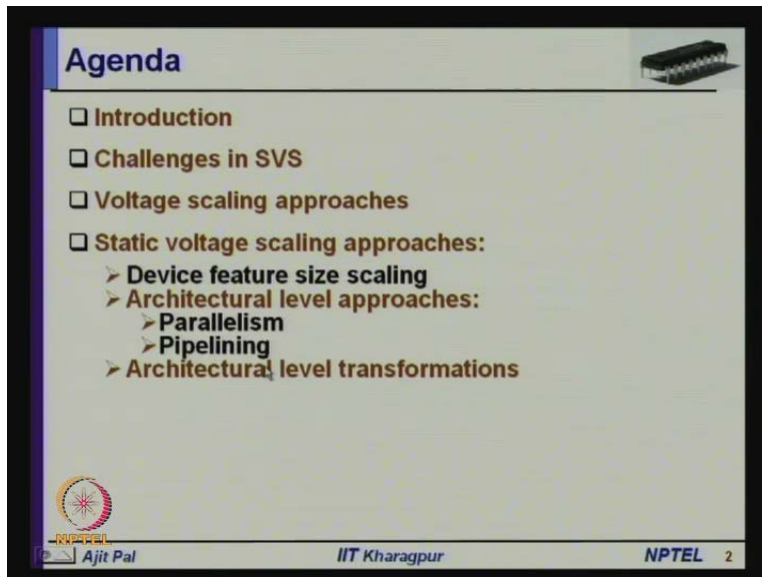


Low Power VLSI Circuits and Systems
Prof. Ajit Pal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture No. #22
Supply Voltage Scaling – I

Hello, and welcome. To today's lecture on supply voltage scaling, this is the first lecture on this topic. In the last three lectures, we have discussed about various sources of power dissipation. And we have identified which parameters we can control, and based on that we shall start our discussion today on supply voltage scaling.

(Refer Slide Time: 00:46)



Agenda

- Introduction
- Challenges in SVS
- Voltage scaling approaches
- Static voltage scaling approaches:
 - Device feature size scaling
 - Architectural level approaches:
 - Parallelism
 - Pipelining
 - Architectural level transformations

NPTEL
Ajit Pal
IIT Kharagpur
NPTEL 2

And here is the agenda of today's lecture after a brief introduction, where we shall recapitulate what we have discussed in the last couple of lectures. We shall discuss about challenges in supply voltage scaling, and then various voltage scaling approaches that we shall discuss, and after that we shall start our discussion on static voltage scaling approaches. Starting with device feature size scaling after that architectural level approaches like parallelism, pipelining, and then architectural level transformations.

(Refer Slide Time: 01:36)

Sources of Power Dissipation in CMOS Circuits

P_{dynamic} : 1. $P_{\text{switching}} = \alpha_L C_L V_{dd}^2 f$
 $+ \sum_{i=1}^K \alpha_i C_i V_{dd} (V_{dd} - V_{t_i})$

2. $P_{sc} = \frac{K W}{L} \cdot \frac{1}{2} V_{dd}^3 \left(1 - \frac{2 V_t}{V_{dd}}\right)^3 \cdot P \cdot f$

P_{static} 2. $P_{\text{glitching}} = \alpha_L C_L V_{dd}^2 f$

$P_{\text{gate leakage}}$ $P_{\text{diode leakage}}$
 $P_{\text{subthreshold}}$ $P_{\text{static}} = V_{dd} \cdot I_{\text{leakage}}$

So, first let us briefly recapitulate what we have discussed in the last three lectures. We have discussed about various sources of power dissipation sources of power dissipation in CMOS circuits, you may recall that the sources of power dissipation can be broadly divided into two categories. One is your power dissipation that is your dynamic and second is static now, under the dynamic sources of power dissipation that means when the input is changing clock is changing. We have three components number one is known as P switching **switching** power dissipation and as we know, this is proportional to $\alpha_L C_L V_{dd}^2 f$ this is the equation $\alpha_L C_L V_{dd}^2 f$ plus you know there are many internal nodes, in which you can say, there are K nodes i is equal to 1 to K it is equal to $\alpha_i C_i V_{dd} (V_{dd} - V_{t_i})$.

This is the equation we derived and second component is p short circuit and that short circuit power dissipation equation, if you **you** may recall that it is equal to $\frac{K W}{L} \cdot \frac{1}{2} V_{dd}^3 \left(1 - \frac{2 V_t}{V_{dd}}\right)^3 \cdot P \cdot f$ that is, equal to beta into one by twelve V_{dd}^3 and then one minus two V_t by V_{dd}^3 and then you have got tau into f and third component was switching short circuit and thirdfourth component was p glitching glitching power dissipation is essentially same as the it is again for the charging and discharging of capacitors so, there also, it will be equal to it will be you can write $\alpha_L C_L V_{dd}^2 f$

d square pot transition and number of transitions that will occur will be it will be dependent on the switching activity.

So, these are the three components P switching P short circuit and then P glitching. These are the three dynamic sources of power dissipation that we have discussed in detail and second is your P static as, we know in CMOS circuits as such, there is no static power dissipation that means there is no static current flow, when the input is 0 or input is high however; there are as many sources of leakage currents as we have discussed gate leakage sub threshold leakage and third one was gate leakage sub threshold leakage and third was because of P gate leakage sub threshold leakage and third was P another source for the your gate sub threshold and third component you know that is, due to diode reverse biased diode you can say these are the three components and the sum total of current we may say this is your $i_{leakage}$ and then P static power dissipation can be equal to V_{DD} into $i_{leakage}$. So, this will be the power dissipation due to static power dissipation and $i_{leakage}$ is the sum total of the gate leakage sub threshold leakage and diode leakage.

So, if we analyse these equations we find that the supply voltage is present in all the equations that means somewhere; it is proportional to V_{DD} square that means quadratic dependence cubic dependence here, also, quadratic dependence and here, it is linear dependence however; this leakage current itself is also, dependent on supply voltage for example gate induced drain leakage and various leakage currents are there which are also, dependent on the supply voltage, so, as a consequence we find that supply voltage is the most important parameter which you can control to reduce power dissipation and that is, the reason. Why we are discussing we are starting our discussion on leakage power reduction starting with supply voltage scaling that means how, we can reduce power dissipation by supply voltage and of course there are other components.

(Refer Slide Time: 06:58)

Introduction

$$P_{switching} = \alpha_L \cdot C_L \cdot V_{dd}^2 \cdot f + \sum_i \alpha_i \cdot C_i \cdot V_{dd} \cdot (V_{dd} - V_t)$$
$$P_{sc} = V_{dd} \cdot I_{mean} = \frac{\beta}{12} (V_{dd} - 2V_t)^3 \tau \cdot f$$
$$P_{leakage} = V_{dd} \cdot I_{leakage}$$

□ **Degrees of freedom inherent in the low-power design space:**

- Supply voltage
- Physical capacitance
- Switching activity
- Threshold voltage

NPTEL Ajit Pal IIT Kharagpur NPTEL 3

As I mentioned, supply voltage physical capacitance switching activity and threshold voltage these are the other components, which we shall consider but right.

(Refer Slide Time: 07:10)

Introduction

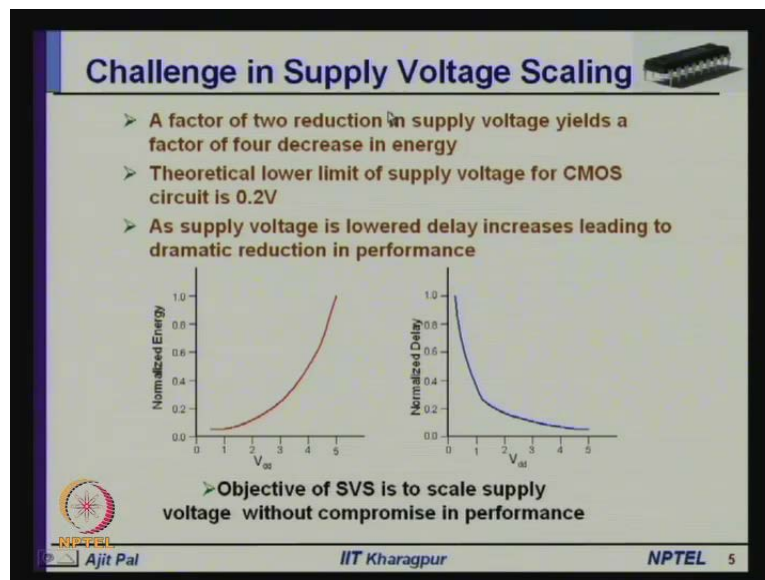
□ **Categorization of the low power approaches:**

- Supply voltage scaling
- Minimizing switched capacitance
- Minimizing leakage power

NPTEL Ajit Pal IIT Kharagpur NPTEL 4

Now, we shall focus on supply voltage scaling, minimizing switch capacitance and minimizing leakage power. This is how we shall continue our discussion one after the other, and first we shall focus on supply voltage scaling.

(Refer Slide Time: 07:33)



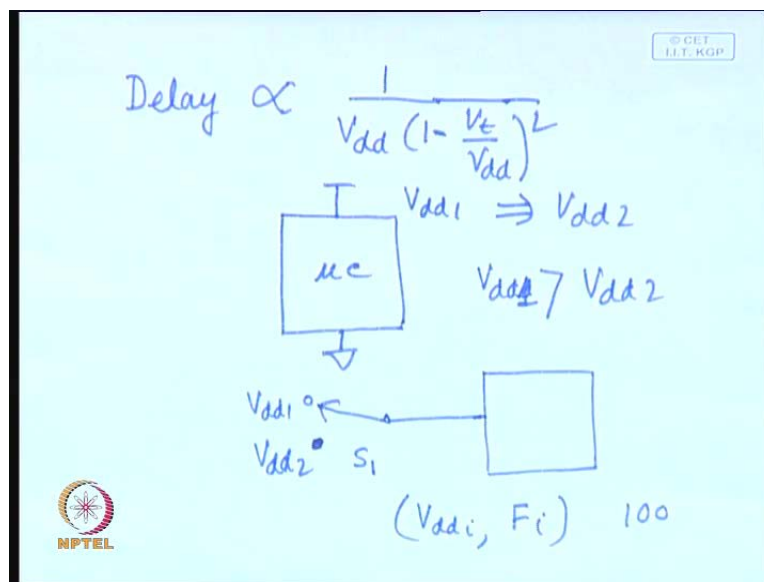
Now, let us, have a look at the challenge in supply voltage scaling. We know that supply voltage scaling is very important and it has the most dominating effect on in reducing the supply in reducing the power dissipation but is there any challenge in it is it so, simple that means say if you say that, if, supply voltage scaling gives you reduction in power dissipation why do not you simply reduce the supply voltage i. I mean as low as to as low as possible but that we cannot really do arbitrarily as we shall see this will be clear from this discussion first of all based on the switching power dissipation you can say a factor of two reduction in supply; voltage yields a factor of four decrease in energy that means energy drawn from the source will be reduced by factor of four assuming that the it has quadratic dependence.

But another point is that what is the lower limit how long we can reduce the supply voltage, it has been established theoretically; that the supply voltage for C mos circuits can be as low as point-two voltage although this is the theoretical lower limit in practice, we cannot really reduce the supply voltage to such a low level. Why the reason is you know you have to either supply the power from battery or from a voltage regulator. So, you cannot really get a power source which

can supply at point-two volt that is, the reason why now need as the supply voltage I mean that is, being used minimum value is about one volt however that is, the theoretical limit. Now, but one very important point is as the supply voltage is lowered the delay increases leading to dramatic reduction in performance.

This is, what you have to take in to account; when we reduce the supply voltage? So, here we see as; we reduce the supply voltage from five volt to one voltage. So, there is reduction note that normalized energy reduces from one to may be less than 0.1 you can see more than ten times reduction in energy occurs as we reduce the supply voltage. So, as I mentioned in the beginning that if, you reduce the supply voltage, by half the energy reduces by fourth times; but what what impact it has got on the delay you can see as we reduce the delay, I mean as you reduce the supply voltage delay is increasing what is the equation for delay earlier?

(Refer Slide Time: 10:25)



We have discussed, the delay in C mos circuits you may recall that delay is proportional to 1 by V d d then 1 minus V t by V d d square. There are other parameters capacitances, and various other things, but we have only considered the component which is related to supply voltage. So, you see assuming that this ratio the is maintained this this you know the ratio of the threshold voltage and supply voltage is maintained that, means as the supply voltage is reduced threshold voltage is also, reduced even then we find if we reduce the supply voltage delay increases

linearly that means delay increases linearly. If, the supply voltage is reduced from some voltage say five volt to three volt to two volt.

But if we consider this parameter suppose, we reduce the supply; voltage without reducing the threshold voltage then what happens in this parameter in this particular part one minus V_t by V_{dd} . This is in the denominator and as the supply voltage approaches the threshold voltage this part becomes very small one-minus this becomes close to one when supply; voltage is close to threshold voltage so, this part will be very small, and square of that will be still smaller and as a consequence delay will be very large that means whenever; we reduce the supply voltage without reducing the threshold voltage and when the supply voltage becomes closer to the threshold voltage there is dramatic increase in the delay as we can see from this curve.

So, here you can see in this part it is definitely; increasing delay is increasing but here, it is increasing at a very high rate because this it has become closer the threshold voltage. So, what is the what is our observation from these two curves from these two curves you see number one is that supply voltage helps us to reduce the power dissipation particularly quadratic reduction occurs reduction in energy but delay also, increases if not linearly I mean in in the in when the supply voltage is away from the threshold voltage the linearly but normally; you know threshold voltage is kept fixed and supply voltage scaling is done then delay increases at a very high rate when it becomes closer to the supply voltage threshold voltage so, that means increasing delay means loss in performance so, if, we do supply voltage scaling there, is loss in performance which will not be accepted for many applications that means many in for many applications people will say yes, you can reduce power **you can reduce power** dissipation by reducing supply voltage, but we are not going to compromise on performance. So, this is the challenge of supply voltage scaling techniques what we have to do we have to reduce the supply voltage in such a way? that there is no reduction in performance we shall not compromise on performance, that means supply voltage has to be done without compromising performance that is, the biggest challenge and how it can be done, we shall discuss when we discuss several techniques.

(Refer Slide Time: 14:18)

Voltage Scaling Approaches

- **Static Voltage Scaling (SVS):** Different blocks or subsystems are given different fixed voltages
- **Multi-level Voltage scaling (MVS):** This is an extension of SVS where the supply voltage is switched between two or few fixed voltages
- **Dynamic Voltage and Frequency Scaling (DVFS):** This is an extension of the MVS where a large number of voltage levels are dynamically applied for different workloads
- **Adaptive Voltage Scaling (AVS):** This is an extension of the DVFS where a control loop is used to adjust voltage and frequency for changing workload

NPTEL Ajit Pal IIT Kharagpur NPTEL 6

Now, here are some of the categories of supply voltage scaling first one is static voltage scaling so, in static voltage scaling normally; you know we develop a technique. I mean supply voltage is reduced and that voltage is maintained say you have got a processor say micro micro-computer you decided to reduce the supply voltage from say V_{dd1} to V_{dd2} and based on the performance requirement. We decide that it will offer it, at V_{dd2} which is where; we assume that V_{dd1} is greater than V_{dd2} so, we scale down the supply voltage and that voltage is maintained at runtime that means; when we are using the device using the system we do not reduce the you do not change the supply; voltage a design design time will fixed voltage and we maintain that so, this is known as static supply voltage scaling that is, the significance of the term static that means; it is not changed while the circuit is in operations and you may have different blocks where you can have different fixed voltages.

Now, second approach which is an extension of this static voltage scaling is known as multi-level voltage scaling so, this is an extension of the static voltage scaling where the supply; voltage is switched between two or few fixed voltages. That means here, what we are doing suppose you have a block logic block may be microprocessor or some other module where you will apply say at one point of time we will apply V_{dd1} and another point of time you will apply V_{dd2} so, there will be a switch which will use either this voltage, or this voltage that means the number of voltages is very few here I have shown only two the the module is applied

with the supply voltage V_{dd1} . So, with that can be changed with the help of switch or it will switch it will use another voltage V_{dd2} that means the supply; voltage can be kept can be kept one or few supply voltages that can be selected with the help of a switch; between two or few fixed voltages that is, the basic idea of multi-level voltage scaling so, you have got multiple voltage levels and between them it will be switched.

And third technique is essentially; an extension of multi-level voltage scaling MVS which is known as dynamic voltage scaling. So, however; whenever, we do dynamic voltage scaling we also, scale the frequency because that will give you more benefit, that means, as you scale down the voltage, you also, scale down the frequency, that is the reason, why this technique is known as $DVFS$ that means dynamic voltage, and frequency scaling, and here, what we do a large number of voltage levels can be dynamically; applied for different work loads? That means you can have may be hundred different voltages, that you can apply, and which you can change dynamically; at runtime so, this is the difference between the first two approaches in the first two cases.

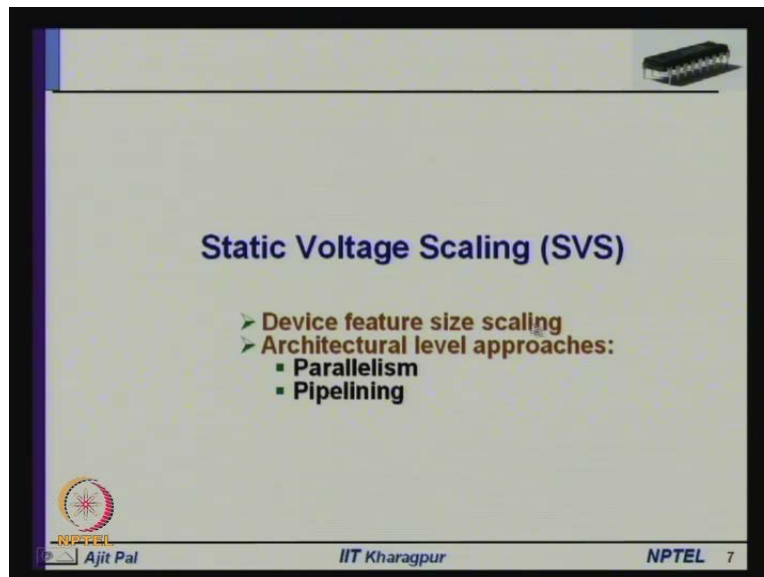
We normally; we do not change at runtime, but in the second case, and the third case that is, your dynamic voltage, and frequency scaling dynamically; at runtime you keep on changing the supply voltage of course depending on the work load of the processor or some other sub-system and also, you use different frequency in other words. You will choose different voltage frequency pair say V_{ddi} and f_i . So, this this pair you will use and the number of such pairs can be may be hundred or more than that so, over a large number of voltage frequency voltages different voltages. You can apply to the module and we shall later on we shall discuss, how exactly; it can be done.

Finally; we have another technique adaptive voltage scaling so, this dynamic voltage and frequency scaling is essentially; a open loop technique by open loop we mean a design time what voltage can be applied what frequency, can be applied these are decided based on you know based on the worst case performance for work case performance requirement. You know whenever; you change the pvt parameter voltage and temperature overall that pvt over when the parameter varies because of different reasons voltage varies at runtime temperature varies at runtime under such conditions. It will it has to work that is, the reason why this approach this dynamic voltage and frequency scaling is very conservative because the design has to be made in

such a way that under worst condition of process voltage and temperature variations. It will work but obviously; whenever you want to do that design has to be very conservative and as a consequence the benefit that, you get is not really; optimised on the other hand in adaptive voltage scaling it is essentially; close loop technique where at runtime the behaviour or the performance of the circuit is monitored with the help of some un-cheap circuit and based on that the voltage and frequency are decided. How, it can be done? That we shall discuss in detail, as a result you can say this is a close loop voltage and scaling voltage and frequency scaling technique and that is, the reason why it is called adaptive voltage scaling.

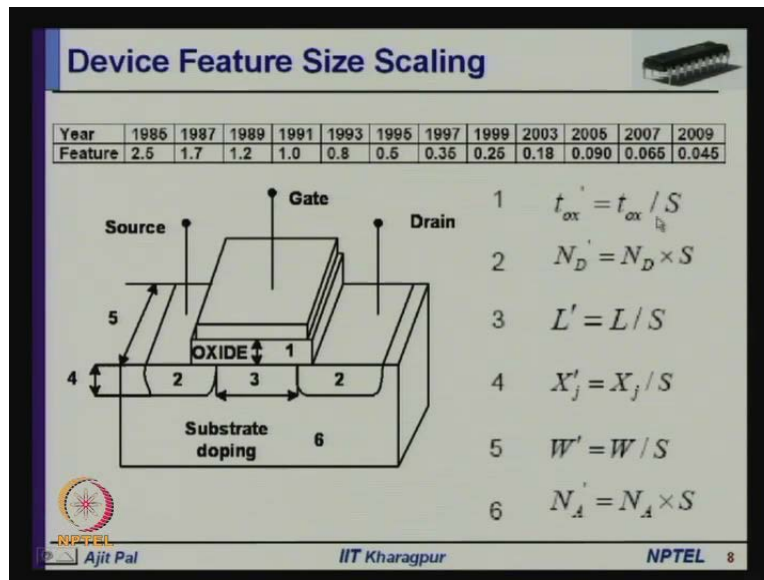
So, it keeps on it adopts the changes that occurs because of changes in parameter value changes in because due to the, change in voltage. Due to the change in temperature and so, it keeps on adjusting the voltage and frequency, because and with the changing of work load. So, with these techniques we shall discuss one after the other of course.

(Refer Slide Time: 21:15)



Today we shall start our discussion, on static voltage scaling so, first we shall start a discussion on device feature size scaling, and then we shall focus on architectural level approaches.

(Refer Slide Time: 21:27)



What do you really mean by device feature size? Scaling as you know because of the advancement of VLSI technology advancement of photolithographic techniques the size of the devices that you can fabricate is gradually reducing for example here, there is a chart and here is the feature size in micron and on first row is giving the year so, in the 1985. The feature size was 2.5 micron that means this this was the lambda value 2.5 micron was the lambda value that was the minimum feature size so, the device's minimum size device was 2 lambda by 2 lambda that means, if you fabricate a device it has to be minimum size will be 5 lambda by that five micron by 5 micron that was the minimum size but you can see over the years how the size has changed in the year 1987 it becomes 1.7 micron.

And in 0.18 micron in 1991 micron and in this way over the years it has reduced and in the year 2009 I mean where we are now, it is now in the deep sub-micron era so, you can see the feature size is 0.045-0.065-0.090-0.18-0.25-0.35-0.5-0.8-1.0-1.2-1.7-2.5 micron that means 45 nanometre so, we are now in the era of 45 nanometres now it is the VLSI chips are implemented using 45 nanometre technology and you may have heard of Moore's law **Moore's law** predicted this kind of reduction based on some simple observation and that Moore's law is still followed until recently; because of increased leakage power there is a problem but you know until recently most law has been followed very closely that prediction.



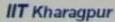

In other words device feature size is responsible for the success of Moore's law what do you really mean by device feature size scaling? Let us look at this diagram, we have got a schematic diagram three dimensional schematic diagram of a MOS transistor, and you have got different features for example this number one is thickness of the silicon dioxide. So, t_{ox} this is the thickness and as you go from one technology generation to the next technology generation the thickness is reduced by a factor S . So, if t_{ox} is the feature size of a particular technology generation next technology generation will have a feature size which is equal to t_{ox} by S and similarly; second parameter N_D that is, your doping concentration in the source and drain regions that is, increased by a factor by S actually this is necessary to maintain constant electric field and what are the features reduced silicon. The thickness of the silicon dioxide the length is also, reduced by factor S so, here the parameter three this is the length from here so, this is the source this is the drain so, the length is L so, whenever; you go for next generation length is reduced by factor S so, L is the present generation L is the next generation.

Similarly the depth of penetration of the device that is, the parameter four that is, also, reduced by a factor S x_j is the next generation depth of penetration of the devices. I mean of the diffusion and then for present generation. It is x_j by S and fifth parameter is the width this is the width and width is also, reduced by a factor S w by S so, you can and finally the last parameter substrate doping that is, that is, increased by factor S . So, we find that all the physical parameters length width thickness of the silicon dioxide depth of penetration of the diffusion these are all reduced however the doping concentrations are increased by factor S so, this is known as feature device feature size scaling. Now, what is the impact of this whenever; we do this what happens?

(Refer Slide Time: 26:26)

Constant Field Scaling

Quantity	Before Scaling	After Scaling
Channel length	L	$L' = L / S$
Channel width	W	$W' = W / S$
Gate oxide thickness	t_{ox}	$t'_{ox} = t_{ox} / S$
Junction Depth	x_j	$x'_j = x_j / S$
Power supply voltage	V_{DD}	$V'_{DD} = V_{DD} / S$
Threshold Voltage	V_{T0}	$V'_{T0} = V_{T0} / S$
Doping Densities	N_A	$N'_A = N_A \cdot S$
	N_D	$N'_D = N_D \cdot S$

So, let us, consider one very important technique that is, your constant field scaling constant field scaling basic idea is that as you reduce the dimension the electric field within the geometry of the device has to be maintained constant that is, the basic idea so, whenever; you want to maintain the electrical field constant as you reduce the dimension of the devices you have to maintain the you have to reduce the supply voltage also, as you can see here as you are reducing the dimension of length l width channel width so, channel length channel width gate oxide thickness t_{ox} junction depth x_j these are all you are reducing and to maintain constant electric field. You have to reduce the supply voltage as well so, it is V_{DD} by S and not only supply voltage you have to reduce the threshold voltage V_{T0} that means V_{T0} is essentially threshold voltage with zero body bias so, with zero body bias we that threshold voltage is also, reduced and in the next generation it will be V_{T0} by S however; as I mentioned the doping concentrations are increased by factor S.

So, whenever you do this kind of constant field scaling then, what what impact it has got on the electrical parameters so, the physical parameters are scaled down the supply voltage is scaled down threshold voltage is scaled down. What impact it has got on the electrical parameters because electrical parameters will decide the performance of the device let us, look have a look at the look at different electrical parameters.

(Refer Slide Time: 28:14)

© GET
I.I.T. KGP

$$1. C'_{ox} = \frac{\epsilon_{ox}}{t'_{ox}} = \frac{\epsilon_{ox}}{t_{ox}} \cdot S = S \cdot C_{ox}$$

$$2. I'_{ds}(lin) = \frac{k'W}{L} \left[(V_{gs}' - V_t') V_{ds}' - \frac{V_{ds}'^2}{2} \right]$$

$$\frac{\epsilon_{ox} \epsilon_{ox}}{t_{ox}} = \frac{I_{ds}(lin)}{S}$$

$$I'_{ds}(sat) = \frac{k'W'}{L} \cdot \frac{1}{2} (V_{gs}' - V_t')^2$$

$$= \frac{I_{ds}(sat)}{S}$$

NPTEL

First let us, consider the capacitance that is, your unit gate capacitance C_{ox} as how it is reduced? it will be reduced by factor of C_{ox} as you know that it is equal to **it is equal to it is equal to** ϵ_{ox} by t_{ox} so, for the next generation this will be the we know that unit gate capacitance is equal to the permittivity of the silicon dioxide by the thickness of the silicon dioxide and this is equal to of course the permittivity does not change as you go to the next generation so, this remains as the same as, the previous generation however the silicon dioxide thickness is reduced as we have seen so, it becomes t_{ox} into S or we can write it that this is equal to S into C_{ox} das that means the the the as we go from one technology generation to the next technology generation so, this is this is the previous generation so, the the unit gate capacitance increases by factor S so, unit gate capacitance increases by a factor S and unit gate capacitance is increasing by factor S obviously, it will have impact on the gate capacitance later on we shall discuss about that.

Second parameter let us consider the current I_{ds} linear so, what is the value of I_{ds} linear this as we know the equation $k'W$ by L into V_{gs} minus V_t into V_{ds} minus V_{ds} by two square. This is the equation for this is the equation for the linear you know that linear current when the circuit is in the mos transistor is in linear region. Now, I mean these are you have to put all das here, this is for the next generation now which parameters are reduced and from this you can say and how it is related to the previous generation as the as you go from one technology generation.

To the next technology generation, how the current change does it increase or decrease? So, in if you consider these parameters you will find that this will be equal to I_{ds} by S that means; it will reduce by factor S that means the supply voltage and all these parameters reduced so, here you will one factor s will come out these two will get neutralized and k has a value that is, your μ and $e o x$ by $t o x$ so, this from this one s will come so, and as a consequence here, you will get I_{ds} by S so, this will be this current will be proportional to I_{ds} by S similarly; that i_{ds} das saturation as we know the expression is $k w$ by l into half and V_{gs} minus V_t square so, here also, you will find that it will be equal to I_{ds} saturation of the previous generation by s that means the current will reduce by factor S as we go from present generation to the next generation current will reduce will reduce by factor S .

(Refer Slide Time: 32:12)

© CET
IIT KGP

$$P' = I_{ds}' V_{dd}' = \frac{I_{ds}}{S} \cdot \frac{V_{dd}}{S} = \frac{P}{S^2}$$

$$C_g' = W'L'C_{ox}' = S^2 W.L \frac{C_{ox}}{S} = S \cdot C_g$$

$$E = P \cdot T = \frac{P}{S^2} \cdot S^L$$

NPTEL

What about power dissipation P that is, equal to i into V_{dd} ? So, let us, consider the power dissipation of the next generation. We have seen that current is reducing by factor of S I_{ds} we can say I_{ds} by S and supply voltage is also, scaled down by factor S so, we can say that P by S square that means the power dissipation is reducing by factor of S square so, as you go from one technology generation to the next technology; generation this power dissipation will reduce by factor S ; what about the gate capacitance C_g C_g das as you know that will be equal to w das into

l das into Cox das and in this **in this** particular case what will happen this will be equal to w and l here, it will be S into w and it will be l into w so, s square w into l and in this it will be Cox and it reduces by a factor S so, this will be equal to S into and this part is your C g.


So, we can find we can see that the gate gate capacitance is also, increasing by a factor S now, from this we can what is our observation and of course another parameter is there that is, your power delay product say power into so, P into delay power delay product so, this you will find that power delay product will not change because power will increased by factor of S square and delay will increase a factor of S square, I mean delay will reduce the delay will reduce so, power delay product that is, energy will remain same energy will remain same as we shall see from this summary.

(Refer Slide Time: 34:21)

Constant Field Scaling

Quantity	Before Scaling	After Scaling
Channel length	L	$L' = L / S$
Channel width	W	$W' = W / S$
Gate oxide thickness	t_{ox}	$t'_{ox} = t_{ox} / S$
Junction Depth	x_j	$x'_j = x_j / S$
Power supply voltage	V_{DD}	$V'_{DD} = V_{DD} / S$
Threshold Voltage	V_{TO}	$V'_{TO} = V_{TO} / S$
Doping Densities	N_A	$N'_A = N_A \cdot S$
	N_D	$N'_D = N_D \cdot S$

Quantity	Before Scaling	After Scaling
Gate Capacitance	C_g	$C'_g = C_g / S$
Drain Current	I_D	$I'_D = I_D / S$
Power Dissipation	P	$P' = P / S^2$
Power Density	P / Area	$P' / Area' = (P / Area)$
Delay	t_d	$t'_d = t_d / S$
Energy	$E = P \cdot t_d$	$E' = \frac{P}{S^2} \cdot \frac{t_d}{S} = \frac{P t_d}{S^3} = \frac{1}{S^3} E$


Ajit Pal
IIT Kharagpur
NPTEL 9

So, we find that, power density not really power delay product so, power density power power is increasing by factor of S; square and you area is reducing by a S square so, it will be it will be a into S square this s square will cancel out so, power density will remain same energy will reduced by factor of S cube.

So, from this here is the summary the gate capacitance is reducing by factor S the drain current is increasing is reducing by factor of S power dissipation is reducing by factor of S. Square density is remaining same delay is reducing by factor of S and energy is reducing by factor of S cube so,




we find that whenever, we whenever; we are doing device size scaling along with supply voltage scaling delay is not increasing there is no loss in performance rather delay is reducing so, we can say this is beneficial to us that means even by doing voltage scaling. The performance is getting upgraded and in fact as a consequence, we have seen how the frequency of operation of the microprocessors has increased over the years as we have gone from one technology generation to the next technology generation. In the eighties you may recall the microprocessors were operating with clock frequency in the range of the megahertz. Now, the clock frequencies in the range of gigahertz and this is the reason for that so, this is constant field scaling.

(Refer Slide Time: 36:18)

Constant Voltage Scaling

Quantity	Before Scaling	After Scaling
Channel length	L	$L' = L / S$
Channel width	W	$W' = W / S$
Gate oxide thickness	T_{ox}	$t'_{ox} = t_{ox} / S$
Junction Depth	X_j	x'_j / S
Power supply voltage	V_{DD}	$V'_{DD} = V_{DD}$
Threshold Voltage	V_{th}	$V'_{th} = V_{th}$
Doping Densities	N_A	$N'_A = N_A \cdot S^2$
	N_D	$N'_D = N_D \cdot S^2$

Quantity	Before Scaling	After Scaling
Gate Capacitance	C_g	$C'_g = C_g / S$
Drain Current	I_D	$I'_D = I_D \cdot S$
Power Dissipation	P	$P' = P \cdot S$
Power Density	P / Area	$P' / Area' = S^2 \cdot P / Area$
Delay ($t_d \propto C_g \cdot V_{th} / I_D$)	t_d	$t'_d = t_d / S^2$

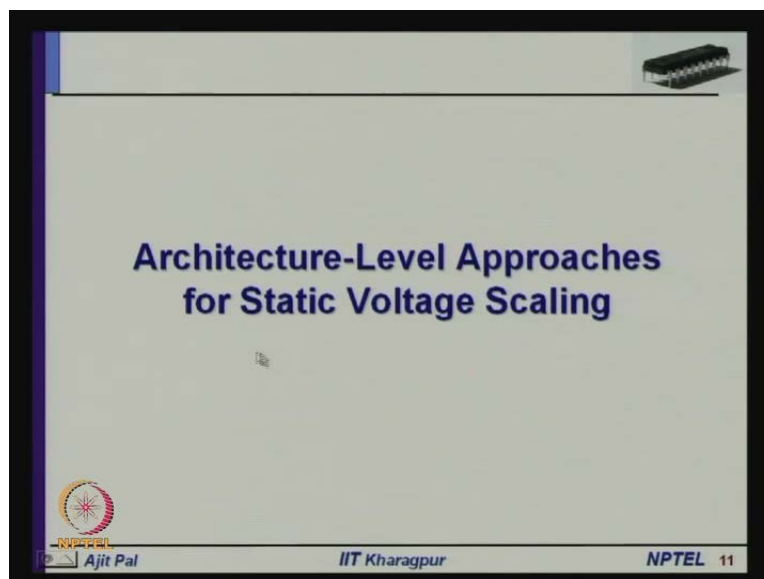




There is another technique which is known as constant voltage scaling you know whenever; we are reducing the device sizes then recharge sizes that means length width. We cannot always reduce the supply voltage because you know there are many other subsistence, which will be operating at supply voltage. We cannot always reduce the supply voltage so, what happens in such a situation. We shall reduce the feature sizes without reducing the supply voltage and obviously; we shall also, not reduce the threshold voltage in such a case so, in such a case as you can see all the parameters physical parameters length width thickness of the silicon dioxide junction depth. All are reduced by factor S however supply; voltage is kept same and threshold voltage is also, kept same however; in such a case you have to increase the doping concentration not not by factor s but by factor of s square.

So, doping concentration is increased the feature sizes different feature sizes are reduced but supply voltage and threshold voltages are maintained same so, this is known as constant voltage scaling and what is the impact of this constant voltage scaling on different physical different electrical parameters. You can see here the gate capacitance also, will reduce by factor S because we are reducing the length reducing the width and the so, the gate capacitance will reduce by factor S and drain current is will also, reduce by factor will increase but here, you can see instead of drain current getting reduced by factor s here it is increasing by factor S because you know the supply voltage is more and dry will be more so, drain current will be more and as a consequence power dissipation is also, increasing by factor S .

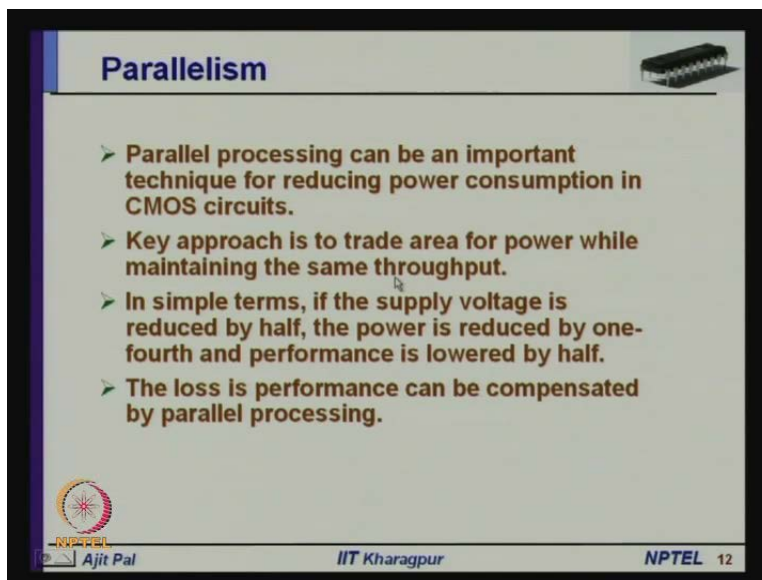
So, it is not really very beneficial from the low power view point and in however; another the most important point is power density. You can see power density is increasing by factor of S cube so, this is not good from the view point of low power because the power density is large then over a small area there will be large power dissipation and that may lead to what is known as hotspots. So, this is not very attractive from low power applications however; for high performance this is good because delay is reduced by factor of S square. So, we have seen that constant field voltage scaling is not very low power, I mean is not good for low power application but whenever; we do constant field scaling then we can reduce the supply voltage and and simultaneously the performance is improved.

(Refer Slide Time: 39:19)




So, with this we now switched to the next topic that is, your architectural-level approaches for static voltage scaling. So, architectural in a architectural level; what we do that is, called r t l level resistor transfer level. We have got different blocks like l u adder multiplier we are we shall be we shall be doing the voltage scaling at this level.

(Refer Slide Time: 39:49)



Parallelism

- Parallel processing can be an important technique for reducing power consumption in CMOS circuits.
- Key approach is to trade area for power while maintaining the same throughput.
- In simple terms, if the supply voltage is reduced by half, the power is reduced by one-fourth and performance is lowered by half.
- The loss in performance can be compensated by parallel processing.

 NPTEL
Ajit Pal IIT Kharagpur NPTEL 12

So, there are two approaches we shall discuss first is parallelism and after that we shall discuss pipeline. First we shall focus on parallelism, we are, you are all familiar with parallel processing normally; parallelism is used to improve performance. We know instead of engaging one labourer, if you engage two labourers obviously the work will increase similarly; instead of using a single processor. If you use double dual processor performance will increase that is, the normal you know normal I mean, there is a convention traditionally; this parallelism is used to improve performance here, we shall see how we can use it for low power.

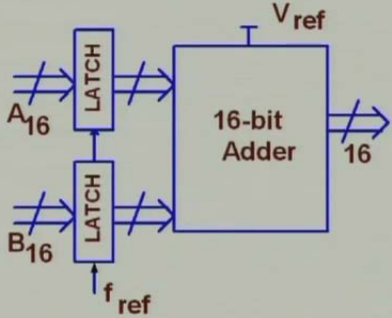
We are not interested in increase in the performance at this point but we are interested in retaining the performance level same and at the same time. We want to achieve lower power, how can it be done by using parallelism.

So, here the key approach is to trade area for power while maintaining the same throughput so, by when we go for parallelism obviously. It will involve use of more than one modules more than one subsistence that will definitely; increase area and we shall see how the throughput can be maintained and that will help us, in reducing power dissipation basic idea is you know as you reduce the supply voltage power dissipation reduces quadratically, that means if, you reduce the supply voltage by say if you make it V_{dd} by two the energy consumption reduces by factor of four but how the area increases. If, you duplicate it to compensate the loss in performance; because of reduction in supply voltage then you know that area will increase by two that means, we shall see that even by duplicating even by increasing the are which will lead to which you may think that, it will lead to more I mean increase in power dissipation you will see that overall there will be overall reduction in energy keeping the performance same. Because the loss in performance will be compensated by parallel processing.

(Refer Slide Time: 42:20)

Example

➤ **Example:** Two 16 bit registers supplies two operands to a adder. Delay of the critical path of the adder is 10 nsec. Operating frequency = 100 MHz



➤ **The estimated dynamic power of the circuit is**

$$P_{ref} = C_{ref} \cdot V_{ref}^2 \cdot f_{ref}$$

13

How let me illustrate, with the help of an example? We have got a two 16-bit resistors and resistors supplies two operands to an adder so, here is a 16-bit adder so, you have got two resistors in the form of latch, and we have got a 16 bit adder and here let us assume the operating frequencies 100 megahertz so, what is the dynamic power dissipation?

(Refer Slide Time: 42:58)

The image shows a whiteboard with handwritten mathematical equations. At the top right, there is a small logo for '© CET I.I.T. KGP'. The equations are as follows:

$$C_L V_{dd}^2 f$$

$$P_{ref} = C_{ref} \cdot V_{ref}^2 \cdot f_{ref}$$

$$P_{par} = 2.2 C_{ref} \cdot \left(\frac{V_{ref}}{2}\right)^2 \cdot \frac{f_{ref}}{2}$$

$$= \frac{2.2}{8} \cdot P_{ref} = 0.227 P_{ref}$$

At the bottom left of the whiteboard, there is a logo for 'NPTEL'.

We know that the power dissipation is $C_L V_{dd}^2 f$ so, f so, for this particular full adder this is considered to be a reference circuit. It is considered that P_{ref} is equal to C_{ref} V_{ref}^2 f_{ref} so, V_{ref} square and f_{ref} this is the power dissipation of this two adder this is the reference power dissipation of this 16-bit adder to input 16-bit adder. Now, when we go for parallelism we shall duplicate it.

(Refer Slide Time: 43:53)

Parallelism

- Here the adder has been duplicated twice, but the input registers have been clocked at half the frequency of f_{ref} .
- This helps to reduce the supply voltage such that the critical path delay is not more than 20 nsec.

The diagram illustrates a parallel adder circuit. Two 16-bit adders are shown, each receiving 16-bit inputs A and B through latches. The adders are clocked at $f_{ref}/2$. Their 16-bit outputs are connected to a 16-bit multiplexer (MUX) which is clocked at f_{ref} . The supply voltage for the adders is $V_{ref}/2$, and for the MUX is V_{ref} . A small image of a chip is in the top right corner.

NPTEL
Ajit Pal
IIT Kharagpur
NPTEL 14

And whenever; we shall do the duplication what will happen the circuit will look like this that means, we shall be having two different 16-bit adders two separate 16-bit adders and they will be applied these two operands A and B and you will be doing some kind of time division multiplexing that means you will be applying to this then to this and also, you will take out the output with the help of a multiplexer. So, this is somewhere; similar to time division multiplexing of the two adders and here, you are applying inputs alternately; and we are taking the output with the help of a mux with the help of a multiplexer. So, here we can see that since, we are not interested in increasing a performance the inputs are applied at the rate of $f_{reference}$ by two to these two parallel to these two adders.

That means we have got two adders which are working in parallel but we are applying inputs at half rate not at the rate of $f_{reference}$ that means, these adders will be operating at may be 50 megahertz earlier it was operating at 100 megahertz so, it will operate about 50 megahertz however; we shall be getting output about the same rate because we are taking this output and this output and in the output is taken out about 100 megahertz rate from the from this multiplexer.

But as you reduce the operating frequency of the 16-bit adders as, we know you can it can operate at a lower voltage so, we are reducing the supply voltage of these two orders by v_{ref} by

2 we can see it is no longer v_{ref} by v_{ref}^2 so, what will be the power dissipations of these parallel adder implementation so, $p_{parallel}$ what will be the power dissipation so, you can say what will be the capacitance. We can see here the capacitance will be at least two times but it may be more than two times, because we have more number of latches and a multiplexer so, capacitance will increase by more than two times and it has been found that the capacitance is increased by may be two with respect to the c_{ref} it is 2.2 that means with respect to the original circuit the capacitance is increasing by factor 2.2 and what about the voltage voltage. We have seen we have divided by 2. We have applied half the voltage in this voltage we can see we have seen that supply voltage has been reduced by a factor of 2 so, v_{ref} will be v_{dd} by $2v_{ref}$ by 2 and frequency of operation is also, reduced by factor of two v_{ref} by 2.

So, for this parallel implementation p_{par} p_{par} is $p_{parallel}$. Which is equal to $2.2 \cdot 2 \cdot c_{ref}$ into v_{ref} by 2 square into f_{ref} by 2 so, we can see it is equal to 2.2 by 8 into p_{ref} so, with respect to the original circuit which is I mean where the power dissipation is $p_{reference}$ of course. We are focusing only on switching power dissipation that is, the one of the most dominant component we find it is reducing by factor of I mean almost one fourth this will be equal to $0.227 \cdot p_{reference}$ so, we we have seen in this particular, case we have used parallelism for low power now instead using parallelism for low power suppose our intention is to increase performance then what will happen.

(Refer Slide Time: 48:04)

Impact of parallelism

➤ The estimated dynamic power is

$$P_{par} = 2.2C_{ref} \cdot \left(\frac{V_{ref}}{2}\right)^2 \times \frac{f_{ref}}{2} \approx \frac{2.2}{8} P_{ref} \approx 0.277 P_{ref}$$

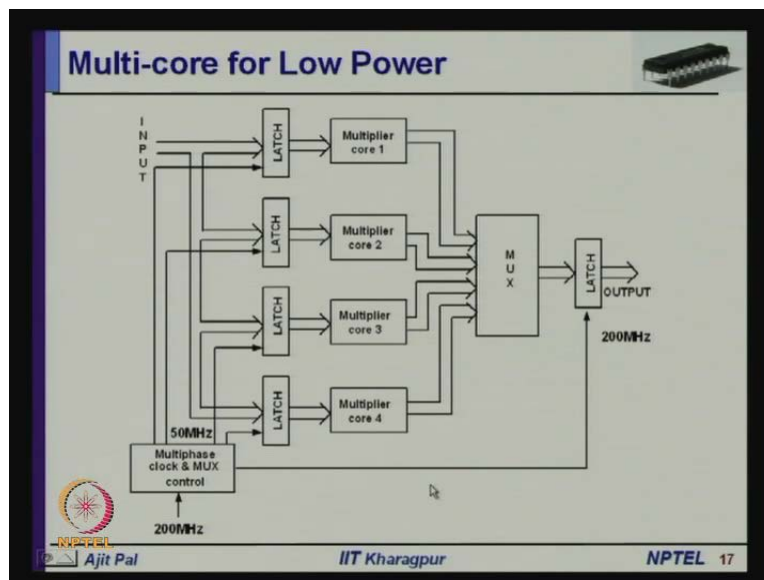
Parameter	Without Vdd scaling	With Vdd scaling
Area	2.2X	2.2X
Power	2.2X	0.227X
Throughput	2X	1X

Ajit Pal
IIT Kharagpur
NPTEL 15

Then in that case what we shall do we shall not go for V d d scaling. So, Without V d d scaling, If, you use parallelism then you can see that area will increase by factor of 2.2 power dissipation also, will increase by factor 2.2 however, the throughput will also, increase by a factor of 2 that means the throughput will double whenever; we go for parallel implementation. So, this is the where we are using parallelism for performance obviously at the cost higher power dissipation and also, about the cost of larger cheap area.

But in the second case, where we are doing V d d scaling? We can see performance level. We have maintained same although you are increasing area the power dissipation is reduced. So, in this particular case we can say we are doing using parallelism for low power that means, we are trading area for low power keeping the performance same so, this is the basic idea behind the parallelism use of parallelism for low power and you will be you will be surprised to know that this is the basic idea behind multicore. Now, it is you are using multicore processors why we have moved to single core to multicore this is the basic idea that, means whenever; we go for multicore the overall power dissipation is reduced how,

(Refer Slide Time: 49:35)

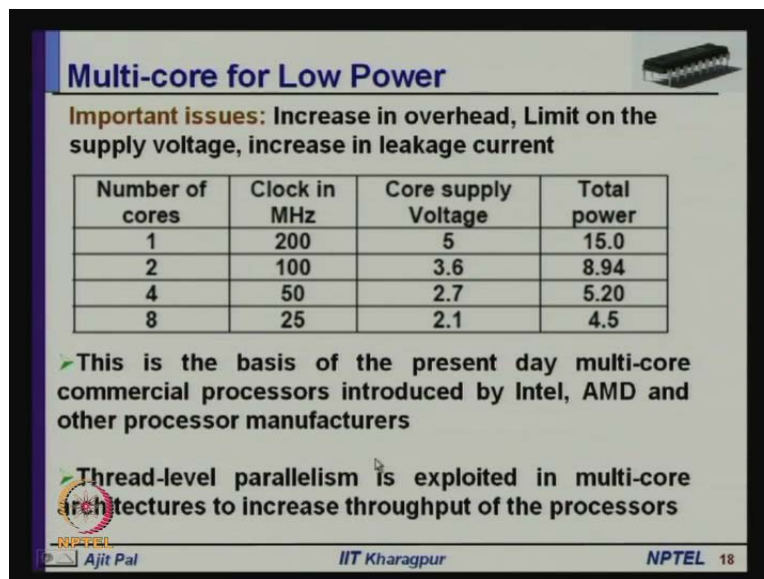


Let me, illustrated with the help of a very simple example multi-core for low power so, here you can see we have used 4 multiplier cores, instead of it may be it may be (0) but I have illustrated

with the help of 4 multiplier cores. We have 4 cores multiplier core number 1 multiplier core 2 multiplier core 3 and multiplier core 4.

So, here you can see we have a multiphase clock and multiplex multiplex are control so, clock is 200 megahertz original clock is 200 megahertz but since, we have got 4 cores to each of them, we are applying in 4 different phases, about the rate of 50 megahertz and then, we have got a multiplexer which of course, which will operate at the rate of 200 megahertz but clock is not applied here, that you have to apply a clock of 200 megahertz to this and take the output to the latch from the output of the multiplexer so, this is the multi-core implementation.

(Refer Slide Time: 50:44)



Multi-core for Low Power

Important issues: Increase in overhead, Limit on the supply voltage, increase in leakage current

Number of cores	Clock in MHz	Core supply Voltage	Total power
1	200	5	15.0
2	100	3.6	8.94
4	50	2.7	5.20
8	25	2.1	4.5

➤ This is the basis of the present day multi-core commercial processors introduced by Intel, AMD and other processor manufacturers

➤ Thread-level parallelism is exploited in multi-core architectures to increase throughput of the processors

NPTEL IIT Kharagpur NPTEL 18

Now, let us, see how the low power is achieved here, the important issues are increase in overhead limit on supply voltage increase in leakage current. You see these are the important issues whenever; you go for multi-core as you go from say single core to dual core then overhead increases because the number of buses increases. You may have to use more number of latches more number of I mean multiplexers of bigger size that means the overhead will increase second is there is a limit on supply voltage. We shall see as you go from multi-core go from you know single core to multi-core we may reduce the supply; voltage to reduce, the power dissipation. But there is a limit up to which you can , we can go for that that means the supply, voltage can be

reduced up to certain limit because each processor will operate satisfactorily, up to certain lower voltage that means, we cannot reduce the supply voltage indefinitely there is a lower limit to that.

And as you keep on increasing the number of cores there is increase in leakage current so, that also, we have to take into consideration particularly, in the present day technology the leakage current will increase significantly; so, these things we have to take into consideration whenever; we go for multi-core for low power and here, you can see; how the power dissipation reduces and how the clock frequency changes as you go from say 1 core to say 2 core or to 4 core to 8 core so, for example for single core it is 200 megahertz operating frequency core supply voltage is 5 volt power dissipation is 15 watt. When it is dual core clock frequency is 200 megahertz 100 megahertz, then supply voltage can be reduced to 3.6 volt and total power dissipation is 8.94 and for 4 core the clock frequency is reduced by 250 megahertz supply voltage can be reduced to 2.7 volt because about this voltage the processor will operate about 50 megahertz and then total power is reduced by to 5.2 watt for 8 core. You can see the clock frequency can be 25 megahertz, with the supply voltage 2.1 volt and power dissipation is 4.5 watt.

So, you see the benefit of using multi-core and how it is how the power dissipation is reduced and this is the basis of the present day multi-core commercial processor introduced by Intel A M D and other processor manufacturers, and you know whenever; we go for multi core actually, thread level parallelism is utilized. We know, you know that whenever; we used pipe lining, we use instruction level parallelism but whenever; we go for multi-core multiple threads will be operating on multiple cores and we use thread level parallelism but those, are the topics which will be discussed in your computer architecture cores.

But here, we can say that the shift towards multi-core is primarily from the for the reason of low power that means whenever; you go for multi-core the power dissipation reduced and later on we shall discuss about, it in more detail and with this. We have come to the end of today is lecture in our next lecture, we shall see how pipe lining can also, be used for low power.