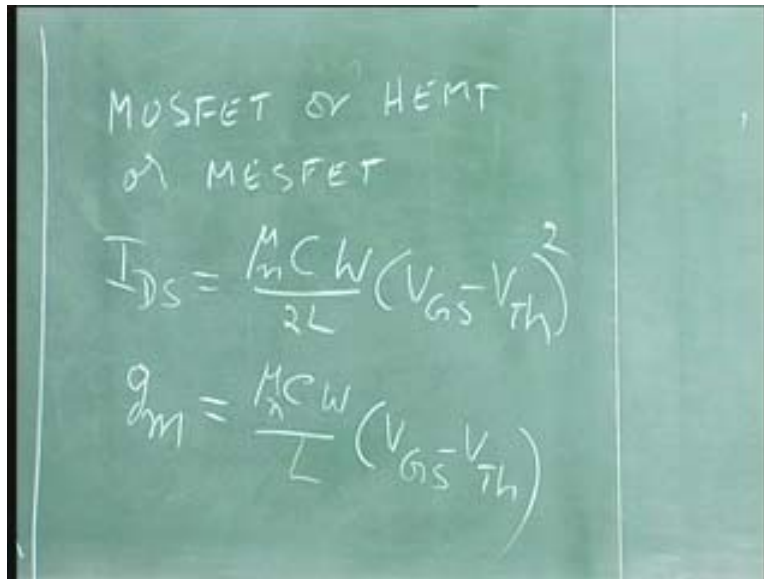
MOSFET OR HEMT  
OR MESFET

$$I_{DS} = \frac{\mu_n C W}{2L} (V_{GS} - V_{TH})^2$$

( $\rightarrow C_{ox}$  OR  $C_S$ )

MOSFET or HEMT or MESFET – all of them have  $I_{DS}$  equals  $\mu_n C$  (C can be  $C_{oxide}$  or  $C_S$ , I have put here C – general term) into W divided by 2 L into  $V_{GS}$  minus  $V_{threshold}$  square. In the case of MOSFET, it is  $V_{threshold}$ ; in MESFET, it is  $V_{threshold}$ ; in the case of HEMT, we call it  $V_{off}$ ; it is just a threshold voltage. This C is  $C_{oxide}$  or  $C_S$ . This depends upon oxide thickness (Refer Slide Time: 29:50). In the case of MESFET, this depends upon the n layer thickness (Refer Slide Time: 29:56). In the case of HEMT, it is the top aluminum gallium arsenide layer thickness, doped layer. We actually want to maximize that by reducing the oxide thickness or by reducing the layer thickness, AlGaAs layer thickness or in the case of MESFET, the gallium arsenide layer thickness. (Refer Slide Time: 30:18)

(Refer Slide Time: 30:21)



MOSFET OR HEMT  
OR MESFET

$$I_{DS} = \frac{\mu_n C W}{2L} (V_{GS} - V_{TH})^2$$
$$g_{dm} = \frac{\mu_n C W}{L} (V_{GS} - V_{TH})$$

What is  $g_m$ ?  $g_m$  is equal to.... (Refer Slide Time: 30:36). That is the transconductance that we have and then, we can see that the transconductance is proportional to width and  $V_{GS}$ . It is a square law device. If we want to improve transconductance, we have to improve this particular.... (Refer Slide Time: 31:08) Of course, this increases only linearly with  $V_{GS}$ . You must have bigger and bigger ratio of  $W$  by  $L$  to have a higher transconductance. Why do we have transconductance? We need higher transconductance because that gives you the ability to charge capacity (Refer Slide Time: 31:23). We must have higher current for a given change in voltage. For a given change in voltage, you can have higher current if  $W$  by  $L$  ratio is large and the device becomes bigger. This was one of the problems in forward bias. We have got big  $W$  by  $L$  ratio. Now, what about this bipolar?



(Refer Slide Time: 31:43)

BJT

$$I_C = I_0 e^{V_{BE}/V_T}$$
$$g_m = \frac{dI_C}{dV_{BE}} = \frac{I_C}{V_T}$$

BJT.  $I_C$  is  $I_0 e$  to the power of  $V_{BE}$  by  $V_T$ . We change the collector current by changing the emitter base voltage. When I change the emitter base voltage, the emitter current changes. Correspondingly, you have the change in the collector current. Now you can see what is  $g_m$ .  $g_m$  is the output current change for a given change in voltage. That is actually equal to  $I_C$  by  $V_T$ .  $\Delta I$  by  $\Delta v$  is again  $I_C$  by  $V_T$ . You can see it depends upon the current. In fact, this changes exponentially with voltage. MOS is or the FETs are swallow devices. These are devices hat **change...** large change in current for a small change in the input voltage, high transconductance devices is bipolar for the given size.

Of course, we can always argue that I can improve the  $g_m$  by making  $W$  by  $L$  ratio large, but that occupies a lot of space. That is not the theme of today's integrated circuit – you want to have smaller devices and that is possible with this thing. That is one of the advantages. For higher speeds, that is why we prefer bipolar. Though it is a slower device by itself, bipolar is faster in this ability to **charge (Refer Slide Time: 33:31) capacity loads**. So, in an integrated circuit **(Refer Slide Time: 33:36) bipolar** and MOS type of devices, bipolar circuit is faster because of its ability to charge all the capacitances.

Ultimately, as I keep on shrinking the device size, it is the capacity load. Where does the capacity load come from? Not merely from the next gate, it has also come from the routing

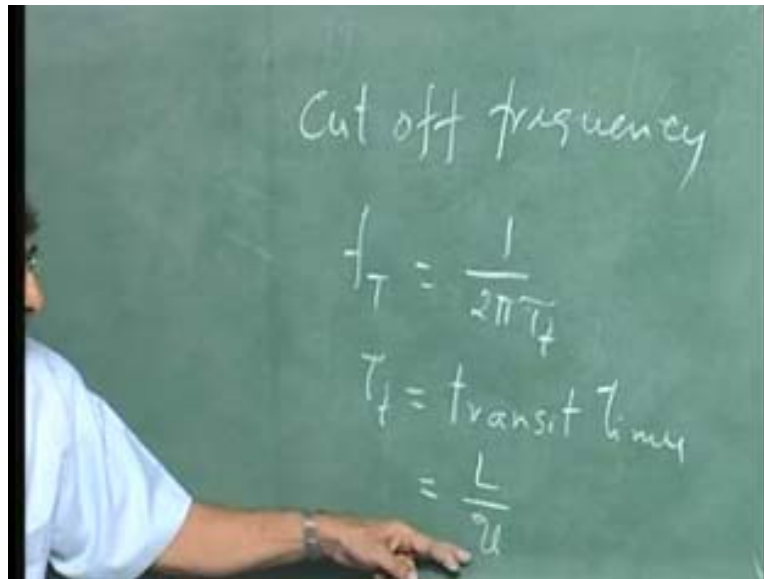
capacitance. That rules the performance completely. That is when it governs the performance and then, we must have to immediately charge these capacitance fast. That is one of the benefits of this. The highest transconductance is in BJT per unit area and it depends upon the collector current, which exponentially depends on  $V_{BE}$  by  $V_T$ .

(Refer Slide Time: 34:20)

Parameter	FETs	BJTs
Cut-Off Freq and Transit time	Channel Length dependent	Base Width dependent
Threshold Voltage	Strongly depends upon doping concentration and thickness of the channel layer in MESFET and that of the doped layer in HEMT	Practically constant (diode cut in voltage) and depends on the $E_g$ of the semiconductor

If we want to talk of the cut-off frequency, what is the cut-off frequency of the transistor?  $1/\tau$  by transit time;  $\omega_c$  is  $1/\tau$  by transit time. The transit time depends on channel length. The transit time is length divided by velocity. We can reduce to get better cut-off frequency.  $1/\tau$  by transit time... here, the base width (Refer Slide Time: 34:52). Now, it is here that the FETs scores a bit because velocities are high. The velocities in the case of FETs are high, because they are transported by drift. We can have even the saturation velocity. That is what we have been seeing all through. Actually, the intrinsic frequency response of the FET is much better compared to bipolar transistor, because for the same channel length, if you take 1 micron channel length, 1 micron base width, this will have better cut-off frequency, because velocities are larger.

(Refer Slide Time: 35:37)



Transit time or the cut-off frequency  $f_T$  is  $1$  by  $2\pi$  into  $\tau_t$  or  $\omega_T$  is  $1$  by  $\tau_t$ . This  $\tau_t$  is the transit time through the channel, through the device. The transit time is actually given by length divided by velocity –  $\text{cm}$  divided by  $\text{cm per second}$ . If we keep on reducing  $L$ , what is the constraint? The constraint is on lithography. Now, the ultimate limit for velocity is of course the saturation velocity or velocity overshoot effect – all those things are there. This will be high in the case of MOSFET and for HEMT, etc., it is very high. That is why we get smaller transit time, higher cut off frequencies of GHz range. **Bipolar... please note** that you must reduce the channel length to get smaller and smaller transit times.

(Refer Slide Time: 37:11)



BJT, this is the MOSFET, this is true for both (Refer Slide Time: 37:19). The BJT transit time is short,  $W$  square divided by  $2D_n$  centimeter square divided by centimeter square per second, so it is second. Here, you can see that transit time is reduced by reducing the base width, but when we write this equation, what is the meaning, implication? I am not going to derive it, because in the first course, we have done it. The implication is that the current transport is by diffusion in the base region. When you have transport by diffusion, the velocities are small, whereas if it is by drift, the velocities are high. You can reduce this transit time by reducing the base width. After all, if this device has to compete with the MOSFET or MESFET or HEMT, we must have the ability to keep on reducing the base width and we must also bring in the electric field.

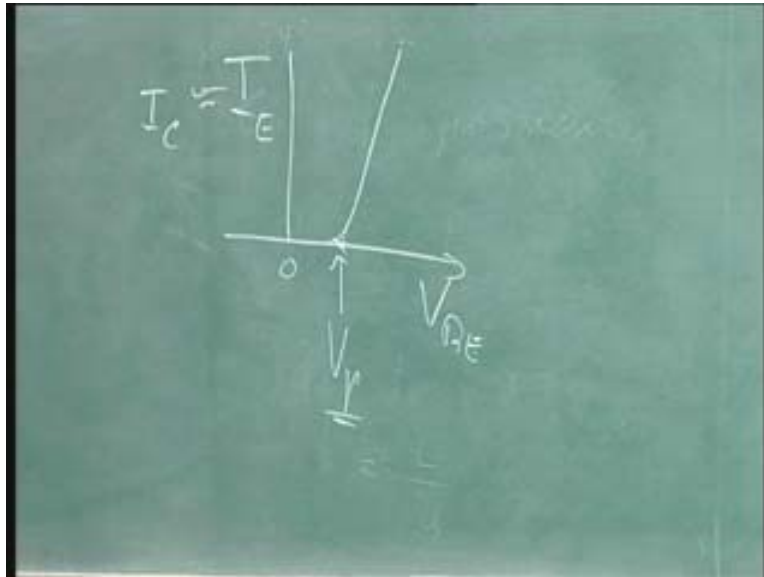
If you bring in the electric field, this is reduced by some factor (Refer Slide Time: 38:40). We can bring in electric field by different methods. Wherever doping concentration variation is there, there is electric field. Wherever there is bandgap variation, there is electric field. This sort of thing is a trick that they have played in the heterojunction silicon germanium transistor, where you do not have to vary doping but you can vary the bandgap to bring in the electric field. I think it will be very interesting to see some of those silicon germanium concepts in this discussion itself.

From the cut-off frequency point of view, you tend to buy the FETs, but when we go into integrated circuits, bipolar has an edge over that. So, you can compete with MOS or HEMT, etc., to some extent at least in terms of intrinsic speed by reducing the base width by transporting the carriers through the base (Refer Slide Time: 39:44) drift – drift plus diffusion. I am just pointing out these things because we want to take a look at that because it performs better in the integrated circuit, because of its better transconductance or smaller size.

The most important of all is the threshold voltage. If you see the threshold voltage of FETs, what does it strongly depend upon? In the case of a MESFET, thickness of that layer, doping of that layer; threshold voltage depends on  $V_{bi}$  minus  $V_{p0}$  and  $V_{p0}$  depends upon doping and thickness. Over the entire wafer, the thickness must be the same and doping must be same. It becomes harder and harder when you are talking of enhancement type of devices, where the threshold voltage is  $V_{bi}$  minus  $V_{p0}$ . We are talking of small  $V_{p0}$ . Then, small change in  $V_{p0}$  brings in a lot of change in threshold voltage. So, it becomes more.

This quantity, the threshold voltage in case of FET depends upon the doping and thickness of the layer. In the case of MESFET, what about HEMT? It depends upon  $V_{bi}$  minus  $V_{p0}$  again. That is the doping of the n plus layer and the thickness of that layer. You want to of course reduce the thickness to improve the transconductance, and then you still have to deal with thinner layers. Thinner layers uniformity over the entire 4-inch wafer or 6-inch wafer becomes more and more difficult. (Refer Slide Time: 41:35) with gallium arsenide 4-inch wafer and silicon, of course, is much better, much bigger wafer. In silicon also, we must worry about the doping, because threshold voltage depends upon (Refer Slide Time: 41:47), threshold voltage depends upon doping substrate, in addition to oxide thickness. Oxide thickness uniformity and doping with substrate both control the threshold voltage. Over (Refer Slide Time: 41:59) wafer, you need that uniformity. This is one of the problems people face. When you go to thinner and thinner oxides, you get perfectly uniform over the entire wafer. What about bipolar? Where do we have threshold voltage in the case of bipolar transistors? People normally do not talk of threshold voltage bipolar transistor.

(Refer Slide Time: 42:23)



In the case of BJT, if we take the I-V characteristics of a diode, emitter base junction,  $I_E$  versus  $V_{BE}$  is like that (Refer Slide Time: 42:44), exponential. What is the threshold voltage here? This is actually equal to  $I_C$ .  $I_C$  is approximately equal to  $I_E$ . What is this quantity? This is 0,  $V_{BE}$  and this is the cut-in voltage  $V_{\text{gamma}}$ . **Cut-in voltage, threshold voltage...** That is the point at which it really begins to conduct. In the case of BJT, you have a cut-in voltage, which we can actually call as threshold voltage.

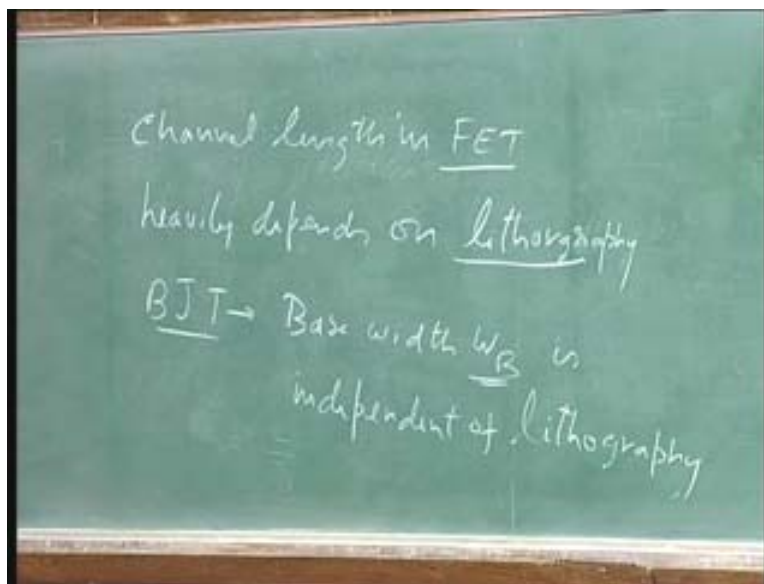
After all, in the case of MOSFET or MESFET also, the current is not 0 when you go to threshold voltage – current is there. Similarly, current is there, which is small compared to the operating point. So, cut-in voltage is 0.65 or 0.7 for silicon, about 1 Volt for gallium arsenide, P-N junction. The moment you fix the material, cut-in voltage is fixed and threshold voltage is fixed. That is one of the biggest benefits of bipolar transistor. You do not have to worry about the change in the threshold voltage over the wafer. One end to other end, it will be the same cut-in voltage and same threshold voltage. That is a big merit of bipolar transistor compared to FETs.

To sum up now, regarding the **benefits of...** and comparative **merits...** If you want to talk about the negative side of bipolar, it is a slower device by itself intrinsically because of stored charges and the switching speed gets hurt because of that. Intrinsically, cut-off frequency is smaller because of the transport of electrons in the base compared to the transport of electrons in the

channel, higher electron mobility (Refer Slide Time: 44:43), the name itself indicates better cut-off frequency. In terms of merits, transconductance is better, so it can charge capacitance (Refer Slide Time: 44:53) better and so, in an integrated circuit, it can perform better. As an individual, it may not perform well, but as a group player, it can perform much better than the FETs. That is why it still survives in the front end. If you want high-speed ICs, then it is the bipolar that we look towards. That is where we have to nurture or take a look at the bipolar transistor.

Added to that, there is one more technological benefit in bipolar transistor. That was one of the biggest scoring points for several years of bipolar transistors – even today, it is. What is that? Apart from this threshold voltage... of course, you worry about threshold voltage when we go into smaller and smaller voltages in devices... You worry about the uniformity of threshold voltage. In the case of bipolar transistor, you do not have to worry at all – it is fixed.

(Refer Slide Time: 45:57)

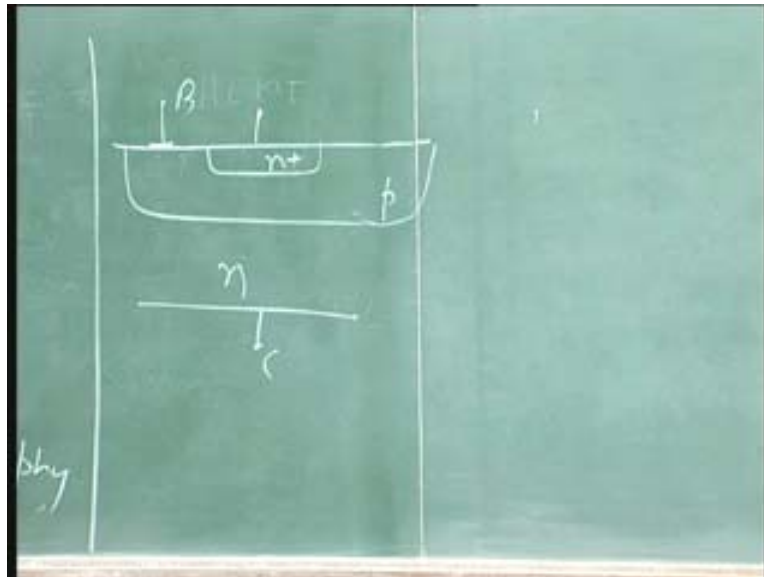


What about the channel length in FET? We have set out on lithography. It depends heavily on lithography, because it can be electron lithography also. We can go in for more and more in (Refer Slide Time: 46:44) lithography techniques to get shorter and shorter channel lengths. What about BJT? FET is that, lithography dependent. For BJT, does the base width  $W_B$  depend upon lithography? Absolutely no – it is independent of lithography; that tells the whole story. When people are comparing MOSFET and bipolar transistor, if you open up 1970s textbooks



and see, it will tell you that MOSFET is slower, because at that time, the lithography was in the range of 5 micron or 10 micron, I think 5 microns and the base width that they could get is 0.5 microns, very comfortably.

(Refer Slide Time: 48:17)



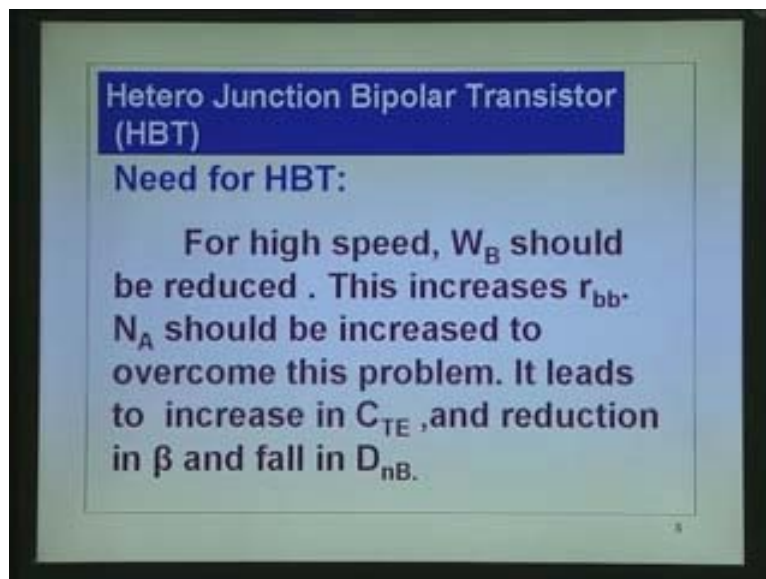
If we take a look at the base width of the transistor, the n-type region, I will just put in this collector. Whether we talk of BJT or discrete BJT, the structure is same. Only we take the contact at the bottom or top. I will put that structure here. This is the p region, n plus, emitter, base. In the integrated circuit, we put the n plus layer here (Refer Slide Time: 48:45) and take the contact of the top. Now, this is the base width. I can make this without resorting to electron lithography or any other thing. I can go down to 0.1 micron. I can get base width of 0.1 or 0.2 microns very easily. Even way back in 1970s, you could do that by double diffusion, whereas there was no way that we could do the MOSFET of 0.1 microns or 0.2 micron technology. Today, we are talking of that, after about 20-30 years, because it is dependent on lithography, whereas this **depends on...**  $W_B$  is independent of lithography. It depends **upon...** if you are doing by diffusion – single diffusion or double diffusion. The diffusion depends upon time and temperature.

If we control this width quite precisely or we can implant this region, implant this region and get the junction. So, we can control this base width precisely by the process, not by lithography; we

do not have to depend upon lithography. That is where actually the bipolar scored in speed in '70s, but it stayed on; even today, it competes with that, because if you do not have high funda lithography, you can depend on that. Now, let us go back and see. All this gives us the motivation to still look at the bipolar device. Now, can we sort out some of the problems that we have set by using compound semiconductors like silicon germanium or gallium arsenide or indium phosphide? What problems? One is of course, you want to go for smaller transit times for base width.

What are the problems if you go to smaller base width? How do we realize the transistors? In the case of compound semiconductors, we will not be **dealing (Refer Slide Time: 51:10) like this.** We will grow an epitaxial layer, base width doped and then, maybe you can implant dopants. This is the way you do things. Let us see the problems that are there in the case of the bipolar transistor and how the problems are overcome.

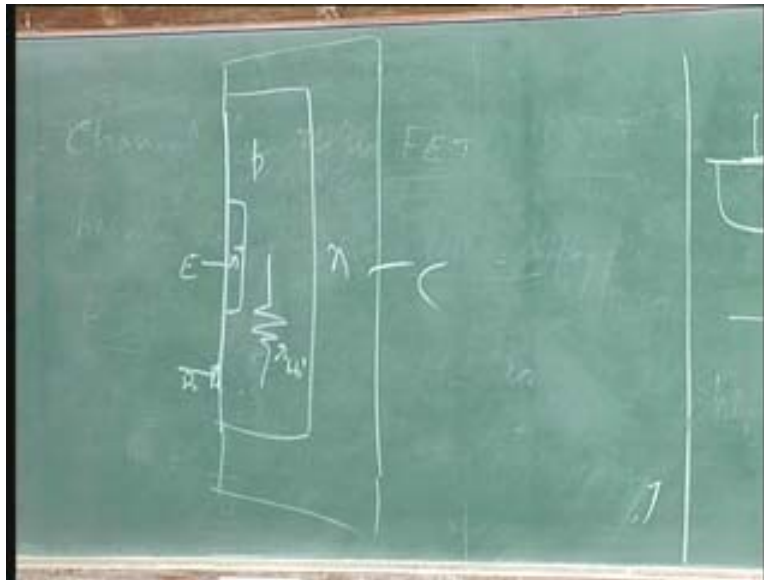
(Refer Slide Time: 51:36)



**What problems have come up one of the two.** We need to go in for a modified structure called heterojunction bipolar transistor to overcome some of these problems that are experienced by the BJT. One of the major problems is actually the transit time. We must cut down the base width. If you cut down the base width, what problems do you have? That is what is put down here. For high speed, base width should be **reduced... intrinsic speed if we want to increase.** What is the

consequence of increasing base width? (Refer Slide Time: 52:10) base width. Let us put that down here.

(Refer Slide Time: 52:14)



If I put a diagram like this, this is n plus, I am still showing the arrangement similar to what we have got in the (Refer Slide Time: 52:25) bipolar junction transistor. This is the base, this is the emitter and base and of course, you have the collector. If I reduce this base width, what is the parameter that is going to be affected? The current flow in the lateral direction goes in this direction. If I reduce this width of the base to reduce the transit time, it is accompanied by (Refer Slide Time: 53:14) cross section to the current flow in the vertical direction, in the lateral direction. The cross section is reduced here (Refer Slide Time: 53:20).

If the cross section is reduced, it increases the resistance. In fact, this resistance for the lateral flow, current flow increases. That is the  $r_{bb}$  or  $r_{bb}$  dash. Conventionally, it is used as  $r_{bb}$  dash. So, that increases. What we have to do is think of the ways of reducing  $r_{bb}$  dash, think of the ways of reducing the base resistance – that can be done by increasing doping. It goes without saying that whenever you have reduction in thickness, you must increase the doping. If we increase the doping in the base, what are the effects that follow? It has tremendous detrimental effects if we increase the doping, but without increasing the doping, we can get better performance by moving on to HBT. I will discuss some of these aspects in the next lecture.