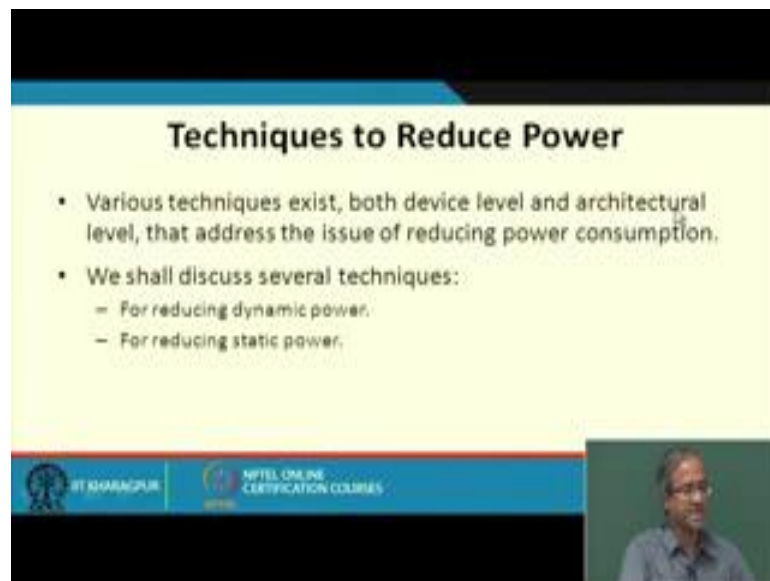


**VLSI Physical Design**  
**Prof. Indranil Sengupta**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 59**  
**Techniques to Reduce Power**

So, we continue with our discussion on low power design. If you recall in our last lecture we talked basically about the different sources of power dissipation in CMOS circuits. And in this lecture we shall be starting some discussion on how we can actually control or reduce power dissipation by making some design changes, design modifications following some design rules, various such techniques are there we shall be looking at them one by one.

(Refer Slide Time: 00:59)



The slide is titled "Techniques to Reduce Power" and contains the following text:

- Various techniques exist, both device level and architectural level, that address the issue of reducing power consumption.
- We shall discuss several techniques:
  - For reducing dynamic power.
  - For reducing static power.

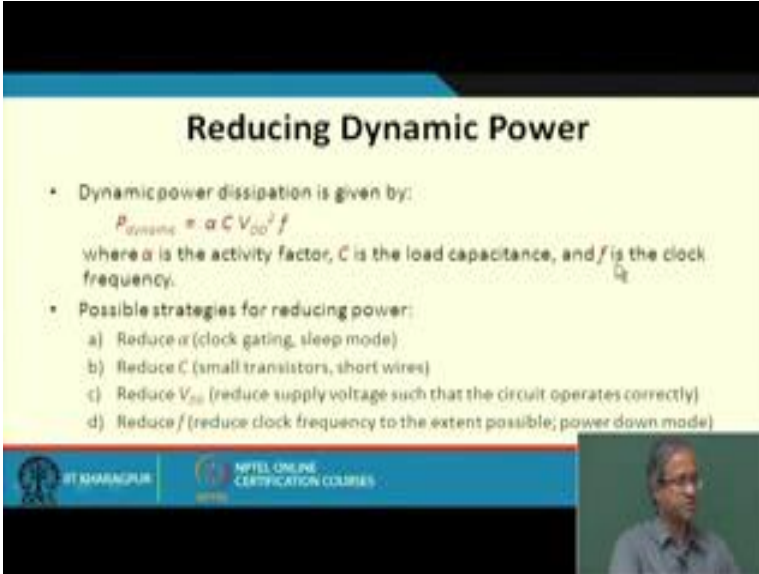
At the bottom of the slide, there are logos for "IIT KHARAGPUR" and "NPTEL ONLINE CERTIFICATION COURSES". A small video inset of the professor is visible in the bottom right corner of the slide frame.

So, the topic of discussion today here is techniques to reduce power. So, we shall see that when I talk about techniques these techniques may work both at the level of the devices means at the transistor level.

And also they may work at much higher design levels may be at the level of the architecture. So, whenever you are designing a circuit even if the higher level, you can take some design decisions that can ultimately impact the power dissipation in a pretty significant way. So, it is not that once we have the transistor level circuit only then you worry about power, not that even at a very high level when you are designing was

circuits may be at the behavioral level also. There are some techniques we shall see talk about some of them, you can follow or adopt so that subsequently when we have the final circuit work in, you have a better control on the power dissipation fine.

(Refer Slide Time: 02:16)



The slide is titled "Reducing Dynamic Power" and contains the following text:

- Dynamic power dissipation is given by:  
$$P_{dynamic} = \alpha C V_{DD}^2 f$$
where  $\alpha$  is the activity factor,  $C$  is the load capacitance, and  $f$  is the clock frequency.
- Possible strategies for reducing power:
  - a) Reduce  $\alpha$  (clock gating, sleep mode)
  - b) Reduce  $C$  (small transistors, short wires)
  - c) Reduce  $V_{DD}$  (reduce supply voltage such that the circuit operates correctly)
  - d) Reduce  $f$  (reduce clock frequency to the extent possible; power down mode)

The slide also features the NPTEL logo and the text "NPTEL ONLINE CERTIFICATION COURSES" at the bottom left, and a small video feed of a presenter at the bottom right.

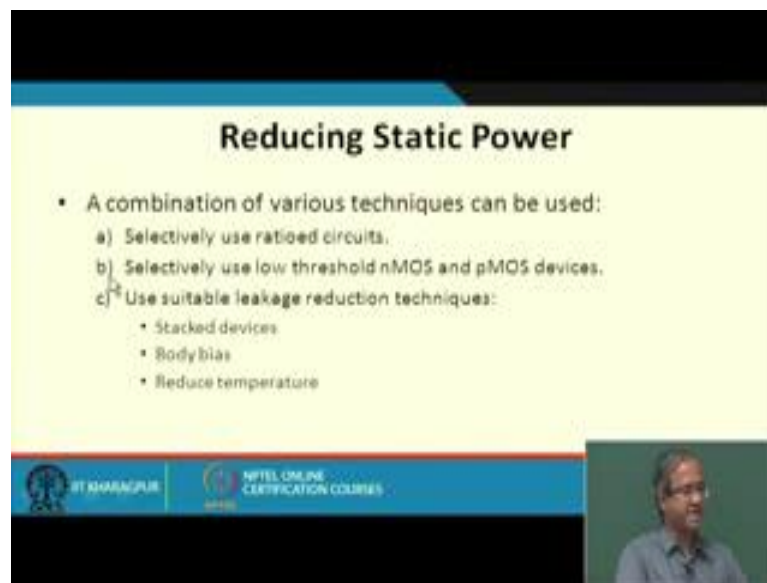
So, the techniques that we shall be talking about they will be primarily for reducing dynamic power, and for reducing static power here. So, for reduction of dynamic power let us recall the expression that you derived during our last lecture, we had said that for a typical CMOS gate the dynamic power dissipation is given by an expression like this. For alpha is the activity factor how much frequently the output is changing state which means it is charging or discharging, C is the load capacitance V DD power supply voltage and f is the frequency of the clock. Now when you are talking about reducing we can reduce any one of these 4 parameters. So, this strategies can include reduction of all this 4.

So, when we say that we want to reduce the activity factor, there are techniques like clock gating, sleep mode we shall see this can be adopted. Clock gating means sometimes we are not required you can stop the clock from coming into a circuit. So, if the clock can be disabled, so the activities will also stop so alpha naturally will go down the value of alpha right. So, similarly you can try and reduce C, small transistors, short wires, smaller fan outs, various techniques can be used to try and reduce the load capacitance. Reducing V DD seems to be very important because it is appearing as the

square. So, reducing supply voltage is also a very common technique, but unfortunately as you reduce the supply voltage your delay also starts increasing. So, it like a trade off fine.

And of course, lastly you can reduce the frequency of course, to the extent possible without sacrificing the performance beyond acceptable levels. So, there can be a power down mode in your circuit, which will automatically reduce the frequency right.

(Refer Slide Time: 04:26)



The slide is titled "Reducing Static Power" and lists several techniques:

- A combination of various techniques can be used:
  - a) Selectively use ratioed circuits.
  - b) Selectively use low threshold nMOS and pMOS devices.
  - c) Use suitable leakage reduction techniques:
    - Stacked devices
    - Body bias
    - Reduce temperature

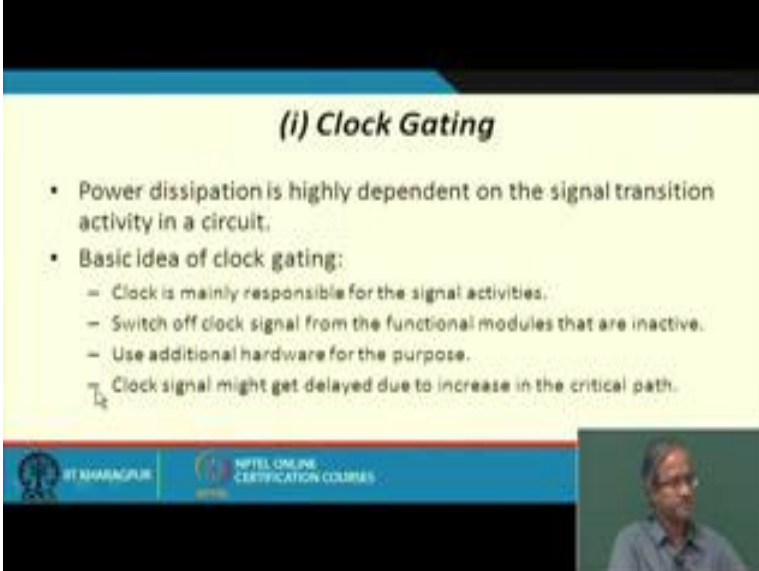
The slide also features logos for "JIT BHARADWAJ" and "NPTEL ONLINE CERTIFICATION COURSES" at the bottom, and a small video inset of a speaker in the bottom right corner.

Now, for reducing the static power, you recall static power is consumed whenever there is a switching of the output. So, momentarily both pull up and pull downs are conducting. So, we can use a combination of several techniques. So, we can use ratioed circuit. Ratioed circuit means the conventional CMOS kind of circuit, where there is a pull up network and a pull down network. So, it is unlike the dynamic CMOS as I said where the pull up network is just a single transistor not like that. So, the output voltage is taken from a circuit which looks like a voltage divider that is typically what we refer to as by a ratioed circuit.

So, whenever we have ratioed circuit there will be several gates in series. So, the chance of this kind of switching currents flowing will become less, so that will automatically reduce. And selectively we can use low threshold devices. So, if you reduce the threshold voltage of the transistor, this static power also reduces this is one of the techniques. And of course static power also includes leakage power; so leakage reduction technique also

includes several ways let us stacked devices means a series of transistors, body bias, we change the substrate bias voltage, reduce the temperature operation, several techniques are there because these effect the leakage current directly, so many techniques can be there.

(Refer Slide Time: 06:12)



**(i) Clock Gating**

- Power dissipation is highly dependent on the signal transition activity in a circuit.
- Basic idea of clock gating:
  - Clock is mainly responsible for the signal activities.
  - Switch off clock signal from the functional modules that are inactive.
  - Use additional hardware for the purpose.
  - Clock signal might get delayed due to increase in the critical path.

NPTEL ONLINE CERTIFICATION COURSES

So, let us look at the methods one by one; first let us go back a little we look at the clock gating approach which is one way to reduce alpha. So, the premises the power dissipation in a circuit is very much dependent on the signal transition activity or alpha. Clock gating what it basically involves. So, we make an observation that clock is a signal which works like a brain of the entire system. So, whenever the clock transitions take place all the circuit elements they work in synchronism. So, it is the clock which is triggering the activities in various parts of the circuit. So, intuitively speaking if you feel that there is one part of a circuit, which is not being used currently why not switch off the clock entirely from that part.

So, that at least of that part the signal switching activity will stop, so the power consumption will become less, this is the basic idea behind clock gating. But one thing we remember. So, if you are test engineer if your task is to do testing, so you will possibly say that well clock gating is not a good idea. If you clock gating then possibly the task of testing will become more difficult; because if due to some error in that gate the clock all together stops from reaching sub circuit, then I will have no way to test that

sub circuit. But from considerations of low power clock gating is considered to be a very simple technique to significantly reduce power. So, it is again a matter of trade off. So, what you really want.

So, what I am saying is that, we switch off clock signal from those functional modules which are not active. This may involve some additional hardware and because of this additional hardware, there can be some additional delay or skew in the clock signal, this also you have to keep in mind.

(Refer Slide Time: 08:38)

- The gate selectively disables the clock.
- Need logic to generate the disable signal.
  - Control logic becomes more complex.
  - The added circuit consumes power.
  - The OR gate delay gets added to clock critical path (skew analysis required).
- Possible strategy for managing skew:
  - The OR gate can replace a buffer in the clock distribution tree.

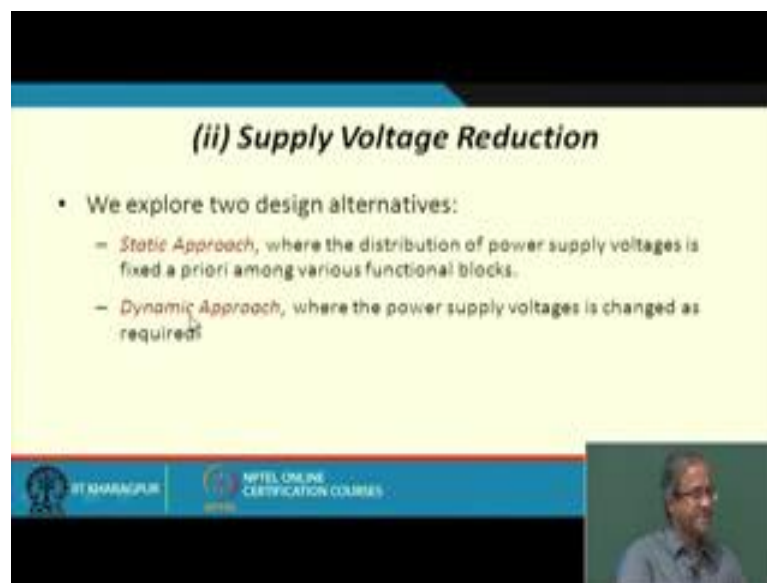
So, let us say, suppose I have a simple scenario like this, where there is a circuit block. So, I have a register where the input is loaded from there it is fed to functional unit, and the register is loaded by a clock. So, instead of this clock directly feeding the register, I am also using another signal called disable to selectively enable or disable the clock.

So, when I know that the functional unit will not be required, I will set the disable signal to 1; so that the output of this or gate will be a constant one, the value of register will not change and therefore, there will be no signal transition in the functional unit. So, the gate can selectively disable the clock fine. But for disabling the clock of course you will need some additional logic, because depending on this scenario you will have to decide and understand that when this functional unit will be required and when it will not be required; so units an additional logic and some additional control unit to generate the signal. So, those additional circuit can also consume additional power and as I said this

clock network when you design were seen this earlier in much more detail, we try to balance the delay so that we have 0 skew clock network, at least up to a level you can ensure that the clock signals reach all the n terminals at the same time.

But here what happens we have inserted an additional or gate. So, this clock signal is getting delayed by an additional gate delay, so there will be an additional skew. So, it may be required to carry out skew analysis to verify the correctness of design. But one thing you can do of course, you see for a clock network you recall you have so many buffers. Now one of the buffers you can replace by this or gate, then at least the delay will be balanced then buffer was taking some delay instead of the buffer, now that now this or gate is coming into the picture, this or gate will be incurring the delay. The delay will remain the same, you are making some changes in the clock network, modify that or gate modify that buffer replacing and by the OR gate.

(Refer Slide Time: 11:26)



The slide is titled "(ii) Supply Voltage Reduction" and lists two design alternatives:

- We explore two design alternatives:
  - *Static Approach*, where the distribution of power supply voltages is fixed a priori among various functional blocks.
  - *Dynamic Approach*, where the power supply voltages is changed as required.

The slide also features the NPTEL logo and the text "NPTEL ONLINE CERTIFICATION COURSES" at the bottom left, and a small video inset of a speaker at the bottom right.

This is one possible strategy you can use, so that skew problem will not arise right. So, this second common approach is supply voltage reduction; this is more impactful, because you see the dynamic power is proportional to this square of this supply voltage. Now there are several approaches which have been explore; static approach, also dynamic approach. Static approach says, well you analysis your circuit. So, I said earlier that you can reduce supply voltage that will make your circuit consume less power, but also your circuit will become slower. So, you can analyze your circuit and find out which

parts of the circuit are not that critical in terms of the delay, you can possibly make them a little slower without impacting the overall performance.

So, if you can identify those sub circuits, reduce the power supply for those blocks or sub circuits' right, this is the basic idea; and you do it statically at the design level itself. This is the so called static approach. Static approach says the distribution of the supply voltages is fixed at priory beforehand among various functional blocks. Dynamic approach is like you can say you have an instruction in a processor, may be the operating system will be authorized to execute that. The operating system can say what will be the current power mode of processor; power up power down, high speed, slow speed, you can define the power mode dynamically in some sense. So, here the power voltages can be changed as required.

(Refer Slide Time: 13:17)

**Static Voltage Reduction – Static Approach**

- The distribution of the voltage is always fixed.
- Additional power delivery network is required.
  - Power routing is more complex.
- Needs special care for interface between power domains.

Static

Low Supply Voltage      High Supply Voltage

Slow      Fast      Slow

The diagram shows three rectangular blocks connected in a line. The leftmost block is labeled 'Slow' and is connected to a 'Low Supply Voltage' rail. The middle block is labeled 'Fast' and is connected to a 'High Supply Voltage' rail. The rightmost block is labeled 'Slow' and is also connected to the 'Low Supply Voltage' rail. The entire setup is labeled 'Static'.

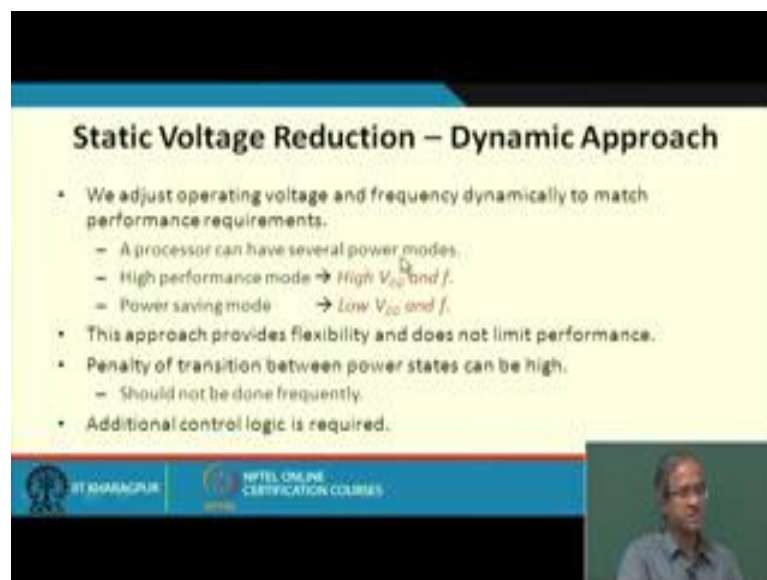
NPTEL ONLINE CERTIFICATION COURSES

So, let us look at some of these approaches pictorially. Static voltage reduction I am illustrating with a simple example here, suppose in a circuit I have 3 functional modules these are shown by this 3 rectangular blocks. Let us assume that I require this central block to be fast, but these two blocks I do not mind if they are running slower. So, what I do? I use pair of supply voltage rails; one is for low supply voltage, other is for high supply voltage.

So, the block which is supposed to run faster they are fed with the high supply voltage, and the blocks which are supposed to be running slower they are fed with the low supply

voltage, and this is done in a static way which means it is always fixed; this distribution is always fixed, but you can see this additional power networks this routing has to be done the I mean instead of our single power supply voltage, now I having two part supply voltages, so additional power delivery networks are required power routing becomes more complex. And we will see this later little bit there are some issue say, so whenever you are driving the output of circuit from this block, to the input of a circuit in this block, you see these two circuits are working at two different voltages. So, sometimes some voltage level translation may be required right. So, you need some special care for these interfaces.

(Refer Slide Time: 15:08)



**Static Voltage Reduction – Dynamic Approach**

- We adjust operating voltage and frequency dynamically to match performance requirements.
  - A processor can have several power modes.
  - High performance mode → High  $V_{DD}$  and  $f$ .
  - Power saving mode → Low  $V_{DD}$  and  $f$ .
- This approach provides flexibility and does not limit performance.
- Penalty of transition between power states can be high.
  - Should not be done frequently.
- Additional control logic is required.

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

*(A small video inset of a speaker is visible in the bottom right corner of the slide.)*

Now, let us look at the dynamic approach. So, what we are saying? We are saying that we are adjusting operating voltage and frequency, not only voltage also frequency both we can adjust dynamically to match the performance requirement.

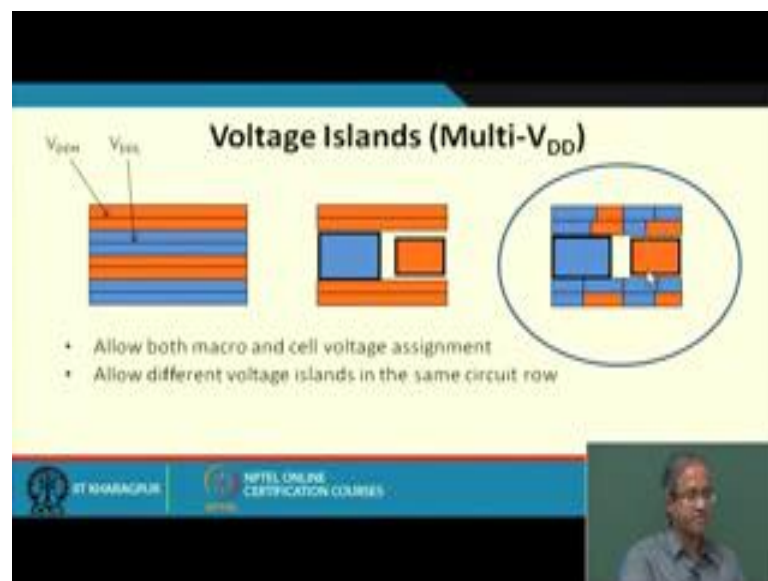
Say typically if you look at the architecture and instructions set of a modern day processor, you will find that the processors have several power modes. See earlier the processors did not worried about power, but now most of the processors which have come to the market, say means in the laptops also you must have seen there are several power modes in the laptop, you can set one of them. This is possible because the processor supports this kind of configuration, you can tell the processor that I want high



performance mode or I want low performance, but high battery mode; that means, I want to consume less energy from the battery; I want to run for longer times like that.

So, there can be two mode there can be several intermediate levels also, say in one extreme it can be high performance mode, where I say higher is power supply voltage and of course, higher frequency of operation. You can have a power saving mode; lower supply voltage, lower frequency of operation of course, there can be several intermediate levels also. So, this dynamic approach provides flexibility; because depending on the user and also the kind of programs you are running you can adjust the power level or the performance level right? But the point to notice that when your reconfiguring switching from one more to the other, it can several milli seconds. So, you should not carry out these transitions very frequently, this you should remember in of course, you need some additional logic for implementing this.

(Refer Slide Time: 17:11)

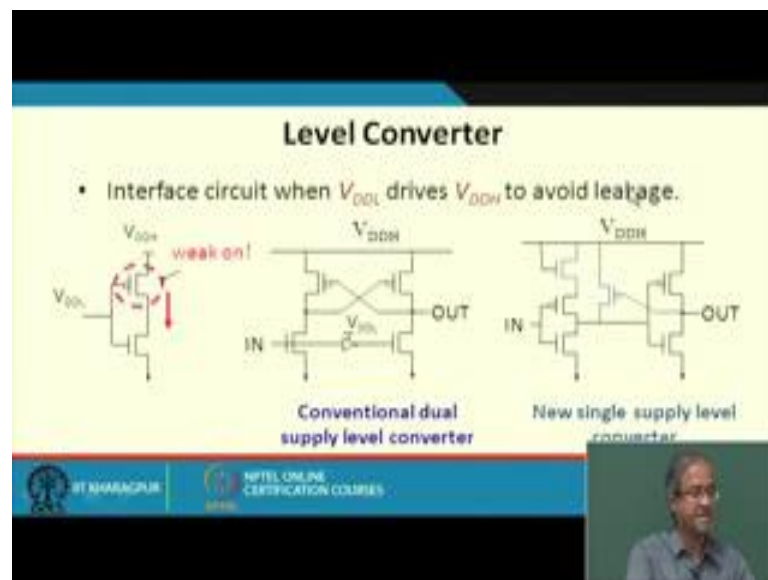


Now, actually here we are saying that how this multiple voltage is can actually be implanted on a chip. You see you can have some voltage islands; like let us look at a standard cell kind of a placement, where these cells are arrange in rows. This orange and blue these may indicate two different voltage levels; orange may indicate  $V_{DD}$  high, blue may indicate  $V_{DD}$  low. So, all the cells which are supposed to high performance they will be placed in these orange rows, and the lower performance one will be placed to the blue rows. Well, in a seek or in a full custom design also you can have the same

kind of philosophy, or you can have arbitrary blocks some of them blue blocks some of them orange blocks, you can put them in your chip wherever you need right.

Now, if you want to generalize this, you can allow different voltage islands these are called islands different voltage islands; you can have different islands even in the same row like this. So, even in the same standard cell row you can have some as low performance, some as high performance varying supplying voltage, but of course, if you do this the problem of routing this apply voltage is will become more difficult right if you mix them together.

(Refer Slide Time: 18:48)

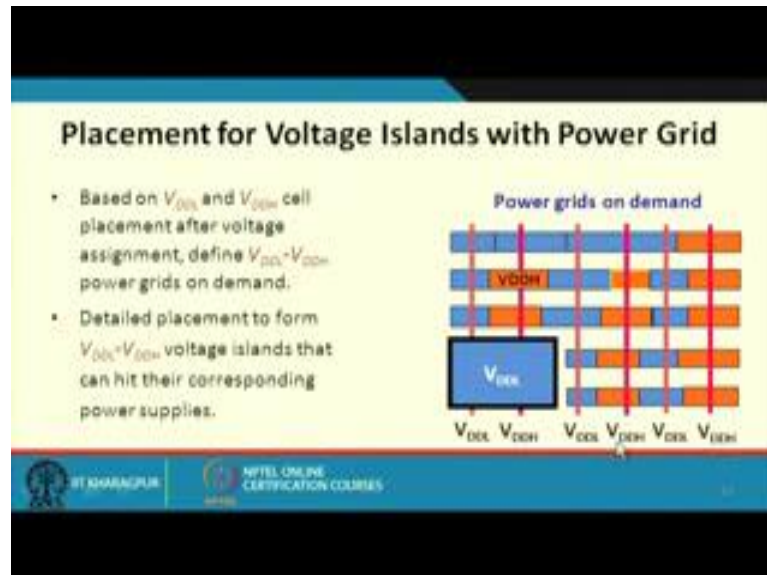


So, as I said earlier that you may need a level converter, so whenever you are driving circuit with a lower supply voltage to higher supply voltage or vice versa. So, I am not going into the detail of this circuit. So, I am just indicating that whenever you have a circuit which is operating at a high supply voltage, but you are feeding it from a circuit which was operating from a lower supply voltage, then even if when this  $V_{DDL}$  is at let us say high 1. So, it is not equal to  $V_{DDH}$  it is less than  $V_{DDH}$ , then this transistor might still be weak on and a small current might be flowing.

So, to carry out the level translations, you have several possible circuits. Like the first circuit says you have a circuit where you need two supply voltages, one gate is working with the lower supply voltage, but in the circuit we have a single supply voltage. So, you can use one of this circuits to carry out the voltage conversion, without incurring this

leakage current, because if we do not take any care this leakage will be a major issue fine.

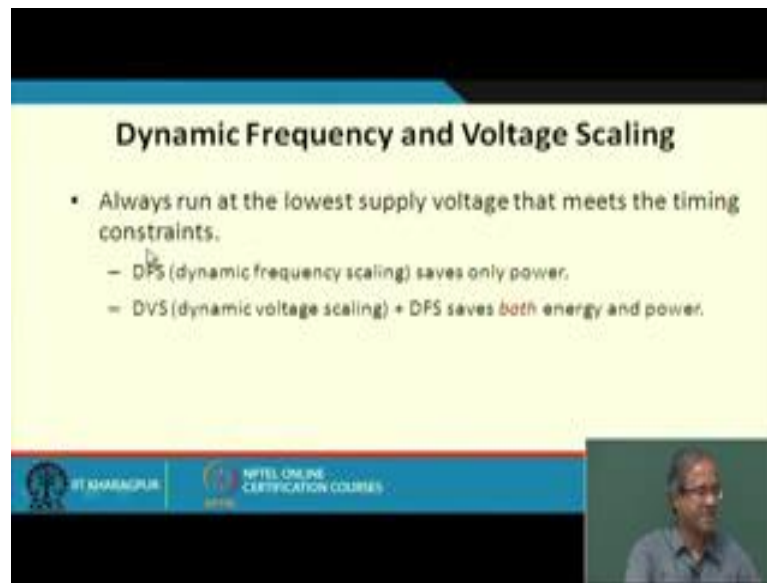
(Refer Slide Time: 20:04)



And here I am giving the solution where you can mix this low and high V DD blocks in the same rows; like you are defining something like a power grid, vertically you are running several wires say alternatively say V DD low high, low high, low high and you place this blocks in such a way they are aligned with the voltage grid with which you want to power them like the orange once, they must have this V DDH rail running through them.

The blue once they must have at least one V DDL this rail running through them and so on right. So, this is strategy where you are embaying or your integrating you can say these power levels multiple voltage levels along with cell placement. You are integrating placement with creation of the well tool creation of the voltage islands. So, at the net list level itself you are defining that which of this cells will be working with a high supply voltage, and which of the cells will be working with a low supply voltage and accordingly you take a decision right.

(Refer Slide Time: 21:30)



The slide features a yellow background with a blue header and footer. The title 'Dynamic Frequency and Voltage Scaling' is centered in bold black text. Below the title, there are three bullet points: a main point 'Always run at the lowest supply voltage that meets the timing constraints.' followed by two sub-points: '- DFS (dynamic frequency scaling) saves only power.' and '- DVS (dynamic voltage scaling) + DFS saves both energy and power.' The footer contains the IIT Kharagpur logo on the left and the NPTEL Online Certification Courses logo on the right. A small video inset of a man is visible in the bottom right corner of the slide frame.

And of course, this is very flexible thing; you want to do the entire thing in a dynamic way. Frequency and voltage scaling you are doing dynamically, and by doing that you can adjust the performance level of the processor according to your requirement.

Now, the factor that is taken into account is something like this, like you see you have some timing constraint in a circuit to be met; like I have a circuit that must produce the output after a delay of  $\Delta$ . Suppose in your design you have designed it in such a way that you are applying a fast clock, the output is generated much earlier than  $\Delta$ ; that is of course, not a problem from the point of your correctness, but from the point of view of power dissipation unnecessarily you are consuming low power, because you are operating that circuit with a higher frequency or a higher  $V_{DD}$ , which is not required because you could have delayed the output a little bit.

So, the idea is that you make the frequencies slower to a level, which is sufficient to produce the output within the required time budget whatever you have right. This is what is mentioned here; that always run at the lowest supply voltage that meets the timing constraint. So, you can have two approaches one is that your only adjusting the frequency, dynamic frequency scaling; see if you adjust the frequency you are saving only power, but not the energy because you see frequency you can increase or decrease.

If you increase the frequency may be you will finish a computation faster, if you reduce the frequency you will finish the computation slower, but the number of transitions

occurring in a circuit will remain the same. So, the total energy will not change may be the instantaneous power of the average power, because you are doing it over a longer period of time the average power will be less, but the amount of power will be drawing from the from the battery that is the energy that will not change right.

But if you do both dynamic voltage scaling plus dynamic frequency scaling; because if you do a voltage scaling then energy is also shift not only power right. Because with frequency you are simply slowing down the process of drawing the current from the battery, but by voltage scaling you are reducing the magnitude of the current, which means you are saving on energy.

(Refer Slide Time: 24:36)

- A DVS+DFS system requires the following:
  - A programmable clock generator (PLL)
    - PLL from 200MHz → 700MHz in increments of 33MHz
  - A supply regulation loop that sets the minimum  $V_{DD}$  necessary for operation at the desired frequency.
    - 32 levels of  $V_{DD}$  from 1.1V to 1.6V
  - An operating system that sets the required frequency + supply voltage to meet the task completion deadlines.
    - heavier load → ramp up  $V_{DD}$ , when stable speed up clock.
    - lighter load → slow down clock, when PLL locks onto new rate, ramp down  $V_{DD}$ .

So, let us look at it. So, system which combines dynamic voltage and frequencies scaling; so I am telling about a typical system, this will require the following. So, it can require a programmable clock generator that is typically based on a phase lock loop control circuit. So, in one of the systems the clock frequency was controllable over a wide range, say 200 to 700 megahertz in increments of 33 megahertz. So, you see there is a white design space available to you; you can set the clock frequency to so many different levels right.

Similarly, you also have a power supply network which is also programmable, that can set the minimum  $V_{DD}$  value; say in that system it has from 1.1 to 1.6 volts in 32 levels, you can said the voltage to many levels in between. So, again here also you can change

the voltage as well as frequency with quite fine control. And thirdly as I said there will be some instructions in a processor that can be used to set this voltage or clock levels, these are some hardware instructions; hardware level instructions, and in a computer system the operating system can be delegated the responsibility of initializing or programming these things, so that it will be setting the required frequency and supply voltage, such that the task completion deadlines are met.

So, if a system is having heavier load the OS detects that, then it can ramp up or increase the  $V_{DD}$  to make it run faster and after  $V_{DD}$  has got stabilized, then it can also increase the clock frequency right. So, the operations can be made faster, but on the other hand if the system is running on a lighter load; then you can slow down the clock first, then when the PLL locks the new frequency you can slow down the  $V_{DD}$ , this is usually the order of changes that is done, right.

(Refer Slide Time: 27:18)

**(a) Using Transistor Stacks**

- Transistor stacks that naturally appear in a design help in reducing leakage current.
  - Stack :: several transistors in series.
  - Effective leakage current decreases.

1.5V  
VG1 = 0  
Node 1: Vq1 = 89mV  
VG2 = 0  
Node 2: Vq2 = 34mV  
VG3 = 0  
Node 3: Vq3 = 14mV  
VG3 = 0

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, talking about the leakage reduction techniques there are various methods which are possible. So, I am not going into the very details of this, I am giving with you idea. So, you can use technique called transistors stacking, dual threshold partitioning, or you can have variable thresholds, let us see what these are.

Transistor stacking means you see in a typical CMOS circuit you have several of this transistors which you get connected in a series, the idea is very simple. So, if you have a circuit where from an output node or a  $V_{DD}$  node, you have few numbers of transistors

to ground then the leakage current will be higher; because each of these transistors even when they are turned off, they will be having some kind of leakage resistance so which typically very high. But if there are such transistors in parallel, this leakage currents will be finding the parallel paths and the effective leakage current will increase, but if you have several of this transistors in series which is called stacking, then the overall resistance will be the some of the offer is instead of all these transistors, that will result in less leakage current of flow simple this is the basic idea.

So, if you have a circuit where there is this kind a transistor stacking in series, then the total leakage current will be reducing, this is pretty obvious.

(Refer Slide Time: 29:01)

**(b) Dual- $V_{TH}$  Partitioning**

- Basic concept:
  - Two types of transistors: low- $V_{TH}$  and high- $V_{TH}$ .
  - Low- $V_{TH}$  is faster, but have more leakage.
  - Matter of design tradeoff.
- We partition the transistors of a circuit into two clusters.
  - Objective: Reduce delays of the critical paths.
  - Try to keep overall power consumption low.

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

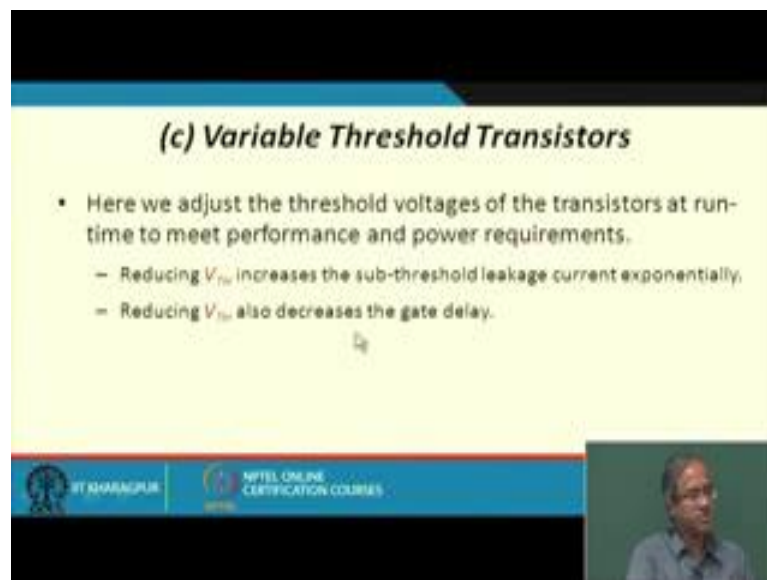
And this is also a very popular technique which is used; you see when the most transistors are fabricated. So, by changing the doping levels of the different regions, you can change the threshold voltage of the transistors. So, the threshold voltage of the transistors have a direct impact on the speed of the devices right, this is the basic idea behind this approach. So, we are saying that our circuit net list can consist of a large number of transistors, some of these transistors we are classifying as low threshold, some of the transistors as high threshold.

Low threshold voltages run faster, but they incur more leakage current. So, there is a trade off. So, normally for those parts of the circuit where I do not need high speed, I will set the transistors to high threshold transistors; because at least my leakage current will

be less, but for the circuits or the parts, which fall in the critical path I want them to work as fast as possible. I define those transistors as low threshold; so in this way I can do some kind a partition of the transistors in to 2 sets; high speed and low speed by analyzing the critical paths and the delay requirements of the circuit, so that my circuit works in a efficient way and yet consume less power right.

So, as I said we define partition of the transistors into two clusters, the basic objective is to reduce delays of the critical paths and of course, to keep the overall power consumption at a low level.

(Refer Slide Time: 31:00)



The slide is titled "(c) Variable Threshold Transistors". It contains the following text:

- Here we adjust the threshold voltages of the transistors at run-time to meet performance and power requirements.
  - Reducing  $V_{TH}$  increases the sub-threshold leakage current exponentially.
  - Reducing  $V_{TH}$  also decreases the gate delay.

The slide also features a logo for "IIT KHARAGPUR" and "NPTEL ONLINE CERTIFICATION COURSES" at the bottom left, and a small video inset of a speaker at the bottom right.

And in general you can have variable threshold transistors; see here you have a technology where not only two thresholds, you can adjust the threshold to a final level to many levels. So, it is not very difficult to do, if you can change the sub stress substrate bias voltage of a transistor, if you have a mechanism to change that voltage then the threshold voltage will also change this is the basic idea behind this approach. So, this point have already said that if you reduce  $V_{TH}$  your transistors become faster, but your leakage current increases quite rapidly, reducing  $V_{TH}$  also decrease makes a transistor faster these are the two things I mentioned already.



(Refer Slide Time: 31:55)

• How to change threshold voltage?

- By changing the body-bias voltage.
- For NMOS, the substrate is normally tied to ground ( $V_{sb} = 0$ ).
- A negative bias on  $V_{sb}$  causes  $V_t$  to increase.
- Adjusting the substrate bias at runtime is called adaptive body-biasing (ABB) or dynamic threshold scaling (DTS).
  - Requires a triple-well fabrication process.
  - More complex.

The slide also includes a circuit diagram of a CMOS inverter with body bias voltages  $V_{sb,p}$  and  $V_{sb,n}$  applied to the substrate of the PMOS and NMOS transistors respectively. A graph shows the threshold voltage  $V_{th}$  (volts) on the y-axis (ranging from 0.4 to 0.9) versus the substrate bias voltage  $V_{sb}$  (volts) on the x-axis (ranging from 0 to -2.0). The graph shows a linear increase in  $V_{th}$  as  $V_{sb}$  becomes more negative.

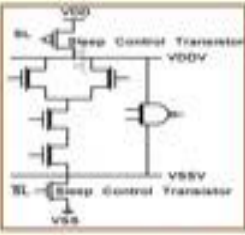
So, the approach works like this. So, you have these transistors this substrate you connect to two voltage sources.  $V_{sb}$  substrate bias you call them substrate bias voltage; one for the P transistor, one for the N transistor. So, we have change in the threshold voltage by changing the body bias voltage. Now this graph actually shows you that by changing the substrate bias voltage, you can change the threshold voltage from 0.4 volt up to about 0.9 volt over a wide range right.

So, during run time you can again have some instructions to change these voltages, so that you can adjust the threshold voltage of the transistors to make them run either faster or slower. So, this again opens up new dimension to the design, the way you are designing your transistor level net list, the way will be programming them partition them, adjust the thresholds; so there is there is a new set of challenges which come it here if you have this flexibility right. But the only problem here is that, if you want to have this kind of a facility your fabrication becomes little more difficult; because traditional CMOS fabrication requires a double well fabrication process, now we need a triple well fabrication process which increases the cost of fabrication, all right.

(Refer Slide Time: 33:32)

### Power Gating using Sleep Transistors

- We can gate the supply rails when the circuit is in sleep mode.
  - In normal mode, **sleep = 0** and the sleep transistors must present as small a resistance as possible (via sizing).
  - In sleep mode, **sleep = 1** and the transistor stack effect reduces leakage by orders of magnitude.
- Or can eliminate leakage by switching off the power supply (but lose the memory state).



IT MARAGUJI | NPTEL ONLINE CERTIFICATION COURSES

And lastly we discussed a method that is called power gating using sleep transistors. So, the idea is fairly simple you see normally we have this  $V_{DD}$  and  $V_{SS}$  these two power supply voltages that is driving the gates. Now we have been auxiliary voltage at a lower level  $V_{DDV}$  and  $V_{SSV}$ . So, what we are saying is that sometimes we can set this circuit to the sleep mode by activating these transistors by setting this  $S_L$  to 1.

So, if you said  $S_L$  to 1 these transistor as well as these transistors they will be turned off. So, now, this circuit will be powered by this lower voltage. So, it is like power down mode. So, if we can put circuits to sleep, where you are not doing any computation right now, power consumption can be significantly reduced. Now see some of the circuits can also be storage elements flip flops. So, you really cannot switch off  $V_{DD}$ . So, if you switch off  $V_{DD}$ , your data stored in the memory elements might get lost. So, it is better to have a power down mode where you just reduce the voltage level, let the data storage be retained, but whenever else you required again you again jack of the voltage and start using it right.

(Refer Slide Time: 35:03)

The slide contains the following text:

- Can reduce power by as much as 1000X.
- Smaller voltage swing (IR drop) on sleep transistors.
- Features:
  - Lower performance.
  - Increased noise coupling.
  - Local power grid design.

On the right side of the slide, there is a diagram of a chip layout with three labels and arrows pointing to specific features:

- Power Switch Control Signals
- Embedded Power Switches
- Rows of Standard Cells

The bottom of the slide features the IIT Kharagpur logo and the text "NPTEL ONLINE CERTIFICATION COURSES". A small video inset of a man is visible in the bottom right corner.

This is basically what is meant by power gating using sleep transistors; and in a chip level you can easily implement this, you can have some power switch is embedded in the standard cells, these are the rows of standard cells you can have some this kind of switch is embedded. So, by using this switch is you can adjust these sleep transistors on and off.

So, this technique if use in a proper way, it has been demonstrated that you can reduce power by as much as 1000 time. So, this is very good design in that sense good design philosophy to control power, but of course, it requires lot of analysis; circuit level analysis, and calculation to find out that which are the blocks where you need to just implement or incorporate this sleep modes.

So, with this we come to the end of this lecture. So, we continue with some other technique for reducing power in our next lectures.

Thank you.