Advanced VLSI Design
Prof. A. N. Chandorkar
Department of Electrical Engineering
Indian Institute of Technology- Bombay

Lecture - 09
Low Power Design Techniques -Part II

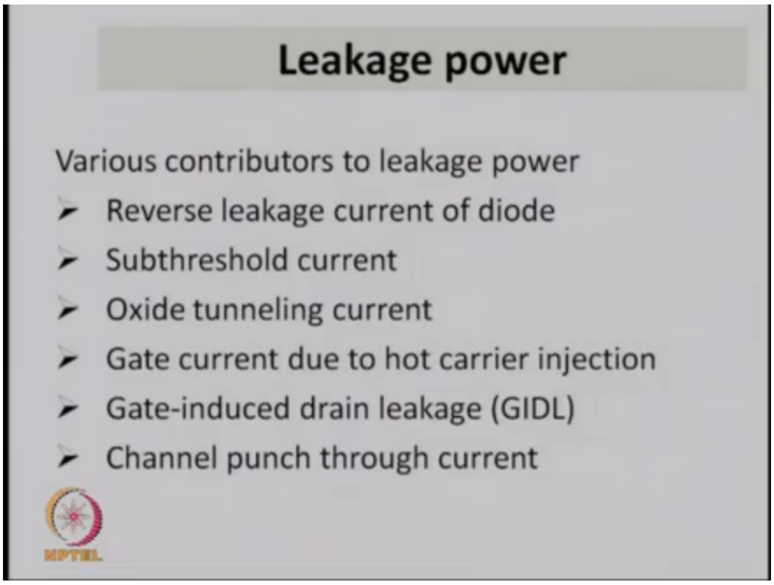
We have been discussing about the power dissipation in CMOS circuits and we also have discussed the relevance of actually designing circuits with low power and the current era when the most of the systems are handled, the power dissipation in this device or in this circuit is very relevant and to make it low power designs, we have actually looked into variety of power aware system designs.

We also looked into saying that what causes the power dissipation in circuits and when we figured out, there are three kinds of power, one is the dynamic power, the other is switching power or short circuit power and the last, but not the least the leakage power. In earlier technologies, it has been found that typically around maybe 75% power goes in dynamic, 20% goes in short circuit power and 5% in case of leakage power.

But as the advancement and technology has taken over from last 10 to 15 years, we went from 0.25 micron process to now almost 28 nanometer process and may soon go into 16 nanometer process. Because of that the devices have become very, very small both in lengths and widths and that has actually created some other problems particularly the power, which is lost in the leakage power.

And I last time shown you that in the 32 nanometer node down it may be found that the standby leakage power maybe larger than the dynamic power and in that case the major research should be done to actually controls the leakage power. Now last time we did discuss, but quickly I will show you what I said last time. This is the last slide I have shown you last time.

(Refer Slide Time: 02:09)



Leakage power

Various contributors to leakage power

- Reverse leakage current of diode
- Subthreshold current
- Oxide tunneling current
- Gate current due to hot carrier injection
- Gate-induced drain leakage (GIDL)
- Channel punch through current

I started saying that leakage power has following contributors. For example, the first and the foremost is the reverse leakage current of the diode of a source or drain junction with the substrate. The second we discussed is subthreshold current and I had last time said that even if the VGS is less than VT we are still in V conversion and because of this state there is a current flowing still in the between source and drain even if VGS is less than VT and this we say as subthreshold current.

Now that means we thought the device is off, but in fact device is not fully off it is still leaking the current. The third possible we say is the oxide tunneling current. This is essentially because as the scaling down of technologies have taken place, the oxide thickness or the insulator in MOS itself has thin down so much so that it is possible for carriers to tunnel through this thin oxide because the large electric field, which vertically it creates.

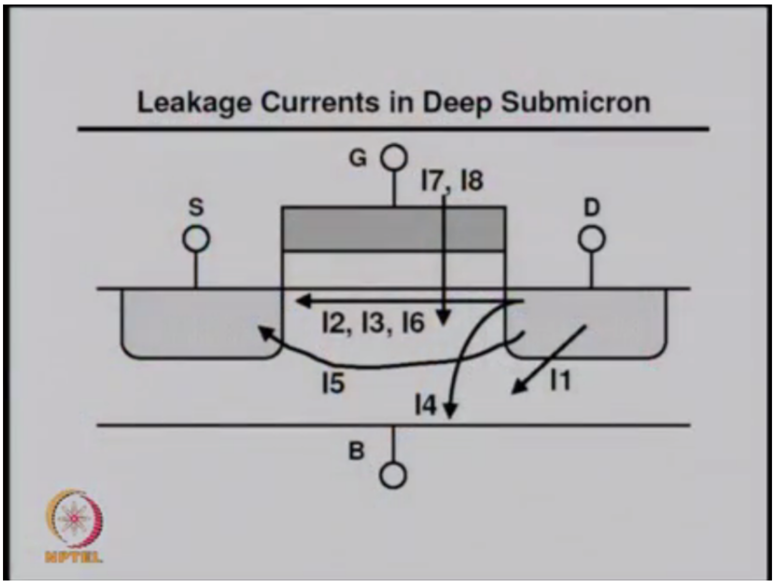
Now the fourth is of course there is a possibility that since the short channels have become very, very small very, very as I say nanometer technologies we are talking about and in that case in the electric field at the drain end is so high because of current field crowding at that time that carriers can get injected across the gate and therefore it is called hot carrier injections.

The fourth rather fifth is essentially occurring because what we call gate-induced drain leakage. This is essentially the problem is happening is that as the source and drain have a stronger doping compared to substrate, there is a large depletion layers both at the drain side as well as the on the source side. Now these large depletion that creates electric field, but even if now let say if your gate is having a zero wires or lower voltage then there is an accumulation layer even if there is no inversion layer.

Prior to inversion there is an accumulation layer, which actually changes the substrate doping at the surface and because the substrate doping is higher there, there is an excessive electric field surround the drain end and then because of this, there is a larger current flowing between source and the substrate through the second accumulation layer and this is called gate-induced drain leakage.

GIDL is very, very relevant now because as the scale down technologies the dopings are anyway increasing to adjust the threshold voltage and the finally since the channel length are becoming extremely small, the source drain depletion layer width can connect to each other without having the gate voltage and that means short circuiting the source drain even without the gate voltage applied we call as punch through and it may actually create a large current because of the short resistive path created between source and drain.

(Refer Slide Time: 05:25)

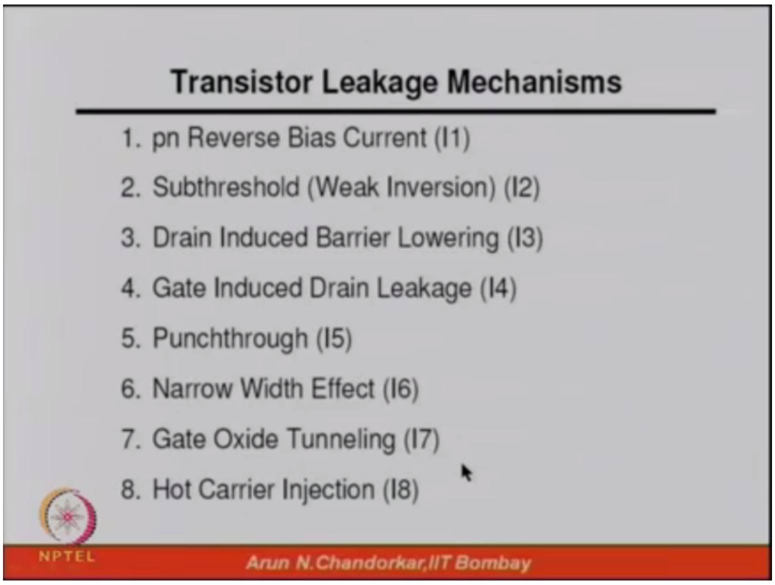

Leakage Currents in Deep Submicron

So having told that this is a typical figure, which essentially says if you look at the I1, which is nothing but the diode leakage current. Same way of course will be true for this side as well so I1 is the diode leakage current between substrate and the drain. I2, I3 and I6 are essential. There are three currents, one essentially is occurring simply because of threshold current is flowing.

The third is essentially because there is a possibility that GIDL current is of course I4, I4 is the GIDL current, I7 and I8 is essentially because of the oxide tunneling happening here and hot carrier effects and I6 is the punch through possibilities and I3 maybe because of drain induced barrier lowering, which occurs because of the larger fields created here because of the source and drains.

So these possible currents, eight types of current, which may essentially contribute to a leakage current even if VGS is less than VT.

(Refer Slide Time: 06:30)



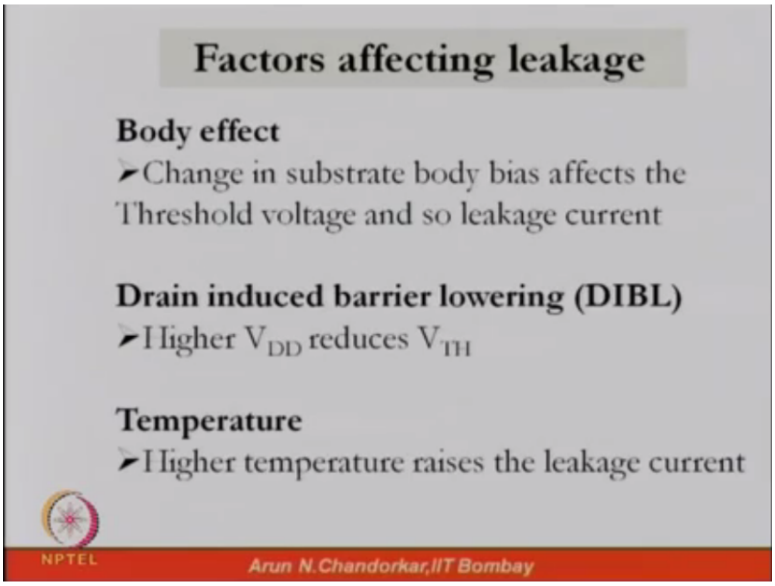
Transistor Leakage Mechanisms

1. pn Reverse Bias Current (I1)
2. Subthreshold (Weak Inversion) (I2)
3. Drain Induced Barrier Lowering (I3)
4. Gate Induced Drain Leakage (I4)
5. Punchthrough (I5)
6. Narrow Width Effect (I6)
7. Gate Oxide Tunneling (I7)
8. Hot Carrier Injection (I8)

NPTEL
Arun N.Chandorkar,IIT Bombay

These are the numbers, which I have given and these are the things which I have already explained to you. Now obviously therefore one can see what affects the leakage, the body effect, change in substrate body bias affects the threshold voltage and so is the leakage current. We shall go into this little detail in few minutes later.

(Refer Slide Time: 06:52)



Factors affecting leakage

Body effect

- Change in substrate body bias affects the Threshold voltage and so leakage current

Drain induced barrier lowering (DIBL)

- Higher V_{DD} reduces V_{TH}

Temperature

- Higher temperature raises the leakage current

NPTEL

Arun N.Chandorkar,IIT Bombay

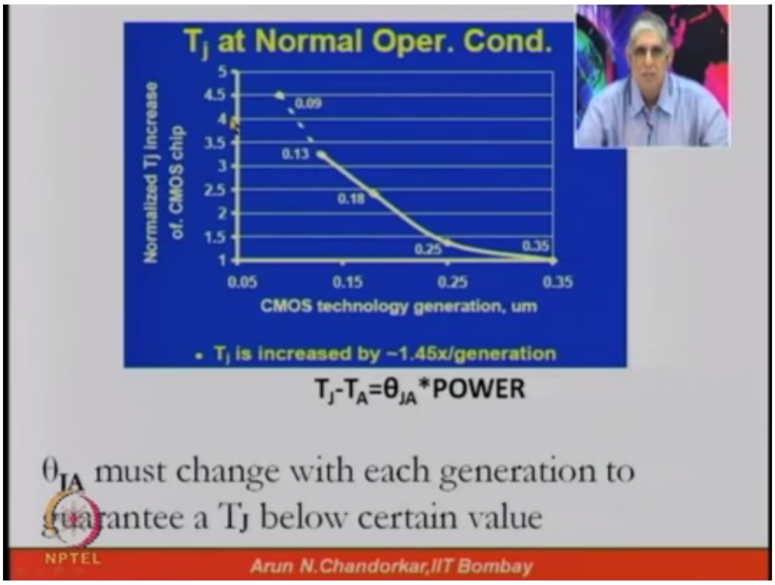
The second issue, which is occurring in the lower channel devices in particular and with technologies as lower the 32 nanometers are even lower because there is VDD is not scaling so much, the higher the VDD we apply the source and drain depletion layer widths are very, very large and because they are very large, the bulk changes, which creates the threshold voltage expression we shall see later.

The bulk charges are already present even without VGS, which means that to create an inversion N-channel now you will require smaller amount of VGS to create inversion N-channel below because already part of the depletion layer is providing with the bulk charge. This that means because of the larger VDD one sees reduction in threshold voltage and finally of course all the currents except the tunneling part currents they are exponential depending on temperatures and therefore larger the temperature, leakage currents are always higher.

Please remember in most of these short channel devices of less than 45 nanometers, the major worry right now is the rise in temperature as we already seen in case of earlier lectures. The temperature may rise to as large as nuclear power, reactor power or rocket nozzles. So the minimum even what we call self-heating experiments we did or people have done shows that the normal temperature on the chip is not 27 degree centigrade as one looks at it is essentially around 70 or 75 degree centigrade.

And which essentially enhances the leakage current proportionately in exponential term. And we already discussed last time to great extent we said larger the power or larger is the CMOS technology generation as we reduce.

(Refer Slide Time: 08:47)



T_j at Normal Oper. Cond.

Normalized Tj increase of. CMOS chip
5
4.5
4
3.5
3
2.5
2
1.5
1
0.09
0.13
0.18
0.25
0.35
0.05
0.15
0.25
0.35
CMOS technology generation, um

- T_j is increased by ~1.45x/generation

$T_J-T_A=\theta_{JA}$*POWER

θ_{JA} must change with each generation to guarantee a T_J below certain value

NPTEL

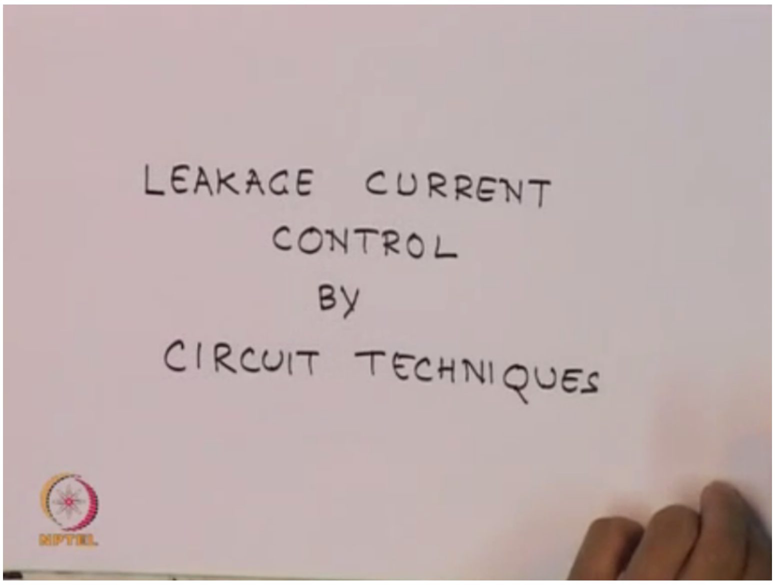
Arun N.Chandorkar,IIT Bombay

There is a temperature rise in the junctions, you can see this is even for 90 nanometers we already at the very high and this CMOS shape is increasing its temperature and it is a normalized temperature so it is four times or five times now already and lower the technology it may become more than 10 times.

Now this means that if I want to dissipate this much power or other if I want to keep temperature below 70 degree or something like this something three or three and a half, so I must remove the heat and that essentially what we call the thermal resistance of the substrate as well as that of the packaging has to be so adjusted that the junction temperature does not rise high enough more than 70 degrees.

Now having told that there is a leakage problem at particularly for subthreshold sub 45 nanometer technologies, we want to know can we reduce this leakage power by circuit technique. One is of course device technique, which we have discussed. Once we have the devices, chips are made. There is nothing much we can do; however, during design and during bulking of the chip, can we actually reduce the leakage current.

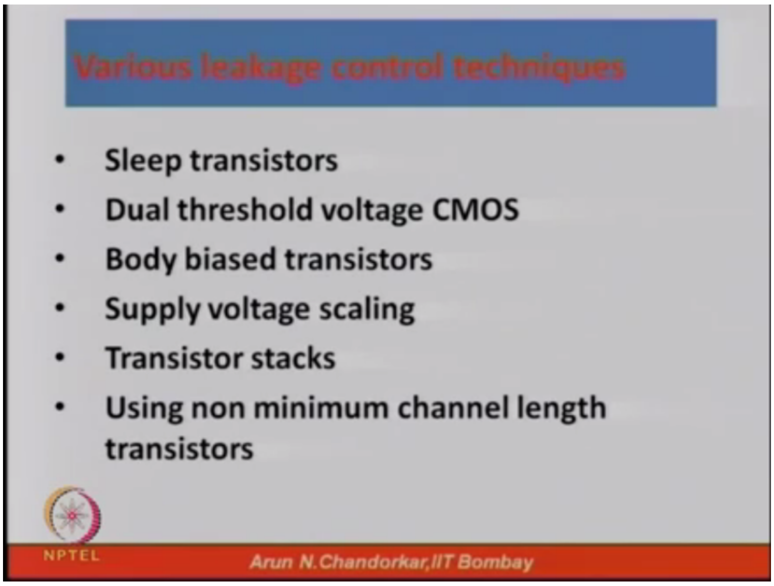
(Refer Slide Time: 10:14)



LEAKAGE CURRENT
CONTROL
BY
CIRCUIT TECHNIQUES

So we now go for the area which we talk about leakage current control by using circuit techniques.

(Refer Slide Time: 10:17)



Various leakage control techniques

- Sleep transistors
- Dual threshold voltage CMOS
- Body biased transistors
- Supply voltage scaling
- Transistor stacks
- Using non minimum channel length transistors

Arun N.Chandorkar,IIT Bombay

Now what essentially is this word I am talking about is the fact there are number of ways. Sleep transistor is one method of reducing the leakage power, the other is dual threshold voltage CMOS instead of dual it can be multi-threshold voltage CMOS that will see later. Dual is the word initially we use, but then it was found that you can have number of devices have different thresholds.

Then there can be also a variable threshold voltage because we can continuously vary the threshold voltage, which is also possible. Then there is a technique of body biased transistors by putting a substrate bias either reverse mostly reverse bias or sometime forward bias. We can actually change the threshold voltage and which essentially controls the leakage currents. One of course the foremost way which can reduce all kinds of power with the reduced dynamic short circuit or even the leakage current into power supply voltage.

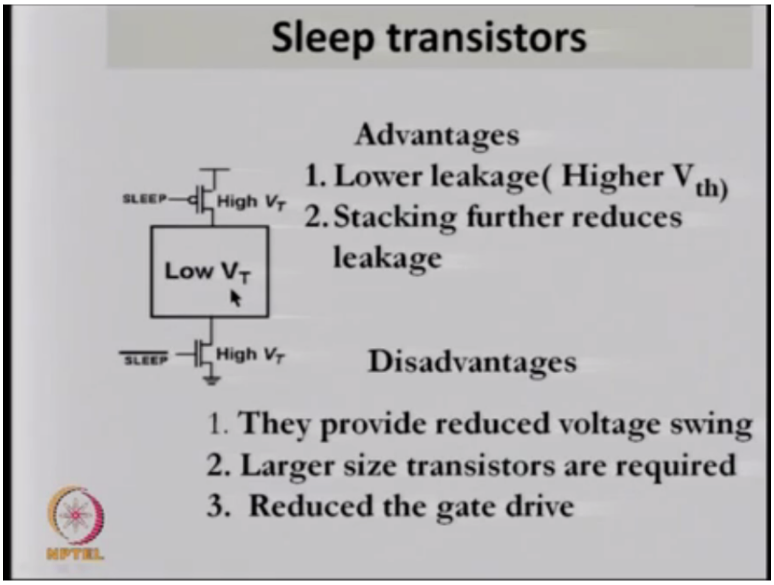
So obviously leakage power to reduce power supply voltage must go down. Say if you scale it then obviously power goes down. So is it possible in circuits to actually reduce the voltages particular voltage at particular times so that the net power reduction is possible. When certainly you are in a standby mode is not in an active mode. Then of course there is another method, which has been tried successfully to many extent is transistor stacks and we will look into this multi-threshold transistor stacks as one way of reducing the leakage power.

And of course if you have a technology possibilities already existing with you that you can modify the technology or other at least during designs you need not use all transistor with the same minimum channel lengths you can have larger length devices and we already seen short channel effects occur only and only if the dimension of the device is within less than say 100 nanometers or something when the short channel effect is very strong.

Though it starts around 0.25, but the effects are very, very strong when it goes below 90 nanometers. So if you have a device, which have larger channel lengths probably much of the problems can be solved; however, the effect of larger channel length will immediately go in the increasing of propagation delay and therefore reducing the speed so can we now adjust the channel lengths devices where the speed is not a criteria what we call the critical paths.

So these are the techniques, which we will use it from the circuit side to some extent from the device side, but mostly from the circuit side and we shall like to see how this leakage power can be controlled. Now before we go to sleep transistors or before we go to the more details about how it occurs, am I quickly go through the list, which I have talked to you.

(Refer Slide Time: 13:17)



Sleep transistors

Advantages

1. Lower leakage(Higher V_{th})
2. Stacking further reduces leakage

Disadvantages

1. They provide reduced voltage swing
2. Larger size transistors are required
3. Reduced the gate drive

One method of reducing the leakage why is the leakage occurring from the power supply to the ground and if the devices are of normal lower threshold. Please remember normal transistor had to have lower threshold voltage because we want to have higher speeds. Now for higher speed, if you keep low VT the leakage is proportional in some sense inversely proportional to threshold voltage value.

Larger the threshold smaller is the leakage we know that. So because of that we now provide additional hardware. We say okay there are one P-channel transistor like a dynamic system we have one P-channel transistor and one N-channel transistor. These are high area high W by L transistors, which has a higher thresholds by design and they are given in signal sleep and sleep bar and we will look into this specifically when we talk about sleep transistors.

Basically you can think when the sleep is 0, P-channel device conducts and when the sleep is 0 sleep bar is 1 so N-channel conducts and therefore and since these are very large W by L transistors the voltage here and here is not very different, voltage here and here is not very different and therefore a circuit when in active mode it behaves like a normal VDD VSS applied here and here and circuit can function at high speeds.

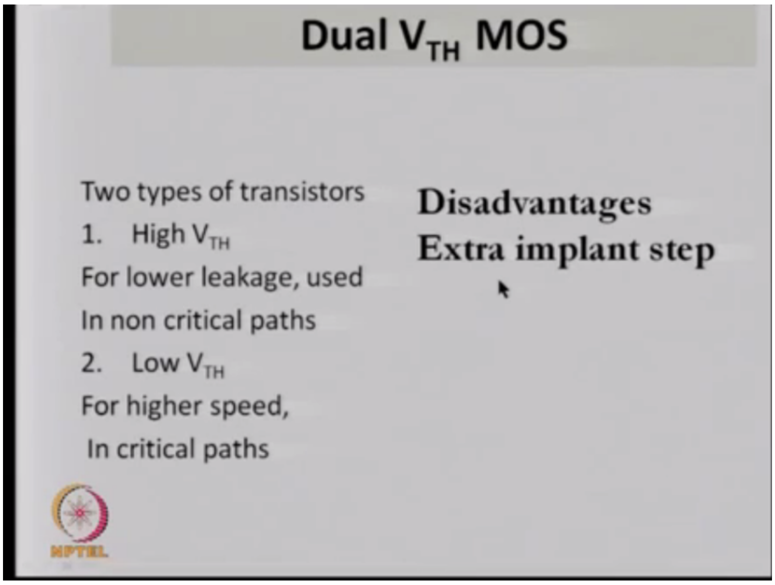
But in the standby mode when you are not operating by a program one can make sleep bar 0 and sleep 1, both P-channel and N-channel can be cut off and during this cut off since there are higher threshold the leakages through these are very, very small and therefore the currents through this circuits will also be small and therefore the leakage power can be minimized. Lower leakage as I said higher thresholds.

This advantage of course one can see once I said larger size devices and there is some finite drop across both P-channel and N-channel. Obviously, there will be smaller VDD and larger VSS here, which means there will be reduced voltage swings. Of course this can be minimized, but there still will require higher penalty on the size itself because if size is larger the area is larger.

And of course that is since your power supply voltage may change to some excess swing is smaller the drive current available from this actual logic will be little smaller. So this is very popular technique; however, this had its own advantages and disadvantages. We will come into it little more detail in the later part. The other I will just first go through the slide and then talk about the theory behind.

The other technique is of course you can have transistors, which you think are the critical paths that means we are slower and you want to improve the speed. We can have those transistors in the critical path as having lower thresholds whereas other areas where the speed was not so important anyway data has to wait for somewhere to reach those path need not run faster.

(Refer Slide Time: 16:22)



Dual V_{TH} MOS

Two types of transistors

1. High V_{TH}

For lower leakage, used

In non critical paths

2. Low V_{TH}

For higher speed,

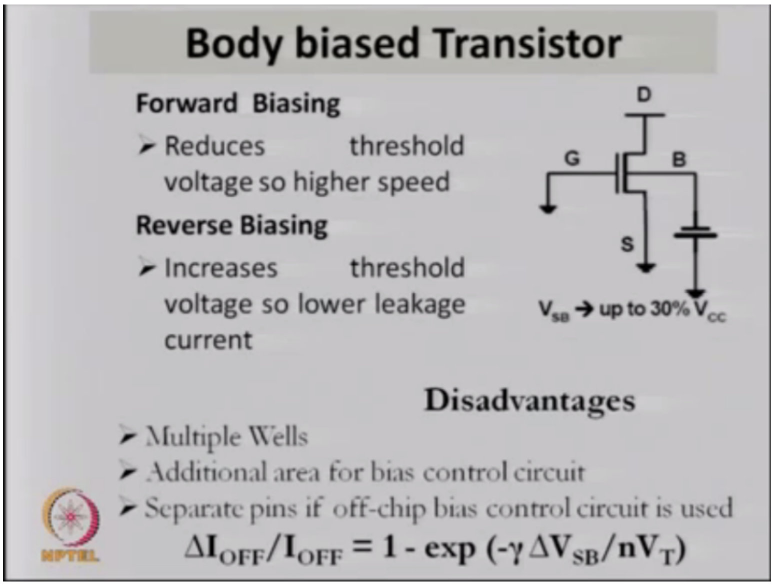
In critical paths

Disadvantages

Extra implant step

And therefore in those cases the transistor can have higher VTH and if those transistor higher VTH please take from me in an off state they will provide lower leakages. To create this different VTH technology wise you will have to do another mask that means an extra implant step. This extra one step creates one mask plus additional process steps, which in a way it is said it cost a million dollars.

(Refer Slide Time: 16:45)



Body biased Transistor

Forward Biasing

- ➢ Reduces threshold voltage so higher speed

Reverse Biasing

- ➢ Increases threshold voltage so lower leakage current

Disadvantages

- ➢ Multiple Wells
- ➢ Additional area for bias control circuit
- ➢ Separate pins if off-chip bias control circuit is used

$$\Delta I_{OFF}/I_{OFF} = 1 - \exp(-\gamma \Delta V_{SB}/nV_T)$$

The third possibility as I say I will go to the theory little later, but quickly look into it. If you see clearly you have a substrate. In a MOS transistor, substrate normally is grounded or either connected to source, but in the case here if I have a substrate bias in right now shown here say for p substrate negative bias it can be even forward bias, but at least negative bias. Then if I apply negative bias, there will be a depletion layer here initially created between drain and source to the substrate because of the applied negative bias.

Now this essentially larger depletion layer at the source and drain end will require smaller gate voltage to create an inversion N-channel, which essentially means that threshold voltage of a back bias device or body bias device will be smaller and those transistor will therefore start acting at lower threshold and therefore higher speeds sorry I made the other way. Since forward biased will reduce the threshold when reverse biased will increase the threshold.

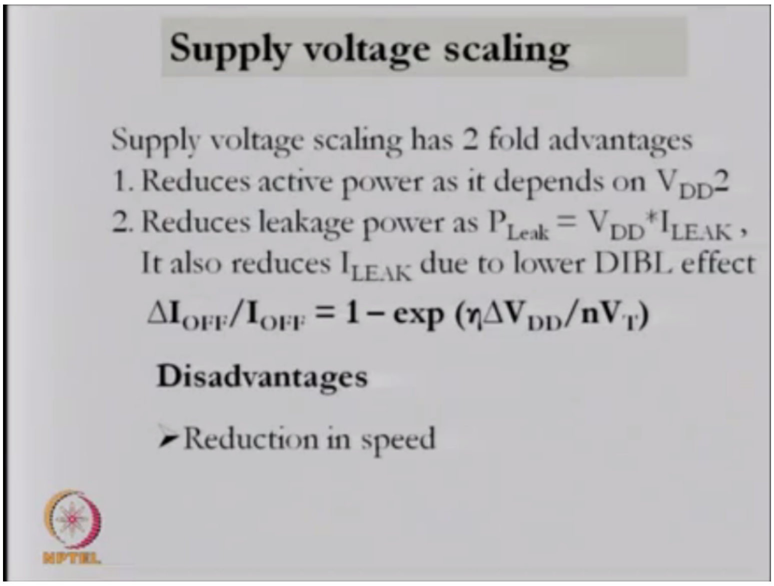
Forward biased will improve the speed and reverse biased will actually increase the threshold and therefore please remember additional charge means larger VT, lower charge in the bulb will have lower VT. So when you have a forward biased, the charges are smaller at the source and drain. Therefore, VT goes down, threshold voltage reduces and therefore one says that essentially you have faster circuits.

So either FBB or RBB can be tried to modulate the threshold shear and because of that one can have that is because this advantage one sees in technology to create such thing you need require separate wells in the CMOS they are related twin wells you may have third wells and sometimes the four wells as well. The biased circuit will require additional area, which you have to give.

And since you have to create a bias control circuit you need additional pins okay and what we see delta I of change in off current to the main off current is proportional to e to the power 1 minus gamma, which is called back body coefficient into delta VSP. So change in this, this is of course kT by q slightly confusing, but it is essentially thermal voltage kT by q. So depending on the VSB value, I can change the off current.

That is the idea behind body biased transistors. So either VSB can be positive or negative and depends on the way bias and therefore I can change in I OFF at my will. The third possibility of reducing the leakage current is supply voltage scaling has two fold advantages, which we always know dynamic power goes as VDD square so there is no question of thinking that a power supply voltage is reduced the dynamic power is going to be reduced because it follows square law.

(Refer Slide Time: 19:51)



Supply voltage scaling

Supply voltage scaling has 2 fold advantages

1. Reduces active power as it depends on $V_{DD}2$
2. Reduces leakage power as $P_{Leak} = V_{DD}{}^{*}I_{LEAK}$, It also reduces I_{LEAK} due to lower DIBL effect

$\Delta I_{OFF}/I_{OFF} = 1 - \exp(\eta\Delta V_{DD}/nV_T)$

Disadvantages

➢ Reduction in speed

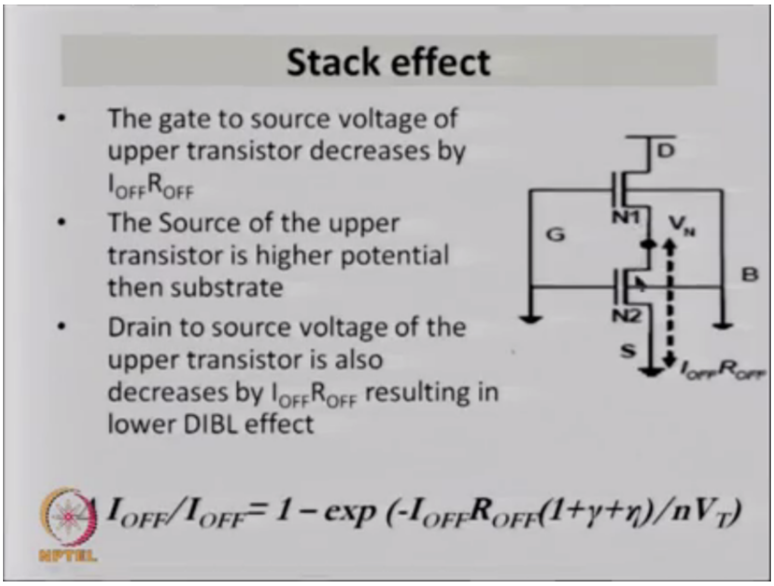
Now we look at the leakage power it is nothing, but VDD into I LEAK and if I can reduce I LEAK as adjusted, I find if we reduce the VDD then one says that the DIBL coefficient we know DIBL or drain-induced barrier lowering occurs because of large VDS available to you.

If your VDS is smaller because of VDD is smaller so obviously DIBL coefficient goes down, drain-induced barrier lowering is going down.

And because of that threshold voltage is actually can be adjusted, which becomes higher and if that becomes higher this becomes smaller and therefore one can improve the leakage power reduction or we can improve the leakage power consumption as well as we can reduce the dynamic power if I scale down supply voltage, but the fact remains that this or whatever scale law we are following as per what more thought, we are unable to scale down supplies in the same node scaling as 0.7 times and because of that the fields are very high and the DIBL coefficient is not very low okay.

However, as soon as I say it increases the threshold voltage the current available to me, which is called I ON to I OFF on state currents; however, active mode currents decreases and therefore the speed goes down. The fourth possibility and as I say I will come back to these each of them once again individually.

(Refer Slide Time: 21:26)



Stack effect

- The gate to source voltage of upper transistor decreases by $I_{OFF}R_{OFF}$
- The Source of the upper transistor is higher potential then substrate
- Drain to source voltage of the upper transistor is also decreases by $I_{OFF}R_{OFF}$ resulting in lower DIBL effect

$$I_{OFF}/I_{OFF} = 1 - exp\ (-I_{OFF}R_{OFF}(1+\gamma+\eta)/nV_T)$$

Here there is an interesting case, if you have a single transistor and then you break them into a series combination. The way I have explaining that two transistors in series is essentially has if they had same channel length then you can see these are to be series that means one upon W is equal to one upon W1 plus 1 upon W2. So if I have two W here, two W here by L then actually it is 1W by L together.

So a single transistor W by L can be actually changed into 2W by L 2 series transistors and now we can see from here that if the leakage current is flowing not necessarily if this is a drain end not VDD if the leakage current flows through down even when the gate voltage is at 0 then you are near subthreshold value slightly higher, but not larger than VT.

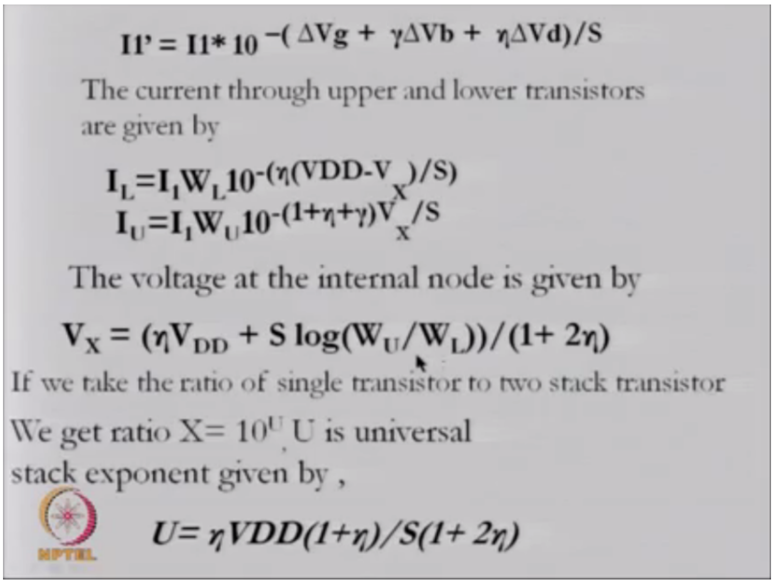
Then one can see from here since the current is flowing N2 source is grounded and therefore there will be a voltage drop across VDS of this into transistor or into N1 order number I will give later and this voltage will rise and if this voltage rises then a lot of interesting effect it will give it to N1 and particular it will increase the threshold voltage of N1. We shall see this little detail later.

And if I increase the threshold voltage of N1 then leakage current through N1 can be smaller has to go down and therefore the leakage power will go down and we shall look into this little detail as I come down. One can see from here in this case, change in off current to the net off current can be proportional to exponential of I OFF into R OFF. R OFF is the resistance in the off state of this transistor into 1 plus gamma, gamma is called back bias coefficient and eta is essential in the DIBL coefficient.

So one can see from here that I can have smaller of current provided I can adjust my values of VTs of N2 transistors much more strongly so that the DIBL effect is lowered and if the DIBL effect is lowered for N1, we shall see that we will reduce the leakage power. This is called stack effect a very, very important method and we shall see in real circuits you need not divide say one transistor into two.

This is only to show you the point I am saying that if there is a series transistor, the leakage current can be lower because of the DIBL coefficient reducing.

(Refer Slide Time: 24:11)



$$I1' = I1 * 10^{-(\Delta Vg + \gamma\Delta Vb + \eta\Delta Vd)/S}$$

The current through upper and lower transistors are given by

$$I_L = I_1 W_L 10^{-(\eta(VDD - V_X)/S)}$$
$$I_U = I_1 W_U 10^{-(1+\eta+\gamma)V_X/S}$$

The voltage at the internal node is given by

$$V_X = (\eta V_{DD} + S\log(W_U/W_L))/(1+2\eta)$$

If we take the ratio of single transistor to two stack transistor

We get ratio $X = 10^U$. U is universal stack exponent given by ,

$$U = \eta VDD(1+\eta)/S(1+2\eta)$$

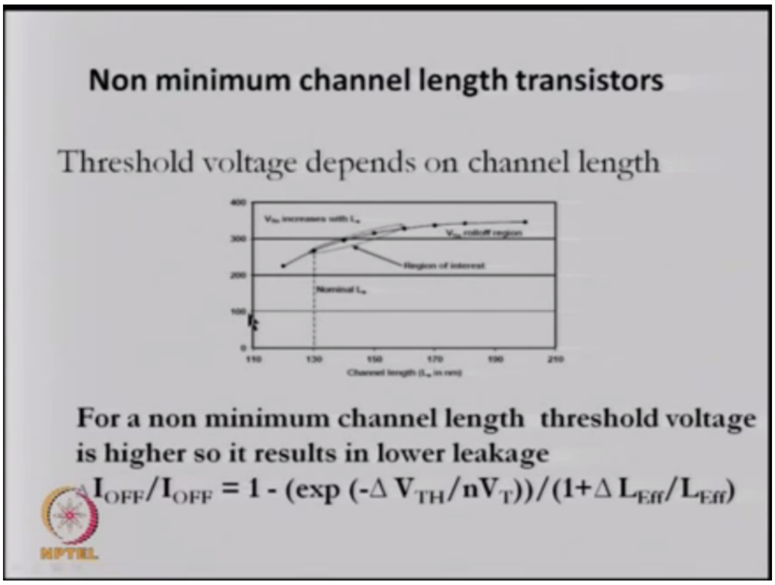
Now this is the expression more details. This is the node voltage VX here at this point say if I see the expressions for I1, which is proportional to 10 to the power delta Vg change in get voltage, change in the substrate bias voltage, change in the drain voltage, this is the DIBL coefficient, this is the back bias coefficient. All three put together this is the gate voltage change as it is called subthreshold slow okay.

The current through upper and lower transistors are given by this expression using the same expression here. This is actually width you can say it is described for the width of the lower transistor, width of the upper transistor and you solving this because these currents are equal in the two transistors one gets what will be the intermediate node voltage VX and one can see if you have a larger VX the DIBL coefficient goes down and therefore threshold rises.

To improve VX one can see from here it has to be larger, S has to be larger not 60 millivolt per decade it should be larger. Upper transistor should have larger width compared to the lower transistors. If we do all this obviously one can see from here that I can increase VX and correspondingly if I find the value to single to this we get a typical ratio of 10 to power U is called universal constant, which is given by from expression from this.

So I can figure it out what should be the size ratio of the W by Ls of the upper and lower transistors and once I get to that value, I will be able to adjust my DIBL coefficient and therefore the threshold increase and therefore reduction in leakage currents.

(Refer Slide Time: 25:55)



Non minimum channel length transistors

Threshold voltage depends on channel length

For a non minimum channel length threshold voltage is higher so it results in lower leakage

$$\Delta I_{OFF}/I_{OFF} = 1 - (\exp(-\Delta V_{TH}/nV_T))/(1+\Delta L_{Eff}/L_{Eff})$$

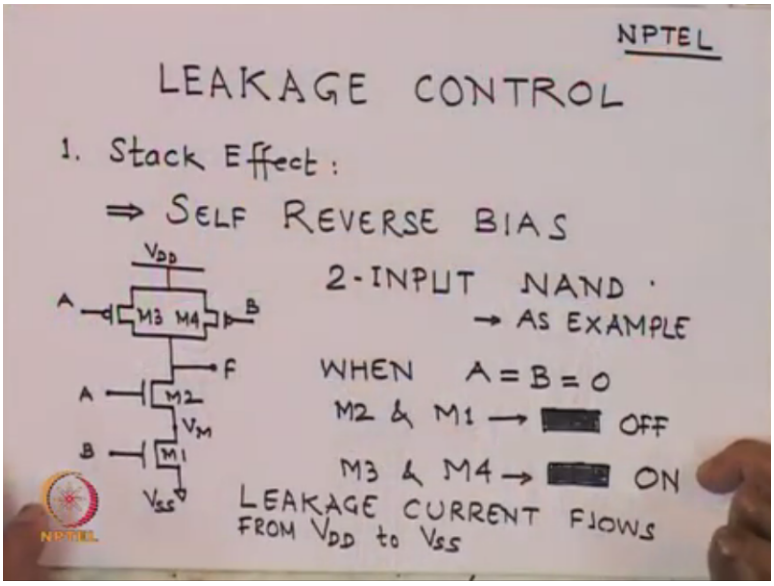
The fifth possibility is actually then if I already said it depends on the channel length, threshold voltage is their function of channel length this is called VT roll-off as it reduce the channel length. This is the old slide, but does not matter this can be further extended down to 45 or 30 nanometers, so VT further goes down. This roll-off of VT essentially means that if the threshold voltage goes down the leakage current will go up.

And therefore somehow we must see that instead of using a short channel or very small channel devices at least for the L circuit transistors, which do not have speed pressure that means they are not in critical paths. Their threshold can be higher and to improve their threshold one possibility is that for almost all possible data in which those transistor will not be in the critical path those transistor can be assigned larger channel lengths and therefore larger threshold voltage.

And because of that change in off current will be adjusted corresponding, so it will have lower leakage correspondingly. So I repeat the basic idea in all the techniques I am suggesting is to improve your threshold voltage whichever way you can for the transistor and larger the threshold voltage the corresponding of threshold slopes increases and because of that also the DIBL coefficient is smaller and because of that the off current starts reducing which in essentially is constituted by subthreshold current, which has become smaller in this case.

So before I go to summarize what I said let me go back again whatever I said so far I will re-talk about the same thing once again.

(Refer Slide Time: 27:46)



NPTEL

LEAKAGE CONTROL

1. Stack Effect:

⇒ SELF REVERSE BIAS

2-INPUT NAND.

→ AS EXAMPLE

WHEN A = B = 0

M2 & M1 → [redacted] OFF

M3 & M4 → [redacted] ON

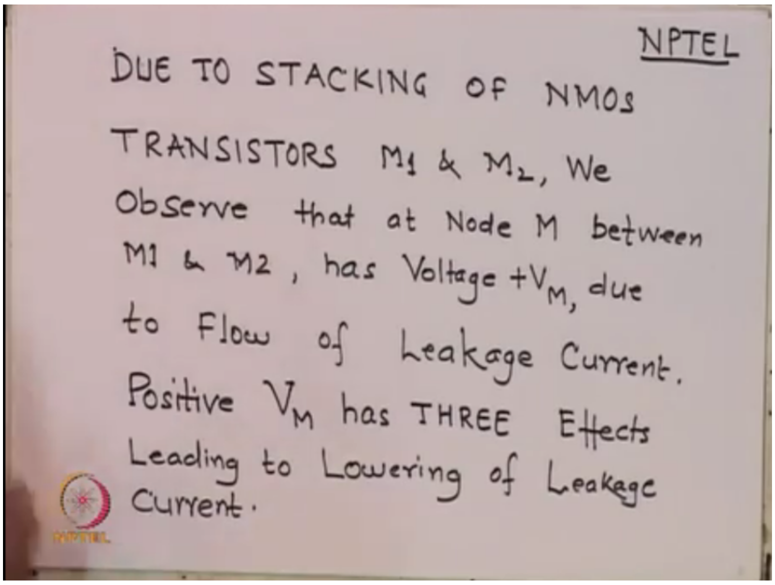
LEAKAGE CURRENT FLOWS FROM VDD to VSS

The first thing first effect we say when you stack the transistors okay. Now here is the NAND gate, which is two input NAND gate as an example shown here. I have a power supply, I have a ground, I have two P-channel devices in parallel for a NAND operation, which has inputs A and B, two N-channel transistors in series, which has inputs A and B and this is of my output for the NAND gate and this voltage, which I am talking is the intermediate node voltage VM.

Now let us take when A and B are 0 so obviously these are on and the output is going high as the NAND function once; however, M2, M1 are switched off as we think so. Since we thought that M1, M2 are switched off they are really in ideal case there should not have been any current, but we have just discussed that there are many amount possibilities of current leakage through M1, M2 even when VGS is less the VT or close to VT.

One sees this path M2, M1 the leakage flows through. So there is a leakage path even if VGS is less than VT. Now what happens due to stacking of NMOS we have just discussed this.

(Refer Slide Time: 29:02)

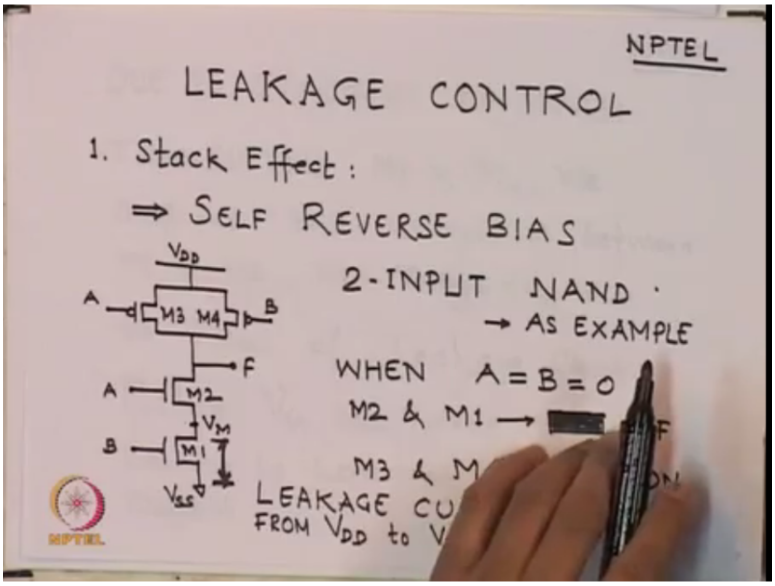


NPTEL

DUE TO STACKING OF NMOS TRANSISTORS M1 & M2, We Observe that at Node M between M1 & M2, has Voltage $+V_M$, due to Flow of Leakage Current. Positive V_M has THREE Effects Leading to Lowering of Leakage Current.

Transistor M1 and M2, which are N-channel transistors, at node N there is intermediate voltage VM, which is essentially occurring because of IR drop across this. I leakage into R off of this will give a voltage here VM so VM is occurring because there is a lower transistor and the upper transistor there is a current flowing in the leakage current flowing, which results in.

(Refer Slide Time: 29:44)



NPTEL

LEAKAGE CONTROL

1. Stack Effect:

⇒ SELF REVERSE BIAS

2-INPUT NAND ·

→ AS EXAMPLE

WHEN A = B = 0

M2 & M1 → F

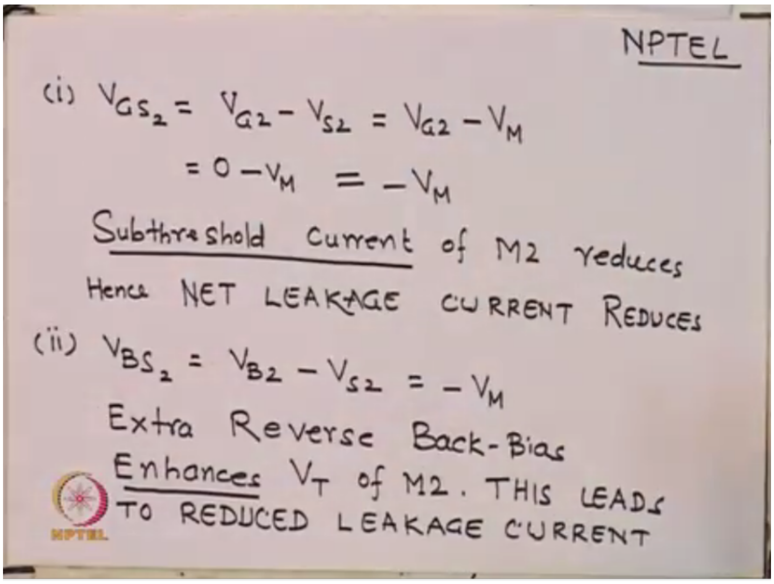
M3 & M ON

LEAKAGE CU

FROM V_{DD} to V

Please remember since the current flow is like this, this potential is always larger than the ground, which means VM is positive. Now what happens if VM is positive and it leads to three major effects. Okay so let me discuss about those three effects. For the transistor two, please remember I am having this as 1 and this as 2 I repeat I have this lower transistor as named M1 and the upper as M2.

(Refer Slide Time: 30:10)



NPTEL

(i) $V_{GS_2} = V_{G2} - V_{S2} = V_{G2} - V_M$

$= 0 - V_M = -V_M$

<u>Subthreshold current</u> of M2 reduces

Hence NET LEAKAGE CURRENT REDUCES

(ii) $V_{BS_2} = V_{B2} - V_{S2} = -V_M$

Extra Reverse Back-Bias

<u>Enhances</u> V_T of M2. THIS LEADS TO REDUCED LEAKAGE CURRENT

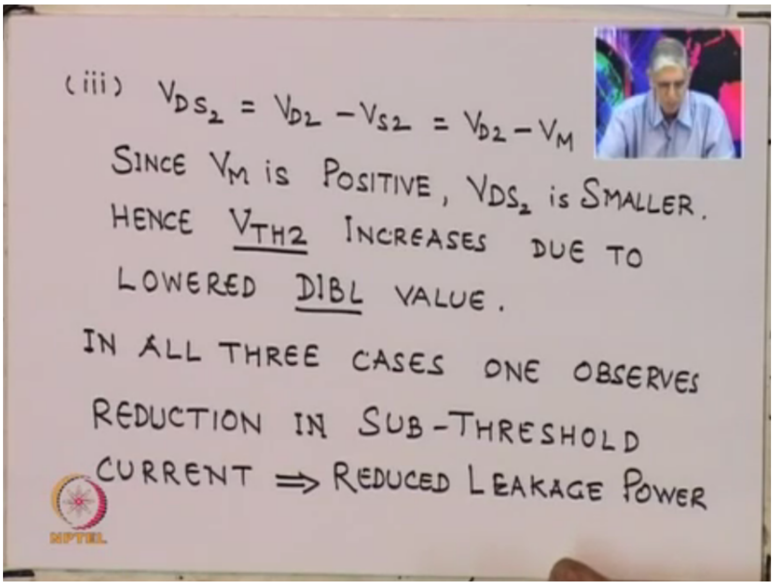
So in those cases the VGS2 that is for the input A VGS, VGS2 is nothing but VG2 minus VS2, which is essentially equal to VG2 minus VM, but since we are keeping VG2 very small close to 0 so we assume the worst case 0, 0 minus VM that means the VGS2 is minus VM. Now we know subthreshold current of M2 will be smaller if the VGS is negative value and since subthreshold current of M2 reduces obviously the net leakage current will reduce.

So the first the foremost simple thing has happened that the VGS effect has taken place, VGS has reduced to minus value and because of that M2 transistors have much smaller leakage and in a circuit only one current can flow if the M2 transistor have a smaller leakage M1 also can flow the same current and no more. Now the second effect, which is equally true for this. If you look at the bulk bias, which I have not shown here.

There is a bulk sitting here this B which is the bulk of this N-channel transistor so between bulk and source of this N-channel M2, there is a voltage now appearing 0 assuming that I actually ground the substrate. Then there is a minus voltage is again occurring as the reverse bias voltage is occurring at the substrate, which is equal to minus VM and we know any reverse bias enhances the threshold voltage.

And since the reverse bias enhances the threshold voltage VT of M2 increases, VT or VTH whichever I think I sometimes said H sometimes VT, but normally I say VTH of M2 enhanced and we know larger the threshold voltage the leakage currents goes down. We have just done that expressions, which shows that the leakage current reduces if VTs are larger. So there are two effects, the first of course is as I said because VGS effect we say second is back bias effect and third, which is not as simple as to be like this.

(Refer Slide Time: 32:47)



(iii) $V_{DS_2} = V_{D2} - V_{S2} = V_{D2} - V_M$

SINCE V_M IS POSITIVE, V_{DS_2} IS SMALLER.

HENCE $\underline{V_{TH2}}$ INCREASES DUE TO LOWERED DIBL VALUE.

IN ALL THREE CASES ONE OBSERVES REDUCTION IN SUB-THRESHOLD CURRENT ⇒ REDUCED LEAKAGE POWER

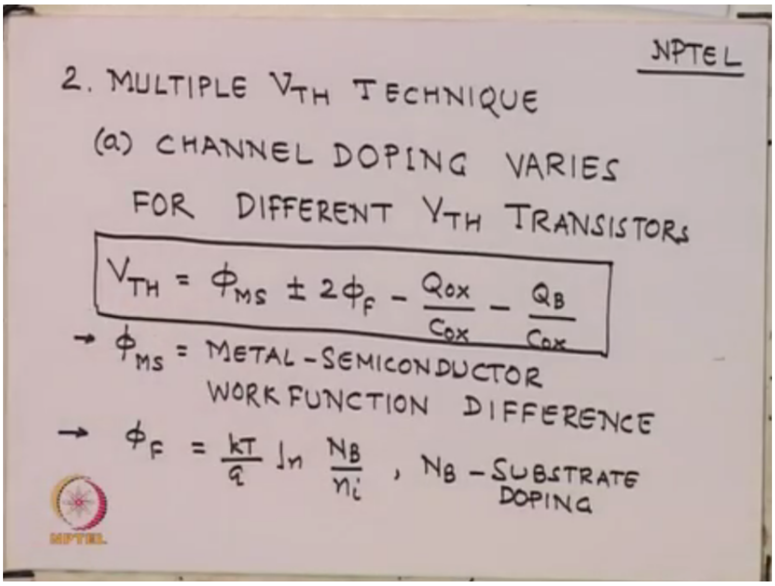
Third of course is if you look at the drain to source voltage of same transistor M2 it is VG2 minus VS2 so it is VG2 minus VM since VM is positive. The VDS2 is now smaller since VDS2 is smaller obviously the effect due to the drain, which is essentially because of the which say drain-induced barrier lowering, drain-induced barrier lowering will go down because your VDS has gone now.

If DIBL coefficient has gone down by the same expression, which I wrote earlier the threshold voltage rises, which means larger the VM you get whatever the three of the reasons I said in either case the threshold voltage will rise and increase of threshold voltage of M2 will actually reduce the leakage current. So one can see from there that the leakage power into subthreshold current can be minimized simply by seeing stacking the two devices.

And in the NAND functions this is natural, two of the transistor will always occur in series and because of that for worst case inputs of 0 0 one will see the smallest current going through it. Now when can see from here in case the situation is that one of the transistor in M2 is 1 and so you may require at actually breaking of the M1 transistor into two series transistor of widths W and still create the stack effect.

The net difficulty will be larger area and therefore probably some penalty you will pay for reducing the leakage power. Now other technique I discussed dual threshold I said dual threshold is a specific example multiple threshold techniques. Now threshold can be varied by number of ways, one of course is by what we call as change in channel doping.

(Refer Slide Time: 34:50)



NPTEL

2. MULTIPLE V_{TH} TECHNIQUE

(a) CHANNEL DOPING VARIES FOR DIFFERENT V_{TH} TRANSISTORS

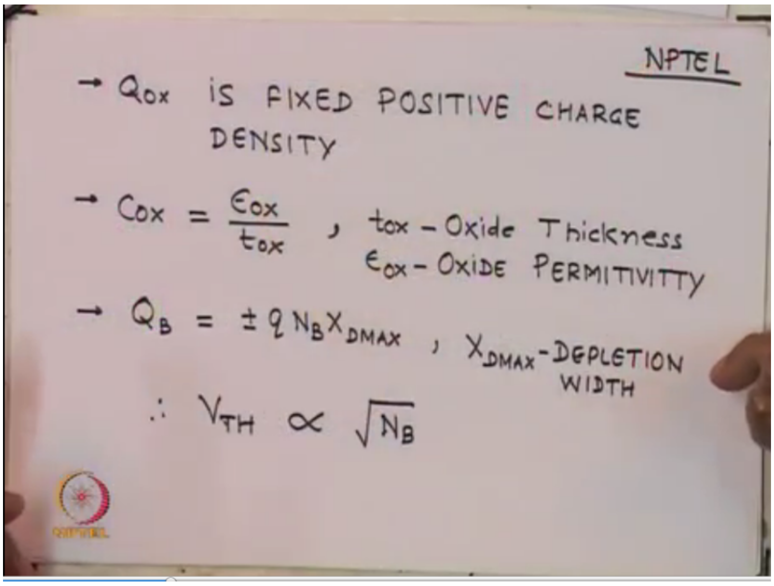
$$\boxed{V_{TH} = \phi_{MS} \pm 2\phi_F - \frac{Q_{ox}}{C_{ox}} - \frac{Q_B}{C_{ox}}}$$

→ ϕ_{MS} = METAL-SEMICONDUCTOR WORK FUNCTION DIFFERENCE

→ $\phi_F = \frac{kT}{q} \ln \frac{N_B}{n_i}$, N_B – SUBSTRATE DOPING

Now one can see from here I have recapitulate for you the expressions. The threshold voltage of N-channel or a NP-channel MOS transistor is phi MS plus minus 2 phi minus Q ox by C ox minus QB by C ox where phi MS is the metal semiconductor or dope poly semiconductor work function difference. Phi F is called the Fermi potential, which is kT by q ln NB by ni it is plus for p substrate and minus for n substrate where NB substrate doping.

(Refer Slide Time: 35:28)



NPTEL

→ Q_{ox} is FIXED POSITIVE CHARGE DENSITY

→ $C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$, t_{ox} – Oxide Thickness
ϵ_{ox} – OXIDE PERMITIVITTY

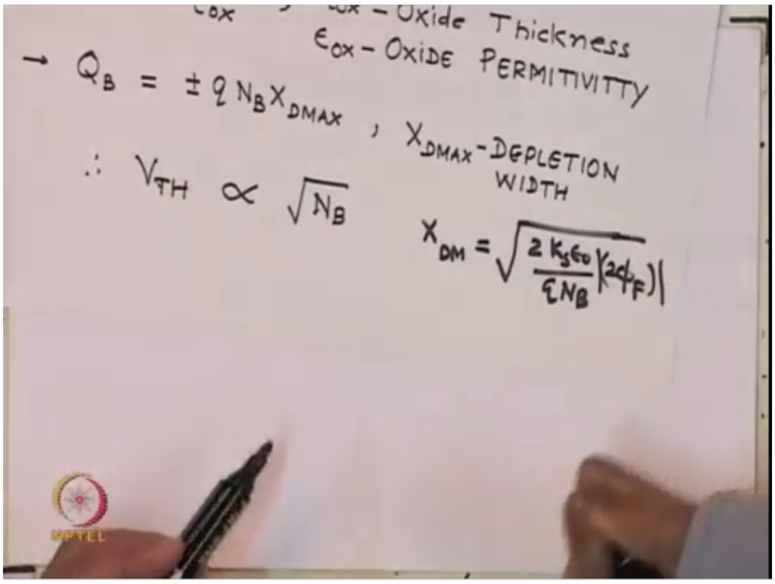
→ $Q_B = \pm q N_B X_{DMAX}$, X_{DMAX} - DEPLETION WIDTH

$\therefore V_{TH} \propto \sqrt{N_B}$

The Q ox is the fix positive charger, these days we are controlling extensively so minus Q ox by C ox I mean not so dominating, but it still existing is called the fixed charged density, Q ox is fixed charge density, C ox have source as oxide capacitance per unit area, which is epsilon ox oxide permittivity and T ox is the oxide thickness. Please remember this epsilon can be different for different dielectric insulators high-k dielectrics will have larger epsilon ox and therefore T ox can be proportionally increased to create the same C ox effect.

And finally the bulk charges prior to the threshold is qNB XDMAX, XDMAX essentially the maximum depletion width, which if you wish I can write an expression for the step junction at least kind of approximation.

(Refer Slide Time: 36:19)



t_{ox} – Oxide Thickness

ϵ_{ox} – OXIDE PERMITIVITTY

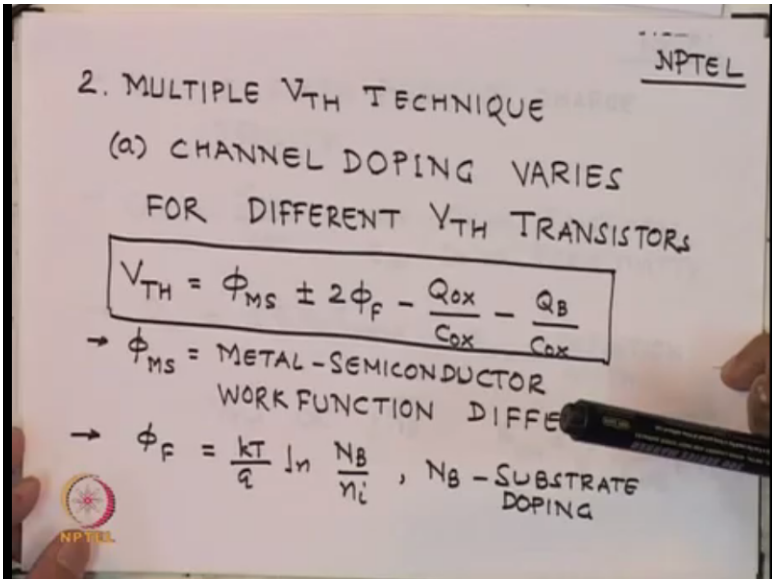
→ $Q_B = \pm q N_B X_{DMAX}$, X_{DMAX} - DEPLETION WIDTH

$\therefore V_{TH} \propto \sqrt{N_B}$

$X_{DM} = \sqrt{\frac{2 K_s \epsilon_0}{q N_B}(2\phi_F)}$

I can write XDMAX is under root of twice KS epsilon naught upon QNB substrate concentration into 2 phi F where 2 phi F is the twice the Fermi potential. Of course if you increase the reverse bias it will become 2 phi plus VHB and therefore it will increase the depletion layer and therefore increase the bulk charge.

(Refer Slide Time: 36:56)



NPTEL

2. MULTIPLE V_{TH} TECHNIQUE

(a) CHANNEL DOPING VARIES FOR DIFFERENT V_{TH} TRANSISTORS

$$\boxed{V_{TH} = \phi_{MS} \pm 2\phi_F - \frac{Q_{ox}}{C_{ox}} - \frac{Q_B}{C_{ox}}}$$

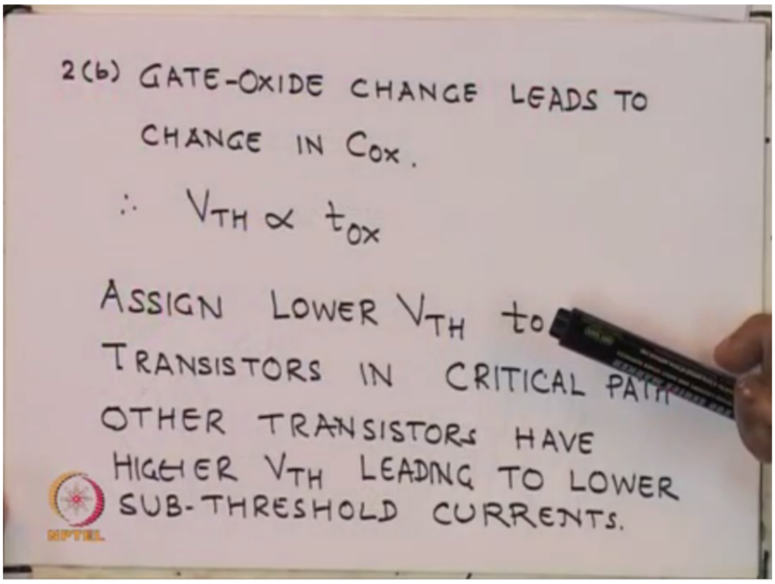
→ ϕ_{MS} = METAL-SEMICONDUCTOR WORK FUNCTION DIFFE

→ $\phi_F = \frac{kT}{q} \ln \frac{N_B}{n_i}$, N_B – SUBSTRATE DOPING

And in the threshold expression if you see if all of it you see the expression this increases with doping and therefore larger the doping larger is the threshold and because it is a root value typically VTH is directly proportional to increase in root of NB. So if you change the doping of the substrate or change near the channel, I can assure you that I can increase the threshold voltage.

So this exactly the first technique we say we can have different transistors have different doping you will remember this is additional masking going on additional cause going on, but all critical paths will have lower thresholds and non-critical paths will have higher threshold and they can be adjusted through the dopings in those transistors.

(Refer Slide Time: 37:48)



2(b) GATE-OXIDE CHANGE LEADS TO CHANGE IN C_{ox}.

$\therefore\ V_{TH} \propto t_{ox}$

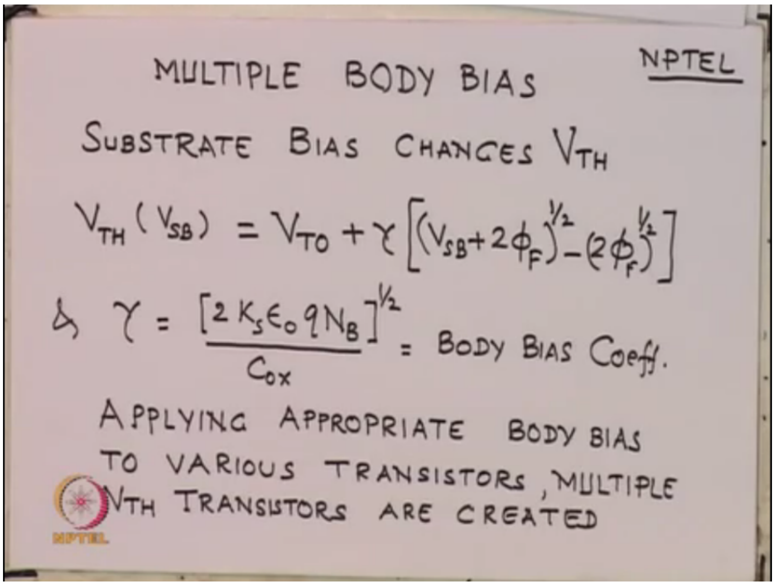
ASSIGN LOWER V_{TH} to TRANSISTORS IN CRITICAL PATH

OTHER TRANSISTORS HAVE HIGHER V_{TH} LEADING TO LOWER SUB-THRESHOLD CURRENTS.

The second possibility is the VT can be changed by C ox value and since C ox is epsilon ox by T ox the threshold voltage is proportional to T ox and since as of now T ox is reducing because of the threshold is reducing. So change in threshold can be adjusted by this so you can have transistors, which have multiple oxide thicknesses of the gate, larger oxide thicknesses will have larger thresholds, thinner oxide thicknesses will have smaller thresholds.

So one can keep as I said the theory is again and again the same besides assign lower threshold to transistors in critical path and assign higher threshold to a barrier, which are not so much speed dependent. So in any case those transistors, which are higher threshold will lead to lower leakage currents.

(Refer Slide Time: 38:42)



NPTEL

MULTIPLE BODY BIAS

SUBSTRATE BIAS CHANGES V_{TH}

$$V_{TH}(V_{SB}) = V_{TO} + \gamma\left[(V_{SB}+2\phi_F)^{1/2} - (2\phi_F)^{1/2}\right]$$

& $\gamma = \dfrac{[2K_s\epsilon_o q N_B]^{1/2}}{C_{ox}}$ = BODY BIAS Coeff.

APPLYING APPROPRIATE BODY BIAS TO VARIOUS TRANSISTORS, MULTIPLE V_{TH} TRANSISTORS ARE CREATED

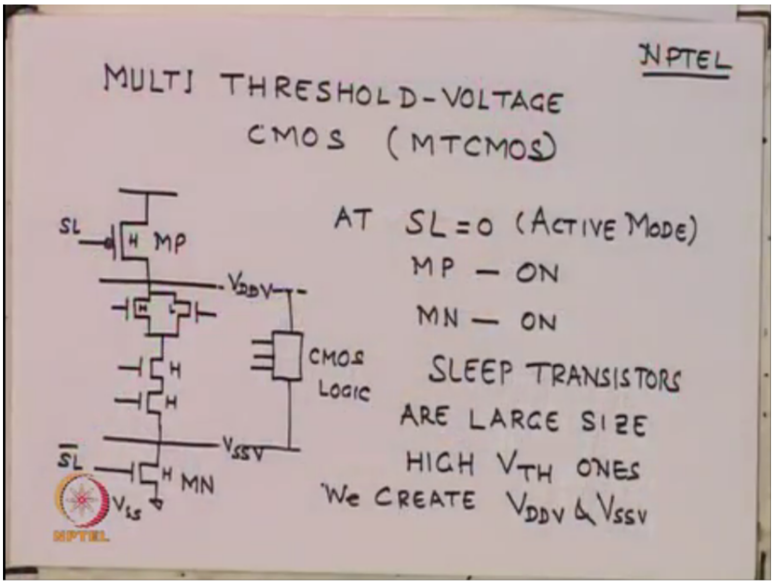
The other possibility we say is called multiple body bias. We already have seen the bias effects. Now here is another technique, which we say okay you can have multiple body bias. We know the threshold voltage with the back body bias if you applied to the substrate bias positive or generally negative.

We can see that the threshold voltage in any substrate bias is zero bias threshold voltage plus gamma times, which is called the back bias coefficient or body bias coefficient VSB plus 2 phi minus 2 phi where phi is the Fermi potential, kT by q ln Na or Nd by ni so one can see larger the VSB value particularly negative value that means plus added to this. One can see from here VT will rise on the contrary of this is negative value this will reduce the VT.

So by forward biasing or reverse biasing, I can change my VT corresponding to VTO initial value and accordingly can assign VT for different transistors. So now you decide should, should have larger thresholds apply larger reverse bias you apply smaller forward bias or even zero bias wherever you require lower thresholds and in that case you can even have multiple body bias for different speeds requirements and one can create number of VTs for different transistors to forego.

But this is something very crucial because if you do this way it is so much data dependent that every time for a different data someone has to do adaptively that technique to control and probably that is what the last I will show you that the end of the day the control will be more adaptive rather than fixed controls okay. Possibility of using multi-threshold CMOS is known and is called MTCMOS.

(Refer Slide Time: 40:42)



NPTEL

MULTI THRESHOLD-VOLTAGE CMOS (MTCMOS)

AT $SL=0$ (ACTIVE MODE)

MP — ON

MN — ON

SLEEP TRANSISTORS ARE LARGE SIZE HIGH V_{TH} ONES

We CREATE V_{DDV} & V_{SSV}

Here is an example shown here to you. You have a P-channel sleep transistor correspondingly you have N-channel sleep transistor and both have higher thresholds and we already said higher threshold means lower leakage currents. Then there is additional circuitry of P-channel and N-channel, which is kept here, which all are high threshold with some kind of a logic, which can be done on this resistance can be adjusted because they are multiple thresholds.

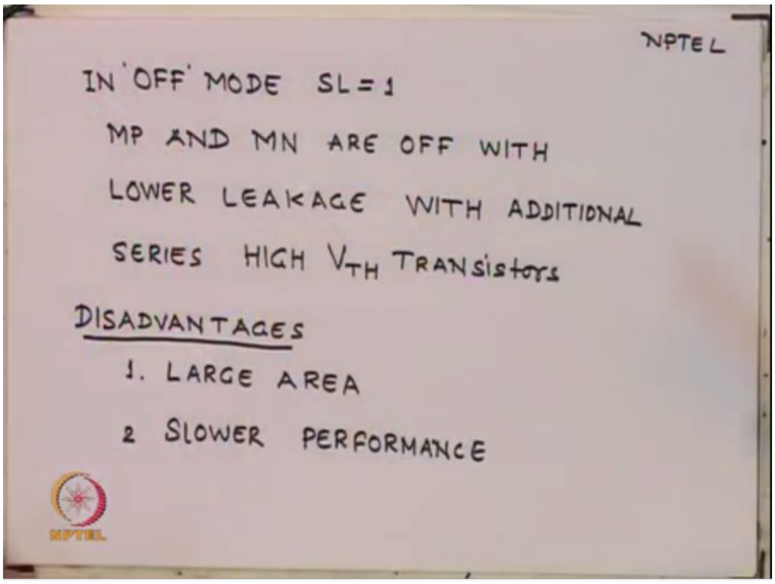
And one can create a typical R across this in fact by putting corresponding 0s and 1s. Now what happens when the sleep is 0, sleep bar is 1 both N-channel and MN and MP, which are P-channel and N-channel sleep transistor they are turned on and since their area is large area transistor W is very large, there are also higher thresholds. There is larger area means larger

W means smaller resistance, so we actually create for this logic, which is the real logic which we are using a VDD which is essentially called virtual VDDV and that is essentially VDD minus drop across the sleep transistor. Similarly there is a ground or VSS voltage, which is virtual ground voltage, which is again this minus this much.

So now I have figured out that I can change of course depending on the size and threshold adjust these voltage drops can be minimal they can be as close to VDD and VS depending on the sizing into correctly. However, all said and done this CMOS can have then adjusted VDDV and width. When the sleep goes to 1 and sleep bar goes to 0, these voltages are anyway available because of the drops across this.

And now this CMOS logic functions with virtual VDD and virtual VSS and therefore in the active mode you have a slightly lower VDD slightly higher VSS for the device so it may reduce speed a bit okay because your swings are smaller, but it will certainly reduce the leakage current when the sleep is on that means when you are not actually using the logic sleep can actually reduce the leakage currents.

(Refer Slide Time: 43:12)



IN 'OFF' MODE SL = 1

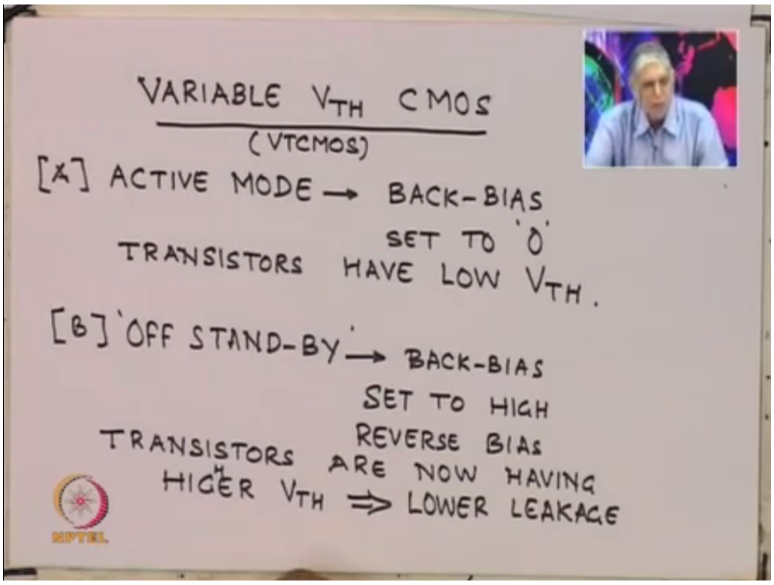
MP AND MN ARE OFF WITH LOWER LEAKAGE WITH ADDITIONAL SERIES HIGH V_{TH} TRANSistors

DISADVANTAGES

1. LARGE AREA
2 SLOWER PERFORMANCE

In a nutshell, what we will say is this advantages of this MTCMOS is it requires larger areas, we already said and to some extent it reduces the performance, speed goes down, we have just now discussed because the swing is smaller so there is reduction in speed.

(Refer Slide Time: 43:35)



VARIABLE V_{TH} CMOS

(VTCMOS)

[A] ACTIVE MODE → BACK-BIAS SET TO '0'

TRANSISTORS HAVE LOW V_{TH}.

[B] 'OFF STAND-BY' → BACK-BIAS SET TO HIGH REVERSE BIAS

TRANSISTORS ARE NOW HAVING HIGHER V_{TH} ⇒ LOWER LEAKAGE

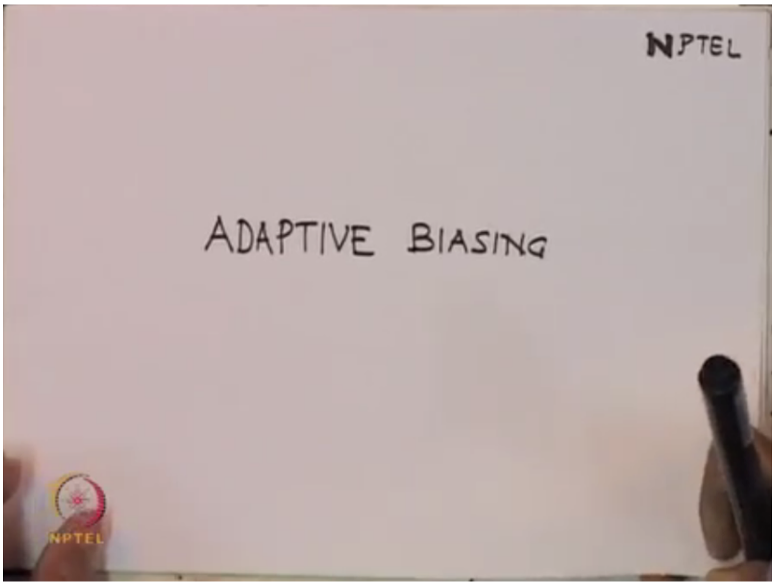
Now modification and so called multiple threshold CMOS is called variable VTHCMOS, it is similar as what we did just now, just not having the sleep transistors we can also have back bias also variable to U. In active mode, the back bias is normally set to 0 and in the case of standby mode back bias is given highest reverse bias and therefore the threshold of those transistors remain smaller in case of active mode, but can increase when the back bias is reverse bias because threshold rises, DIBL coefficient goes down and because of that the leakage current goes down.

So the problem is if I have a variable threshold, which I can do by biasing I can have different reverse biases for different transistors and therefore I can adjust the leakage current differently or essentially if I connected with my MTCMOS additionally with this then there is an advantage that I can have a sleep mode along with this back bias and two together can minimize the leakage current and can probably continue to have higher speeds in the active modes.

So this is way of doing so dual threshold is also among the multiple or variable threshold. If you only choose two values of threshold you say it is a dual threshold system, if you use as we have done there MTCMOS, if you have a variable then you say it is variable VTCMOS, all these are essentially similar, basically we are doing two techniques, one is using sleep transistors, the other is using back bias and if you do this together we have now a control on the leakage power.

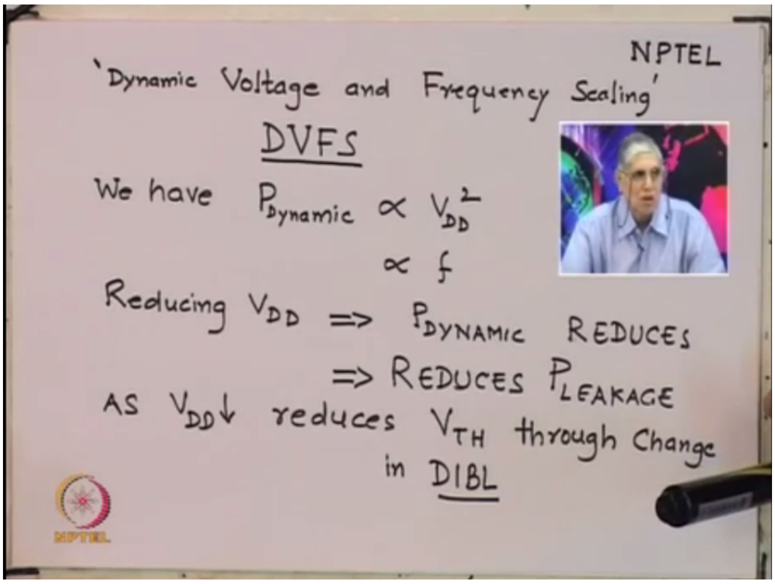
I just now said to you that to do this different data will have a different transistor on and off situation so the critical path may not be all the time same for all kinds of data in a data path particularly if you are looking a processor, you know with 64-bit inputs available to you, a design cannot be for a fixed threshold, yeah if you keep fixed threshold on and average power, leakage power will go down, but the better technique which essentially is now followed in the most processer like ARM and also in new Intel processors, which are trying to reduce the low power or even the 686 equivalent from the AMD.

(Refer Slide Time: 46:17)



All these are low power processors or low standby power in particular, but low power processors and since the effort all across the world is to reduce low power, the technique which is now being adopted from the based on whatever I have discussed so far is called adaptive biasing.

(Refer Slide Time: 46:32)



NPTEL

'Dynamic Voltage and Frequency Scaling'

DVFS

We have $P_{Dynamic} \propto V_{DD}^2$

$\propto f$

Reducing V_{DD} => $P_{DYNAMIC}$ REDUCES

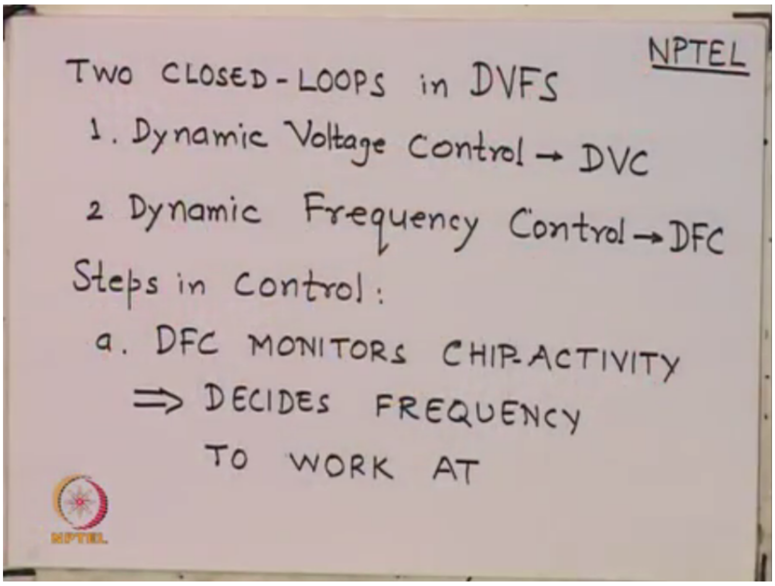
=> REDUCES $P_{LEAKAGE}$

AS $V_{DD}\downarrow$ reduces V_{TH} through Change in DIBL

The most common method which was first tried way back is called dynamic voltage and frequency scaling. We know that dynamic power is proportional to VDD square and dynamic power is also proportional to the clock frequency CV square alpha C effective VDD square into f. So if dynamic power increases with f and also increases square of VDD.

So one is quite clear that I will not like to reduce f because clock frequency I want to improve because I want to have performance. If f is not to be scale down, then the dynamic power is only this, but reducing VDC and of course will reduce leakage power also, but we know if we you reduce, VDD threshold can also be changed through a DIBL coefficient okay. This is the technique which we apply.

(Refer Slide Time: 47:28)



NPTEL

Two CLOSED-LOOPS in DVFS

1. Dynamic Voltage Control → DVC

2 Dynamic Frequency Control → DFC

Steps in Control:

a. DFC MONITORS CHIP-ACTIVITY

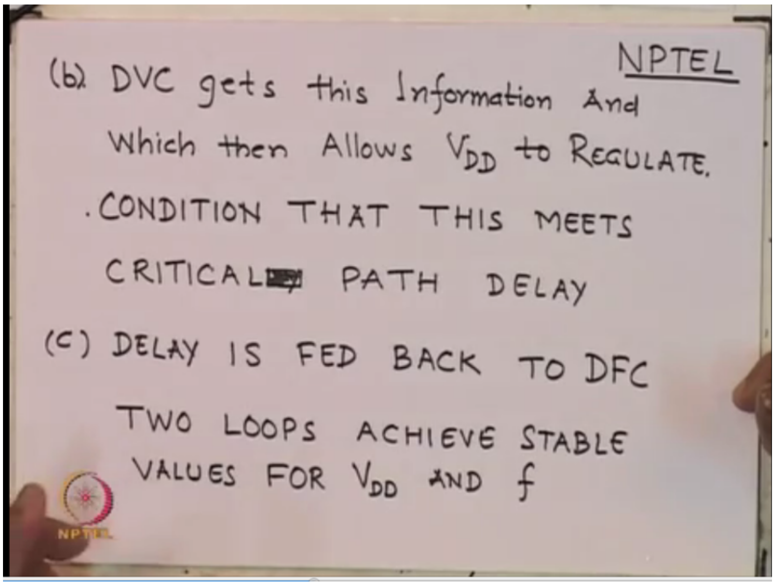
=> DECIDES FREQUENCY TO WORK AT

So we say two closed loops are available in control, one in the DVFS system direct dynamic voltage and frequency scaling, the one is dynamic voltage control. So depending on the data requirement and the power you are setting up, the voltage can be scaled down okay you are scaled up so is called DVC loop. The dynamic frequency control in which the speed which you already assign f is fixed for U.

So it actually fine for given voltage the speed for that speed even calculates what voltage is required goes back and keeps doing the two loops. So typically in a nutshell I can say there are few things and few steps in this control may be one, two, three, I will say few of them. I am not giving full detail, they are available in mini papers of recent origin of 2009, 10, 11, 12 last three four years.

The DFC monitors chip activity so what is frequency after all it decides when the data is different how much one zero transitions are taking place. Since it finds out the chip activity so it decides the frequency to work at.

(Refer Slide Time: 48:46)



NPTEL

(b) DVC gets this Information And Which then Allows V_{DD} to REGULATE.

. CONDITION THAT THIS MEETS CRITICAL PATH DELAY

(c) DELAY IS FED BACK TO DFC

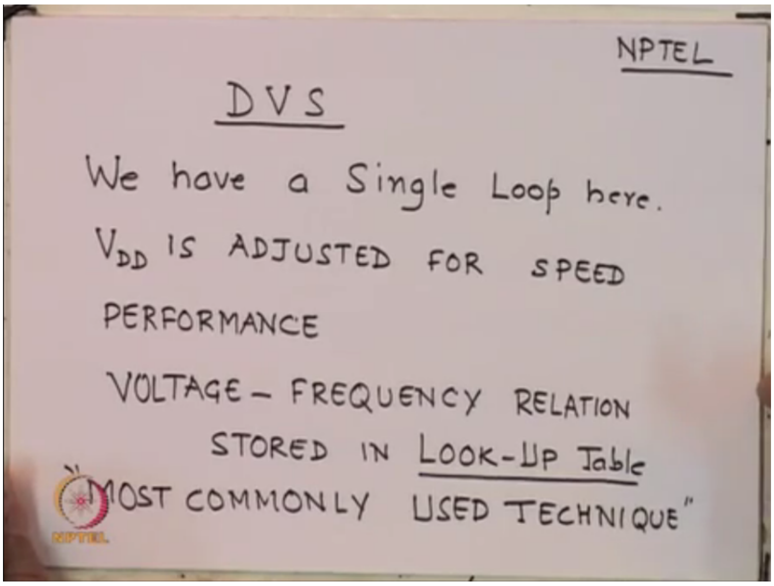
TWO LOOPS ACHIEVE STABLE VALUES FOR V_{DD} AND f

Now if you decide frequency to work at then the DVC that is the dynamic voltage control loop gets this information and which then allows VDD to change to corresponding to that frequency and the condition that it actually meets at least the critical path delay, which is your slowest path. Now if that meets for this voltage change (()) (49:02) this delay is again fed back to DFC.

And again the frequency it finds out from the activity and till the two loops get stable for a one value of VDD and f you automatically the system will work at some lower VDD on a given speeds. So this was the technique which was quite popular; however, please remember the cost here is they have two loops here and therefore little hardware cost and time. It will slightly slow down because it has to go through at least three or four times in the two loops to adjust to its value.

And the clock frequency therefore has to be reduced because it has to happen within that. Now we have a other technique, which is similar to what we said just now. So we say it is called dynamic voltage scaling okay DVS.

(Refer Slide Time: 50:01)



NPTEL

DVS

We have a Single Loop here.

V_{DD} IS ADJUSTED FOR SPEED PERFORMANCE

VOLTAGE – FREQUENCY RELATION STORED IN LOOK-UP Table

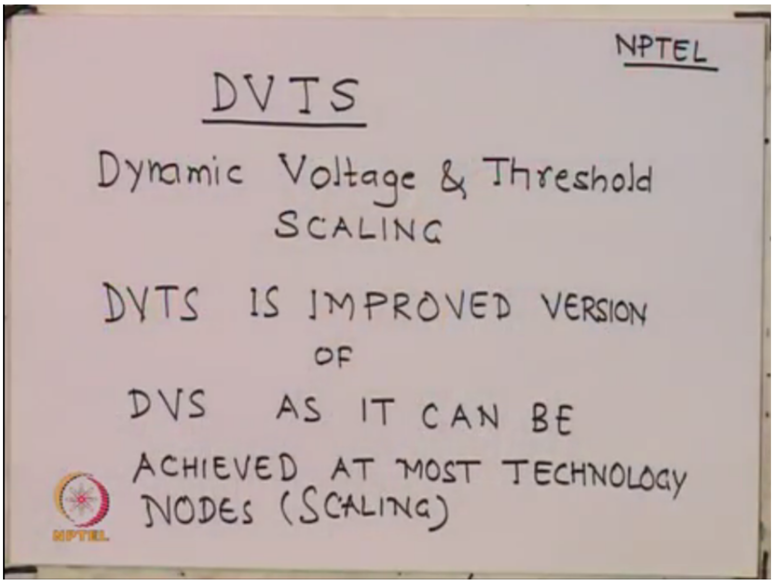
"MOST COMMONLY USED TECHNIQUE"

Now in this dynamic voltage scaling, we have a single loop of the voltage VDD is adjusted for speed, given speed is perfect so VDD is just adjusted and voltage frequency relation how do we know then. Then what do we do is instead of online doing the frequency power supply readjustment we actually create a look-up table initially okay and for both voltage and frequencies VDD and frequency values.

And as I change the VDD, I go into the look-up table and find what is the frequency I am operating and using this technique we can probably arrive at a reasonable value of VDD which will have given frequency requirements. Now this is called DVS, which is till date very commonly used technique for adaptive power biasing okay. Please remember this voltage is for two ways, one is the voltage we are giving it to power supply, the other voltage is into the back bias.

So both voltages are talked about though I am only giving in one word both back bias as well as power supply voltage are actually modulated as per the frequency requirements.

(Refer Slide Time: 51:26)



NPTEL

DVTS

Dynamic Voltage & Threshold SCALING

DVTS IS IMPROVED VERSION OF DVS AS IT CAN BE ACHIEVED AT MOST TECHNOLOGY NODES (SCALING)

Finally there is a not really new last three four years maybe I said dynamic voltage and threshold scaling. Now it is improved version of what we just talked about DVS that is dynamic voltage scaling as it can be achieved at why this was tried that in this technique the VDD also changes the threshold using the look-up table we know what voltages are required for substrate bias as well as for VDD to get the frequency of operation.

And for those voltages we know thresholds are varying so you adjust your threshold for leakage power okay. For those you figure out what frequency range you can attain and for this how much voltages you should apply between substrate bias and VDD. Of course this is again a loop system you need to go I say LAL hardware is called power management unit, which essentially creates different VDDs, different voltages and therefore allow you to add different thresholds at different points dynamically depending on the data as well as the architecture one uses.

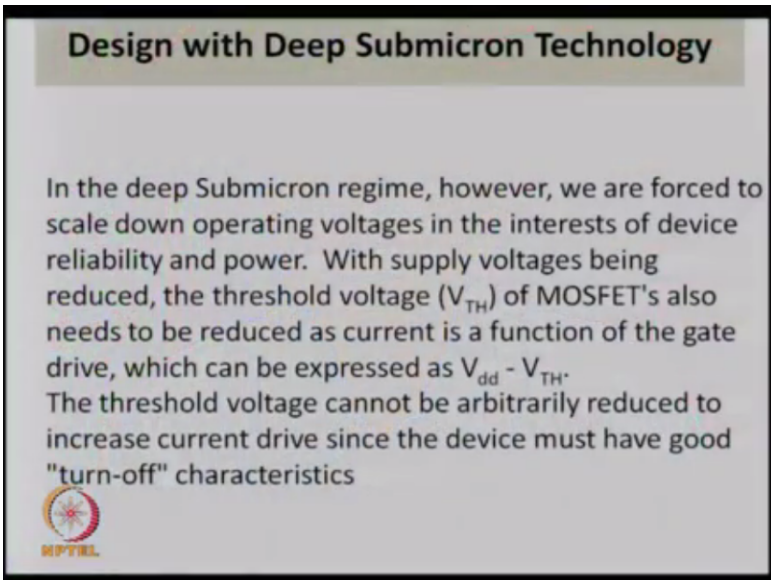
The biggest advantage of this algorithm DVTS algorithm, which does this you need a small processor to do this or small controller unit to do this, but if you achieve that, that gives something great advantage because it is then becomes independent of technology note, this technique can be applied for almost any kind of technology you go from 45, 32, 28 or 22 or 16.

So this is more likely of course whenever things are very good your additional hardware and the catch there is the cost of or the power dissipation in this additional control hardware should not exceed the power you are trying to save in the whole hardware otherwise if that happens then the whole purpose gets defeated. So having shown you variety of circuit techniques to reduce this leakage power as well as to see reduction in dynamic power.

Please also feel that this short circuit power is also proportional to VDD minus VT and also proportional to this W by L of N-channel to P-channel ratio. So one figures out that the short circuit current or short circuit power can be also minimize if the threshold rises. So short circuit power is not separately controlled. If you adjust your VDD or you increase your threshold then in either case both dynamic power and switching power can go down.

Of course because there is something to see that the rise and fall of the input pulse should be fast enough compared to the propagation delay, which of course is the device dependent phenomena, which is essentially decided by the full threshold technology control okay.

(Refer Slide Time: 54:24)

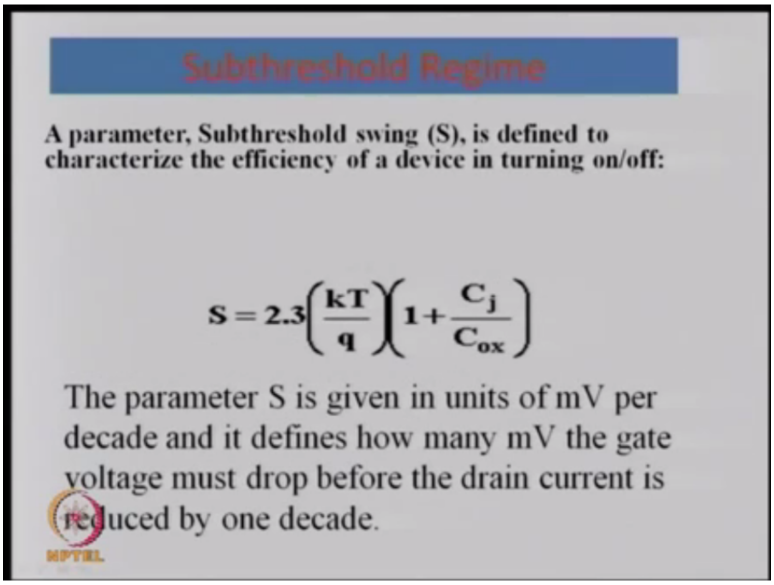


Design with Deep Submicron Technology

In the deep Submicron regime, however, we are forced to scale down operating voltages in the interests of device reliability and power. With supply voltages being reduced, the threshold voltage (V_{TH}) of MOSFET's also needs to be reduced as current is a function of the gate drive, which can be expressed as V_{dd} - V_{TH}.
The threshold voltage cannot be arbitrarily reduced to increase current drive since the device must have good "turn-off" characteristics

So coming back in nutshell what we can say if we are designing deep submicron that is below 45 or 65 nanometers, we are forced to scale down voltages in interests of device reliability and power with supply voltage being reduced, threshold voltage is also needs to be reduced as currents in their function of the drive. The threshold voltage cannot be arbitrarily reduced to increased current drive since the device must have good turn-off characteristics.

(Refer Slide Time: 54:47)



Subthreshold Regime

A parameter, Subthreshold swing (S), is defined to characterize the efficiency of a device in turning on/off:

$$S = 2.3\left(\frac{kT}{q}\right)\left(1+\frac{C_j}{C_{ox}}\right)$$

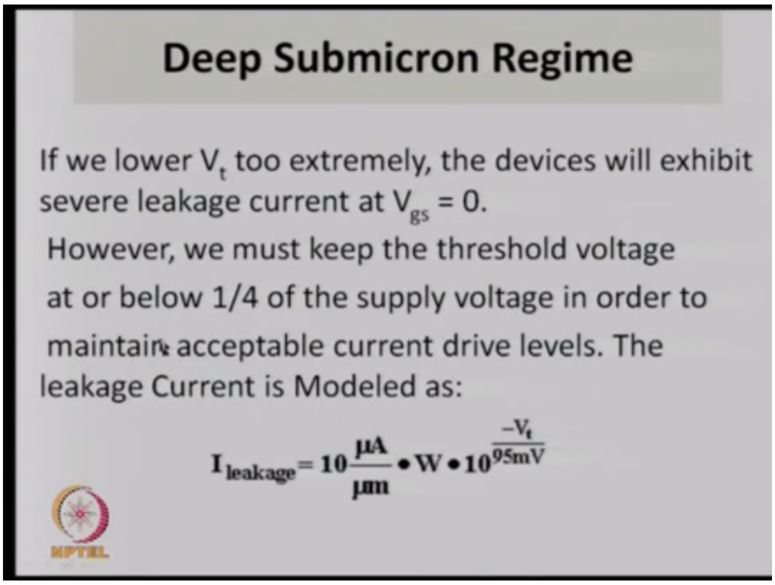
The parameter S is given in units of mV per decade and it defines how many mV the gate voltage must drop before the drain current is reduced by one decade.

The other possibility of worry which essentially one sees in turn-off is a parameter subthreshold swing is defined as the efficiency of a device to turning on to off is called subthreshold swing or slope S, which can be given as 2.3 kT by q1 plus CJ by C ox where CJ is the junction capacitor source drain to substrate, C ox is oxide capacitor. Typically this S is 60 millivolt per decade of voltages and millivolts, 60 millivolt per decade of current change.

One can see from here if I want faster turn-on to turn-off ratio I must have S larger and if that occurs one has to have some better device to because otherwise this short circuit currents will be larger in the case because turn-off to turn-on is not very fast. Please remember this larger is also means that the subthreshold current actually also is reduced at lower voltages and therefore the leakage power further goes down.

So this is another issue where one asked to worry about in DVFS or DVS or DVTS either of the regimes one must take care of in our algorithms how to adjust this S values.

(Refer Slide Time: 56:17)



Deep Submicron Regime

If we lower V_t too extremely, the devices will exhibit severe leakage current at $V_{gs} = 0$.
However, we must keep the threshold voltage at or below 1/4 of the supply voltage in order to maintain acceptable current drive levels. The leakage Current is Modeled as:

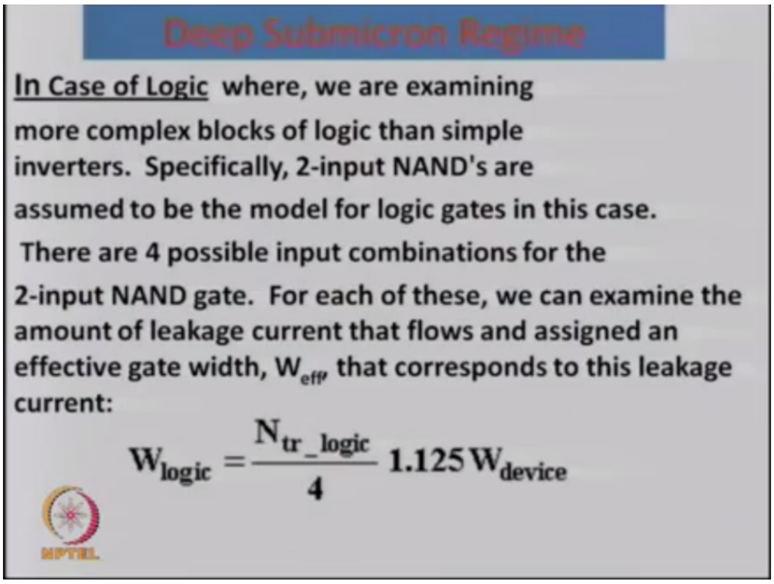
$$I_{leakage} = 10\frac{\mu A}{\mu m} \bullet W \bullet 10^{\frac{-V_t}{95mV}}$$

If we lower threshold to extremely the device will exhibit severe leakage currents at VGS is equal to 0. We have just discussed this; however, we must keep the threshold at or below one fourth of the supply voltage in order to maintain acceptable current rise because on current is also required so you cannot reduce too much or you cannot increase too much because if you want drive which is on current then you must have some amount of VGS minus VTP made available to you.

The leakage current model I have given in that expression but in simpler model I give it to you is I leakage is 10 micro m per micron into width of the transistor into tenth power VT threshold by 95 millivolt. Now what essentially is I am trying to say that by adjusting the widths of the transistors one probably can have an threshold voltages one can have both threshold voltage as well as the width one can adjust the leakage currents.

This is how you will require these models in your algorithms, so I thought I should provide you some models.

(Refer Slide Time: 57:23)



Deep Submicron Regime

In Case of Logic where, we are examining more complex blocks of logic than simple inverters. Specifically, 2-input NAND's are assumed to be the model for logic gates in this case. There are 4 possible input combinations for the 2-input NAND gate. For each of these, we can examine the amount of leakage current that flows and assigned an effective gate width, W_{eff}, that corresponds to this leakage current:

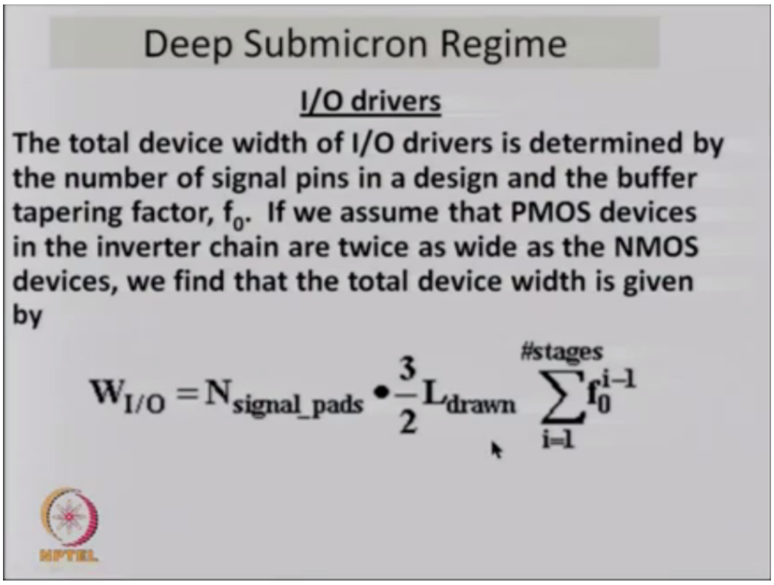
$$W_{logic} = \frac{N_{tr\_logic}}{4} 1.125 W_{device}$$

In case of logic where we are examining more complex block of logic rather than simple invertors let say for the case of two input NAND, which is not very complex, but which is much complex than in invertors. There are four possible input combination even for 2-input NAND gate 0 0 0 1 1 0 1 1 and for each of these we can examine the amount of leakage current that flows and is assigned in effective gate width.

Because for the worst case, we must find what should be the effective W which corresponding to a leakage current of which logically one can express as transistor logic. Number of transistors in this case is 4, 2 inputs and number of this is 2, this is 4 and this is actually derived from actual expressions, actual graphs this is a fit function system. So one can see from here the width of the logic is proportional width of device by this kind of expression.

So adjust in your series combinations the correct width so that you can have relatively good drive current at the same time you may have lower leakage currents.

(Refer Slide Time: 58:49)



Deep Submicron Regime

I/O drivers

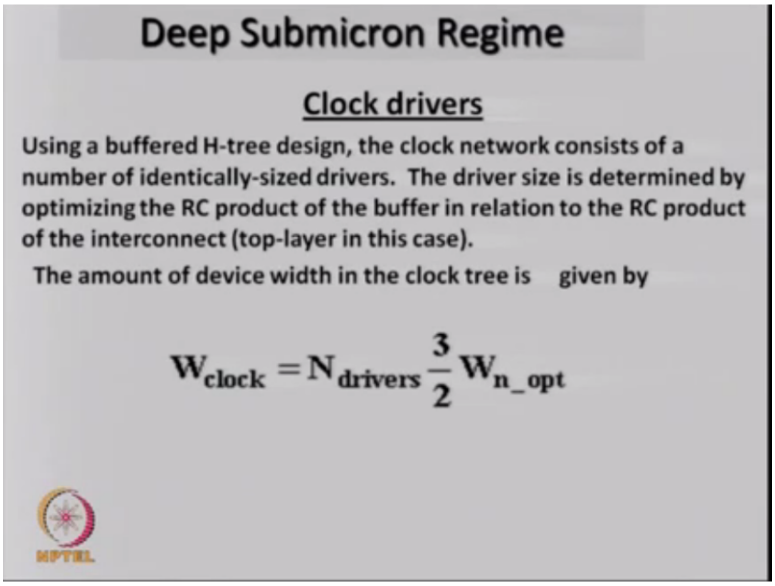
The total device width of I/O drivers is determined by the number of signal pins in a design and the buffer tapering factor, f_0. If we assume that PMOS devices in the inverter chain are twice as wide as the NMOS devices, we find that the total device width is given by

$$W_{I/O} = N_{signal\_pads} \bullet \frac{3}{2} L_{drawn} \sum_{i=1}^{\#stages} f_0^{i-1}$$

particularly for I/O drivers we can use the logical effort technique to go into the chain of them, but if you are using a single driver by similar arguments for N-channel and P-channel device operating a factor of which buffer is called the ratio of capacitor higher to low one can find the width of this proportion to how many paths you have, what is the actual length drawn from this to the device and how many stages you are going through for this buffer factor okay.

So I/O design is very, very difficult design though it said very trivially here, but one can see from here the at least the input and output buffers consumes a very large amount of power so many at times in our hurry to design a circuit we keep forgetting that there will be a huge loss of power at the two ends at the input and output and one must take care much more than the normal circuit design in this case so that the net power is minimized.

(Refer Slide Time: 59:30)



Deep Submicron Regime

Clock drivers

Using a buffered H-tree design, the clock network consists of a number of identically-sized drivers. The driver size is determined by optimizing the RC product of the buffer in relation to the RC product of the interconnect (top-layer in this case).

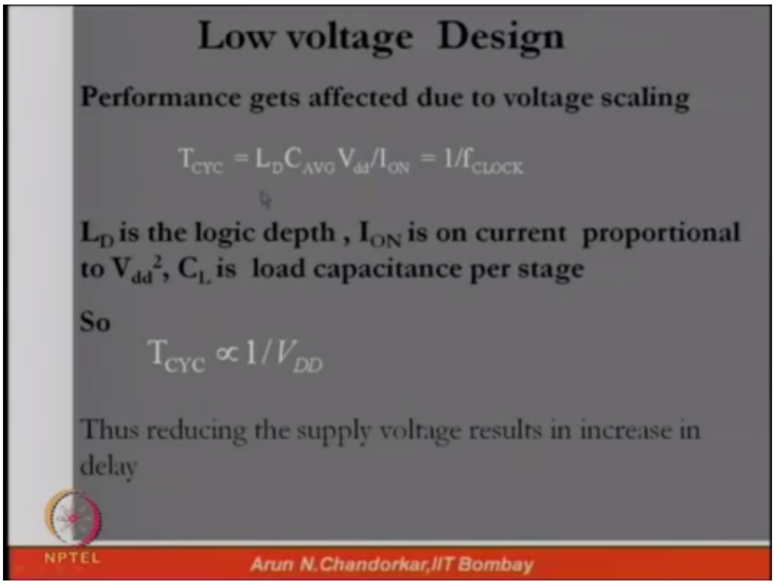
The amount of device width in the clock tree is given by

$$W_{clock} = N_{drivers} \frac{3}{2} W_{n\_opt}$$

Then another issue which is coming into clock drivers, the powers if they are too longer clock then you are putting a lot of capacity load so dynamic power is very large so the RC time constant of N interconnect through which the clock is moving as to be so adjusted and it should be at the highest layer of metal layer in the case of normal technology the maximum clock frequency which is allowed should be proportional so you figure out for that driver what should be optimal N-channel width.

So that it can give the required clock and you must always create H-tree for the case of clock distributions. All these are shown you to somehow to reduce the dynamic power in the case of additional circuits, which normal circuit designers at times do not realize that they may be the ones who may actually create large power dissipations. The other technique as I say where the limitations are coming.

(Refer Slide Time: 01:00:39)



Low voltage Design

Performance gets affected due to voltage scaling

$$T_{CYC} = L_D C_{AVG} V_{dd}/I_{ON} = 1/f_{CLOCK}$$

L_D is the logic depth , I_{ON} is on current proportional to V_{dd}^2, C_L is load capacitance per stage

So

$$T_{CYC} \propto 1/V_{DD}$$

Thus reducing the supply voltage results in increase in delay

NPTEL

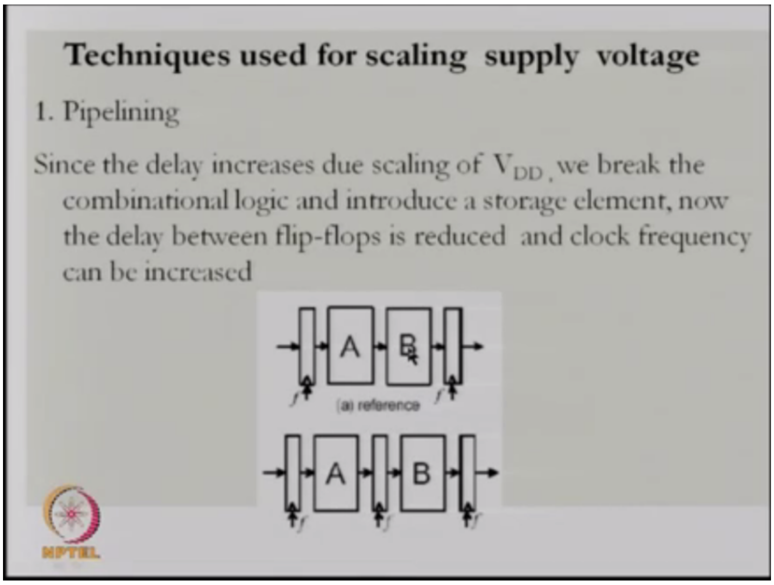
Arun N.Chandorkar,IIT Bombay

We are said that if you reduce VDD everything goes well fine. So here is something which one must look into before we do VDD reduction. The cycle time which is nothing but the logical depths that is in a chain of logic is called logical depth 1, 2, 3. If there is 3 gates are driven by each other then it is called depth of three. So LD is the depth of logic, C average is the average capacitance seen, VDD is power supply on current which essentially T cycle means the period of this one transition to go is one upon f clock.

C average is average capacitance of the load so one can see the T cycle is inversely proportional to VDD if I substitute correspondingly because we know I ON is proportional to VDD square so if I see T cycle of this. So obviously if you reduce VDD you want to have low voltage design, the first effect is your speed is going to be lower because your delay is going to increase do what you.

So the first thing when we said that you must do some kind of adaptive simply for this reason because reduction in VDD may be helpful in some way, but it does actually change your speed itself to a lower value.

(Refer Slide Time: 01:02:31)



Techniques used for scaling supply voltage

1. Pipelining

Since the delay increases due scaling of V_{DD}, we break the combinational logic and introduce a storage element, now the delay between flip-flops is reduced and clock frequency can be increased

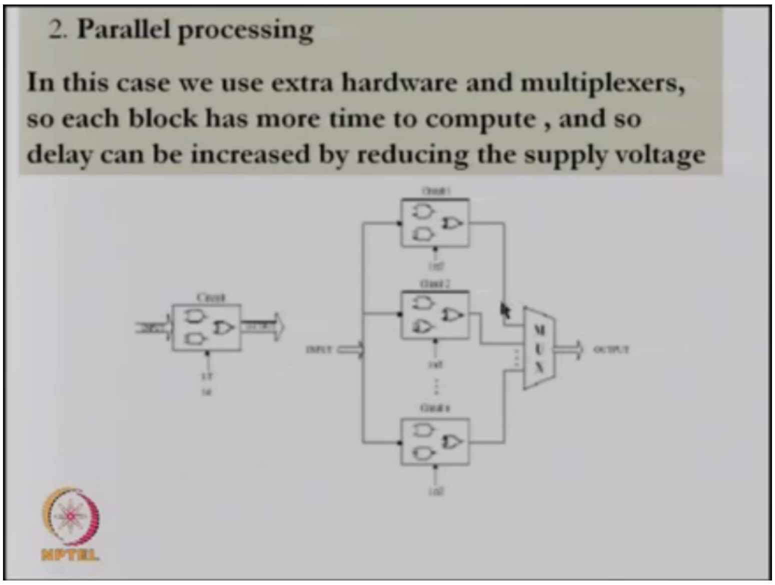
Now the other techniques of doing the low voltage this is essentially supply is called pipelining. Since the delay increases due to scaling of VDD, we break the combinational logic and introduce a storage element. For example these are your latches or flip-flops register, these are essentially this is your logic. So you actually this is your normal you have a data coming through a register.

And the logic is A and B may be NAND or whatever functions and or planes like FPGA or PLAs and then finally the output is given at another frequency which clocked at F2 or F same either non overlapping clocks or otherwise. So you have an input output registers and you are in logic. This is what normal reference is. Instead you put a input through a register through a logic A, put additional register in between A and B and do this.

Now it is quite obvious now the delay between flip-flop is reduced because you know this delay is essentially governed by there are two flip-flop delays; however, by increasing the F here please remember this is like putting in a pipe line at first data met a time to come out, but once data comes in every clock cycle you have the data and therefore one can see from here that here you can always have after only two clock cycles and in this case every clock cycle will have a data.

And since the VDD is reduced so your logic is going slow, but anyway does not matter because now the data flow is governed essentially by the pipelining system and every clock you have the output. This is essentially what is called throughput rate is available to you. So one method even if you have a scale VDD, the logic should be more like a pipe line data flow. The other possibility of course if you have lower VDD, you can have an extra hardware and many multiplexes to support you okay.

(Refer Slide Time: 01:04:16)



2. Parallel processing

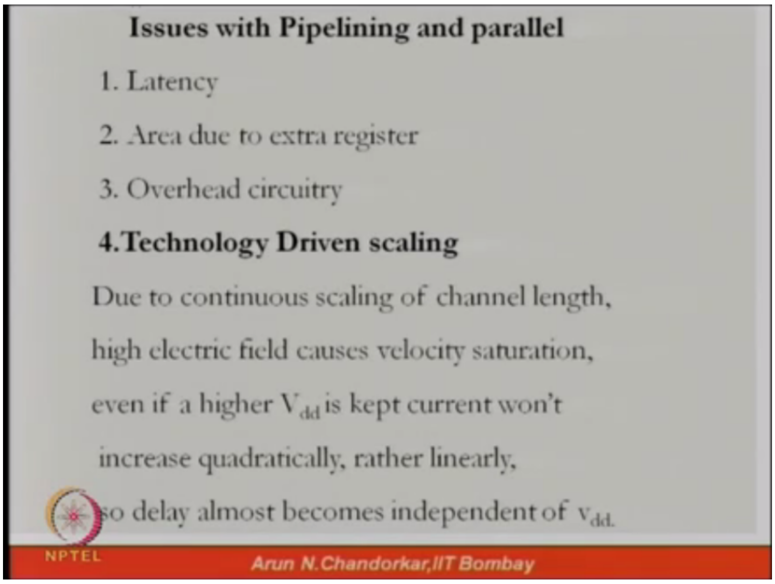
In this case we use extra hardware and multiplexers, so each block has more time to compute , and so delay can be increased by reducing the supply voltage

You can now divide your work into number of parallel paths and every clock cycle depending on the select here one of them may come. So this is essentially similar like pipelining, but only thing is that the data is partitioned by you and it needs a great effort to partition it equally okay at least their time for which this select signal changes this must occur correctly in this each of them should not exceed that.

So essentially the critical path among them will decide the select signal here; however, each can run at its own supply voltage because at the end of the day this MUX is going to decide the throughput and theretofore again the clock signal, which is driving the select signals and therefore in paralleling also you can have a lower voltage supply and we already said lower the voltage supply power is minimized in all three cases dynamic, switch as well as leakage.

And therefore you can do architectural thinking, these are called architectural thinking either use pipe lines or use parallel processing. What is the penalty you are paying? Additional hardware so the cache is that when you put any additional hardware can you afford it because if they consume power then you have the whole game lost. So one has to worry keep finding out what is the additional power you are going to spend to reduce the net power okay.

(Refer Slide Time: 01:05:59)



Issues with Pipelining and parallel

1. Latency
2. Area due to extra register
3. Overhead circuitry
4. **Technology Driven scaling**

Due to continuous scaling of channel length, high electric field causes velocity saturation, even if a higher V_{dd} is kept current won't increase quadratically, rather linearly, so delay almost becomes independent of $v_{dd.}$

NPTEL
Arun N.Chandorkar,IIT Bombay

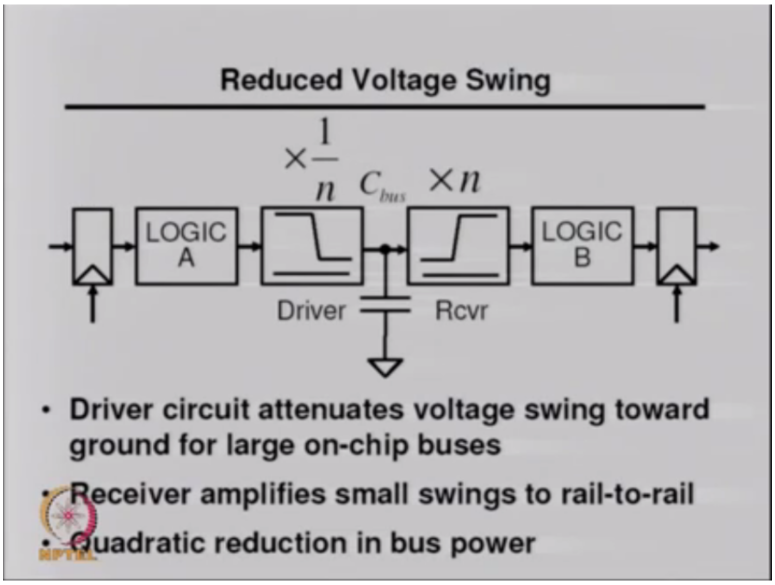
If any one of us are doing architectural power reduction, so here is the problems which you may see issues which are related to pipe line or paralleling. There is an issue of latency, latency essentially the net delay between input to output. So one must see to it that they for how many cycles data will not be available to you that is the depth of your register. So is that acceptable to you okay.

Because you know you cannot wait for even the first data output to arrive for any length of time so there is some latency issue and we also know from our general understanding of a system that the throughput rate and latency are somewhere related okay and therefore when newly designed any circuit using either of these architectural techniques one must take care of latency as well as throughput rate.

Obviously you are putting extra registers everywhere you may put the depth of register may not be one flip-flop or more than one flip-flop in that case your additionally area and this over rate circuitry will consume power and therefore how much one has to worry about. However, VDD is getting worrisome as I say is called technology driven scaling is creating a problem and which is why this pipe line was thought of.

Due to continuous scaling of channel length because VDD is not scaling in the same way the electric fields are increasing, velocity saturation effects are seen, even if higher VDD is kept, current will not increase quadratically rather than linearly so the delay almost becomes independent of VDD, which is very fantastic. That technology scaling actually is helping you to reduce power because you can then reduce the VDD anyway.

(Refer Slide Time: 01:07:57)



Reduced Voltage Swing

- Driver circuit attenuates voltage swing toward ground for large on-chip buses
- Receiver amplifies small swings to rail-to-rail
- Quadratic reduction in bus power

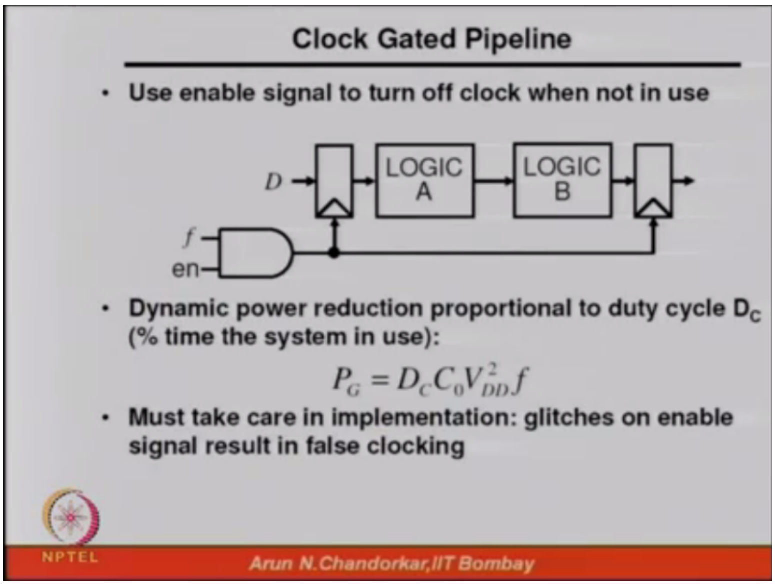
But there have to be combination of all kinds of techniques I discussed along with pipelining and paralleling and see to it that the net power is minimized at no cost of increase of delay. I just told you one of the method of reducing power if you have a smaller swing that is 0 to VDD instead of you may have VDD to 0.1 voltage swing, VSS is little above. Then one can say so for example your level may go down one upon n times this is your register pipe line, this is your logic, then you have a driver, which does not swing fully.

Then this is capacitor end then the receiver end you again have into n kind of this. Now you can see this, this nice margin should be sufficient it may come back to this old level to make logic B pass through register. Divert circuit attenuates voltage receiver amplifies back to the rail-to-rail and it can be found that the net power reduction in the bus. This is essentially used in the bus data transmissions.

This is essentially the bus part. Bus has the larger capacitance these days. Please remember I have not discussed so much, but interconnect is the major worry as of now for both speed and power, speed essentially because capacitance is larger so is R is larger these days because of the technology I am using, RC time constant is very high for the interconnect or the bus and because of that delay is going down.

So to improve this somehow and since C is larger obviously the power dissipation is larger. So to do this low power in the bus one technique is called reduced voltage swings. Please remember additional registers, additional driver receiver circuits how much power you consume on that, that decides whether to use this kind of structures.

(Refer Slide Time: 01:09:51)



Clock Gated Pipeline

- Use enable signal to turn off clock when not in use

- Dynamic power reduction proportional to duty cycle D_C (% time the system in use):

$$P_G = D_C C_0 V_{DD}^2 f$$

- Must take care in implementation: glitches on enable signal result in false clocking

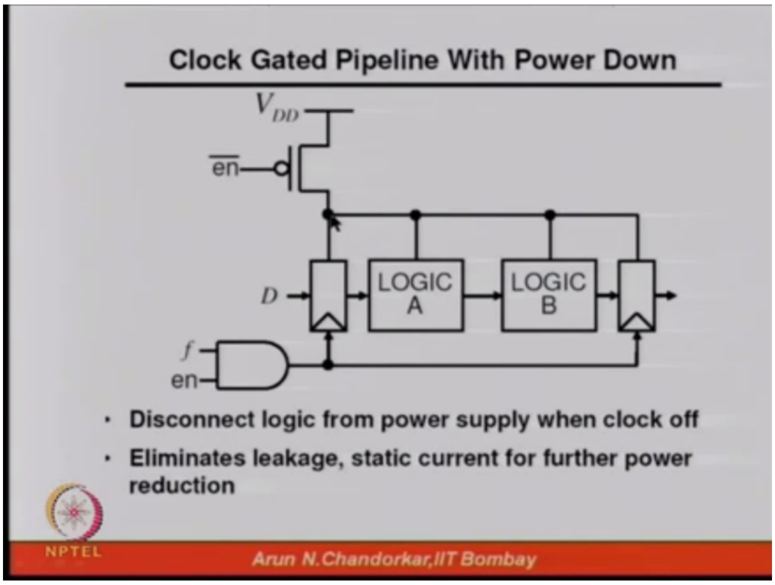
NPTEL

Arun N.Chandorkar,IIT Bombay

Then there is a possibility of clock gated pipe line use you know enable signal to turn off clock see after all clock is driving it all the time okay. So when the clock is not required, the clock is fed through a signal and en gate which one input is enable. If your enable goes 0 the clock does not go up. So essentially the logic can be held to this both input output register can be switched off when no data is expected to go through.

The other is dynamic power reduction is proportion to duty cycle how long the system is used or activity coefficient. So if I stop working here when I am not needed, then obviously I am reducing the power. On the next part we shall see whenever we do on off the major worry is what we call glitches okay. This may result in what we call false clocking.

(Refer Slide Time: 01:10:45)



Clock Gated Pipeline With Power Down

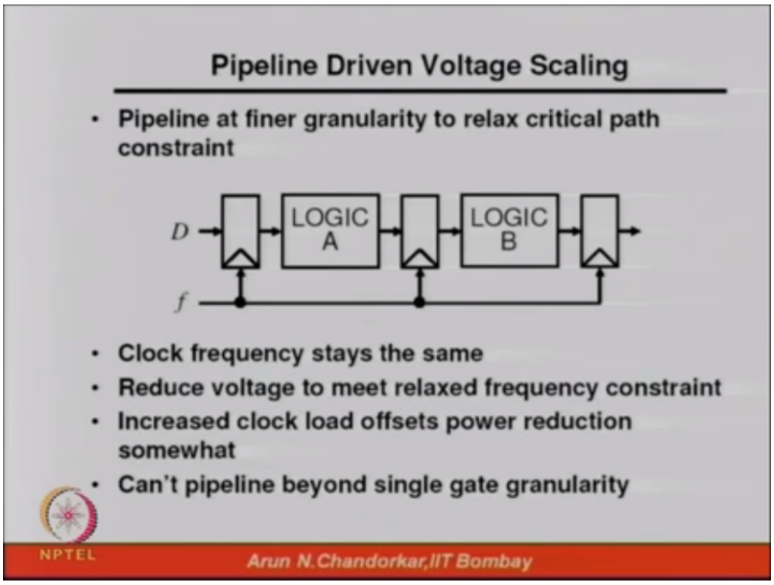
- Disconnect logic from power supply when clock off
- Eliminates leakage, static current for further power reduction

Arun N.Chandorkar,IIT Bombay

There is a clock gated pipe line with further power down so you can have this enable signal coming through a P-channel gate, enable bar coming here and all this power supply voltage of themselves can be reduced as I said you this is like a sleep transistor a voltage drop there is a lower VDD made available to you, power down, VDD is VDD virtual here and enable of course will stop the signal as if enable bar is 0 enable is 1, it is like an active mode with lower VDD okay.

And when enable is 0, enable bar is 1, this is switched off and the full logic is off. So essentially you disconnect the logic from power supply when the clock is off eliminates the leakage as it cause the sleep transistor is very large so P-channel device leakage is very small and because of that one sees that the leakage power also goes down. So it reduces the net power.

(Refer Slide Time: 01:11:50)



Pipeline Driven Voltage Scaling

- Pipeline at finer granularity to relax critical path constraint

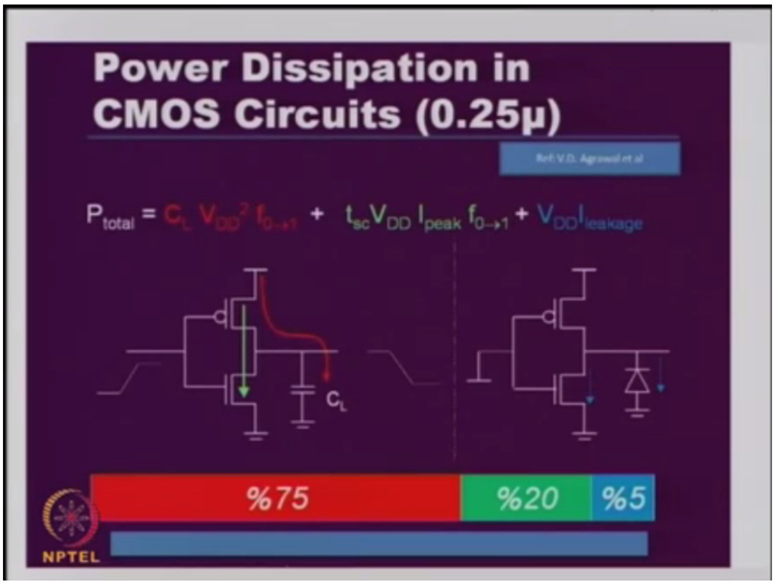
- Clock frequency stays the same
- Reduce voltage to meet relaxed frequency constraint
- Increased clock load offsets power reduction somewhat
- Can't pipeline beyond single gate granularity

Arun N.Chandorkar,IIT Bombay

The next of course is now this is what I just discussed. If I put the register in between the advantage that I am saying clock frequency stays the same in the case of pipe line driven voltage scaling, reduced voltage to meet relaxed frequency constraints, increased clock load offsets power reduction somewhat, cannot pipe line beyond a single gate granularity this is requirement, but this is how one can probably use pipe lining.

You cannot have too long depths of pipe line because then the delay will never be correct, the latency will be very, very large okay.

(Refer Slide Time: 01:12:26)



Power Dissipation in CMOS Circuits (0.25μ)

Ref: V.D. Agrawal et al

$P_{total} = C_L V_{DD}^2 f_{0\to1} + t_{sc} V_{DD} I_{peak} f_{0\to1} + V_{DD} I_{leakage}$

Now the last part of this power which is not really last last, but one may be. We are talking now the power dissipation in a CMOS circuit. An example is taken from a VLSI design conference and other conference papers through Prof. VD Agrawal's group at Auburn

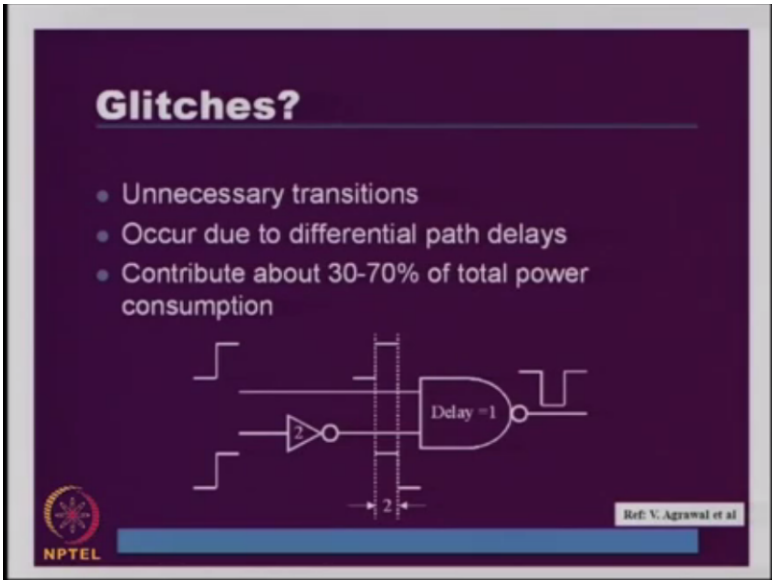
University and these slides are provided by him to me. Of course they are probably available on website as well okay. You can go Google at name VD Agrawal or Vishwani Agrawal and probably you may get some of those.

A typical CMOS charging discharging transient is shown here and we see dynamic power is because of charge discharge, but even if this transistor is on or off if this is 0 or this is high, there is a static current going through it initially at least and during that this device has to supply current here and here, same way discharge has to come through here. So the net power dissipation we discussed so far if going from 0 to 1 transition is CL VDD square f 0 to 1.

Then short circuit current is time for what short circuit occurs VDD into I peak into f going from 0 to 1 at what times and finally of course the leakage when the device so called is switched off and this is really not off so the leakage power and in old technologies of 0.25 this what I said you 75% dynamic power, 20% in short circuit power and only 5% in leakage power.

And I have shown you earlier some table, which shows in 32 nanometer node this is becoming 60% or 70% of the power and this is 30%. So now one is worry about because if this power increases what do we do, but that apart which we already seen how to reduce leakage power; however, the worry which is not so much shown in the dynamic power is right now this issue.

(Refer Slide Time: 01:14:36)



Glitches?

- Unnecessary transitions
- Occur due to differential path delays
- Contribute about 30-70% of total power consumption

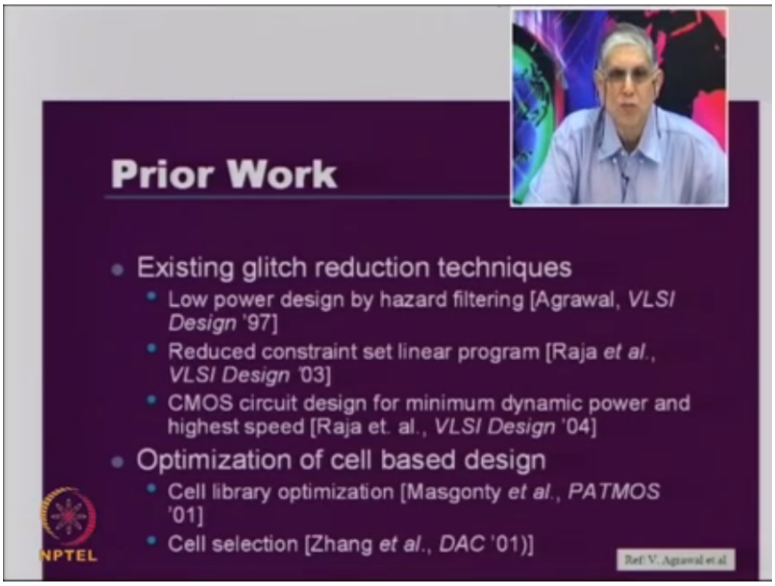
Ref: V. Agrawal et al

Here is the circuit you know some unnecessarily transition is essentially called glitch. If you have a logic which this probably if you are remember I already discussed these issues particularly in the case of logical effort equal delay system, but if in case there are no equal delays between this and this, one can see there will be one additional transition occurring here which may result in the long output and you need not have actually switching over, but it may switch.

And this switching may occur which is not expected and that may lead to as high a power consumption as much as about 30% to 70% and this glitch power is actually coming up now very much essentially because your frequency of operation is going to give you hertz and smallest line delays itself can actually cause the glitches. Please remember these are after all a metal lines and line delays.

Different lengths of metal lines or poly lines can create the self-delays and that may give a huge glitch power. Now these are some papers, which Vishwani Agrawal has stated about so one can go and look at this.

(Refer Slide Time: 01:15:59)

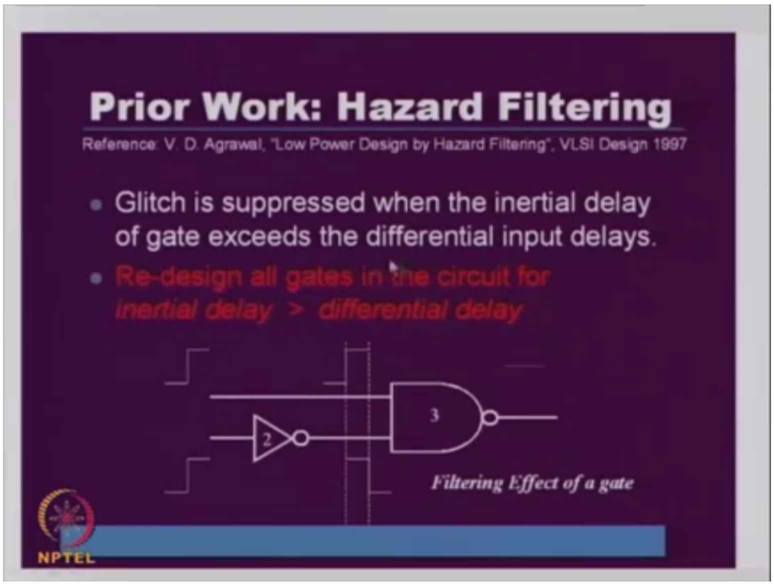


Prior Work

- Existing glitch reduction techniques
  - Low power design by hazard filtering [Agrawal, *VLSI Design* '97]
  - Reduced constraint set linear program [Raja *et al.*, *VLSI Design* '03]
  - CMOS circuit design for minimum dynamic power and highest speed [Raja et. al., *VLSI Design* '04]
- Optimization of cell based design
  - Cell library optimization [Masgonty *et al.*, *PATMOS* '01]
  - Cell selection [Zhang *et al.*, *DAC* '01)]

Their essential effort was for optimization of cell based designs how to improve the cell selection etc, etc for low power glitch powers. So this is their earlier work okay. This sheets you can always collect as you will have that. So this is essentially techniques, which are available in this.

(Refer Slide Time: 01:16:25)



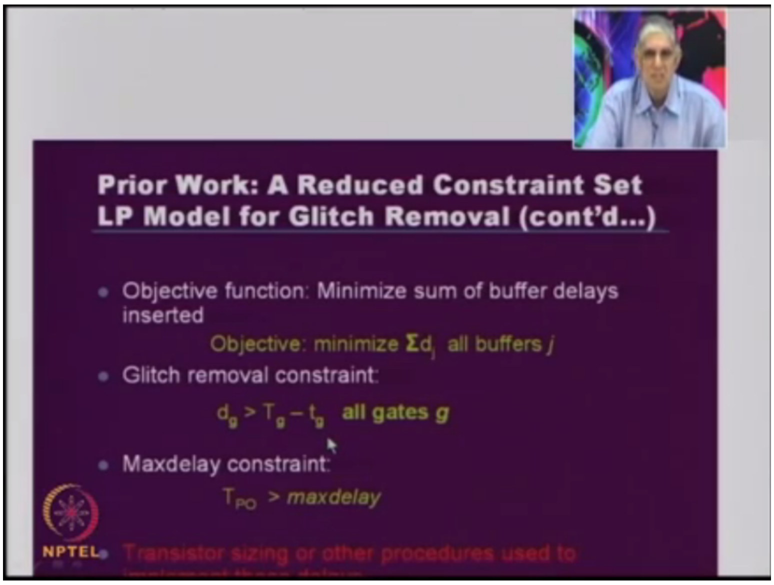
Prior Work: Hazard Filtering

Reference: V. D. Agrawal, "Low Power Design by Hazard Filtering", VLSI Design 1997

- Glitch is suppressed when the inertial delay of gate exceeds the differential input delays.
- Re-design all gates in the circuit for *inertial delay > differential delay*

You can see here the redesign all gates, glitch is suppressed when the inertial delay of gate exceeds the differential input delays. So redesign all gates in the circuit for inertial delay, which is greater than the differential delay. If you do that that is essentially what this is a old paper which is available in VLSI design conference is called filtering effect okay.

(Refer Slide Time: 01:16:50)



Prior Work: A Reduced Constraint Set LP Model for Glitch Removal (cont'd...)

- Objective function: Minimize sum of buffer delays inserted

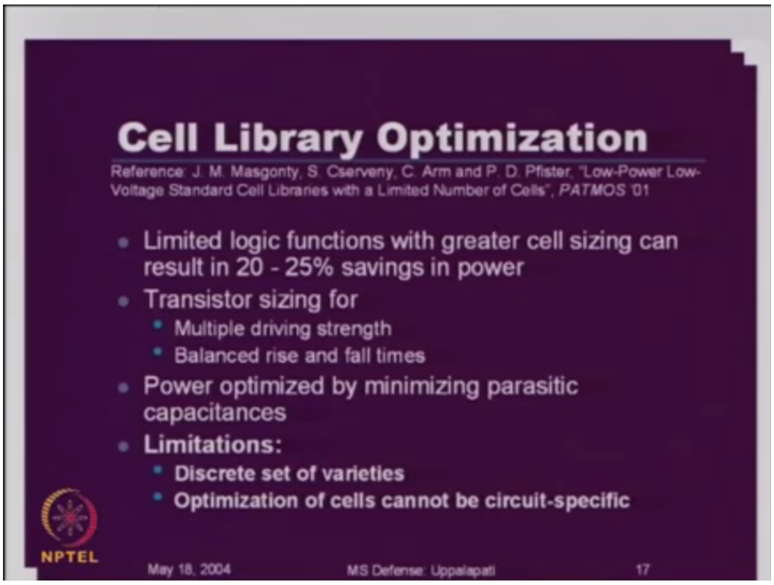
  Objective: minimize Σd_j all buffers *j*
- Glitch removal constraint:

  $d_g > T_g - t_g$ **all gates *g***
- Maxdelay constraint:

  T_{PO} > *maxdelay*
- Transistor sizing or other procedures used to

So as I said there is already prior work done by many others including Vishwani. The method is objective function minimize some of the buffer delay is inserted, objective minimized net delay for all buffers J, glitch removal constraint Dg should be greater than Tg minus tg for all gates g, maximum delay should be smaller than the net propagation delay. Therefore new transistor sizing and procedures are to be used.

(Refer Slide Time: 01:17:24)



Cell Library Optimization

Reference: J. M. Masgonty, S. Cserveny, C. Arm and P. D. Pfister, "Low-Power Low-Voltage Standard Cell Libraries with a Limited Number of Cells", *PATMOS* '01

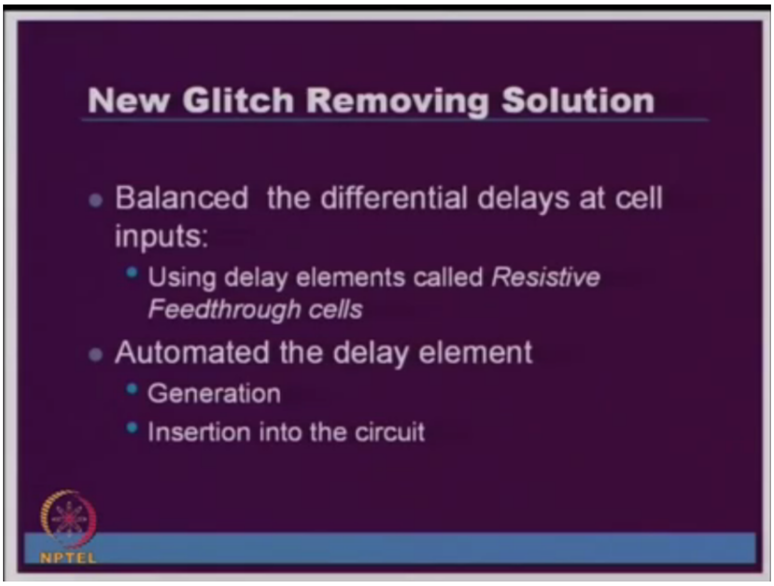
- Limited logic functions with greater cell sizing can result in 20 - 25% savings in power
- Transistor sizing for
  - Multiple driving strength
  - Balanced rise and fall times
- Power optimized by minimizing parasitic capacitances
- **Limitations:**
  - **Discrete set of varieties**
  - **Optimization of cells cannot be circuit-specific**

May 18, 2004 MS Defense: Uppalapati 17

We can see you can do cell optimization as they are suggested transition sizing can be again multiple driving strengths, balanced rise and fall times, power optimized by minimum parasitic capacitances. Of course there is a discrete set of variety is possible. You create a different cells, which can give options then in the case of normal design and then you know particularly the cells are not very circuit specific.

For all the possible hardware number of cells available may not be sufficient so that may be large cost.

(Refer Slide Time: 01:17:59)

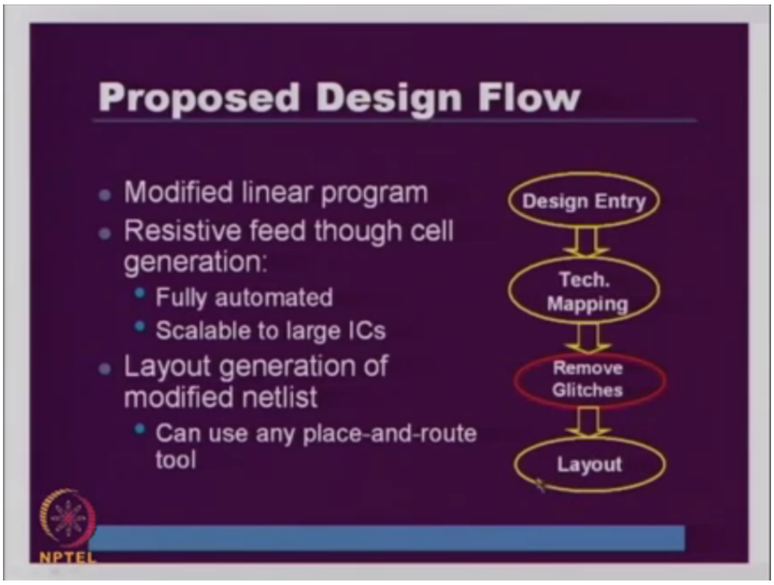


New Glitch Removing Solution

- Balanced the differential delays at cell inputs:
  - Using delay elements called *Resistive Feedthrough cells*
- Automated the delay element
  - Generation
  - Insertion into the circuit

New glitch removing solutions balance the differential delays at cell input itself, which is called feedthrough cells, automate the delay element, generate and insert into the circuits and if you do this then probably a glitch can be minimized.

(Refer Slide Time: 01:18:13)



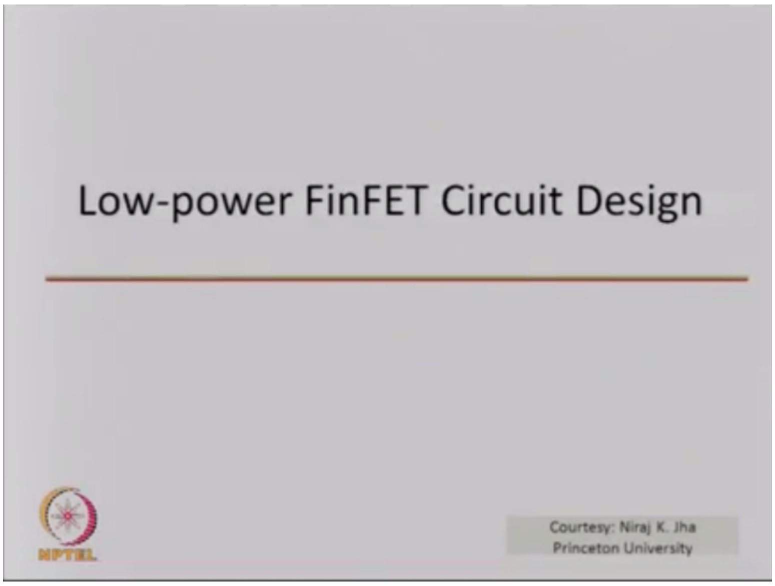
Proposed Design Flow

- Modified linear program
- Resistive feed though cell generation:
  - Fully automated
  - Scalable to large ICs
- Layout generation of modified netlist
  - Can use any place-and-route tool

And this is typically a design flow, which they have suggested, we start with design entry, we do technology mapping, remove glitches by the techniques, which we have suggested and then go for the layout. Resisted feedthrough cell generation fully automated, scalable to any ICs, layout generation are modified netlist can use any place to place route tool. So this is essentially a work from a computer science persons.

They want to actually not play so much with the technologies and that is available whatever available to you on the spice or any other models available. So given the design entry can we do still glitch power reduction or at least create the cells IPs for different such requirement for drive currents and this but having equal delays at least differential equal delays so that the glitches have minimized.

(Refer Slide Time: 01:19:06)



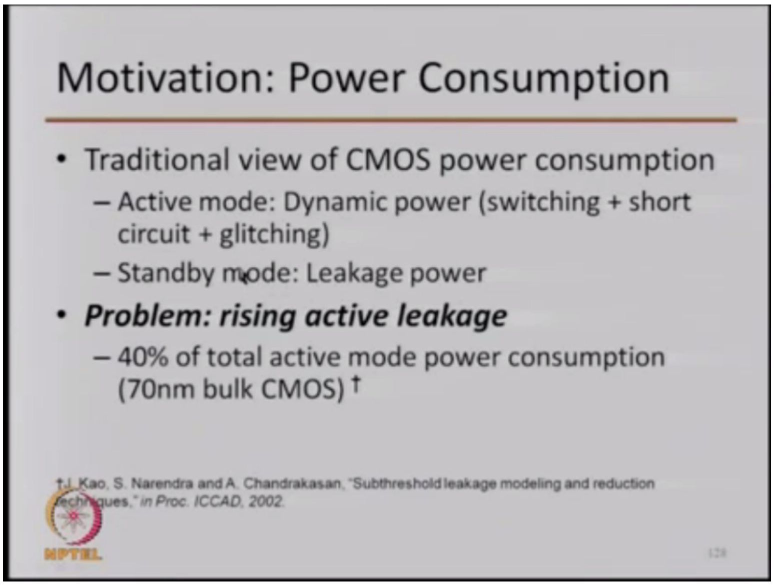
Low-power FinFET Circuit Design

Courtesy: Niraj K. Jha
Princeton University

The last but not the least maybe many more things maybe I will quickly go into this. The new structure of a MOSFET, which has appeared in last 10 15 years is called FinFET, which is essentially the new MOSFET, which is going to be used in almost every low power circuit. The slides which I am presenting to you here is from courtesy of Niraj K. Jha at Princeton University.

Of course these slides are available on web page I trust but anyway Jha being our good old friend from the VLSI design conference was kind enough to give me many years ago. So I am first time showing you here.

(Refer Slide Time: 01:19:47)



Motivation: Power Consumption

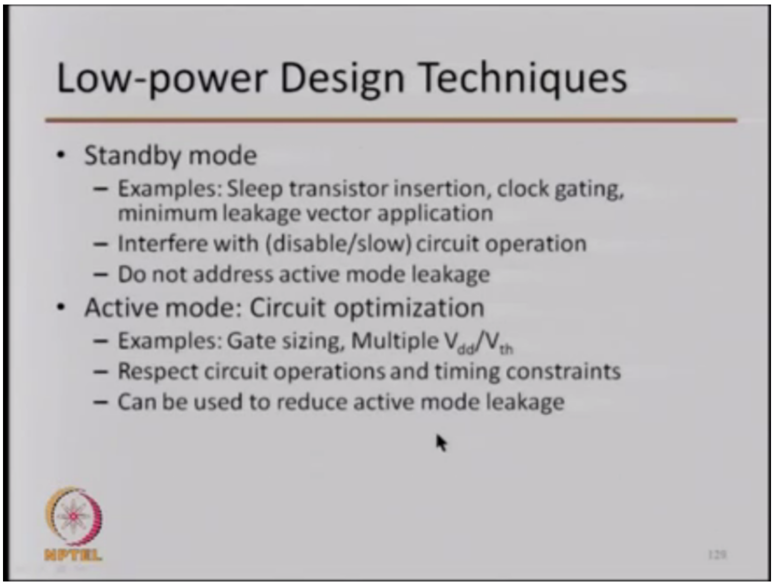
- Traditional view of CMOS power consumption
  - Active mode: Dynamic power (switching + short circuit + glitching)
  - Standby mode: Leakage power
- ***Problem: rising active leakage***
  - 40% of total active mode power consumption (70nm bulk CMOS) †

† J. Kao, S. Narendra and A. Chandrakasan, "Subthreshold leakage modeling and reduction techniques," *in Proc. ICCAD, 2002.*

128

So what is the motivation the traditional view of CMOS power consumption, active mode this is called dynamic mode which includes nano switching and short circuit plus glitching and the last is the standby mode, which is the leakage power. The problem as we all of us had just seen that active mode power is 40% even at 70 nanometer bulk CMOS, 60% is really going to the leakage power. This is essentially due to the old paper of 2002 by J Kao, S Narendra and A Chandrakasan.

(Refer Slide Time: 01:20:18)



Low-power Design Techniques

- Standby mode
  - Examples: Sleep transistor insertion, clock gating, minimum leakage vector application
  - Interfere with (disable/slow) circuit operation
  - Do not address active mode leakage
- Active mode: Circuit optimization
  - Examples: Gate sizing, Multiple V_{dd}/V_{th}
  - Respect circuit operations and timing constraints
  - Can be used to reduce active mode leakage

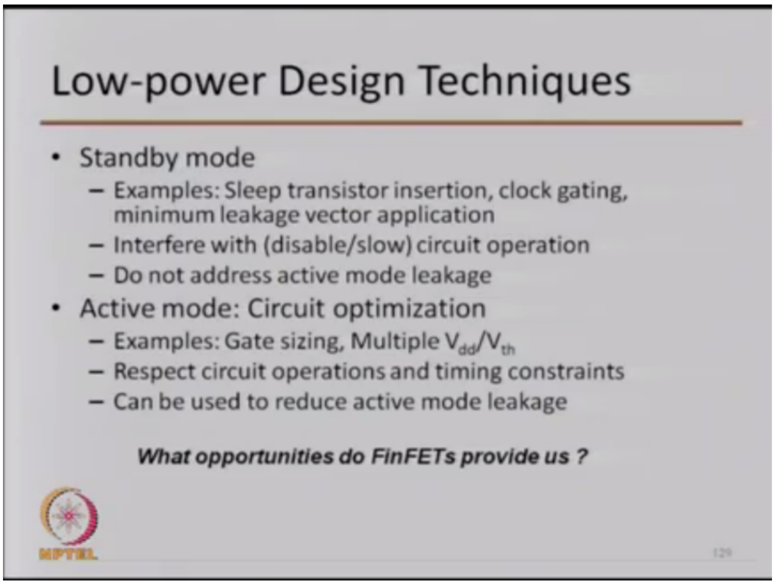
NPTEL

128

So what is the techniques for the leakage that is standby, sleep transistor we are just talked, clock getting we have seen, we can have leakage vector applications, glitching of course we shall see later. Interfere with disable as say switch off switch on possibilities in circuit operation and do not address active mode leakages, do not play too much about these you know during active mode do not try to play VTs and anything on the leakage because during that mode let higher current be possible.

Active mode circuit optimization with include gate sizing, multiple VDD to threshold ratios, being both multiple VDD and multiple threshold I already discussed all of them, respect circuit operations and timing constraints, can be used to reduce active mode leakage okay. So we can have now techniques in which this however this assumes as standard transistor, which is a normal N-channel or P-channel MOSFET or CMOS in general case what we have used so far.

(Refer Slide Time: 01:21:26)



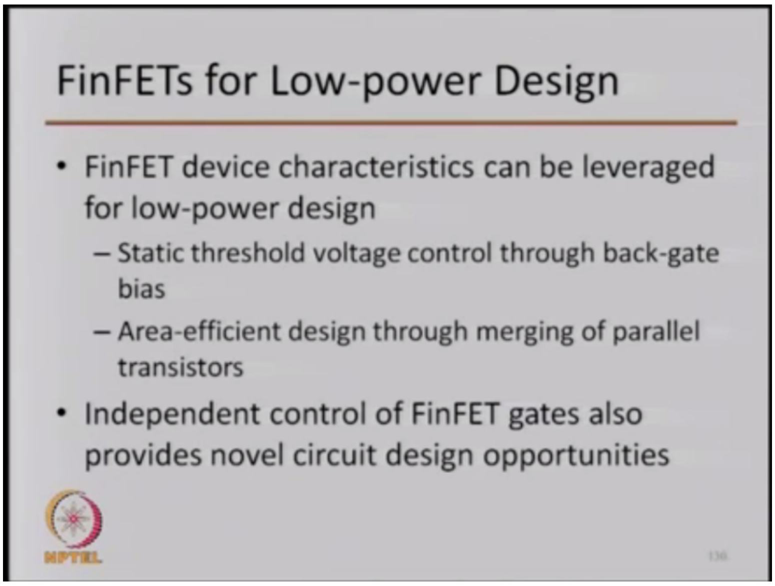
Low-power Design Techniques

- Standby mode
  - Examples: Sleep transistor insertion, clock gating, minimum leakage vector application
  - Interfere with (disable/slow) circuit operation
  - Do not address active mode leakage
- Active mode: Circuit optimization
  - Examples: Gate sizing, Multiple V_{dd}/V_{th}
  - Respect circuit operations and timing constraints
  - Can be used to reduce active mode leakage

What opportunities do FinFETs provide us ?

What opportunities therefore a typical structure called FinFET over a normal MOSFET. A FinFET is a device, which characteristics can be leveraged for low power design.

(Refer Slide Time: 01:21:32)



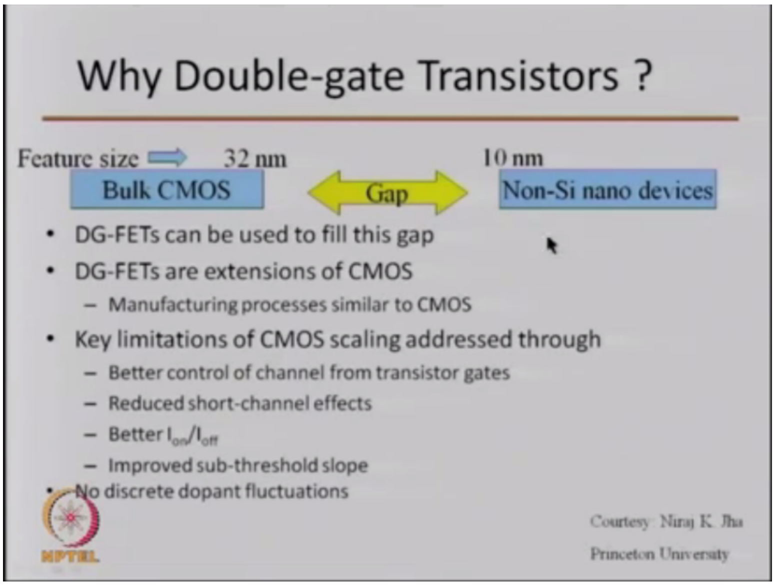
FinFETs for Low-power Design

- FinFET device characteristics can be leveraged for low-power design
  - Static threshold voltage control through back-gate bias
  - Area-efficient design through merging of parallel transistors
- Independent control of FinFET gates also provides novel circuit design opportunities

The static threshold voltage control through back-gate bias as we could do in normal DVS kind of techniques, area efficient design through merging of parallel transistors, this is another feature of FinFET that you can have reduced area compared to multiple MOS transistors and normal transistors.

Independent control of FinFETs gates either you can have connecting all the gates or you can have independent control and you can have therefore different novel circuit design opportunities. So this is how FinFETs were thought as a replacement for normal MOSFET and we believe that they can also be have since your threshold can be this, area efficiency can be done, capacitance can be minimized, one probably can have low power design using FinFETs.

(Refer Slide Time: 01:22:32)



Why Double-gate Transistors ?

Feature size ⇨ 32 nm 10 nm

Bulk CMOS ⟷ Gap ⟷ Non-Si nano devices

- DG-FETs can be used to fill this gap
- DG-FETs are extensions of CMOS
  - Manufacturing processes similar to CMOS
- Key limitations of CMOS scaling addressed through
  - Better control of channel from transistor gates
  - Reduced short-channel effects
  - Better I_{on}/I_{off}
  - Improved sub-threshold slope
- No discrete dopant fluctuations

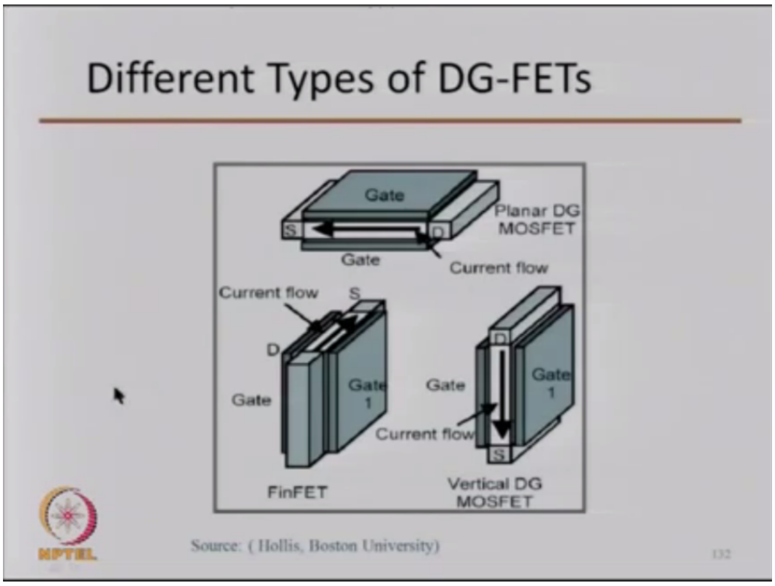
Courtesy: Niraj K. Jha
Princeton University

Here is some typical up to say 32 nanometer case you already said that bulk CMOS and you have non-silicon nano devices, which may come into 10 nanometer, you are still away from here. Many things are tried here, but as I am the stronger supporter of silicon next 30 years we are with silicon come what may. So let us look this to reach this we have still a gap so what can be done. DG-FETs can be used to fill this gap instead of bulk CMOS you have the double gate or multiple gates as FinFETs as they call DG-FETs are extension of CMOS manufacturing process is same as CMOS.

The key limitations of CMOS scaling address through better control of channel from transistor gates, reduced short-channel effects, better I on to I off, one thing I discussed for a good high performance circuit is higher on to off ratio of currents this is possible and of course because of the variety of parameters under your control now additional parameters the subthreshold slope can be improved, which essentially will reduce the leakage power.

And one can probably get away from the problems of dopant fluctuations as they occur in the normal MOSFETs.

(Refer Slide Time: 01:23:44)



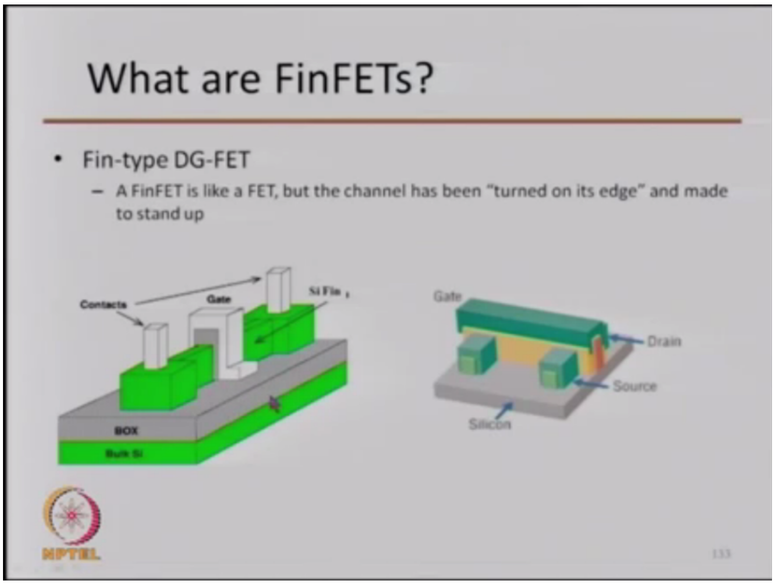
Different Types of DG-FETs

Source: (Hollis, Boston University)

132

There are structures this is called planar DGA MOSFET, this is called multiple fins connected, which is called FinFET and there is also same structure in the vertical mode possible is called vertical DG MOSFET. These are standard figures.

(Refer Slide Time: 01:24:00)



What are FinFETs?

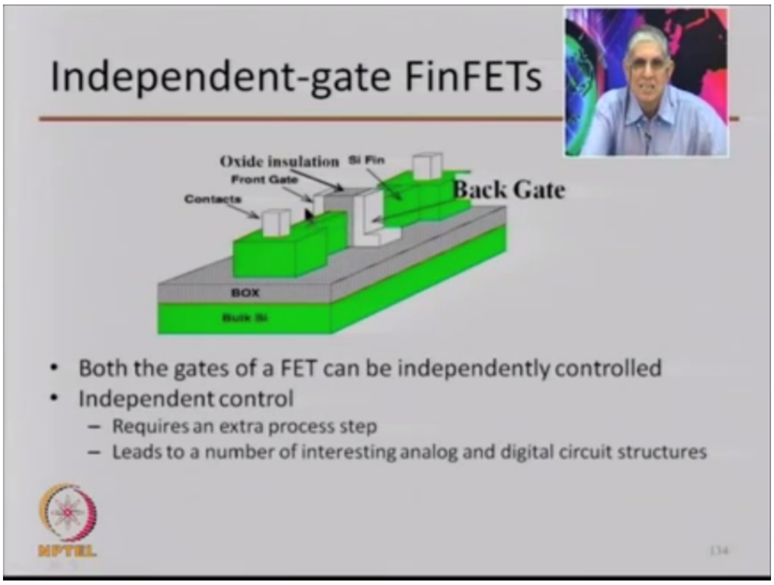
- Fin-type DG-FET
  - A FinFET is like a FET, but the channel has been "turned on its edge" and made to stand up

133

A typical fin-type DG MOSFET can be shown here to you to size. Please remember this is your gate and these are your source drains, these are two sources and back side are two drains. If you see this figure this is your gate which is shown here, this is source and this is your drain contacts to this. Now this is one FET now one can see from here why we say it is double gate because one can have control from this side, this side and the top side.

So the channel in this is not only, this is the channel length, but even the channel width which may act like a transistor for this kind of this. So we have now as if additional control possible, there is a gate here, there is a gate here and that gate on the top, so essentially you have double gate okay.

(Refer Slide Time: 01:25:01)



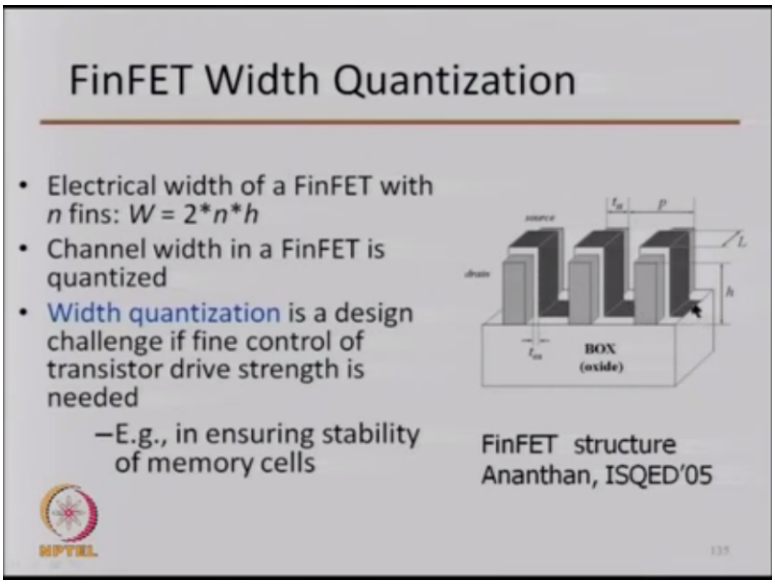
Independent-gate FinFETs

- Both the gates of a FET can be independently controlled
- Independent control
  - Requires an extra process step
  - Leads to a number of interesting analog and digital circuit structures

So one can see if you have independent control this gate and this gate has separate bias as possible then we say it is called IG gate, IG FinFET. Both gates of FET can be independently controlled and therefore requires of course unique and extra process step. This is called back gate and this is called the front gate okay and in between is the oxide thickness, this is your source drain okay.

Please remember this is called thickness of this here is called the fin thickness, thickness of green line here is essentially called silicon fin and that is most important. That is why it was named FinFET okay.

(Refer Slide Time: 01:25:44)



FinFET Width Quantization

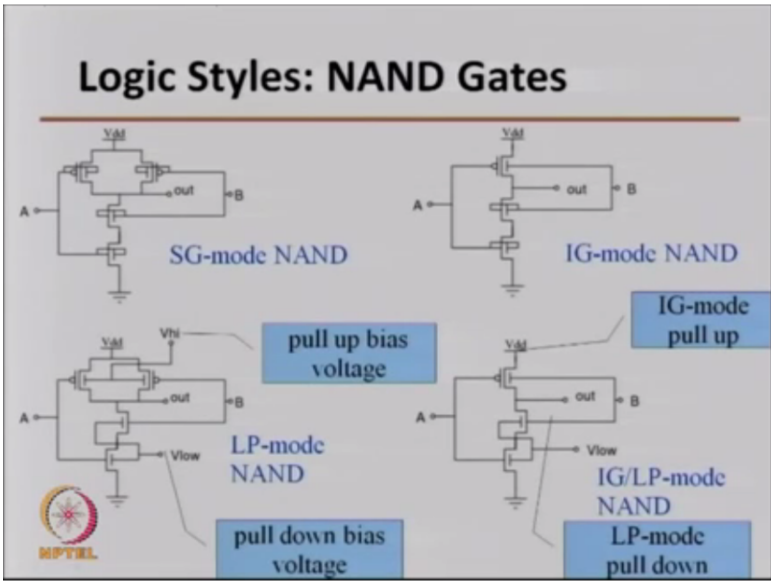
- Electrical width of a FinFET with *n* fins: $W = 2*n*h$
- Channel width in a FinFET is quantized
- Width quantization is a design challenge if fine control of transistor drive strength is needed
  - E.g., in ensuring stability of memory cells

FinFET structure
Ananthan, ISQED'05

In the case of FinFET, these are number of FinFET. You can actually connect all gates like this shown here. So this will become a common gate 1 single gate structure is called SG FinFET single gate, if you are independent control, then you say independent FinFET or IG FinFET. Typically if you are connect then it depends on how many such n fins are there n fins they are two times n.

Please remember h is the height of this because that is where the channel is going to form, so width of the channel is two times n means 1 into 3 here, channel width in a FinFET is quantized. Width quantization is a design if the fine control of transistor drive strength is needed. Now this certainly very helpful in having a good memory. We will not look into this in this course.

(Refer Slide Time: 01:26:39)



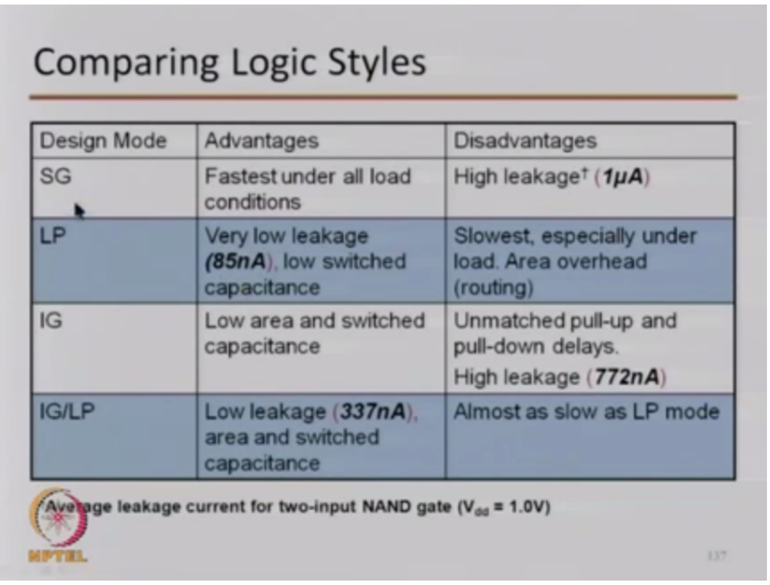
Logic Styles: NAND Gates

Here four possible structures shown here one is single gate, other is independent gate, and one can see from here, here in the normal single gate FinFETs the back gate bias is connected to the gate itself and this is a standard NAND gate since the FinFET we can have lower leakage current because of the normal single gate we do not have control but in the case of low power this you may have a separate power supply for the back bias substrates okay.

For both P-channel we have a high voltage here pull up bias voltage, this is pull down bias voltage, which is one can be forward bias other can be reverse bias and one can adjust the threshold of this, one can adjust the threshold of this and therefore one can have some clock going on this. So when in active mode, they behave normally in off mode, they increase VT so much then the leakage currents goes down.

So similar technique was tried in the case of independent gates and either of these four techniques have been tried for the implementation of NAND gates. It becomes very difficult in the case of very complex logic to use the IG mode gates because the connectivity is too many places; however, this has been tried and this is one of the major technique of reducing the low power in the newer circuits of below 32 nanometer nodes.

(Refer Slide Time: 01:28:17)



Comparing Logic Styles

Design Mode	Advantages	Disadvantages
SG	Fastest under all load conditions	High leakage† (***1μA***)
LP	Very low leakage (***85nA***), low switched capacitance	Slowest, especially under load. Area overhead (routing)
IG	Low area and switched capacitance	Unmatched pull-up and pull-down delays. High leakage (***772nA***)
IG/LP	Low leakage (***337nA***), area and switched capacitance	Almost as slow as LP mode

Average leakage current for two-input NAND gate (V_{dd} = 1.0V)

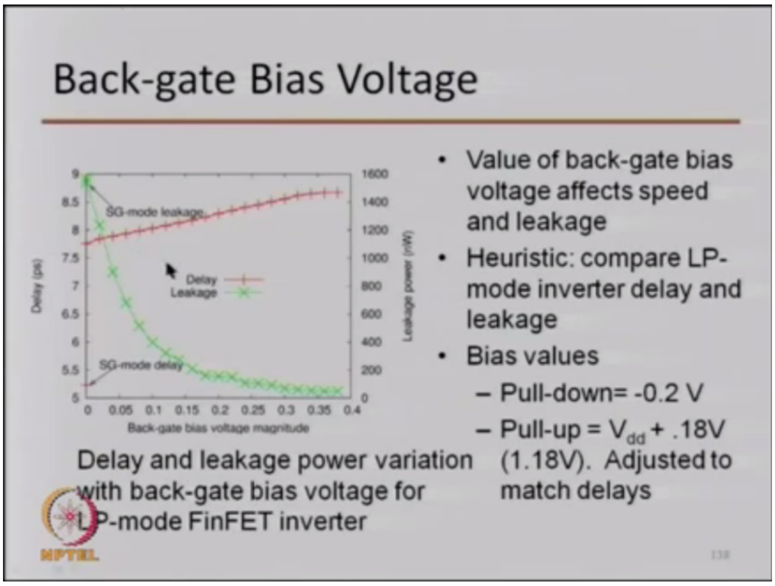
137

This is a comparison as far ISG, LP, IG, IG/LP just to get you an idea very low leakage current 85 nano amp okay in the case of SG very high leakage because you are connecting gate to the substrate so 1 micro amp whereas in the case of independent gate LP you have larger than this; however, because the width is very small comparatively and disadvantage of course is low leakage and not so low, but this is very, very low leakage.

So now you can see the speed of a circuit essentially can be better with SG, it also gives you the SGs version with low power can give better leakage; however, many other switch capacitance analog circuits or any other blocks can be best attain for low power using IG/LP okay. You can have higher or lower leakage depending on you match the pull up or pull down.

So there are advantages and disadvantages in SG and please remember SG has the worst thing is that it has normal SG has very high leakage but LP version of this has however once as soon as you say low leakage the speed has gone down for a FinFET. So depending on only low power, only high performance or a standby one of the possible combinations can be chosen.

(Refer Slide Time: 01:29:45)



Back-gate Bias Voltage

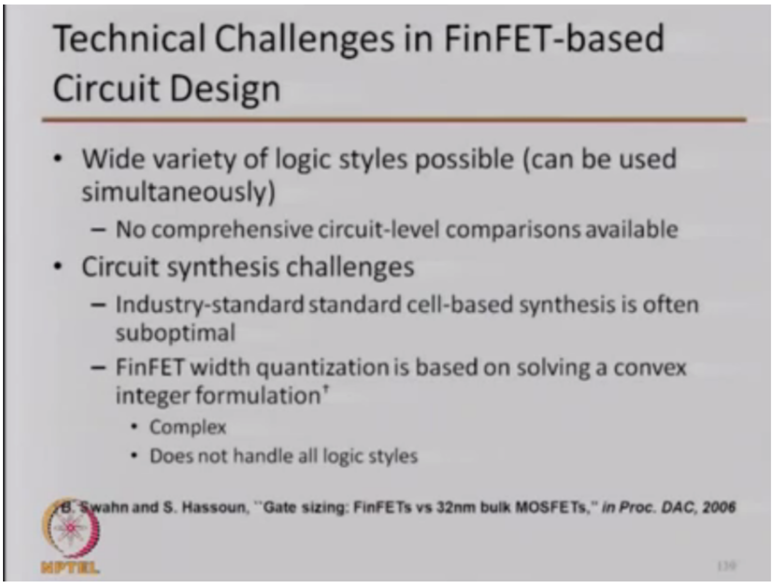
Delay and leakage power variation with back-gate bias voltage for LP-mode FinFET inverter

- Value of back-gate bias voltage affects speed and leakage
- Heuristic: compare LP-mode inverter delay and leakage
- Bias values
  - Pull-down= -0.2 V
  - Pull-up = V_{dd} + .18V (1.18V). Adjusted to match delays

This is typically what I am trying to show you, red shows the delay and green shows the power so this is only shown for SG kind single gate, you adjust your back gate bias with a low power and in that case one can see this is the back gate bias as you increase it, the leakage goes down, but if you increase it the delay also rises. So now you can adjust someway how much back gate bias, how much leakage, how much delay or speed you want and correspondingly tell your biases.

So that the on current to off current ratio is of your choice correspondingly and the low power the power is minimized.

(Refer Slide Time: 01:30:33)



Technical Challenges in FinFET-based Circuit Design

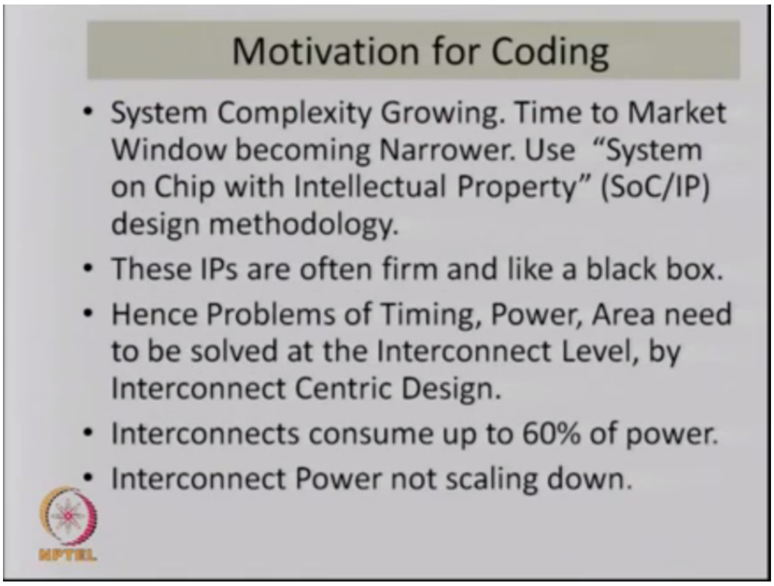
- Wide variety of logic styles possible (can be used simultaneously)
  - No comprehensive circuit-level comparisons available
- Circuit synthesis challenges
  - Industry-standard standard cell-based synthesis is often suboptimal
  - FinFET width quantization is based on solving a convex integer formulation†
    - Complex
    - Does not handle all logic styles

†B. Swahn and S. Hassoun, "Gate sizing: FinFETs vs 32nm bulk MOSFETs," *in Proc. DAC, 2006*

There are variety of challenges in FinFET-based circuit design is no comprehensive circuit level comparisons are available, there are not enough tools to control design tools available at higher levels, there are not enough standard cells available so that you can synthesize for optimal or suboptimal operations. FinFET width quantization is based on solving a convex integer formulation which though I solved it very simply, but it is not so.

It is extremely complex so you doing lot of variability issues also adds to it, it becomes extremely complex and it does not count as I say handle all logic styles you cannot have Dominos and every other style in the FinFETs; however, I mean you can have something and you may not have all of it.

(Refer Slide Time: 01:31:22)



Motivation for Coding

- System Complexity Growing. Time to Market Window becoming Narrower. Use "System on Chip with Intellectual Property" (SoC/IP) design methodology.
- These IPs are often firm and like a black box.
- Hence Problems of Timing, Power, Area need to be solved at the Interconnect Level, by Interconnect Centric Design.
- Interconnects consume up to 60% of power.
- Interconnect Power not scaling down.

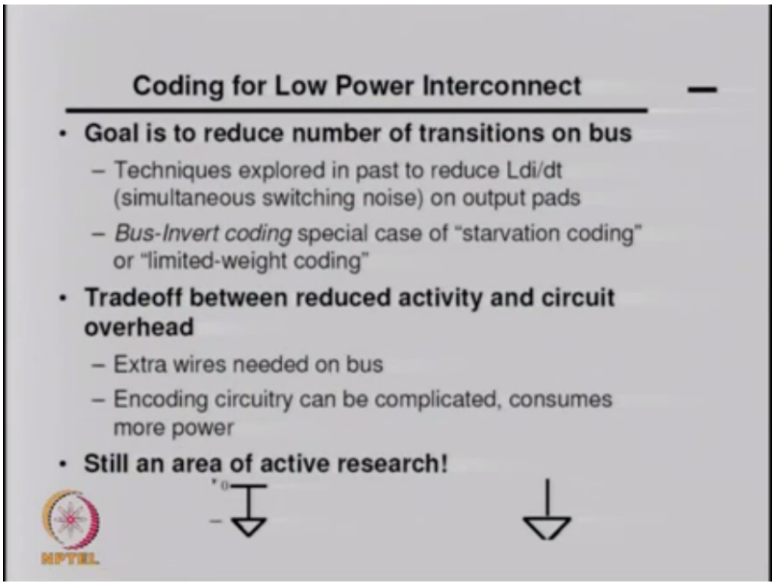
The last part quickly will go I think already I am running short of time, but let me finish this. One other technique of reducing power is essentially coming because of the interconnect power consumption. We already has this if we have a system on chip okay which is nothing but with intellectual property SoC/IP there are lot of SOPs, SoCs or IPs are getting marketed and they are normally firm and they like a black box.

There are issues of timing, power, area to be solved for any interconnect layer for each of them. So this particular part I am going to talk about more about interconnects in an SoC or any in circuit per se. Interconnect consumes large power 60% of the current processor DSP processor in particular is consumed through a interconnects. So obviously all the techniques we discussed for the devices and all techniques of architectures everything they were valid only for device related performance.

But if the power is additional to those is coming more from the interconnect one should worry more about the interconnect power and one of major worry is that this is not scaling down because the RC time constraints are not scaling down. So how do you reduce the power at least the dynamic power? So we know dynamic powers has something to do with the activity coefficient.

Dynamic power is alpha times CV VDD square into f so can I reduce alpha, which is the activity coefficient so what we say our goal is there to reduce number of transition on the bus.

(Refer Slide Time: 01:33:15)



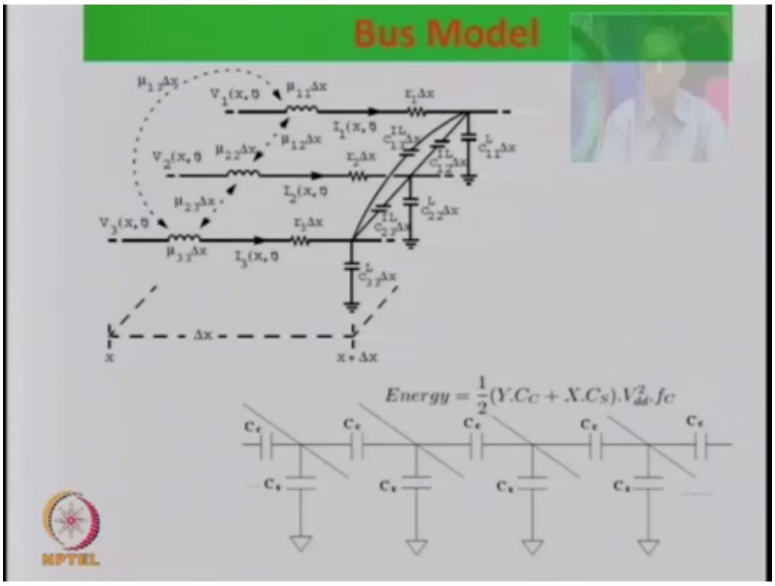
Coding for Low Power Interconnect

- **Goal is to reduce number of transitions on bus**
  - Techniques explored in past to reduce Ldi/dt (simultaneous switching noise) on output pads
  - *Bus-Invert coding* special case of "starvation coding" or "limited-weight coding"
- **Tradeoff between reduced activity and circuit overhead**
  - Extra wires needed on bus
  - Encoding circuitry can be complicated, consumes more power
- **Still an area of active research!**

So techniques explored in the past to reduce Ldi by dt, which is called switching noise on the output pads. This is of course is always present even now, Ldi/dt problems cannot be easily solved this is always be present. However the other power reduction could be which is called Bus-Invert coding as in the case of what we call starvation coding or limited-weight coding. Now between there has to be tradeoff between reduced activity and circuit overhead.

So we say you reduce alpha and to do this if you put additional circuit to do that how much is the power on the additional circuit. So over head is what most important. You need extra wires to do this so additional power on the bus. Encoding circuitry can be complicated sometimes and decoding is also equally complicated and it may consume large power sometimes.

(Refer Slide Time: 01:34:18)



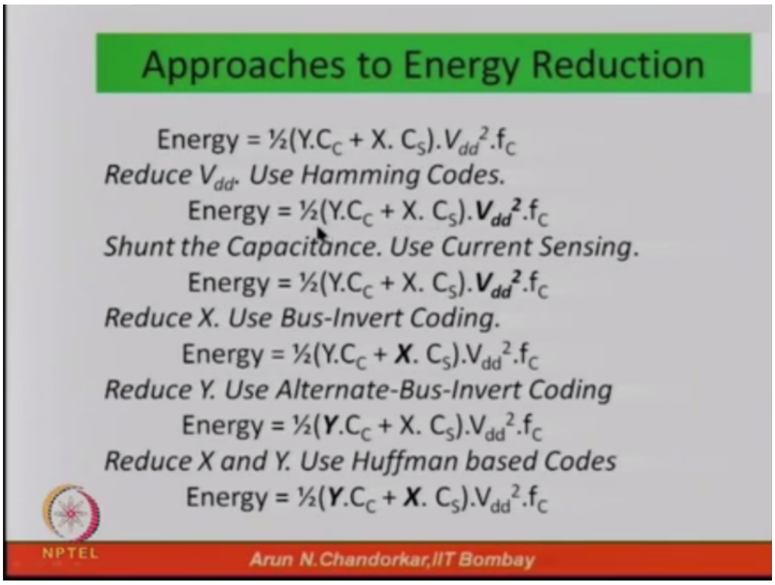
Bus Model

$$Energy = \frac{1}{2}(Y.Cc + X.Cs).V_{dd}^2.fc$$

And one has to worry about this power, low power interconnect. A typical bus can be modeled into LCR circuit like a transmission line and our kind of model which I am showing you is something one of my student in 90s to 2000 time has worked on. Similar model has been chosen by many people and they say a typical between the two lines of a bus, there will be a capacitance and between the bus and each ground, there will be substrate capacitance.

So this is each line in a bus so there is a capacitance between the wires and there is a capacitance for wire to the substrate. So there is something called lateral and vertical capacitances. This is called Cs and this is called Cc. So we say energies half Y times Cc X times Cs times Vdd to describe this. Now what are this X and Y, how many sections we have is deciding that.

(Refer Slide Time: 01:35:30)



Approaches to Energy Reduction

$\text{Energy} = \frac{1}{2}(Y.C_C + X.\ C_S).V_{dd}^2.f_C$

Reduce V_{dd}. Use Hamming Codes.

$\text{Energy} = \frac{1}{2}(Y.C_C + X.\ C_S).\boldsymbol{V_{dd}^2}.f_C$

Shunt the Capacitance. Use Current Sensing.

$\text{Energy} = \frac{1}{2}(Y.C_C + X.\ C_S).\boldsymbol{V_{dd}^2}.f_C$

Reduce X. Use Bus-Invert Coding.

$\text{Energy} = \frac{1}{2}(Y.C_C + \boldsymbol{X}.\ C_S).V_{dd}^2.f_C$

Reduce Y. Use Alternate-Bus-Invert Coding

$\text{Energy} = \frac{1}{2}(\boldsymbol{Y}.C_C + X.\ C_S).V_{dd}^2.f_C$

Reduce X and Y. Use Huffman based Codes

$\text{Energy} = \frac{1}{2}(\boldsymbol{Y}.C_C + \boldsymbol{X}.\ C_S).V_{dd}^2.f_C$

NPTEL

Arun N.Chandorkar,IIT Bombay

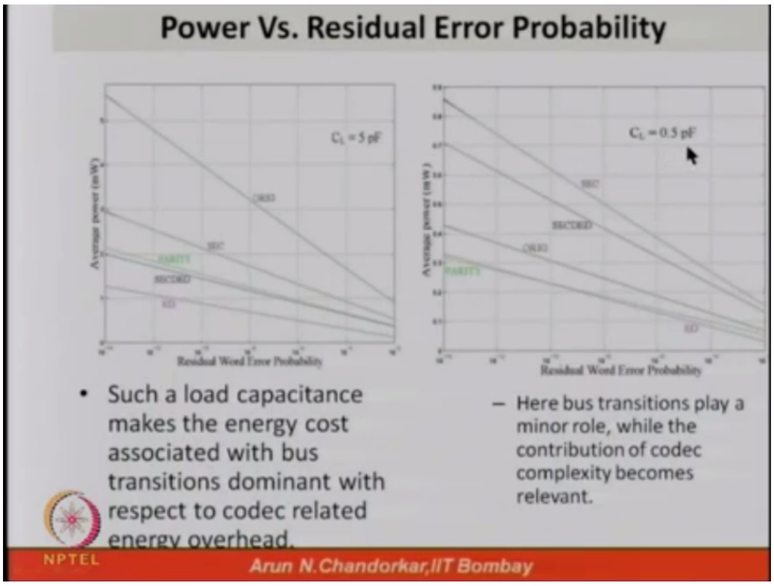
So this is the energy so let see how do I reduce the energy reduction. If we reduce Vdd then we can reduce the energy so what I do is use Hamming Codes. If you use reduced Vdd and use Hamming Codes, this can be reduced at YCcVdd square fc, shunt capacitance use current sensing so I have different techniques you do Bus-Invert Coding you get this power, so you adjust only X, reduce Y use alternate Bus-Invert Coding.

So you can see either reduce Y, and reduce X and reduce this, fc is not controlled because if you control fc then your speed goes down so I am only interested in transition to Cc, transition to Cs power supply. So depends on I can do coding on the data arriving on a bus and we may use different methods to actually reduce the power. So particularly I am interested in reduction in X and Y so different techniques we have suggested in literature.

And if you got many people do if you only do Vdd scaling then we say they use having codes we shall see what they are. If you do shunt capacitance this use a Bus-Invert for X there is alternate Bus-Invert for reduce Y what we have done we have used a new technique, which is based on modified Huffman Codes in which both X and Y can be minimized and if you can minimize both X and Y the net energy lost on interconnect can be minimized further.

Okay I will skip this because this will require lot much effort to explain but let us look into power versus residual error probability.

(Refer Slide Time: 01:37:17)


Power Vs. Residual Error Probability

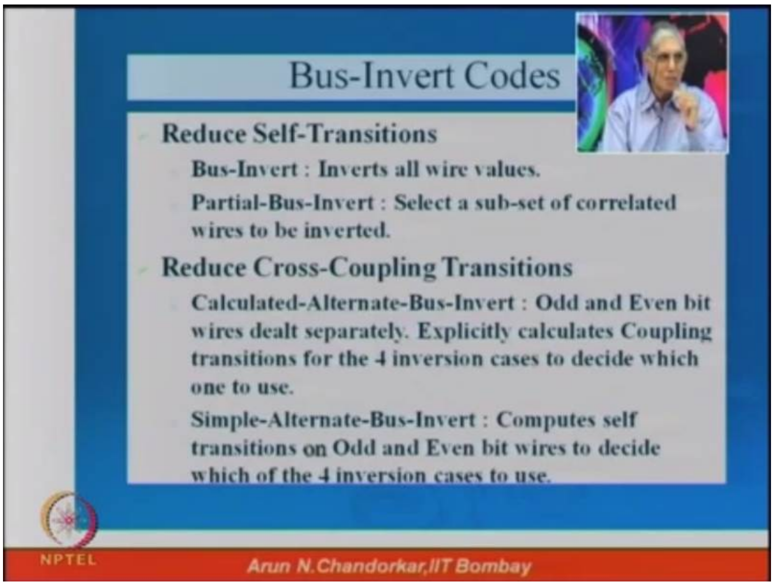
- Such a load capacitance makes the energy cost associated with bus transitions dominant with respect to codec related energy overhead.
  - Here bus transitions play a minor role, while the contribution of codec complexity becomes relevant.

Arun N.Chandorkar,IIT Bombay

So depending on the capacitance value, this is half PUF and this is a 5 PUF depends on the different kinds of coding used this is single event error. For example, this is multiple so one can see from here bus transitions play a minor role in the case of lower capacitance whereas they play a larger role if the load capacitance is very high. So if there are larger transitions on a higher capacitance buses then you have large power lost.

This particularly occurs in a normal code existence where the switchings are constantly whereas the data goes with very high speeds.

(Refer Slide Time: 01:37:57)


Bus-Invert Codes

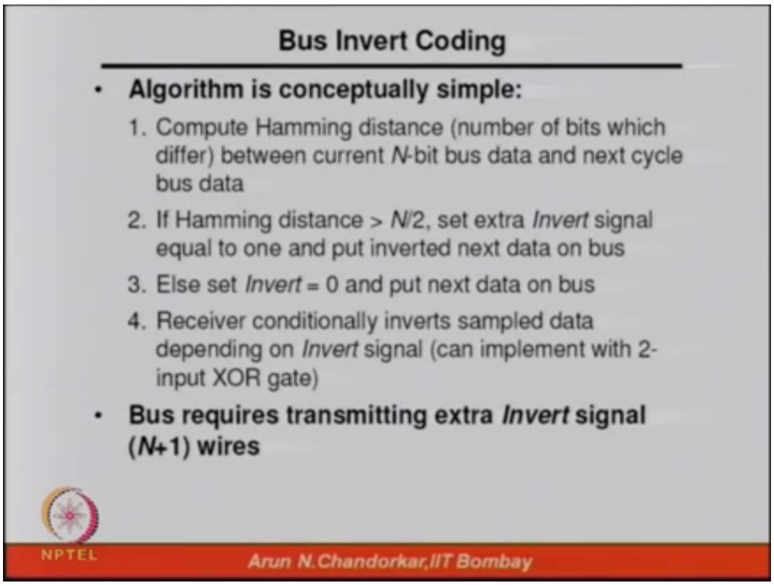
- Reduce Self-Transitions
  - Bus-Invert : Inverts all wire values.
  - Partial-Bus-Invert : Select a sub-set of correlated wires to be inverted.
- Reduce Cross-Coupling Transitions
  - Calculated-Alternate-Bus-Invert : Odd and Even bit wires dealt separately. Explicitly calculates Coupling transitions for the 4 inversion cases to decide which one to use.
  - Simple-Alternate-Bus-Invert : Computes self transitions on Odd and Even bit wires to decide which of the 4 inversion cases to use.

Arun N.Chandorkar,IIT Bombay

The typical idea in Bus-Invert Code is to simple, invert all wire values say if you have one you make it 0 okay this is called Bus-Invert. Now what is that advantage, the advantage of Bus-Invert is the assumption is that if the data is coming and the lost data is let say on a one wire there is 0 so next data should come 0, when there is no transition, but the next bus should also should have one which must receive one again so there are no transitions.

So the coupling as well as this if I change it I will probably feel that at least one of them will be reduced transitions okay and in that case for a particular data I will have lesser number of transitions, so this is what essentially Bus-Invert does. In the case of alternate Bus-Invert we actually see odd and even bit wires dealt separately explicitly so how do we find this. So the better method I may give the very simple technique okay here is the one.

(Refer Slide Time: 01:38:59)



Bus Invert Coding

- **Algorithm is conceptually simple:**
  1. Compute Hamming distance (number of bits which differ) between current *N*-bit bus data and next cycle bus data
  2. If Hamming distance > *N*/2, set extra *Invert* signal equal to one and put inverted next data on bus
  3. Else set *Invert* = 0 and put next data on bus
  4. Receiver conditionally inverts sampled data depending on *Invert* signal (can implement with 2-input XOR gate)
- **Bus requires transmitting extra *Invert* signal (*N*+1) wires**

NPTEL

Arun N.Chandorkar,IIT Bombay

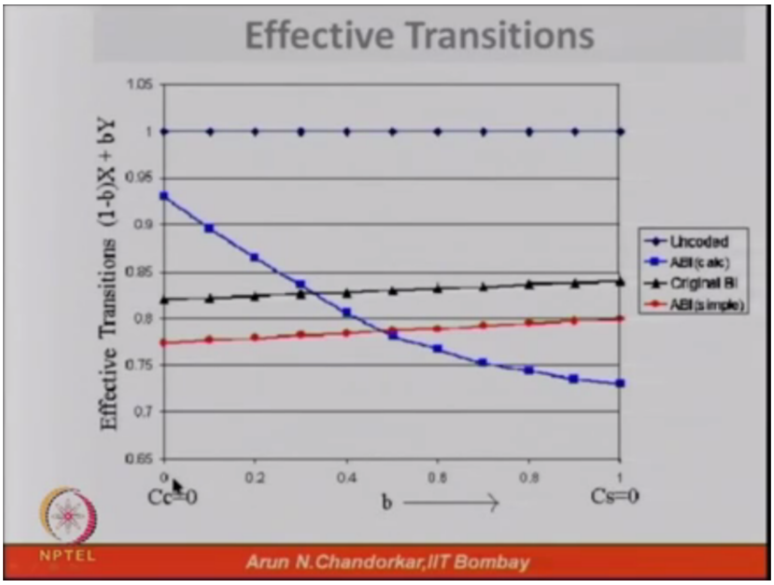
What is essentially the way we are doing it? What we do is we figure out how many ones and how many zeros are in a data okay. The difference between them is called the Hamming distance, number of ones minus number of zeros. Once you know they are Hamming distance that is how many they differ and if the Hamming distance is larger than N by 2 then you must invert extra invert signal equal to 1 and put inverted next data on the bus okay so that the transitions are minimized.

If else you put invert to 0 that is whether to change the data on the bias if you find the Hamming distance is less than N by 2 what does that mean, 0s are larger okay. If the 0s are larger, you do not invert them because anyway you will require more 0s to available no transition so you actually prefer to have those.

So every time first figure out that means there must be a circuitry which must find the Hamming distance for each data arrival from the last data to new data how much is the new

Hamming distance and correspondingly give invert signals whether to wire should be inverted or non-invert code it like that. Once you code it you pass through data on a line and at the end you decode it.

(Refer Slide Time: 01:40:29)



Effective Transitions

Arun N.Chandorkar,IIT Bombay

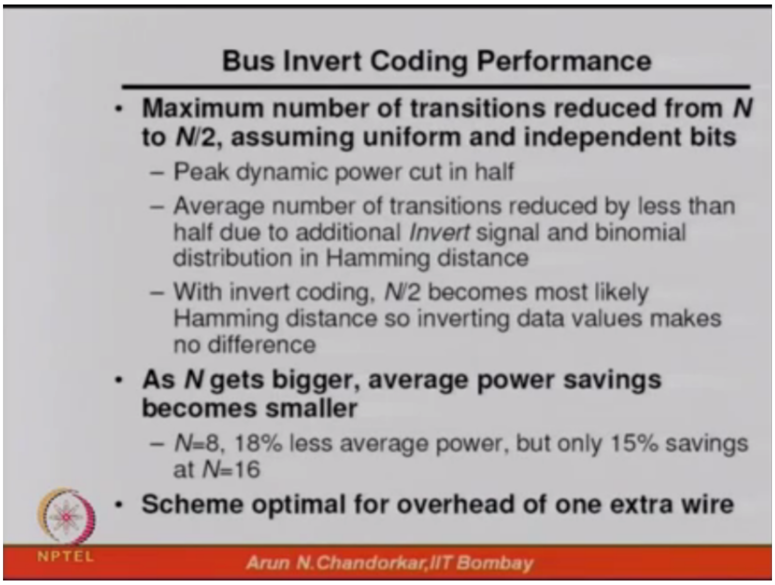
So that the original data is stored okay and that is essentially what is called Bus-Invert technique. This is the graph shown here this is called effective transitions versus the ratio of Cc to Cb. You can see if Cs is 0, here Cc is 0 means b is 0, b is 0 means Cc upon Cs plus Cb. So Cc upon coupling capacitor divided by the substrate capacitance plus net capacitance if I say is called B.

Please remember I am again saying Cc that is the coupling capacitance between wires divided by the net capacitance, which is Cc plus C substrate. So at this point Cc 0 and here Cc is 1 means Cs is 0. So for the three cases, this is un-coded you can see large activity okay if you do not code it if you do alternate bus coding it is something like this ABi. If you do original Bus-Invert Coding, which is the triangle this is lower than this. Red is ABi as is implemented okay.

This is of course actually implemented this is theoretical. One finds the red ones is what is most important so one can see from un-coded to the coded implemented this is theoretically one feels that ABi should be better when the Cc is roughly half equal to Cb, but this is not true if we figure it out this is anyway increasing this is theoretically evaluations we did and we implemented it on assimilated actually hardware.

Then you figure out alternate Bus-Invert Coding actually reduces from 1 to 0.75 to 0.8 it is increasing, but not very much so is this original Bus-Invert Coding but still it is higher than alternate Bus-Invert. So now I say the best transition is that you do alternate Bus-Invert Coding correspondingly for almost any interconnect and you may have activity coefficient reduced.

(Refer Slide Time: 01:42:40)

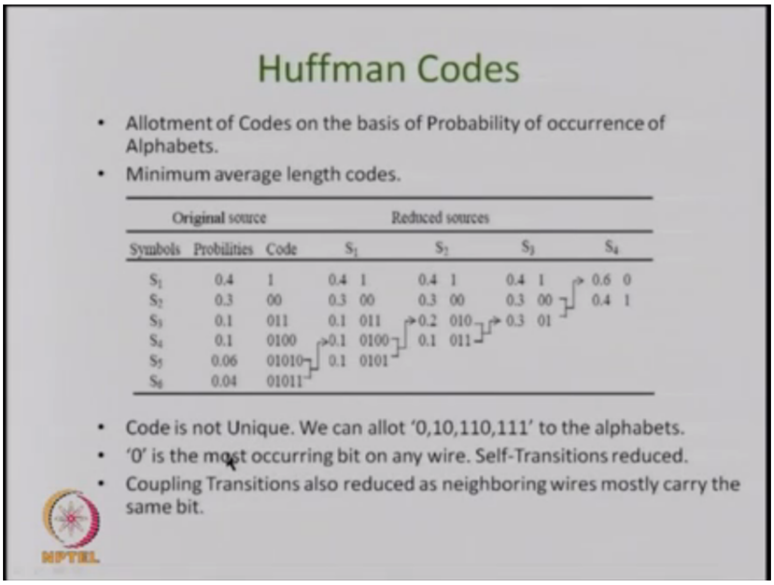


Bus Invert Coding Performance

- Maximum number of transitions reduced from *N* to *N*/2, assuming uniform and independent bits
  - Peak dynamic power cut in half
  - Average number of transitions reduced by less than half due to additional *Invert* signal and binomial distribution in Hamming distance
  - With invert coding, *N*/2 becomes most likely Hamming distance so inverting data values makes no difference
- As *N* gets bigger, average power savings becomes smaller
  - *N*=8, 18% less average power, but only 15% savings at *N*=16
- Scheme optimal for overhead of one extra wire

Now the problem is as there is a maximum number of transition therefore reduced from N to N by 2 assuming uniform and independent bits. Peak dynamic power is therefore cut off to half but with invert coding N by 2 becomes most likely Hamming distance so the inverting data values makes no difference and gets bigger average power saving becomes smaller and larger than that the saving is lower because the other power starts increasing, scheme optimal for overhead one extra wire is also required to create invert signal.

(Refer Slide Time: 01:43:09)



Huffman Codes

- Allotment of Codes on the basis of Probability of occurrence of Alphabets.
- Minimum average length codes.

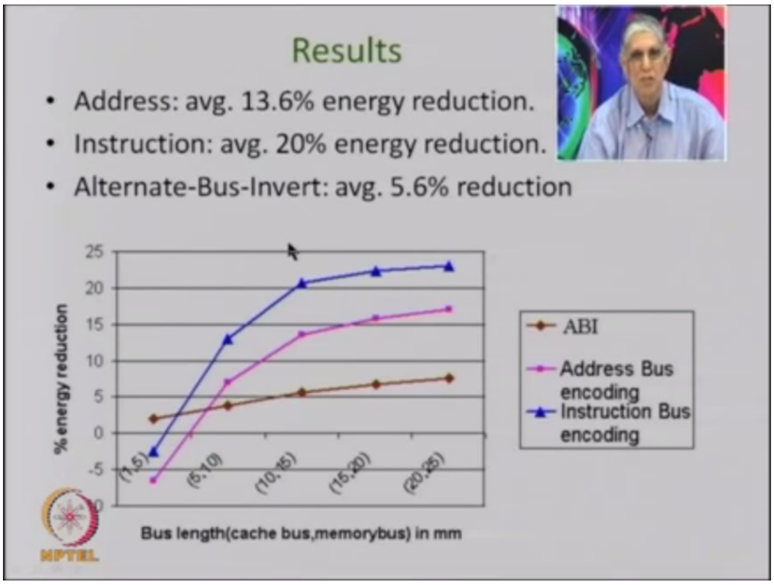
Original source			Reduced sources							
Symbols	Probilities	Code	S_1		S_2		S_3		S_4	
S_1	0.4	1	0.4	1	0.4	1	0.4	1	0.6	0
S_2	0.3	00	0.3	00	0.3	00	0.3	00	0.4	1
S_3	0.1	011	0.1	011	0.2	010	0.3	01		
S_4	0.1	0100	0.1	0100	0.1	011				
S_5	0.06	01010	0.1	0101						
S_6	0.04	01011								

- Code is not Unique. We can allot '0,10,110,111' to the alphabets.
- '0' is the most occurring bit on any wire. Self-Transitions reduced.
- Coupling Transitions also reduced as neighboring wires mostly carry the same bit.

We suggested another technique, which this is you can go into probability theory of information theories on this area. One of the very famous code of data transmission is called based on probability of occurrence of alphabets is called Huffman's Code, S is called the symbols, this is probability, this is the code which it creates, code is not unique we can allot 0 1 0 1 1 0 1 1 1 to each alphabets.

However, this number keeps rising you can see the number of bits rising as you increase the symbols, but we did we actually truncated that only three bits which is called modified Huffman's Code. We actually applied it to standard buses, we have compared our result with this.

(Refer Slide Time: 01:43:45)



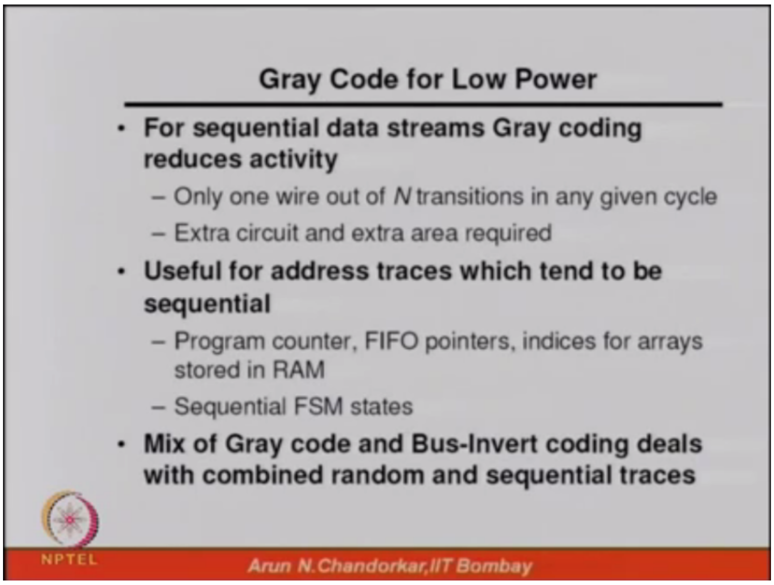
Results

- Address: avg. 13.6% energy reduction.
- Instruction: avg. 20% energy reduction.
- Alternate-Bus-Invert: avg. 5.6% reduction

For a typical on processor with a memory on with a bus length going for in a cache from say this much to this much in millimeters one can see this is roughly this is cache bus and this memory bus this is large enough you feel it, but this is essentially in M on processor we have actually. So blue one essentially is normal instructing bus encoding, this is addressed bus encoding and ABi is which is what is shown here which is what we have tried okay.

And using our Huffman's Code and we figured out that we can actually minimize for any lengths of buses the power activity coefficient or energy reduction is almost 5% to 10 % more and maybe 6% less I mean more reduction on an average in our codes.

(Refer Slide Time: 01:44:56)



Gray Code for Low Power

- For sequential data streams Gray coding reduces activity
  - Only one wire out of *N* transitions in any given cycle
  - Extra circuit and extra area required
- Useful for address traces which tend to be sequential
  - Program counter, FIFO pointers, indices for arrays stored in RAM
  - Sequential FSM states
- Mix of Gray code and Bus-Invert coding deals with combined random and sequential traces

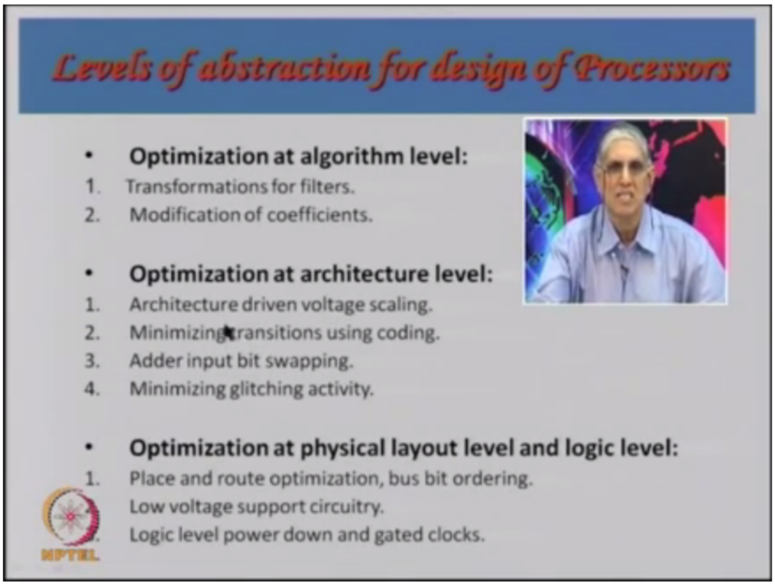
NPTEL

Arun N.Chandorkar,IIT Bombay

The last but not the least one of the simplest technique, which initially everyone tries in a sequential data stream is to go Gray coding okay. Only one wire out of N transition in any given cycle is on extra circuit only changing, you know in Gray code only one bit changes so if you have a data converted to Gray codes only one wire out of N transitions in any given cycle will change.

Extra circuits and extra areas are therefore required, useful for address traces, which tends to be sequential likes a program counter, FIFO pointers indices for arrays and stored in the RAM this may be useful and many of the sequential fine state machines said the states you require the state transitions there at that time maybe use Gray codes. Mix of Gray code and Hamming code based Bus-Invert probably can do both random and sequential traces power reduction and one can have low power circuits.

(Refer Slide Time: 01:46:12)



Levels of abstraction for design of Processors

- **Optimization at algorithm level:**

1. Transformations for filters.
2. Modification of coefficients.

- **Optimization at architecture level:**

1. Architecture driven voltage scaling.
2. Minimizing transitions using coding.
3. Adder input bit swapping.
4. Minimizing glitching activity.

- **Optimization at physical layout level and logic level:**

1. Place and route optimization, bus bit ordering.

Low voltage support circuitry.

Logic level power down and gated clocks.

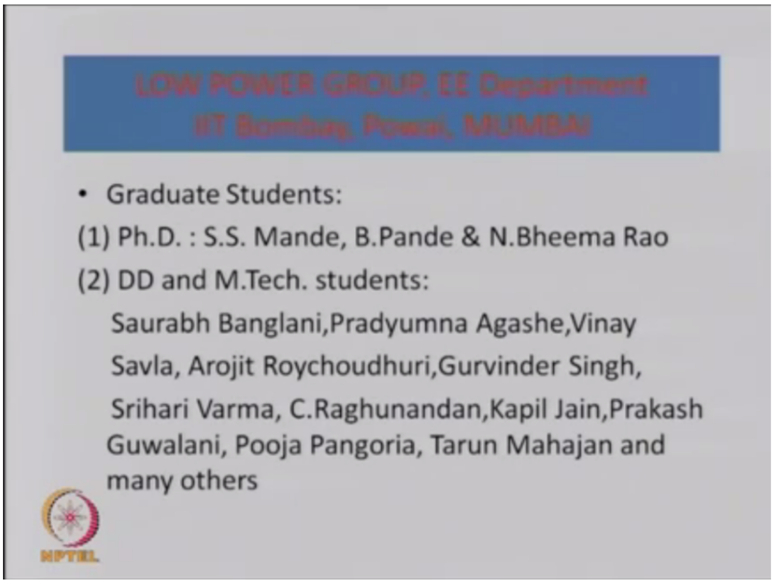
At the end of the day in conclusions, one can reduce for any design of a processor optimization and algorithm level you can have transformation for filters like glitch can be reduced, modification of coefficients we can do lot of things in algorithm itself so that the amount of computation itself is minimized.

At the architectural level, you do architectural level voltage scaling, minimize the transition using coding just now I have discussed, adder input bit swaps, which are discussed arithmetic later will show you, minimize glitch as much as possible by equalizing the delays. At the layout level and logical level place and route optimization bus bit ordering, you do all low voltage support circuitry, logic level power down and gated clocks can be used.

If all the techniques discussed by me are applied, one can design a low power processor which requires all kinds of architectures, all kinds of circuits in different architectures and can have a low power. Please remember one of the major effort all across the world right now is to create a very, very low power processor or has one of them maybe Intel is also having in some tablets.

So we are trying to reduce the power because this I pads or tablets everyone wants to have extremely low power circuits. Many of them are not very high performance circuits, but some part are high performance so you need to control threshold for those architecture parts; however, overall the effort is to create a low voltage, low power processors and all my such techniques, which I discussed with you so far can lead to such designs.

(Refer Slide Time: 01:48:04)



LOW POWER GROUP, EE Department
IIT Bombay, Powai, MUMBAI

- Graduate Students:

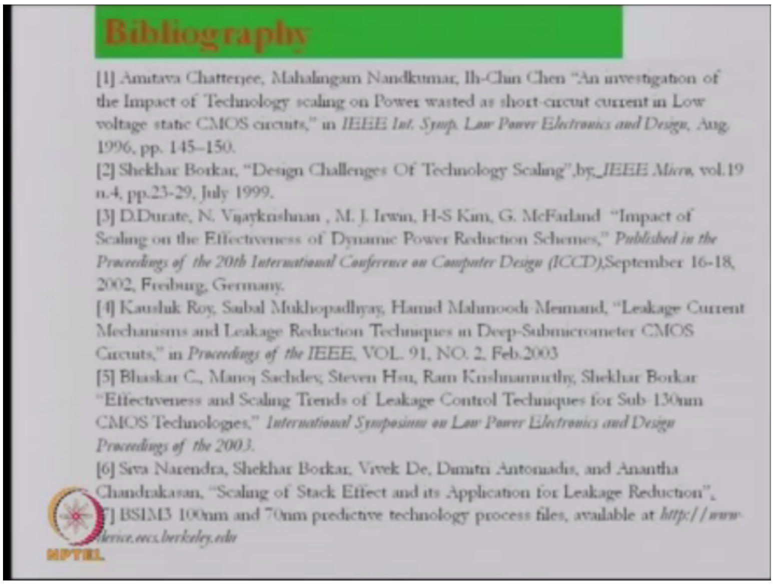
(1) Ph.D. : S.S. Mande, B.Pande & N.Bheema Rao

(2) DD and M.Tech. students:

Saurabh Banglani,Pradyumna Agashe,Vinay Savla, Arojit Roychoudhuri,Gurvinder Singh, Srihari Varma, C.Raghunandan,Kapil Jain,Prakash Guwalani, Pooja Pangoria, Tarun Mahajan and many others

Some of my graduate students who helped me in this long time are stated here, Mande, Pande, and Bheema Rao are my Ph.D students. There are many M. Tech students, which I have not listed many those who have directly worked with me on the low power area are listed here, Saurabh Banglani, Agashe, Savla, Roychoudhuri, Guruvinder, Srihari Varma, Kapil Jain, Raghunandan, Guwalani, Pangoria, Mahajan many others.

(Refer Slide Time: 01:48:40)



Bibliography

[1] Amitava Chatterjee, Mahalingam Nandkumar, Ih-Chin Chen "An investigation of the Impact of Technology scaling on Power wasted as short-circuit current in Low voltage static CMOS circuits," in *IEEE Int. Symp. Low Power Electronics and Design*, Aug 1996, pp. 145–150.

[2] Shekhar Borkar, "Design Challenges Of Technology Scaling",by, *IEEE Micro*, vol.19 n.4, pp.23-29, July 1999.

[3] D.Durate, N. Vijaykrishnan , M. J. Irwin, H-S Kim, G. McFarland "Impact of Scaling on the Effectiveness of Dynamic Power Reduction Schemes," *Published in the Proceedings of the 20th International Conference on Computer Design (ICCD)*,September 16-18, 2002, Freiburg, Germany.

[4] Kaushik Roy, Saibal Mukhopadhyay, Hamid Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," in *Proceedings of the IEEE*, VOL. 91, NO. 2, Feb.2003

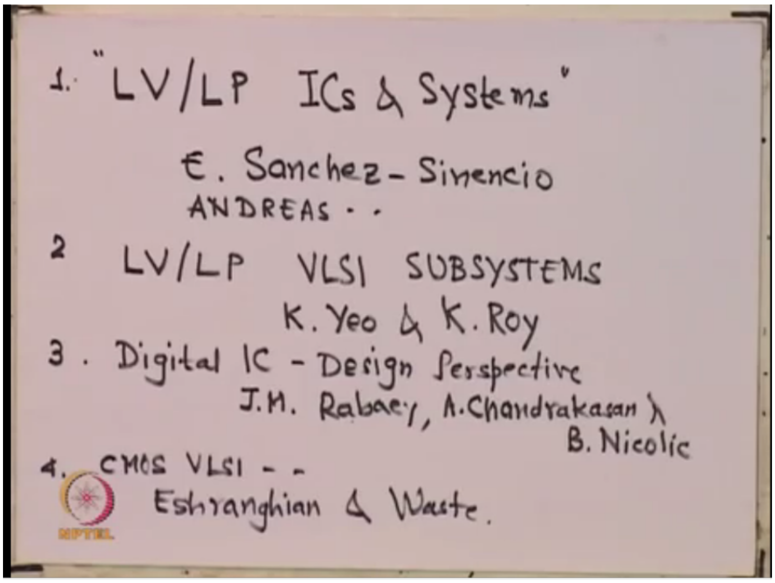
[5] Bhaskar C., Manoj Sachdev, Steven Hsu, Ram Krishnamurthy, Shekhar Borkar "Effectiveness and Scaling Trends of Leakage Control Techniques for Sub-130nm CMOS Technologies," *International Symposium on Low Power Electronics and Design Proceedings of the 2003.*

[6] Siva Narendra, Shekhar Borkar, Vivek De, Dimitri Antoniadis, and Anantha Chandrakasan, "Scaling of Stack Effect and its Application for Leakage Reduction",

[7] BSIM3 100nm and 70nm predictive technology process files, available at *http://www-device.eecs.berkeley.edu*

I forgotten some of them, I apologize if their names are missing here and thank you. These are some of the references, which I will be providing to you.

(Refer Slide Time: 01:48:49)



1. "LV/LP ICs & Systems"
E. Sanchez - Sinencio
ANDREAS . .
2 LV/LP VLSI SUBSYSTEMS
K. Yeo & K. Roy
3 . Digital IC - Design Perspective
J.M. Rabaey, A.Chandrakasan &
B. Nicolic
4. CMOS VLSI - -
Eshranghian & Waste.

The three books which I may actually advise few books for this low power design you may like to have is low voltage/low power ICs and systems by E. Sanchez his full name is I think Sinencio or something I do not know exactly but may be Sinencio. This is one book this is the name low voltage low power ICs and systems. I think the other author is Andreas I am not sure but just check Andreas something, something.

The second is again low voltage/low power ICs VLSI subsystems by K. Yeo and Kaushik Roy from Purdue University. Third book, which is very popular is digital IC design perspective by very famous people, best text book for basic VLSI design, which in my first course I discussed is by J.M. Rabaey, Anantha P Chandrakasan from MIT, this is from Berkeley and third author is B. Nicolic from Berkeley and fourth which is one of my most standard book, which I keep using in my VLSI design is CMOS VLSI design system etc by Eshraghian and Neil Weste.

If you use these four books most of the things, which I discussed this last two will give the basics of power design or power consumptions and how this. Then of course number of papers, which I gave you, will give you the actual details of the things, which I showed in this work and you can see why power reduction is becoming so very relevant in the present era. Thank you for the day.