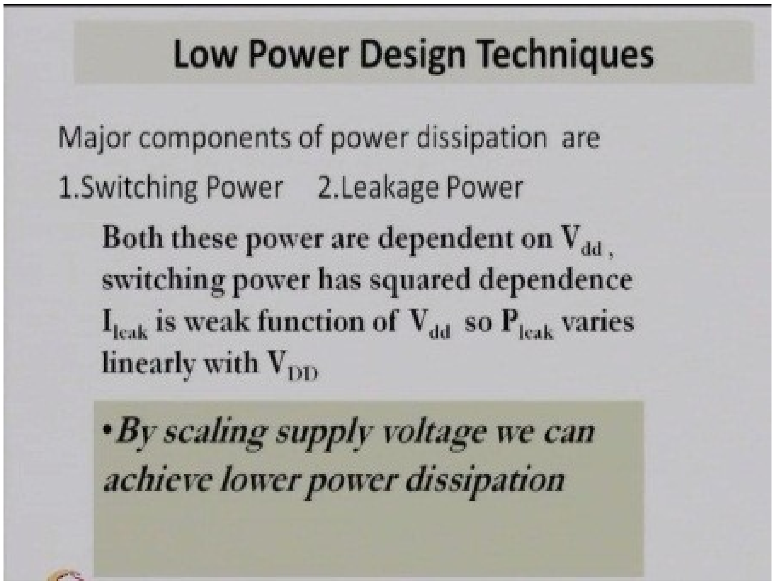
Advanced VLSI Design
Prof. A. N. Chandorkar
Department of Electrical Engineering
Indian Institute of Technology – Bombay

Lecture – 08
Low Power Design Techniques

Once again welcome to our advanced VSLI design course and we have been discussing right about the Power Design and essentially as I discussed earlier in my earlier two talks I did say that the current need of every system is to have low power dissipations, a variety of users particularly for the handrail or mobile systems. The power is-- low power is very, very important. But so is the power requirement of a higher MIPS requirement of normal microprocessor as well.

(Refer Slide Time: 00:57)



Low Power Design Techniques

Major components of power dissipation are

1.Switching Power 2.Leakage Power

Both these power are dependent on V_{dd} , switching power has squared dependence I_{leak} is weak function of V_{dd} so P_{leak} varies linearly with V_{DD}

- *By scaling supply voltage we can achieve lower power dissipation*

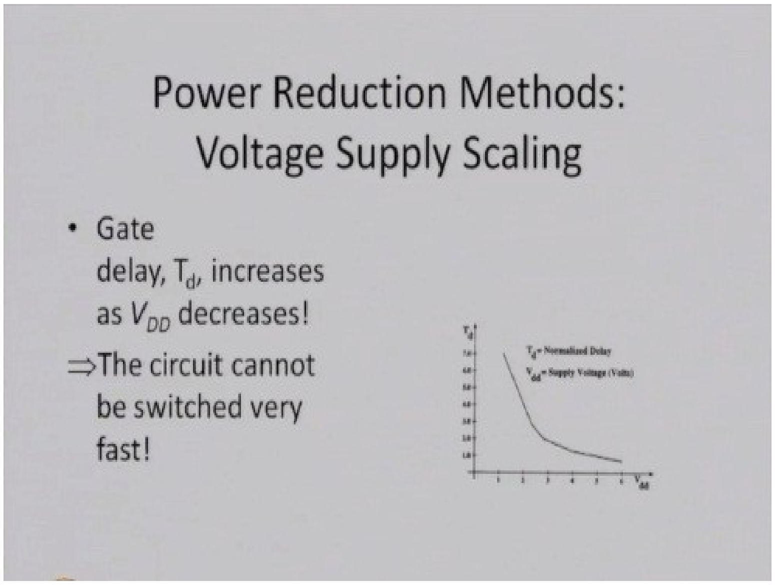
So with this the low power microprocessor is required in variety of applications on the field including for example wireless networks, wireless nodes. One is really worried about these days about the low power design, so I continue with what I have said earlier discussed about generalities then I discussed about the power dissipation in CMOS circuit and particularly we said there are three possible powers, one is due to the capacity charging and discharging.

The other is short circuit power which essentially is also can be clubbed into the Switch Power. And then finally we talked about the third one which is the most important worrying right now is the Leakage Power. Now, if you see the two powers switching power or the leakage power what we observe now that both are these power are dependent on the power supply voltage Vdd and therefore switching power has squared dependence.

And the power, I mean switching power is propositional to Vdd squared CVDD squared F and therefore Vdd is a term appearing in the dynamic power so is this appear in the case of switching power that is the short circuit power. However, in the case of leakage power the since the power is nothing but I leakage power into Vdd so obviously P leak is linearly proposition to VDD. But the fact about all these whether it is proposition to VDD or VDD square.

If you want to reduce the power dissipation one is it is quite obvious to us that we must reduce we must scale down on power supply. And as we scale down the power supply voltage we can achieve a low power dissipation.

(Refer Slide Time: 02:39)



Power Reduction Methods:
Voltage Supply Scaling

- Gate delay, T_d, increases as V_{DD} decreases!

$\Rightarrow$The circuit cannot be switched very fast!

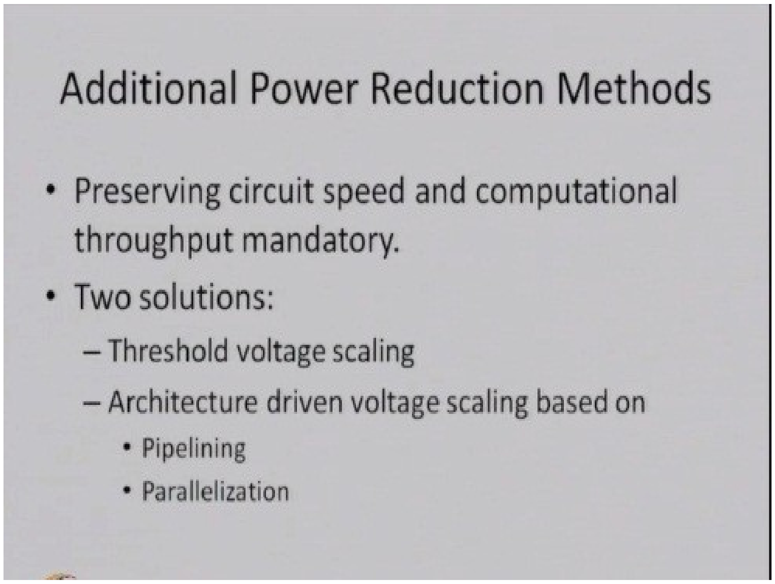
There are number of ways in which power reduction can be done by voltage scaling. The one of the techniques shown here is Gate delay, please remember if I reduce the power supply voltage we know that the Gate Delay Td increases, we also know how this occurs because if the power supply voltage goes down the charging and discharging current available for capacitance become smaller.

And since that become smaller obviously the time taken to charge or discharge the load capacitance will be larger. So if you reduce the power supply then the speed of course is the something which you have to give up. So obviously this is figure which I shown you here is essentially is talking about delay versus power supply and once you lower power the delay starts rising.

And if you going to have a lower than around 1.5 voltage lower the delay actually rises very, very shortly. Now this fact has to be understood that one cannot scale down power supply voltages so very easily because the speed is also one of the major criteria. However, there are circuits as I discussed earlier which are only called low power or low standby power circuits which actually are not really worried about the speed.

They may be working on less than 500 mega vaults or even lower sometimes and those circuit power maybe the major criteria in that case power supply voltage can certainly be reduced.

(Refer Slide Time: 04:18)



Additional Power Reduction Methods

- Preserving circuit speed and computational throughput mandatory.
- Two solutions:
  - – Threshold voltage scaling
  - – Architecture driven voltage scaling based on
    - Pipelining
    - Parallelization

The additional power reduction methods which will allow this low power design possible is the preserving circuits speed and computational throughput mandatory. Now if you have a criteria that you have to have same speed and also you want the data to be available to you at the given throughput rate, in that case what can be done. There are two resolutions, one of course threshold voltage scaling.

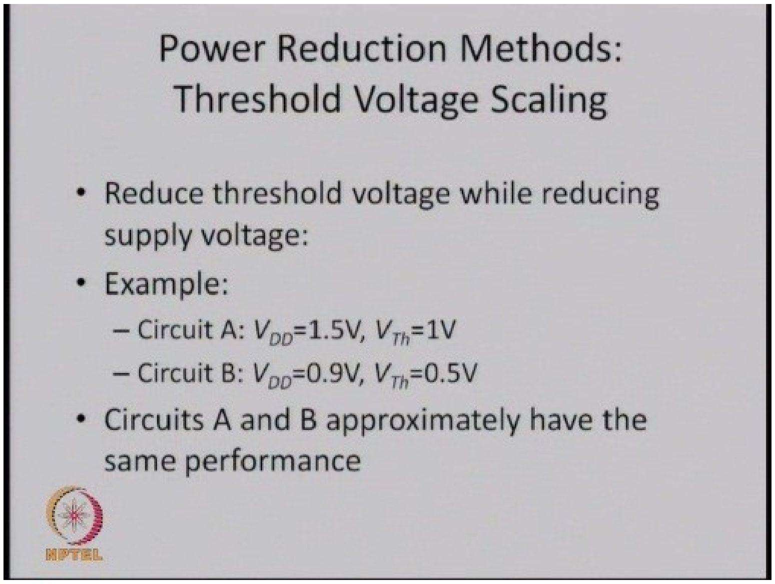
So you can reduce the threshold voltage because after all the current in the mass transistor is propelled to Vgs minus Vt since it is proposition to VGS minus VT if you reduce VT and VGS can go even lower VDD but VDD minus VT then can be higher and therefore current can be made higher and if higher the current larger is speed we know and therefore this speed can be persevered by keeping lower power supply voltage but also reducing the threshold voltage.

The second possibility or second solution to get the same speed and same throughput for the logic you are implementing is the architectural driven device voltage scaling based. There are two possibilities in which one can have pipelining or one can have parallelization and either of this architecture technique. Though the speed of the net circuit does not remain to be lowered then what you are expecting.

But the individual components do as if running at a higher-- lower clock rates or you can say running higher clock but the net circuit speed is what is desired and individually therefore by pipelining or parallelization we may be able to reduce the net power. And this is very interesting because that is the only way circuit we can actually reduce the power dissipations. Of course some cost-- whenever you achieve something you will have to give up something.

And let us see what happens when we go through such architectural driven voltage scaling. Now, first and the foremost method which most people believe is to reduce the power reduce threshold voltage while reducing supply voltage.

(Refer Slide Time: 06:21)



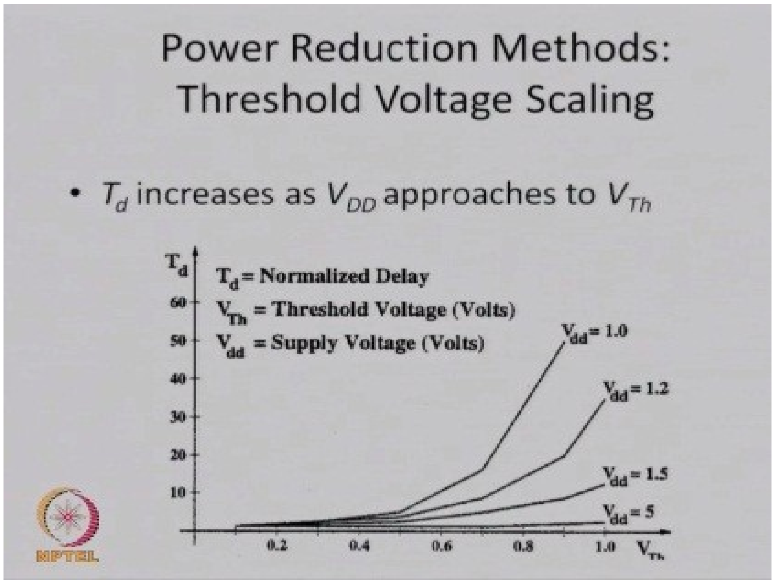
Power Reduction Methods: Threshold Voltage Scaling

- Reduce threshold voltage while reducing supply voltage:
- Example:
  - Circuit A: V_{DD}=1.5V, V_{Th}=1V
  - Circuit B: V_{DD}=0.9V, V_{Th}=0.5V
- Circuits A and B approximately have the same performance

Say example, there can be two possibilities the circuit A may actually have VDD 1.5 Volt and threshold of 1 Volt. And a circuit B may have VDD 0.9 Volt and threshold of half a Volt. And if you see the; if you calculate the using this VDD and threshold voltage values; if we calculate the propagation delay and therefore the speed and hence we call the performance they seem to be almost identical.

So now we can see that I reduce V DD, I reduce threshold and I can still attain the speed.

(Refer Slide Time: 06:54)



Power Reduction Methods: Threshold Voltage Scaling

- T_d increases as V_{DD} approaches to V_{Th}

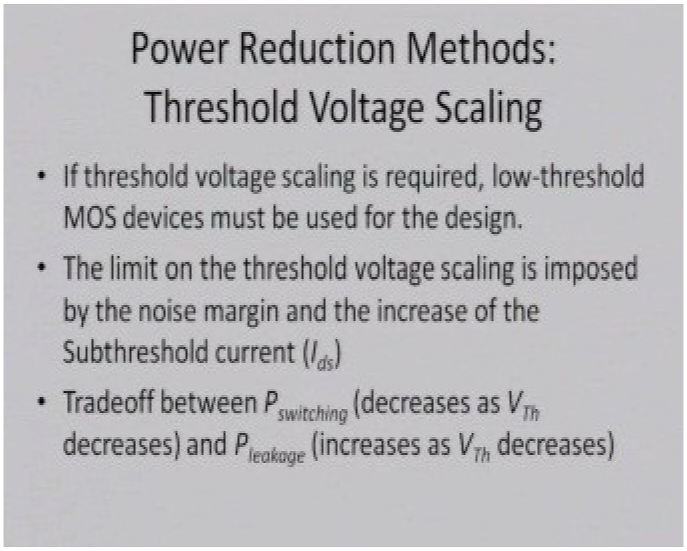
There is a this is some same similar graph shown here Td increases as V DD approaches threshold that is where how much VDD can be reduced. So here is a figure shows threshold voltage versus delay. And we are varying a threshold voltage from say 0.2 Volt to 1 Volt and we are also varying VDD from 1 Volt to 5 Volt. Of course 5 Volt hardly anyone is working but at the same time if you see that if your threshold voltage is anything less than 1 volt.

And if you are power supply voltage is large the delay of course a normalized delay is very, very small around one or two normalized delays. Whereas, if you start reducing the power supply voltage from 5 Volt down-- upwards you can see as VDD goes to 1.5 we already under normalized delay of around say even at say threshold voltage is reduced to say 0.4 or 0.5, so it is around still slightly higher.

But if you go when the VDD approaches threshold voltage which is your say point this is 1 Volt and making it 0.4. So one can see that, as I approach this threshold voltage, power supply voltage near to threshold voltage then the Td starts enhancing. And this is very, very important because essentially which means that the difference between threshold voltage and the power supply voltage had to be maintained higher so that the speed is not lost.

Alternatively, speed will be lost if VGS minus VT reduces. What I am going to say about this is during this threshold voltage scaling.

(Refer Slide Time: 08:40)



Power Reduction Methods:
Threshold Voltage Scaling

- If threshold voltage scaling is required, low-threshold MOS devices must be used for the design.
- The limit on the threshold voltage scaling is imposed by the noise margin and the increase of the Subthreshold current (I_{ds})
- Tradeoff between $P_{switching}$ (decreases as V_{Th} decreases) and $P_{leakage}$ (increases as V_{Th} decreases)

A threshold voltage scaling voltage is required low- threshold MOS devices must be used for the design. The limit on the threshold voltage scaling is imposed by the noise margin and the increase of the Subthreshold current as we some of these problems we shall see little later. Means the noise margin can be understood let us say I have a-- I reduced my threshold voltage to 0.2 volt or 0.5 volt.

And I always supply a 0.8 volt or even less than 0.8 say 0.6 volt so I my VGS minus VT sufficiently high to create a good amount of current. But when I go below point 0.2 volt which is my threshold one can see from here that 0.2 voltage is hardly 8kt by your room temperature and therefore anything below this voltage-- if you have to reduce your output to lower than Vt which will be around close to 15 millivolt or so you are well within the noise levels.

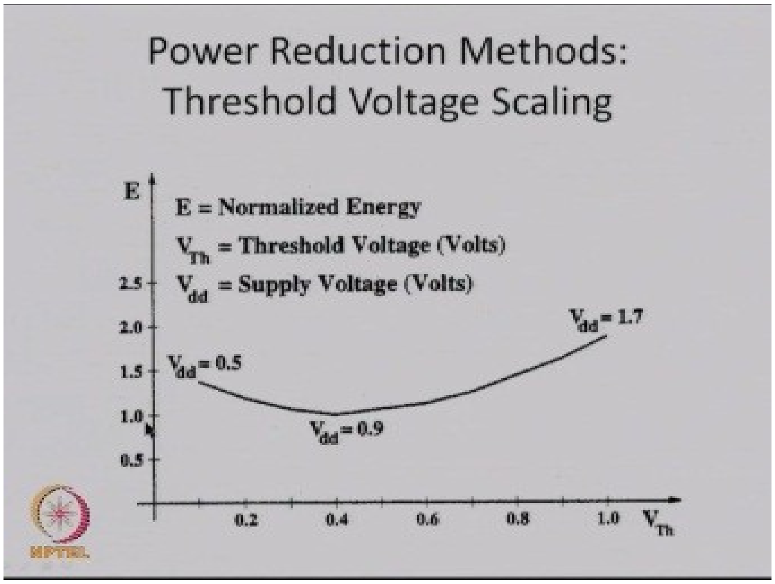
And therefore it will be very, very difficult for us to maintain noise margins because the thermal noise or even other noises may actually start dominating at those points and therefore the noise may actually dominate if threshold voltage is very, very low. However, please remember as long as your VGS minus VT or therefore VDD minus VT is large the speed certainly can be met.

However, we also worried later as we shall see that the threshold current that means when the VGS goes below VT we know the current continuous to flow and you know the reduction in voltage less than threshold has a slope of IV characteristic there, ID VGS characters there and one finds that a small change anywhere in threshold actually changes sub-threshold currents.

So obviously, if there 16 milli volt per decade as what we say so one can see from here that large subthreshold current can also lead to leakage currents. And therefore, too much reduction in threshold voltage is not very much advisable unless otherwise you do some more tricks to really ward off this issue. So therefore tradeoff says P switching or dynamic power we certainly P switching decreases if threshold decreases.

And P leakage which increases as threshold increases. So when I ask to tradeoff between the dynamic power and the leakage power in the regime in this threshold voltage is scaled down.

(Refer Slide Time: 11:10)



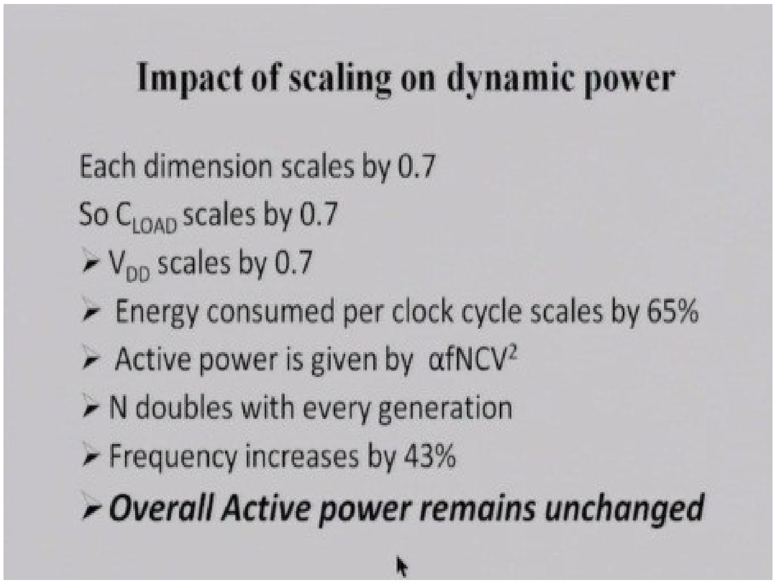
Power Reduction Methods:
Threshold Voltage Scaling

So in the same thing what I said about power reduction method in threshold voltage scaling technique. Instead of showing you power now I am showing you energy, which is normalized energy versus threshold voltage, plotted at three power supply voltage VDD 0.5, 0.9 and 1.7 and we believe that as you reduce your threshold voltage down for the energy minimal to occur and energy we can show you the energy is nothing.

But power in time energy power into time energy per unit time-- energy power unit time is the power. So if you integrate over the time for which this period for which you want to know energy it can be figured out by writing such an expression for energy minima that around 0.4 volt of threshold and at that time the VDD being 0.9 will give you the minimum energy.

So for a given technology and given device structure which you use you may get an energy minimum, and for which there would be some kind of an optimal threshold voltage which will give you the lowest power dissipation.

(Refer Slide Time: 12:18)



Impact of scaling on dynamic power

Each dimension scales by 0.7

So C_{LOAD} scales by 0.7

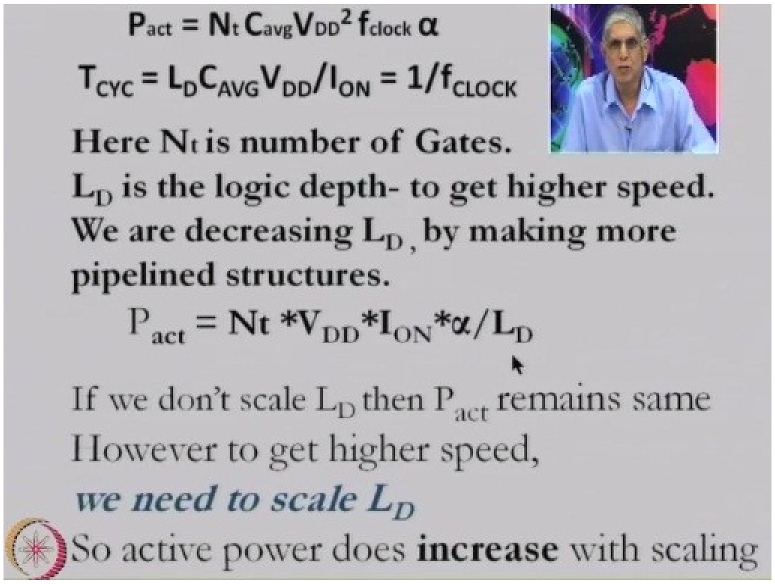
- V_{DD} scales by 0.7
- Energy consumed per clock cycle scales by 65%
- Active power is given by $\alpha fNCV^2$
- N doubles with every generation
- Frequency increases by 43%
- ***Overall Active power remains unchanged***

Now, people keep saying that I also said in my first few talks we keep saying that as far as moves law every technology node is nothing but the reduction of lengths and widths and every other thing by a scale factor by 0.7 and if we reduce scale down all the parameters 0.7 to follow moves laws that every year components doubles because 0.7 into 0.7 is 14.49 which is half. So obviously if you reduce the component density will double which is very obvious by moves law.

However, what it impacts on is C LOAD scales by by 0.7, VDD also scales by 0.7. Let us assume right now we are scaling by saying what is called constant VD scale. Energy consumes per clock cycles scales by 65%; active power given by we have already derived it alpha NCV square and n is the number of components for this so it doubles every generation. So frequency increases by 43%.

However, if you do this calculation at the end of the day one interesting result one observed that if you scale down by 0.7 volt everything the overall active power remains unchanged. So you are trying to reduce the power you thought you have active power reduced but this alpha NCV square term actually does not change very much when you scale down. Okay.

(Refer Slide Time: 13:57)



$P_{act} = N_t C_{avg} V_{DD}^2 f_{clock} \alpha$

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

Here N_t is number of Gates.

L_D is the logic depth- to get higher speed.

We are decreasing L_D, by making more pipelined structures.

$P_{act} = Nt * V_{DD} * I_{ON} * \alpha / L_D$

If we don't scale L_D then P_{act} remains same

However to get higher speed,

we need to scale L_D

So active power does **increase** with scaling

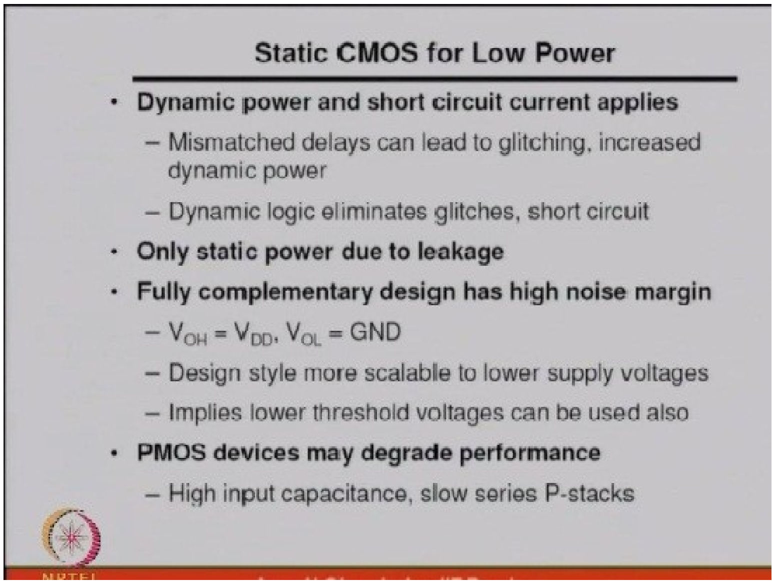
So the P active power is essentially is equal to Nt is a number of Gates; C average is the average capacitance of the Gate, or load which is; VDD square is the power supply voltage square; F clock is the frequency at which data is loaded, alpha is activity coefficient. This we already derived earlier, so if we write the period for one data cycle, then it can be written as LD which is logical depth that is if you have a three series component three series dates in one deriving the other then LD is three so LD is called the logical depth.

So if you see this term LD C average VDD by is the ION is the cycle time which is nothing but one by F clock. So one can see if I substitute this part if I reduce the-- this is interesting if I reduce the logical depth from here my clock frequency goes up. However, if I substitute this in case of power which is nothing but Nt, VDD, I on upon alpha, I just substituted this term inside here for F clock 1 upon LD C average VDD I ON substitute here.

And I get a term Nt-- P active power is Nt that is number of Gates VDD which is power supply voltage, on current which is the on current from the transistor, alpha is the activity coefficient and LD is essentially the logical depth. So if you reduce the logical depth the power does increase and therefore one interesting feature is if you do not scale LD then P active remain same. However, to get higher speed we need to scale down LD.

And if you show active power, so you will actually to improve this, you reduce the LD because then only your-- by logical effort we are seeing that occurs and the speed will go up and if you scale down L D then obviously the P active power will increase of scaling.

(Refer Slide Time: 16:02)

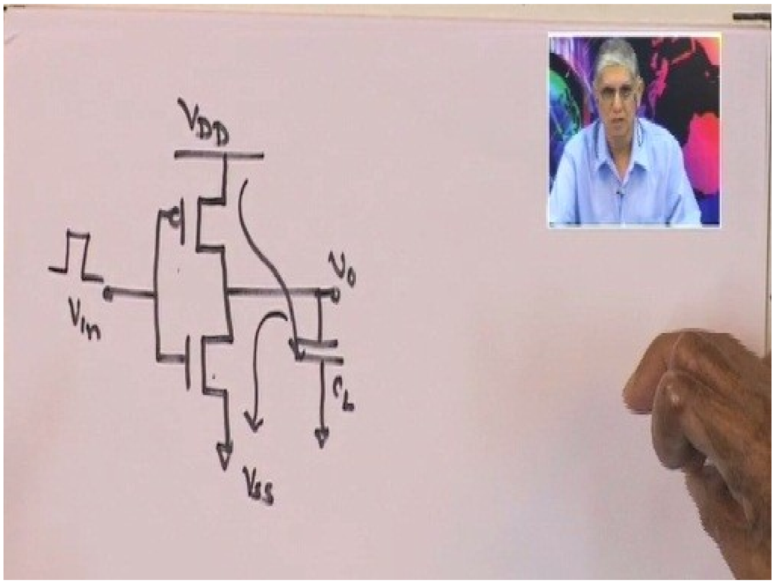


Static CMOS for Low Power

- Dynamic power and short circuit current applies
  - Mismatched delays can lead to glitching, increased dynamic power
  - Dynamic logic eliminates glitches, short circuit
- Only static power due to leakage
- Fully complementary design has high noise margin
  - $V_{OH} = V_{DD}$, V_{OL} = GND
  - Design style more scalable to lower supply voltages
  - Implies lower threshold voltages can be used also
- PMOS devices may degrade performance
  - High input capacitance, slow series P-stacks

Now there are variety of CMOS design styles and we will like to see each of them in light of their low power performance. Now if you take a static CMOS which is the most standard CMOS circuit which essentially.

(Refer Slide Time: 16:25)



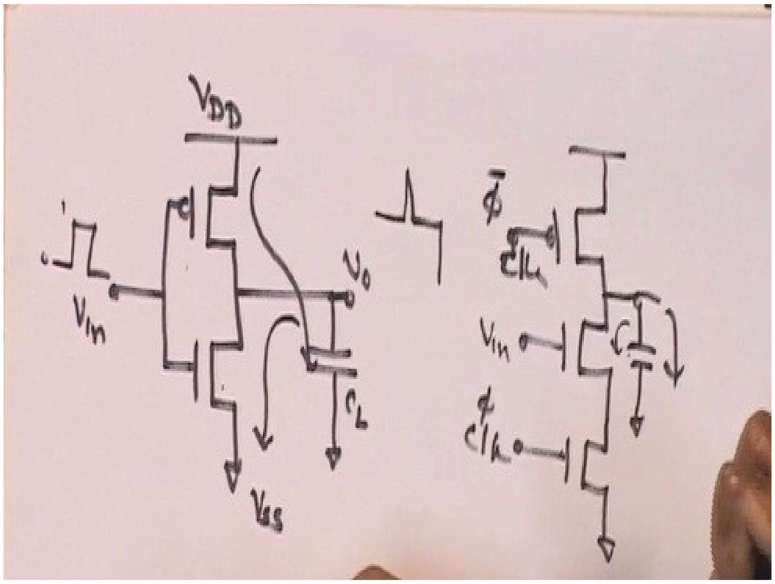
Let say represent inverter which shows you have a P channel transistor and you have a n channel transistor which is connected and this is your power supply this is your VSS ground; this is your input and this is deriving a net load of capacitive L which is load capacitance and this is your V0. So if you look at this the logical power we know it is proposed.

Because the charging through the power supply or discharging through the ground essentially leads to power dissipation and in between when the way input changes state high to low or low to high both transistors on for a while which is essentially gives the static power on that and that is and also because of the scale down devices both in N channel and P channel never really turned on fully and which means there will be a leakage path throughout.

Now if you see dynamic power in short circuit current applies mismatched delays can lead to glitching. So the problem is this when we shall see it depends on this logical input you give which is 0 to 1 in a pulse form and depends on the sizing you do and depending on the currents you are able to switch in this transistors there may be a possibilities of switch occurring something like this sorry glitch occurring something like this.

And any such glitch can lead to a high power high dynamic power dissipation. However, if you do-- this is occurring because transistor this was initially at 0 on and this was off when this goes to 1 this became off and this became on they switch actually.

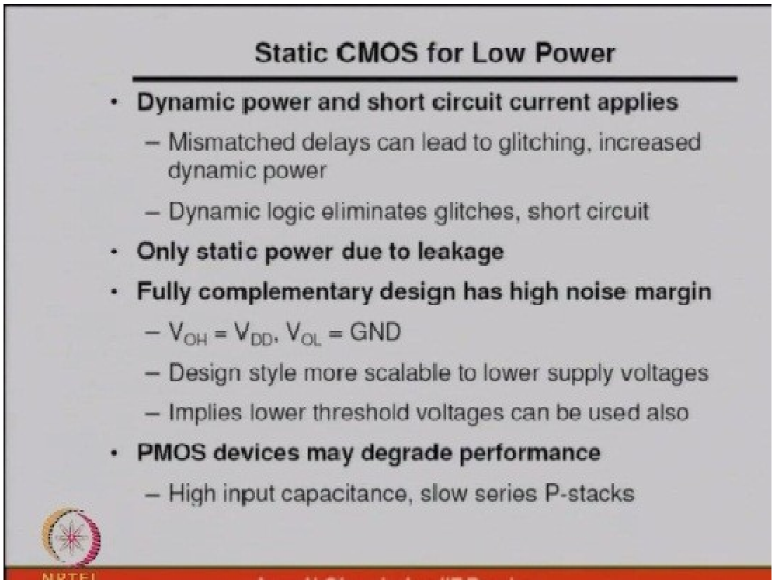
(Refer Slide Time: 18:24)



But if they do not switch as in the case of dynamic let us say I have a P channel device which is given to a clock or phi bar and I have a dynamic logic in which, I may keep another transistor or I may not I may actually put this also at clock phi phi bar let us say, this is independent. But if you see in a dynamic system when the clock bar is 0 or when the clock is 1 -- clock is 0 clock bar is-- sorry this is phi so if this 1 this is 0 and that P channel conduct and charges-- this is called P charge.

But when phi goes to 1 P channel turns off and this capacitor discharges or does not discharge depend upon the input on this logic which essentially means there is no short circuit path between power supply and ground and therefore dynamic logic eliminates glitches because there is no time they really switch on the same time. So the short circuit power and glitches can be minimized if you use dynamic logic. That is what I said when we say only static power due to leakage.

(Refer Slide Time: 19:43)



Static CMOS for Low Power

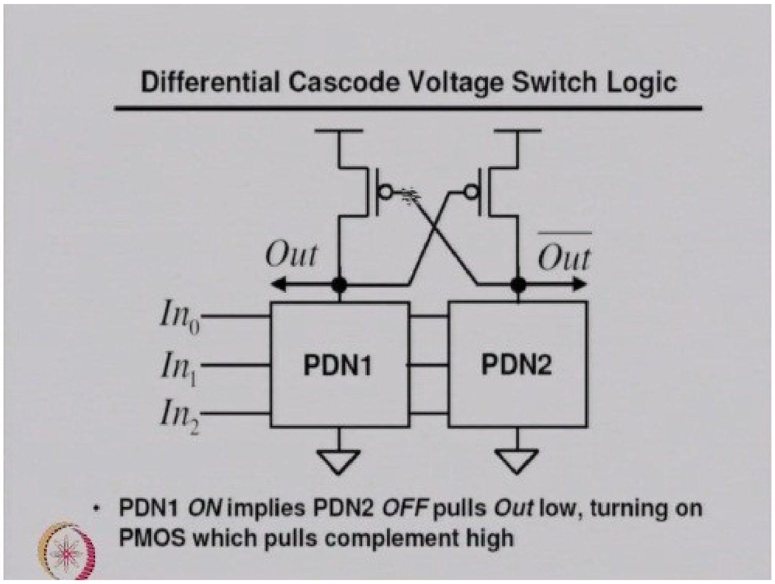
- **Dynamic power and short circuit current applies**
  - Mismatched delays can lead to glitching, increased dynamic power
  - Dynamic logic eliminates glitches, short circuit
- **Only static power due to leakage**
- **Fully complementary design has high noise margin**
  - $V_{OH} = V_{DD}$, V_{OL} = GND
  - Design style more scalable to lower supply voltages
  - Implies lower threshold voltages can be used also
- **PMOS devices may degrade performance**
  - High input capacitance, slow series P-stacks

Now fully complementary design has high noise margin, we know if we keep VOH = VDD and V OL is ground in which case the design style more scalable to low power supply voltages because one can then reduce because any of one ground where there is power supply, so scaling down is much easier and the full swing is also possible, which essentially implies that lower the threshold voltage can be used if we have fully complementary designs static CMOS.

The PMOS devices may degrade performance simply because high input capacitance, slow series P-stacks. Actually if you have large numbers of P-stacks in this they can actually slow down because of you know their mobility being lower in a stack case the net capacitance may increase and also you may have a slower series P-stacks and therefore they may degrade the performance in the case of static CMOS as far as performance I am talking about speed.

The other possibility of you know looking for a power-- low power design is use differential Cascode voltage which is logic, this is very interesting logic which create both out and out bars simultaneously.

(Refer Slide Time: 20:57)



Differential Cascode Voltage Switch Logic

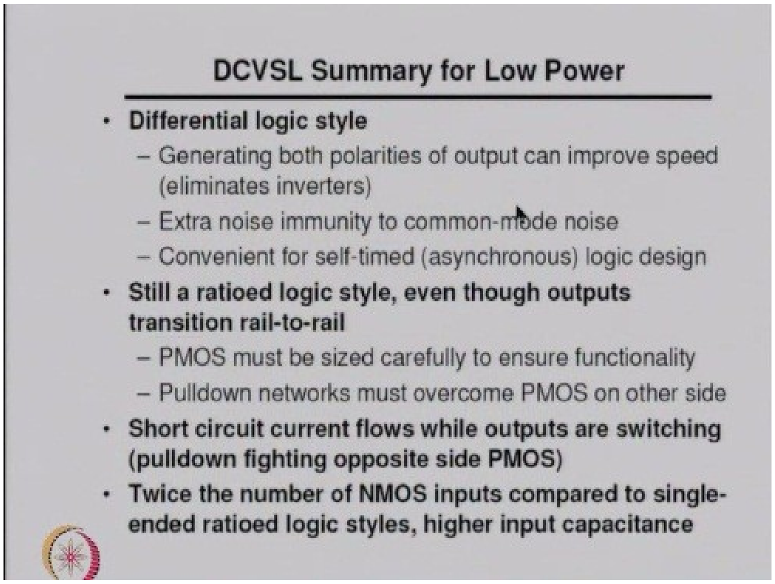
- PDN1 *ON* implies PDN2 *OFF* pulls *Out* low, turning on PMOS which pulls complement high

One can see from here there are two P channel devices is cross connected to the output. And PDN1 and PDN2 are N channel logical this-- only thing is the criteria is that when this is on PDN2 is off or when the PDN2 is on therefore PDN1 is off. So if this let us say for a given inputs PDN1 requires this output to go to 0 in which case at that time anyway this was not receiving any data.

So one can see from here that since this is zero this turns on the output becomes higher which is complementary in a same goal. In case this PDN1 is essentially is requires this to be one that is this transistor turns off and this transistor last stage if this much this high this must be acting this must be 0 so 1, 0 is retained. Now please remember PDN2 OFF pulls Out low and turning on PMOS which pulls complement high, this pulls complement high.

So this Differential Cascode Voltage Logic as one advantage essentially that both out and output and there is no additional power required in actually getting both 2 and complement outputs. The only problem which I see here is you have two such b locks to work on additional hardware and the power wise it does not active power but it has an additional area and additional area additional component density which you have to pay for.

(Refer Slide Time: 22:46)



DCVSL Summary for Low Power

- **Differential logic style**
  - Generating both polarities of output can improve speed (eliminates inverters)
  - Extra noise immunity to common-mode noise
  - Convenient for self-timed (asynchronous) logic design
- **Still a ratioed logic style, even though outputs transition rail-to-rail**
  - PMOS must be sized carefully to ensure functionality
  - Pulldown networks must overcome PMOS on other side
- **Short circuit current flows while outputs are switching (pulldown fighting opposite side PMOS)**
- **Twice the number of NMOS inputs compared to single-ended ratioed logic styles, higher input capacitance**

Now what I said I may repeat again differential logic style generating both polarities of output can improve the speed because they are no inverters, see basic problem in all the time taken as we define the propagation delays going from 0 to 1 and 1 to 0 average delays TPHL + TPLH since there is an inversion, in this case there is no inversion and since there is no inversion the speed is very high as does not have inverters.

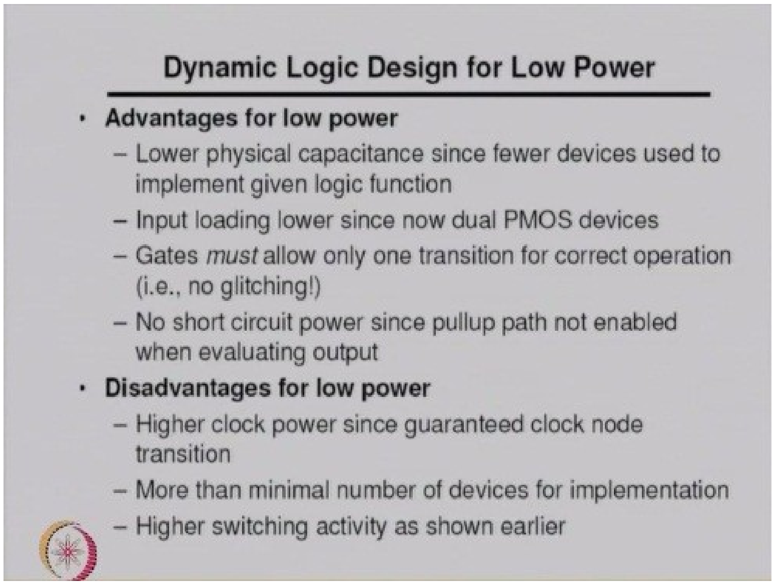
Since it is a extra noise immunity to common-mode noise. So there is a differential circuit obviously common-mode noise is eliminated which is standard for your differential amplifier kind of system. The third possibility advantage one see in the case of DCVSL is convenient for self-timed asynchronous to logic style, because the PDN1 and PDN2 can concerned because there is no really clock requirement.

However, please remember it is still a ratioed logic style even the output transition is rail-to-rail, PMOS must have to be sized carefully to ensure going to one and zero properly or pulldown--pull up ratio has to be such that 1 and 0 can be fully at the full rail voltages. And therefore to pull down network must overcome the PMOS on the other side because you know now they will actually the other side PMOS will actually will also draw a current from the PDN1.

And therefore one has to pull down network has to really ensure that it receives that much current to make it 0 or 1 as desired. The third point which is of interest short circuit current flows while outputs are switching, pull down fighting opposite PMOS. When the outputs are switching pull down fighting you can see from there if I go back. If you see here, when this is pulling down this is essentially loading it. Okay. Or for this-- this is essentially loading it.

So one can see from here that the short circuit current flows while outputs are switching, pull down fighting opposite side at PMOS. Twice the number of NMOS puts because you want both out bar you have PDN1 and PDN2 compare to single ended ratioed logic styles and therefore this two higher capacitance. But it certainly gives you low power. So what are the advantages of DCVSL logic?

(Refer Slide Time: 25:20)



Dynamic Logic Design for Low Power

- **Advantages for low power**
  - Lower physical capacitance since fewer devices used to implement given logic function
  - Input loading lower since now dual PMOS devices
  - Gates *must* allow only one transition for correct operation (i.e., no glitching!)
  - No short circuit power since pullup path not enabled when evaluating output
- **Disadvantages for low power**
  - Higher clock power since guaranteed clock node transition
  - More than minimal number of devices for implementation
  - Higher switching activity as shown earlier

It is the following that for the low power the advantages are lower physical capacitance since fewer devices used to implement given logic function. Input loading lower since now sense dual PMOS devices. Gates must allow only one transition for correct operation and therefore no glitching. No short circuit power since-- sorry, this is we are talking about dynamic logic and not DCVSL.

In a dynamic logic as we discussed here in this figure you can see in a dynamic logic lower physical capacitance since fewer device only you have all N logic and P channel N channel as a complimentary for phi bar so we have are fewer devices. So obviously the net capacitance seen here will be smaller. Input loading lower since new now dual PMOS device so actually you do not if this is only loaded here so one does not require additional loading.

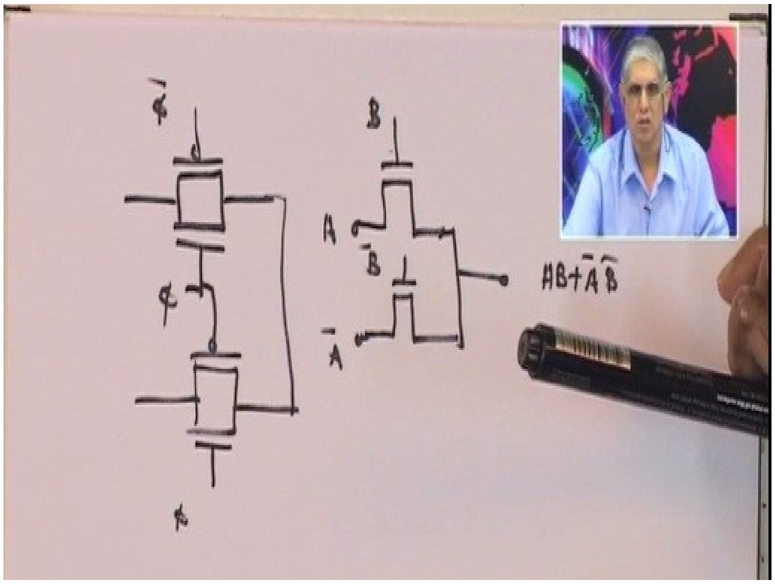
Case allow one transition for correct operation since it is only one input should occur when the clock bar is available or clock is not one at that time since your input as to settle which has half the cycle roughly you can say therefore no glitching requirements. No short circuit power since pull up part can path not enabled when – see when this is on so this is on-- this is off when this is on this is off.

And therefore short circuit power is minimal since pull up path is no enabled when evaluating output. However, every good thing has this disadvantages well. There is disadvantage in the light of a low power requirement. Higher a clock power since guaranteed clock more transitions you know since you are using phi bar you need a power which will guaranteed make this on-off of these two devices.

So that means you require driver for clock and this driving of the clock essentially means that you are consuming additional power. More than minimal number of devices for implementation because you can see if we have static PMOS I would have required only two required three and of course it does not mean currently what I meant was that if you need extra clock and clock bar and that means the net number of devices are not smaller the logically number or devices maybe smaller.

Higher switching activity as shown earlier and because of that the dynamic looks to be good in some cases though may have because of clocking system may actually show higher switching activity and we know dynamic power is proposition to switching activity. Now the other possible way of implementing logic we all know is essentially coming from CPL.

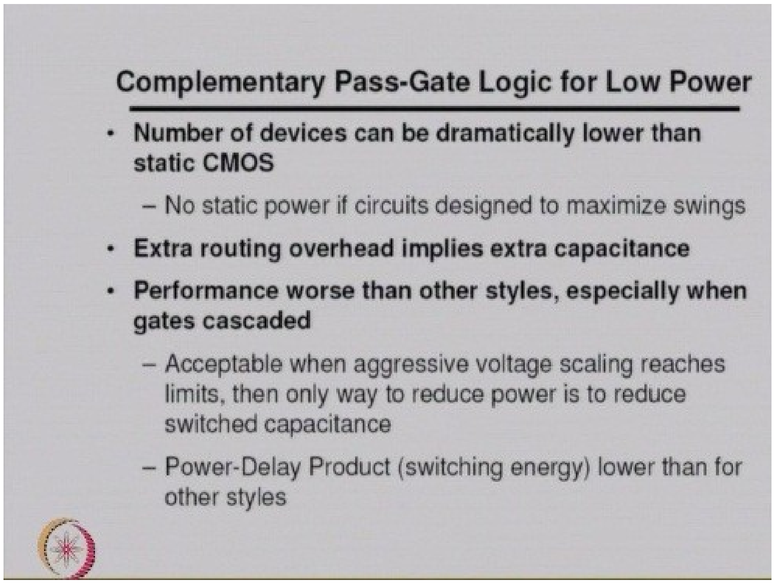
(Refer Slide Time: 28:12)



So for example if you have one complementary this is CMOS pass gate another CMOS pass gate okay. This may be phi this may be phi bar and this may be input or even normal pass transistor logic without CMOS can be shown equivalently something like this. If I connect this one can see from here if it is a A, B, A bar, B bar this output could be AB+A bar B; this is essentially its now which we are trying to implement from here.

So what essentially passed transistor allows is to create any communitarian logic can be implemented and using feedback pass or using a loop paths even we can create sequential blocks using pass gates. Now since there you can see from here there is no power supply; there is no inversions going on so obviously one feels that it should create a much low power and it will also have much lower transistor count okay.

(Refer Slide Time: 29:41)



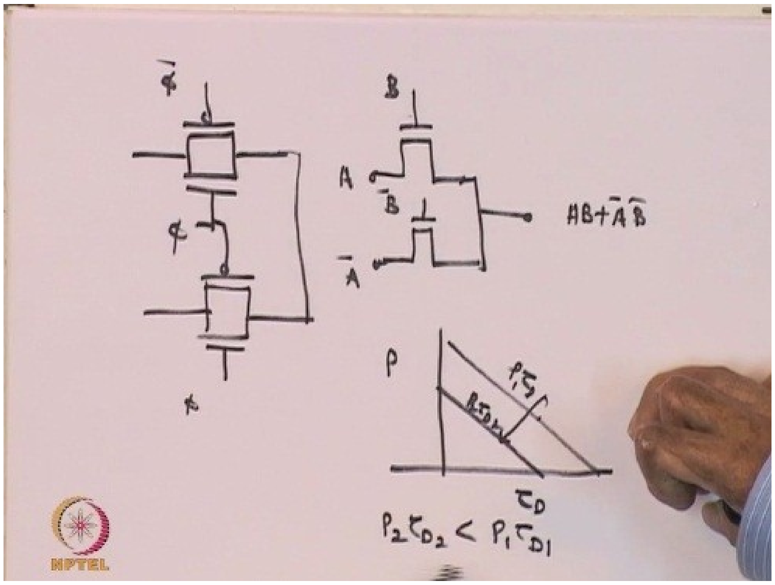
Complementary Pass-Gate Logic for Low Power

- **Number of devices can be dramatically lower than static CMOS**
  - No static power if circuits designed to maximize swings
- **Extra routing overhead implies extra capacitance**
- **Performance worse than other styles, especially when gates cascaded**
  - Acceptable when aggressive voltage scaling reaches limits, then only way to reduce power is to reduce switched capacitance
  - Power-Delay Product (switching energy) lower than for other styles

Since there are number of devices are dramatically lower than static CMOS then one can say no static power is circuit design to maximize swings, extra routing overhead implies extra capacitance you will have to run AA bar and BB bar and you need some extra capacitance to derive. Performance worse than other styles, especially when gates are cascaded. Of course these are acceptable when aggressive voltage scaling reaches limit then only one way to reduce power is to reduce switched capacitance.

So in some case when you are scaling down technologies to 45 or 32 or down CPL may become very dominant design style or CMOS style of implementation of logic. One of the big advantage it will give power delay product lower than for any other style we had earlier discussed to you in my earlier figures that it is.

(Refer Slide Time: 30:39)

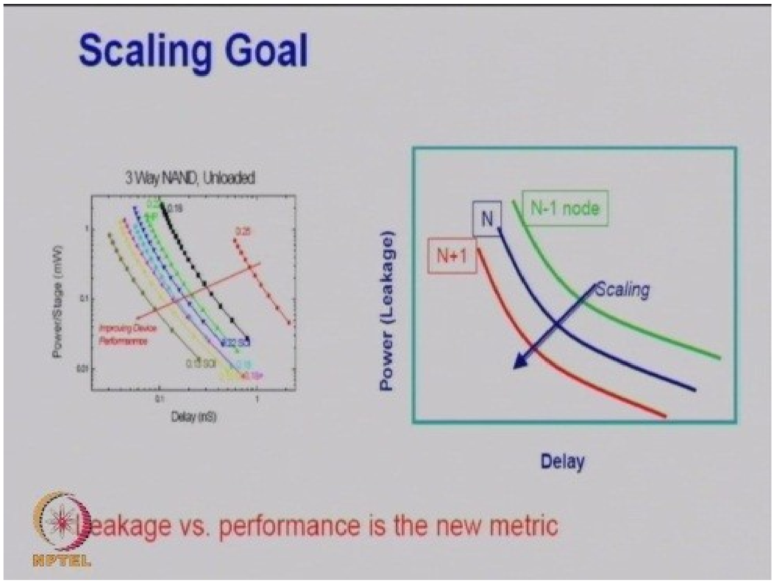


If I plot tau over 6 power, okay this line is called power delay line and if you want better design you should go from here to here which has let us say this is P1 this is P2 tau D2 such that P2 tau D2 is less than P1 tau D1 tau D1 is a delay and power is the delivered by the gate, so if you can see come down lower than this at higher speed you have a low power and CPL does improve power delay product which is less than or which is called also switching energy is always less than any other design style using CMOS or NMOS circuits.

So essentially this is good, however it has extra routing price; extra capacitance power is somewhere there; it is not very easy for a switch to act ideally and therefore there are problems for KT by C noise or sharing of charge or feed forward of course CMOS takes care of feed forwards but otherwise it does have its own problem as a pass gate. So when I am scaling down over the years that is what mostly I like suggesting.

So we look into the problem which we are now getting, we already discussed earlier that the larger node technologies 0.25, 0.5 or even higher the channels links were very high since the channel links were very high since the channel links were dopings were also smaller relatively between the substrate and the ratio. So what use to happen then that the diode leakage currents or any other leakage mechanism leading to the leakage power was much smaller even the subthreshold slope that the net leakage current does not change drastically.

(Refer Slide Time: 32:48)



Scaling Goal

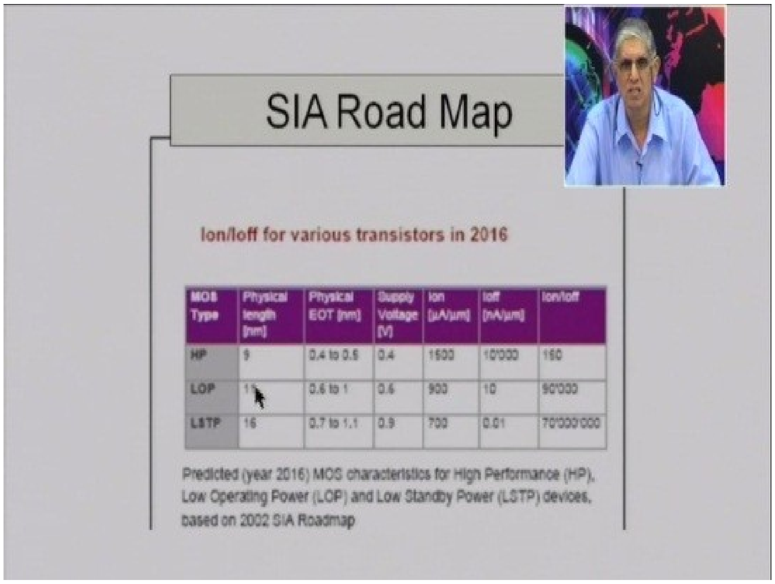
Leakage vs. performance is the new metric

However, what has happened over the years. As I am scaling down I see now if I plot powers per stage shown here millivolts this and I plot against delay here in another seconds. So if I change the technology say from 0.25 down to say 90 nanometers or like this 0.13 or even below if we go below one is finding that the delay is decreasing definitely okay, delay is decreasing but the leakage power at any node you go for any delay say for example for 90 nanometer or 0.13 for the same decade the leakage power per stage is higher.

So essentially what is happening here the same thing can be shown here, if this is your n-1 node, this is your nth node and this is your future node as you scale down okay the power will keep on increasing your respective which node you are working at. Lower this-- the though it is not very obvious but the slope of rise of leakage power is much higher compare to these two. If you go even further it will actually rise much faster here.

Now this means the new metric or performance one should now worry about if you are looking for gigahertz applications of your CMOS ship you must also be equally worried about the leakage power which is visited without being utilized in any sense.

(Refer Slide Time: 34:37)



SIA Road Map

Ion/Ioff for various transistors in 2016

MOS Type	Physical length [nm]	Physical EOT [nm]	Supply Voltage [V]	Ion [µA/µm]	Ioff [nA/µm]	Ion/Ioff
HP	9	0.4 to 0.5	0.4	1500	10'000	150
LOP	11	0.6 to 1	0.6	900	10	90'000
LSTP	16	0.7 to 1.1	0.9	700	0.01	70'000'000

Predicted (year 2016) MOS characteristics for High Performance (HP), Low Operating Power (LOP) and Low Standby Power (LSTP) devices, based on 2002 SIA Roadmap

A typical SIA Roadmap shows. There are two currents of interest one is called the on current in which when the transistor is on VGS is which is greater than VGS is greater than threshold, once it is on current particularly the on current is normally defined when the device is instauration. And off current includes all the currents like subthreshold current, the diode leakage current and also the other any kind other depended leakage currents which will shall show you later, seven of them in fact, they all contribute to off states.

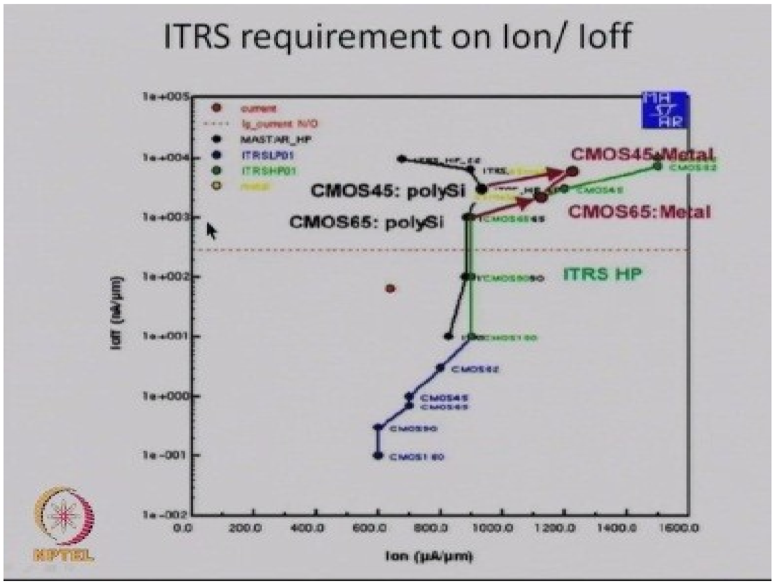
So off current to on current is very important to maintain larger on current and smaller off current is what is ideally required you need every large ratio of I on to I off so that you have no leakage power, very, very low leakage power can adjust our on current to suit your performance requirement. We also know there are three kinds of circuits I have already discussed one is high performance then other is low power and the last is low standby power.

For the high performance, low operating power and low standby power devices of course this is 2002 roadmap I could not get the latest one from my ITRS in this format. However, this shows the trend from this it is the channel length physical length so you are going from in nanometers, okay. You can see what is requirement of this, EOT essentially called Equivalent Oxide Thickness because great the thickness also is going to be reduced as you scale down.

So if you see this is 0.4 to 0.6 micron 0.6 to 1 and 0.7 to 1. equivalent, okay in really life it will be less than 10Angstroms. On current is defined as per micro ampere per microns it requires 1500 this for LOP phi 900 this is 700; off current required is 10 micro amperes per micron this is one and this is 0.1 so ratio is we want larger for as you go from high performance to low standby power. This is something we are requiring in the all circuit which we are going to design for three such applications.

As I said this is one typically mobile or a -- even what we could say is the iPads or tablets which are running on a battery power you are looking for LSTP circuit. Many of them are LSTP circuit.

(Refer Slide Time: 37:32)



ITRS requirement on Ion/ Ioff

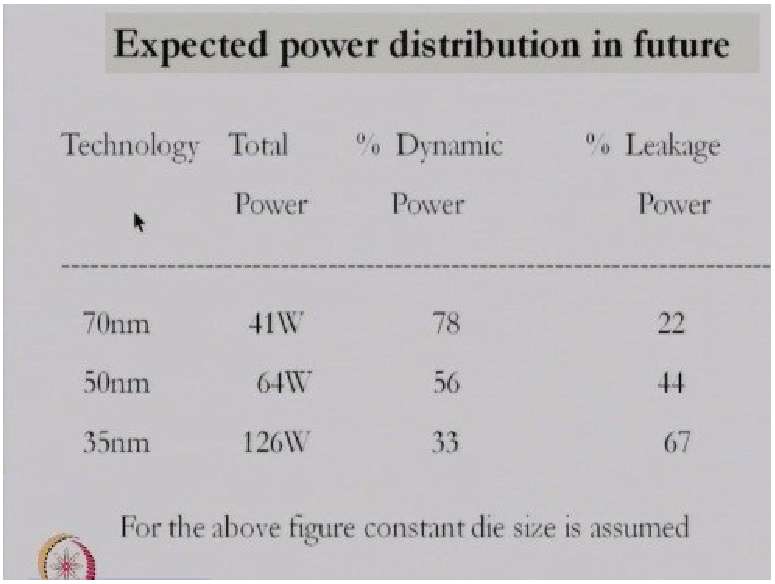
Now looking at the technology side just to give some quick glimpse of this-- this is a figure which according to ITRS which is the technology roadmap society which shows on current micro ampere micron sorry off current versus the on current which is micro and micron and this shows you a figure of a different technologies CMOS 180 nanometer down and this green one is high performance circuit in which the on current is very high, okay and the correspondingly.

Sorry, on current is constant and off current is-- so if you really look for high performance circuit in this figure somewhere around 900 or 1000 micro ampere per micron of this the off current has gone from this two orders up to this. So our idea is somehow in technology to stem this height going suddenly jumping this is the actual technology graph CMOS 65 nanometers, okay. If you go for newer technologies we must do some kind of what you would say work function engineering or something.

And you use metal gates instead of poly silicon gates you can see poly silicon gates are not as good as metal gates because what we are really trying now is to increase the on current okay at least this size but should not increase you can off current, if you can improve the this at least the high performance circuit data can be required and similar argument can be given for low power and low standby power.

So now also the another issue of circuit design for low power or low standby power is that technology which is going to be used for all individuals for each of those process will not be identical you have to tailor your technologies once skin for what the kind of application you have in mind for the particular design.

(Refer Slide Time: 39:40)



Expected power distribution in future

Technology	Total Power	% Dynamic Power	% Leakage Power
70nm	41W	78	22
50nm	64W	56	44
35nm	126W	33	67

For the above figure constant die size is assumed

There is an interesting figure-- a table shown to you here this is called (()) (39:38) university. They have shown of course some when they showed it the technology has not scaled down to 35 nanometers then but they still valid actually. They evaluated as a scaling was going down they expected that the net from 70 nanometers which is essentially 65 nanometers 45 and 32 they are just rounded up of the calculations for by them.

This is 65 nanometer node; this is 45 nanometer node; this is 35 nanometer node. At those nodes technology the total power they say 41 Walt at 65 nanometers and 120 almost triple at 32 nanometer node. This is one and half times. So 45 to 65 nanometer that is not very much but when you go 0.7 of that further then we find almost doubles of 64 and triples compare to this.

Now of this total power dynamic power or total power of dynamic 41 watts at 65 nanometer nodes the dynamic power is 78%, whereas the leakage power is just 22%. But if you scale down to say 50 nanometer or what called 45 nanometer node, the dynamic power reduces to 56% and the leakage power enhances to 44%. And if you go further down which 35 nanometers, this become 33% the leakage power is 67%.

And if you further scale down if you can extend this value it may happen that the leakage power will be 80% and this may be useful power to you would be dynamic power will be on current power would be only 20%. What it essentially is trying support -- of course I assume 3mm by 3 mm fix chip size fix chip size the chip size that may change in real life and therefore this data is not the actual data or exact data for real measurement but it does show you the trend.

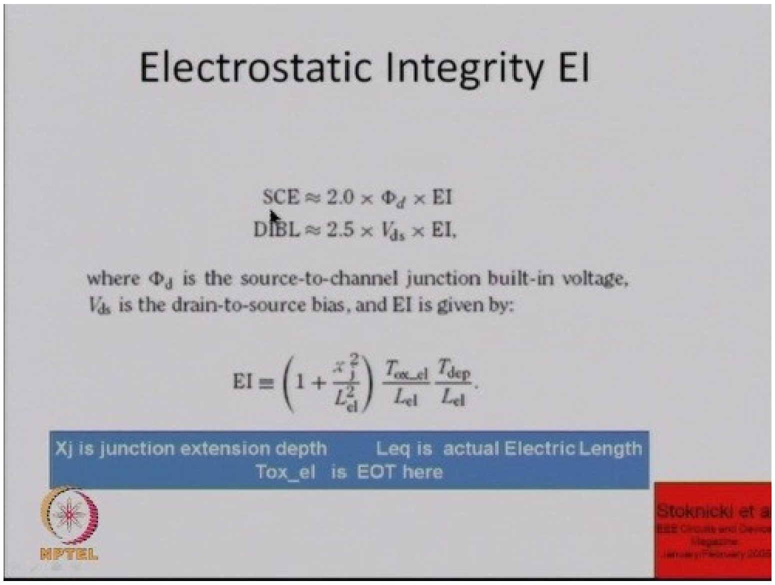
Now just to make this little lecture little interesting I mean now see you that as most of the newer IPods or iPhones or including some Samsung or Nokia or Sony or anyone you name even the Chinese one they are already working on 32 or below the 32 nanometer processors, on processors or any other processors. Since they are working on low technology nodes the power dissipation is major worrisome for all of this circuits, mobile phone for example.

Now if this leakage power becomes higher which means the battery will dream faster, if you are not really operating it the mobile phone is not on. The advantage of this things not happening good for you is that if you keep your mobile on all the time that means you keep talking on your mobile your battery may drain less power compare to if you would have just kept it in the standby mode.

And I think one such reason now I figure why younger colleagues of mine who are under graduate or graduate studies they keep talking on mobile phone insanely instantly when they talk on roads or they are in the room or wherever they are probably they knew this foil better than me, they realized that battery drains slower if they keep talking and if they just keep the mobile on the last hand by more the power dissipates faster, dissipation is higher.

And they have to recharge the mobile or this is the fund but that maybe possible maybe reasons why people talks so much on mobile these days. The another problem in technology when I scale down the technologies there is something a word which IT analysis said which is called the Electrostatic Integrity E1 or EI—sorry EI.

(Refer Slide Time: 43:41)



Electrostatic Integrity EI

$$\text{SCE} \approx 2.0 \times \Phi_d \times \text{EI}$$
$$\text{DIBL} \approx 2.5 \times V_{ds} \times \text{EI},$$

where Φ_d is the source-to-channel junction built-in voltage, V_{ds} is the drain-to-source bias, and EI is given by:

$$\text{EI} \equiv \left(1 + \frac{x_j^2}{L_{el}^2}\right) \frac{T_{ox\_el}}{L_{el}} \frac{T_{dep}}{L_{el}}.$$

Xj is junction extension depth Leq is actual Electric Length
Tox_el is EOT here

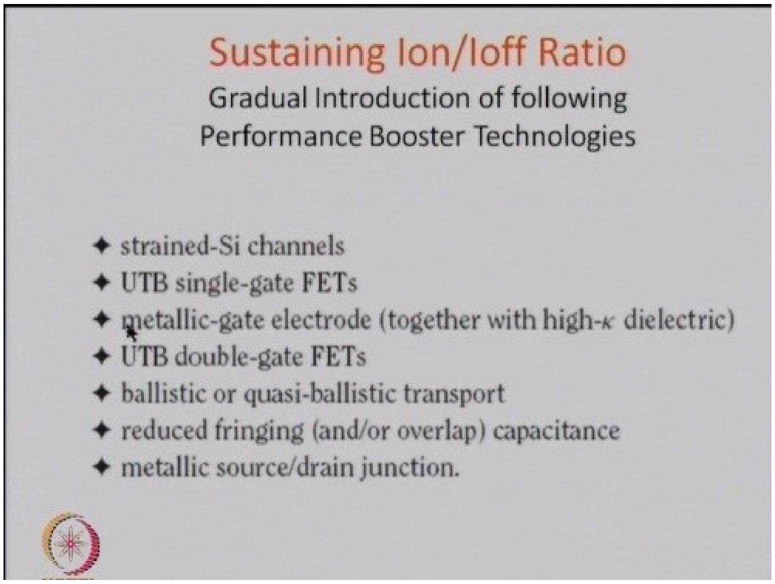
Stoknicki et a
EEE Circuits and Device
Magazine
January/February 200

If you look at Short Channel Effects and the DIBL which is essentially controls the threshold voltages and the currents in the mass transistor therefore the mass transistor a currents in the mass transistor. It can be shown that this phi d which is called the Short to Channel Build-in voltage and Vds of course is drain to source bias, okay. Then we see that the short channel effects is can be start when it is 2 time phi d EI, DIBL is 2.5 time Vds time EI.

So if where EI is essentially given by 1+ x j square L el square T ox equivalent oxide thickness by channel length T depletion by L channel length electron actual equivalent channel lengths. So if you see from the technology side if you scale down this okay equivalent channel length and since you are scaling down it goes down and this is square term which is increasing. Now this means the short channel effect and the DIBL coefficient will be directly propositional to the technology you use.

And since they both effects the threshold in some way they actually reduce the threshold voltage, while the leakage paths will be much stronger subthreshold path will be very strong and therefore lot of leakage power will happen if you have lower nodes of technology.

(Refer Slide Time: 45:27)



Sustaining Ion/Ioff Ratio

Gradual Introduction of following Performance Booster Technologies

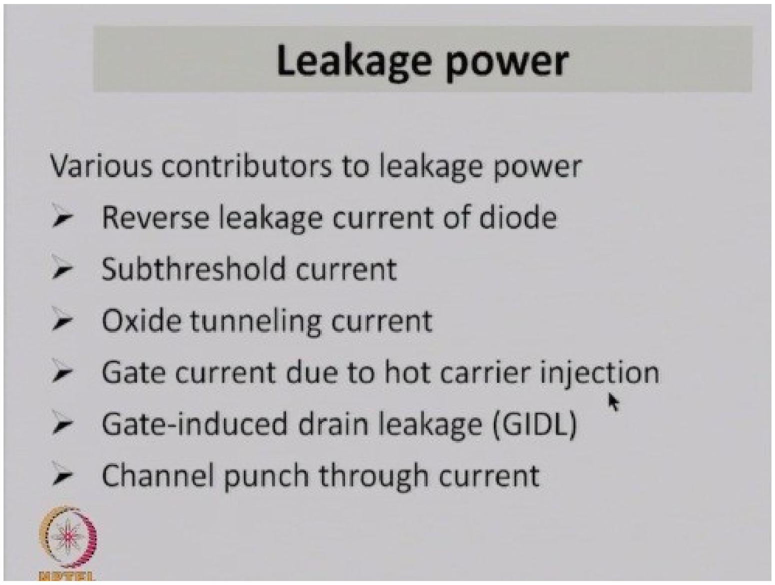
- strained-Si channels
- UTB single-gate FETs
- metallic-gate electrode (together with high-κ dielectric)
- UTB double-gate FETs
- ballistic or quasi-ballistic transport
- reduced fringing (and/or overlap) capacitance
- metallic source/drain junction.

Now there are other ways for high performance circuit in particular if you wish to-- I am talking for fun because we are saying their power is not a criterion as much as the speed. So for this there are performance booster technologies available and they have been gradually introduction the following way. Intel has started working on strained-Silican channels, you have Silicon germanium as instead of source in silicon.

You have ultra-thin body single-gate FETs, UTB as they called; then you are using metallic-gate electrode with high-k gate oxide extra silicon dielectric; your Ultra Thin Body double-gate FETs which is dual gate-FETs which will call this. The modified version is nothing but Fin-FETs and we will look into the power using Fin-FET little later. Then there are other divisor possible which are quasi-ballistic or ballistic transport possibility reducing the fringing or and/or overlap capacitances and metallic-source junction instead of silicon source/drain junctions.

These are of course technology details.

(Refer Slide Time: 46:23)



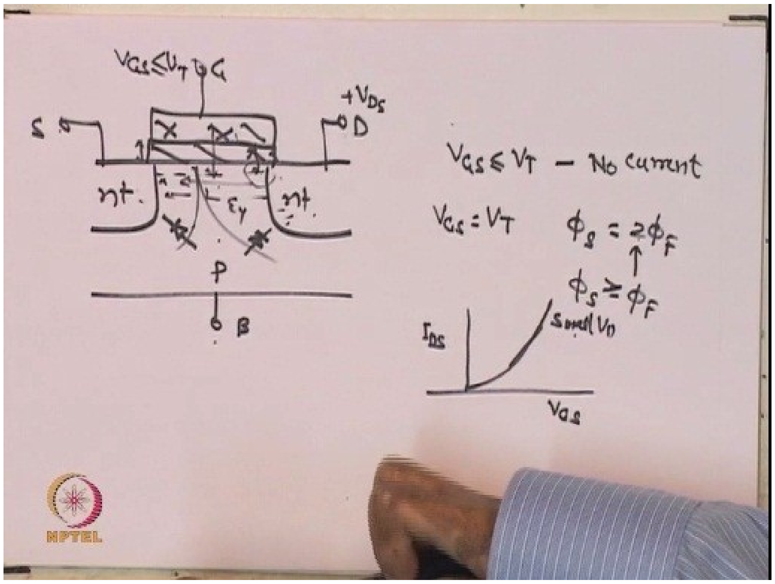
Leakage power

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

However, the part worrisome which I just now discussed in the mass transistor as you scale down them to technology nodes particularly below 45, 32, 22, 16, 11, 7 and where I do not know lower will be. These are the various contributors to leakage power. There is of course-- I will show you figure little later because two things are not showable here to me, I did not prepare properly.

(Refer Slide Time: 46: 57)



The first and the foremost transistor if you have a mass transistor shown here this your source this is your drain a same channel and this is your substrate this is your bulk; this is your insulator and this is your gate. This is your drain this is your source this is your n channel mass FET, so if you see this-- this is essentially a diode same of course here. There is a diode between source and substrates.

So particularly source is grounded bulk is grounded even then the diode has a leakage current very small. If you see this in n channel this will become V DS which is higher; this is heavily reversed by diodes and therefore leakage current is very high in this. So the first contribution to leakage power is the reverse leakage current of these two diodes then we believe that when VGS is less than Vt there is no challenge here, okay.

The channel does not exist VGS < VT no current because we say there is no electron channel. But in reality we know that the threshold voltage is defined as VGS = VT when the band bending we say phi s = 2 phi f where phi s is the surface potential and phi f fermi potential. Now in reality the inversion starts when phi s is just greater than phi f a less than sorry I mean yeah its greater than phi f value wise.

So which means the band has already bend for this so we say that reverse leakage of the current of the diode essentially may be – sorry subthreshold current essentially will occur when phi s lies between phi f and 2 phi f and that means even if VGS goes below VT the ID VGS characteristics IDS VGS characteristic all small, small VGS do show some characteristic like this and that means this current may not be very, very small.

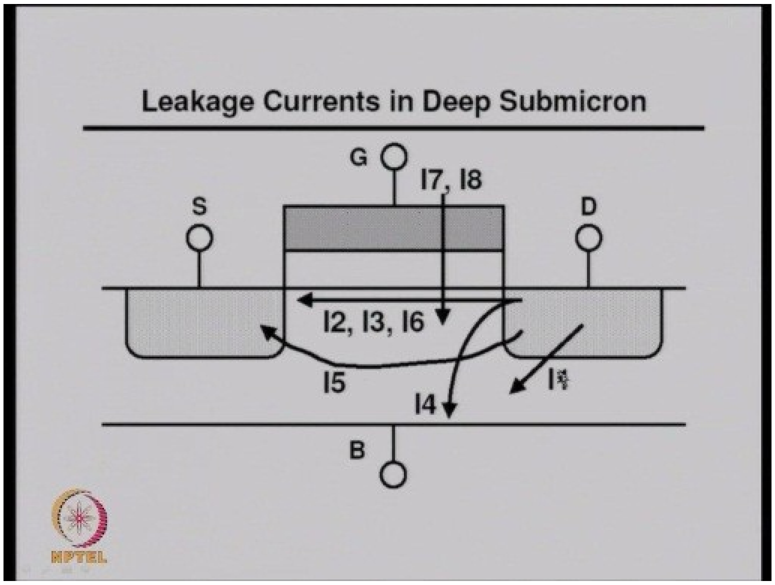
And in fact the slope which we are going to use here which is called subthreshold slope may actually increase that means the number the change in current change in voltage for large change in current will be different and in which case one has to worry about leakage currents being stronger at lower technology nodes. Then since this also thickness of oxide is also scaling there is an oxide internally directly carriers which is called internal normal tunneling band to band tunneling one can say.

Then there is a possibility that if this channel length is very, very small as the scaling goes the electric field across this Ey-- this is Ey direction is very high; carrier somewhere at the drained and may actually pump inside the oxide and in which case there will be hot carrier injection in the gate and which may constitute the current.

There is also if you increase if the channel length are very larger this electric failure larger this is also larger, the gate can influence the drain electric fields, we believe that it is only gate the induces the channel but the drained together may induce the leakage path much stronger and this we say gate induced rain leakage. The channel punch through current the two depletion layer with associated with this diode.

And this diode may merge in this case punch through can occur. So these are possibilities of leakage power and these are what they are shown here.

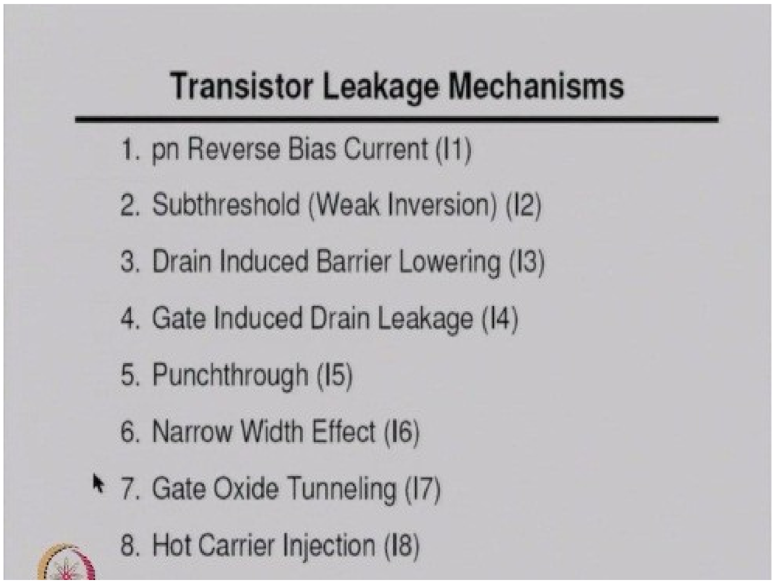
(Refer Slide Time: 51:10)



Leakage Currents in Deep Submicron

I want essentially the diode leakage I2 essentially I2, I3 and I6 are all are between this I17, tunneling and gate induce the leakage are through this and then there is of course there is a current part of this current band may come here and this hot carrier effect is also part of this, so one can see in a dip submicron part of nanometer down technologies these currents may dominate already written the names which are the currents.

(Refer Slide Time: 51:42)

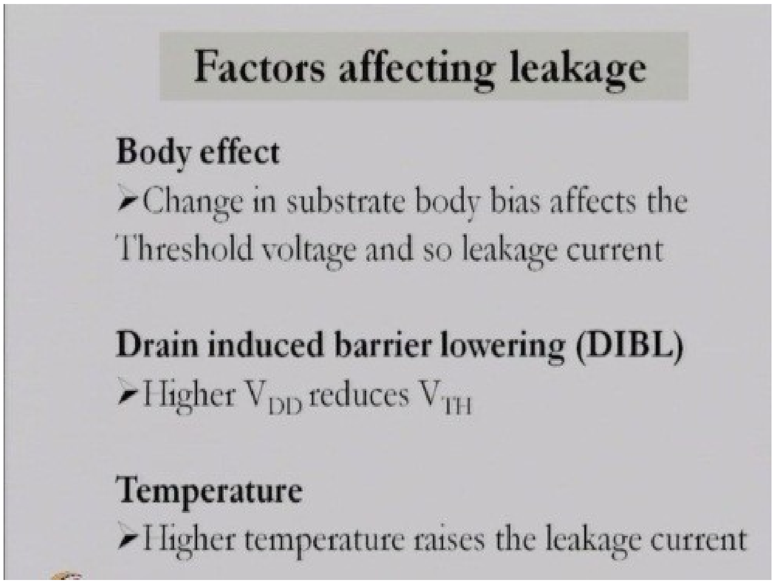


Transistor Leakage Mechanisms

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)

And each current is essentially contributing to transistor leakage mechanism.

(Refer Slide Time: 51:51)



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

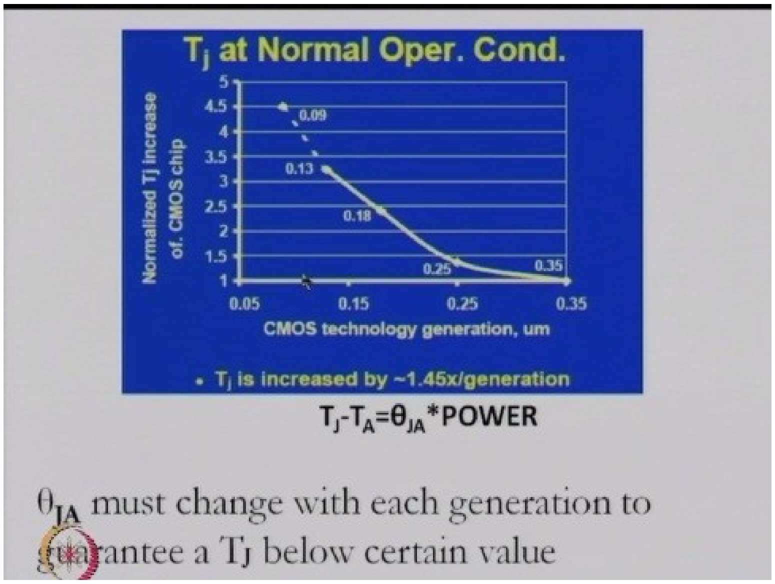
➢Higher temperature raises the leakage current

So, what is the factors of leakage? Change in substrate body bias affects the threshold voltage. We are very interested to know how control leakage so we are now looking into the—we say change in substrate body bias affects the threshold voltage and so the leakage current. There is another issues which is very interesting is called DIBL Drain Induced Barrier Lowering. We always thought that gate controls the channel depletion layer.

But essentially if the channel lengths are smaller the source/drain also have large depletion layers and drain can actually therefore contribute to electric fields and reduce the barrier at the drained and therefore currents can increase. Essentially the threshold changes because of the DIBL coefficient changing. However, we know the higher the threshold of VDD, VTH will be smaller reduces the threshold voltage because than in that case since threshold is lower currents are larger one can say.

Then there is a last but not the least, higher temperature raises the leakage currents which is obvious diode currents are e to the power Q by and Kt. So temperature is very strongly increasing the currents. And then increase in currents leads to increase in nature and since increase the temperature at different -- increasing as we already discussed over the years.

(Refer Slide Time: 53:09)



T_j at Normal Oper. Cond.

- T_j is increased by ~1.45x/generation

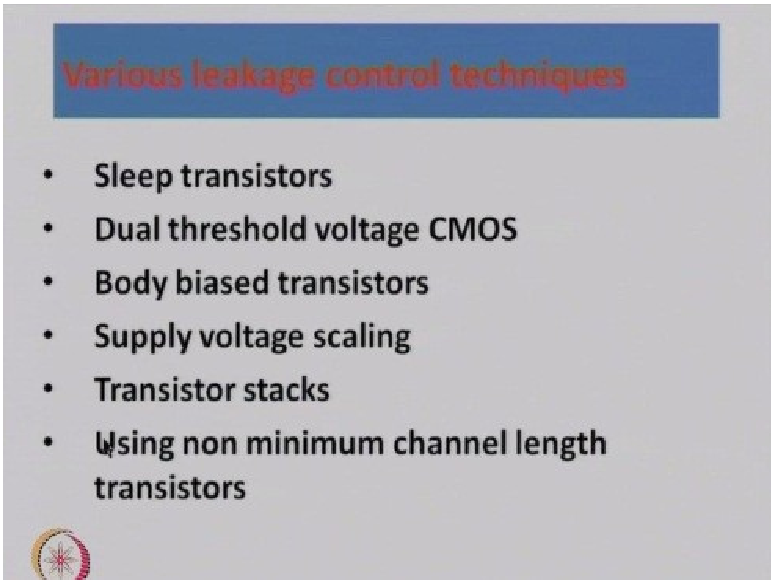
T_J-T_A=θ_{JA}*POWER

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

Now if junction temperature arises over the unbend temperature which is essentially called thermal resistance into power, please remember junction temperature increases 1.45 per generation times the last one and therefore if you scale down one node then you actually increase the one and half times the leakage thermal temperature Junction temperature.

And therefore to maintain junction temperature within a limit so that the junction remain junction one has to worry about cooling which is called we must change the thermal resistance from junction to ambient such that heat is dissipated faster.

(Refer Slide Time: 53:47)



Various leakage control techniques

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

We will come back to it next time and we will show you that if we want to control so called these leakage currents what circuit techniques-- lastly of course it is not circuit lastly we will say different channel length which is also to some extent of technique; we will say if you use these techniques probably one can still play with reduced leakage currents. Thank you for the day.