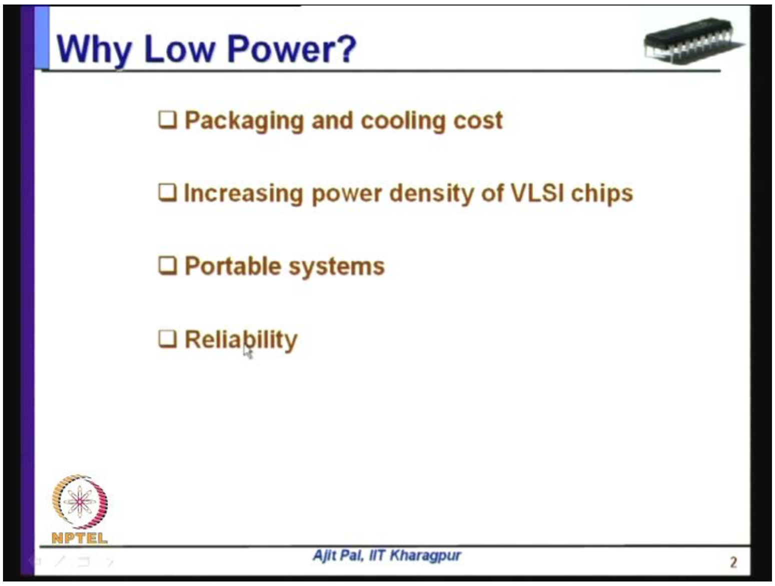
Low Power VLSI Circuits and Systems

Prof. Ajit Pal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture No. # 40
Course Summary

Hello and welcome to the last lecture of the lecture series on low power circuits and systems. In this lecture, I shall give an overview of the entire course. In other words, it will give a summary of the various topics that I have covered in this particular lecture series.

(Refer Slide Time: 00:43)



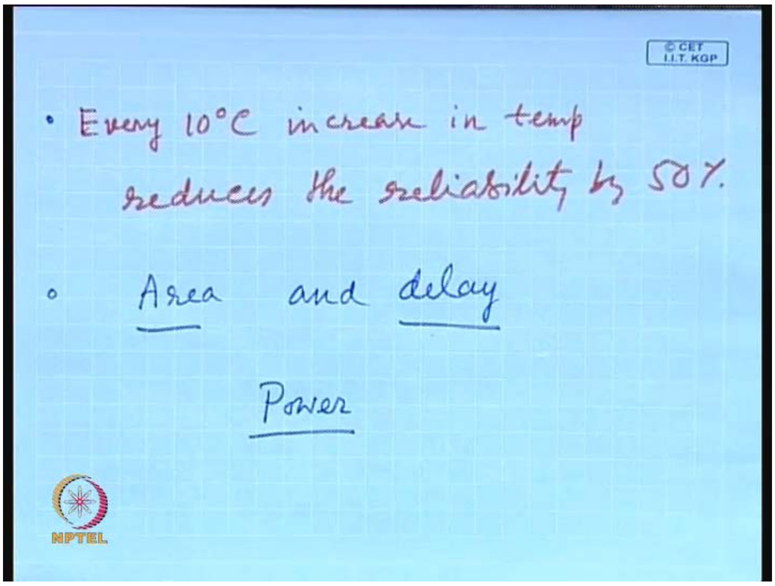
Why Low Power?

- ❑ Packaging and cooling cost
- ❑ Increasing power density of VLSI chips
- ❑ Portable systems
- ❑ Reliability

You may recall that first I address the question, why low power? Why at all low power circuits and systems are necessary? And we know that there are various reason, number one is packaging and cooling cost. If the circuit is not designed for low power, the power dissipation is high, so the packaging and cooling cost of the chip is higher. So, to reduce the packaging and cooling cost, it is essential to go for low power design. So, this is one reason. Second is as you know as the device size is sinking, because of the advancement of VLSI technology the power density of the VLSI chip is increasing; that means, the amount of power that is being dissipated per unit area that is increasing; that lead that will lead that leads to what is known as hotspot.

So, to avoid that I mean to reduce the increasing low power density it is also necessary to go for low power design so that the power density is not very high and they know hotspot is developed. Then another reason is the use of VLSI circuits in portable systems. As you know there is a proliferation of portable systems, portable battery operated systems in the form of laptops, cell phones, PDA's and so on. So, in for these devices, as you know they are running from battery. So battery life is the important criteria which decides its salability and you know if the power dissipation is not is not low then, you cannot really sale this products. So for portable systems low power is extremely important than reliability. It has been found that the power dissipation is directly related to low power.

(Refer Slide Time: 03:04)



© CET I.I.T. KGP

- Every 10°C increase in temp reduces the reliability by 50%.

○ Area and delay

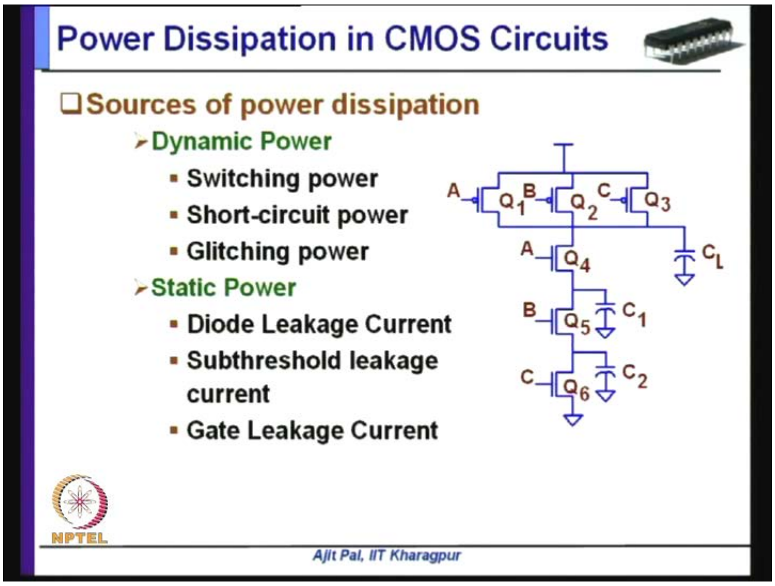
Power

NPTEL

And it has been observed that every 10 degree increase in temperature 10 degree centigrade increase in temperature reduces the reliability by half by 50 percent. So you can see, why the low power is important before reliable operation of the circuit. Is it necessary to have go for low power design so that the temperature within the core of the chip is not very high.

Then coming to the environment, as you know as the processors are dissipating power, that power is ultimately discharged to the environment; so, it is leads to ultimately leads to global warming. So, it has strong impact on the environment; that is the reason another reason why it is necessary to go for low power. And as a consequence, there is a changing trend. You know earlier the whenever the chip design is considered then area and delay were the two parameters which were considered for the design of a circuit. But in addition to area and delay, another factor is now has become very important that is your power; that means, in the present day VLSI chip design apart from area and delay of the circuit, it is essential to have low power; that means, power has become another important design criteria. So, these are the various reasons why low power is important in VLSI circuits.

(Refer Slide Time: 05:09)



Power Dissipation in CMOS Circuits

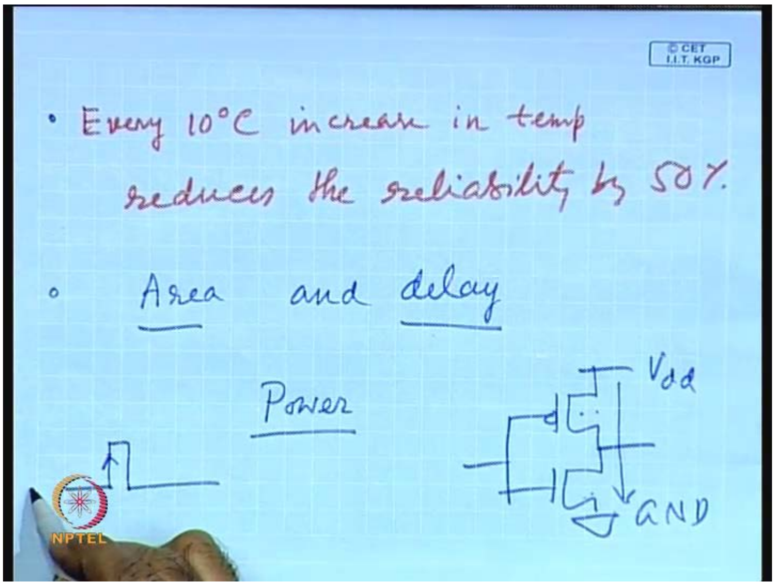
❑ Sources of power dissipation

- Dynamic Power
  - Switching power
  - Short-circuit power
  - Glitching power
- Static Power
  - Diode Leakage Current
  - Subthreshold leakage current
  - Gate Leakage Current

Ajit Pal, IIT Kharagpur

Now, let us see how we can really go for low power. To to to design circuits by using low power; I mean to have low power, it is essential to identify the sources of power dissipation. I mean only then we can we can go for reducing the power dissipation or go for design for low power. So as we know, the power dissipation can be broadly divided in to two types; dynamic power. Dynamic power is the power dissipation, when the circuit is in operation; that means, input is changing; clock is applied to it; output is changing. In such a case, whatever is the power dissipation that is known as dynamic power, and dynamic power has got three components; switching power, short circuit power and glitching power. I have already discussed in detail, what do we really mean by switching power? Switching power is essentially used for charging and discharging the capacitors; various capacitors that is present within the circuits. So, as you charge a capacitor then, there is some power dissipation and these transistors and also whenever you discharge then there is power dissipation. Then, short circuit power dissipation is whenever you know in the middle part of the voltage, when the supply input voltage is less than V dd by V t and greater than the threshold voltage of the pMOS transistor then, there is a then the nMOS and pMOS transistors are shorted.

(Refer Slide Time: 06:55)



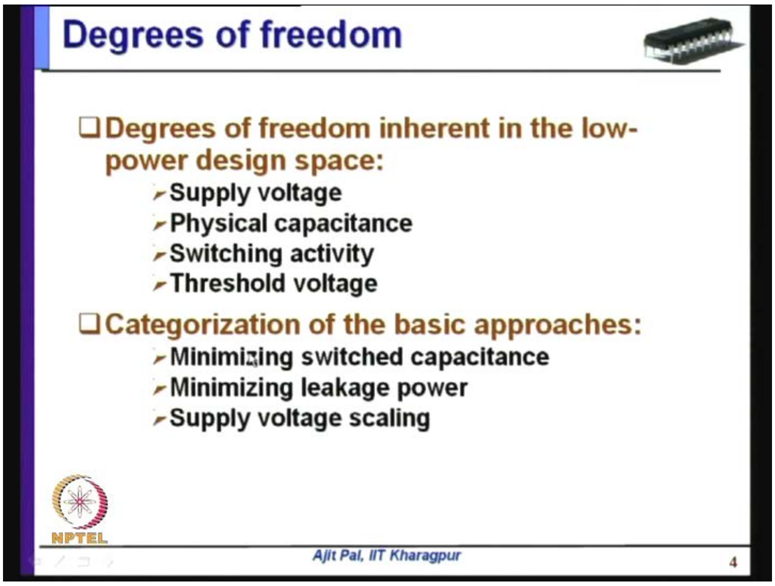
- Every 10°C increase in temp reduces the reliability by 50%.
- Area and delay

Power

That means, you have got two transistors in series. This is the simple inverter; the both the transistors turn on for for the input voltage range for certain input voltage range and during that there is a shorting between the V dd and ground and this is known as the short circuit short circuit power, and glitching power dissipation occurs because you know, there are undesired transitions within the circuit because of the delay. Ideally, the gate should not have any delay, but unfortunately there is finite delay of the gates and that delay leads to power dissipation leads to glitches and whenever glitch occurs undesired transitions like this type of spikes occurs they lead to power dissipation; that means, again for charging whenever it goes from zero to V dd and again and also power dissipation occurs when the glitch signal goes from V dd to ground. So, this is glitching power dissipation.

Then, there static power dissipation when the circuit is in standby. Normally as we know, we assume that when the circuit is in standby, there is no power dissipation, but unfortunately that is not true in deep sub-micron circuits because there is still power dissipation, because of you know leakage currents. And there are different types of leakage currents like diode leakage current, sub threshold leakage current and gate leakage current, which I have discussed in detail, and these leakage currents lead to power dissipation even when the circuit is not in operation; that means, it is in the standby conditions. So, when the circuit is operation both dynamic power and static power dissipation occurs. On the other hand, when the circuit is in standby condition not in operation then static power dissipation occurs. So, these are the sources of power dissipation.

(Refer Slide Time: 09:07)



Degrees of freedom

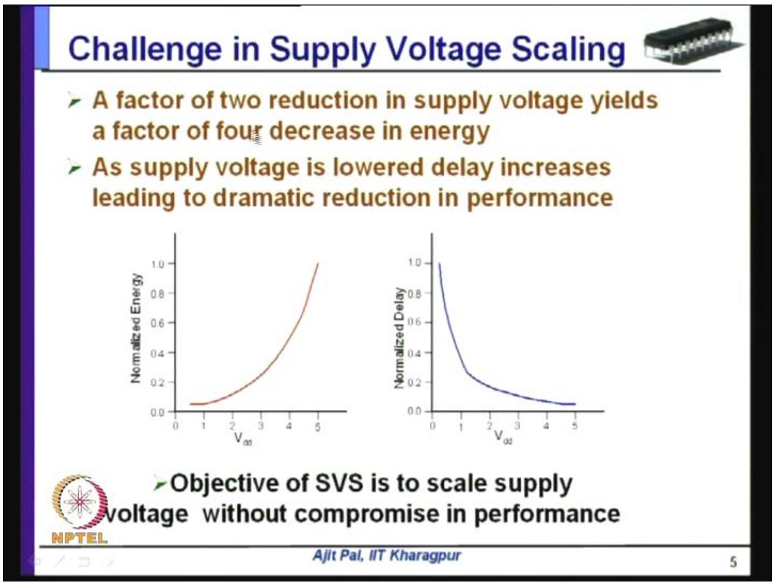
❑ Degrees of freedom inherent in the low-power design space:
- Supply voltage
- Physical capacitance
- Switching activity
- Threshold voltage

❑ Categorization of the basic approaches:
- Minimizing switched capacitance
- Minimizing leakage power
- Supply voltage scaling

Now let us see, how we can really identify what are the degrees of freedom inherent in low power design space. So, various sources of power dissipation we have discussed. Now, on what parameters these power dissipations depend that we have identified discuss in detail. And we have seen that, the power dissipation is dependent on these parameters like supply voltage, physical capacitance, switching activity; these three are responsible for you know the switching power dissipation. As we know that power dissipation is proportional to alpha C L V dd square and f. So, alpha is a switching activity; V dd is a supply voltage; C L is the physical capacitance and f is the frequency of operation. So, these are the parameters which we can control to reduce the dynamic power. On the other hand, for reducing the leakage power, threshold voltage is the key parameter which can be controlled to reduce the leakage power that is your when the circuit is in stand by condition.

So, these are the parameters which we can which we can control. And based on this, we have categorized various low power design approaches and these are these can be categorized in to three types; number one is minimizing capacitance, minimizing leakage power and supply voltage scaling. So, these are the three categories which have been discussed in detail in this lectures in this lecture series, and and I shall try to give an overview of these techniques in this lecture.

(Refer Slide Time: 11:07)



Challenge in Supply Voltage Scaling

- A factor of two reduction in supply voltage yields a factor of four decrease in energy
- As supply voltage is lowered delay increases leading to dramatic reduction in performance

- Objective of SVS is to scale supply voltage without compromise in performance

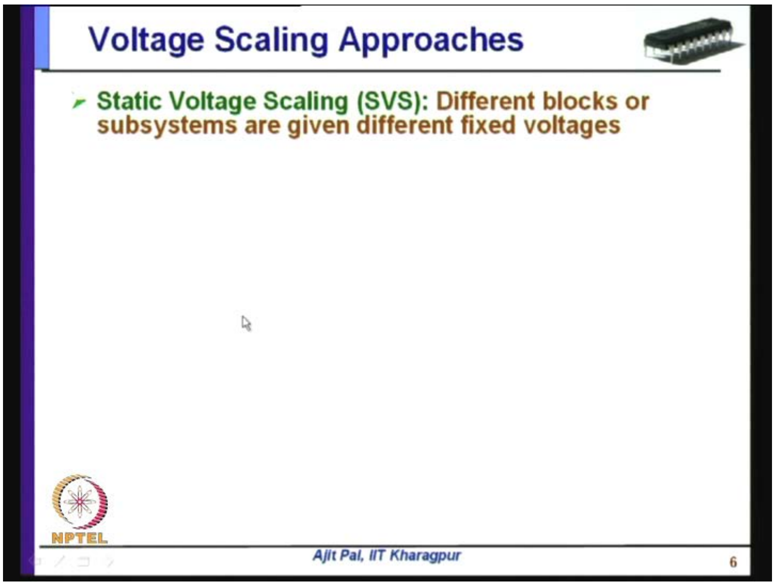
Ajit Pal, IIT Kharagpur

5

Now as we have seen, supply voltage is the most important parameter because there is a I mean square law; I mean dependence V dd square is proportional to power dissipation. So, supply voltage has strong dependence on power dissipation, and that is the reason why supply voltage is very, very important and reduction in supply voltage is used to reduce the power dissipation. Now, there is a big challenge; whenever we use supply voltage scaling for reducing power dissipation indeed there is I mean, a factor two reduction in supply voltage yields a factor of four decrease in energy. Unfortunately, as the supply voltage is lowered, delay increases. As you can see in this curve, there the normalized energy is plotted taken as the supply voltage. As the supply voltage is reduced from 5 volt to 1 volt, we can see the how the energy is getting reduced; energy dissipation is getting reduced.

Unfortunately as we do so, we find that as the supply voltage is reduced from 5 volt to 1 volt, the delay is increasing sharply. So, what it means? It means that, we can reduce the power dissipation by reducing the supply voltage, but that leads to increase in delay. In other words performance degrades. So the challenge is we shall reduce the power dissipation without compromising performance that is the challenge. So, objective of supply voltage scaling is to scale supply voltage without compromising performance. That is the basic idea of supply voltage scaling techniques used for achieving low power.

(Refer Slide Time: 13:14)



Voltage Scaling Approaches

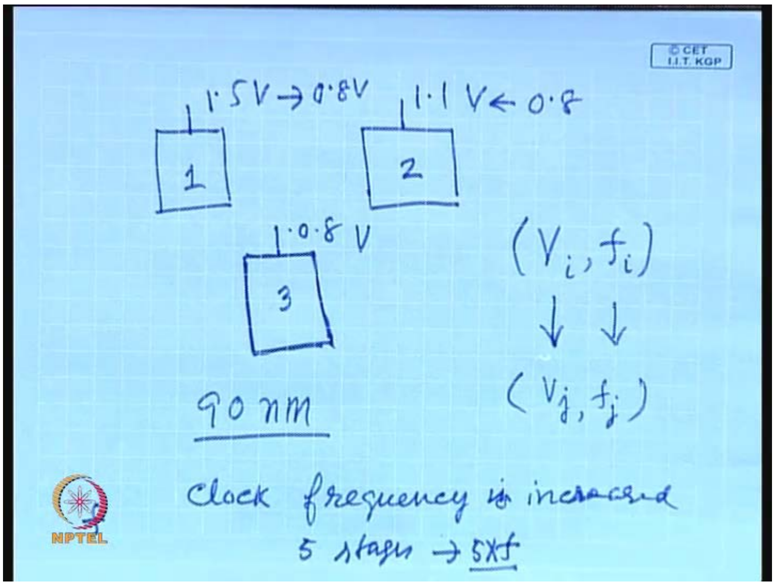
- Static Voltage Scaling (SVS): Different blocks or subsystems are given different fixed voltages

Ajit Pal, IIT Kharagpur

6

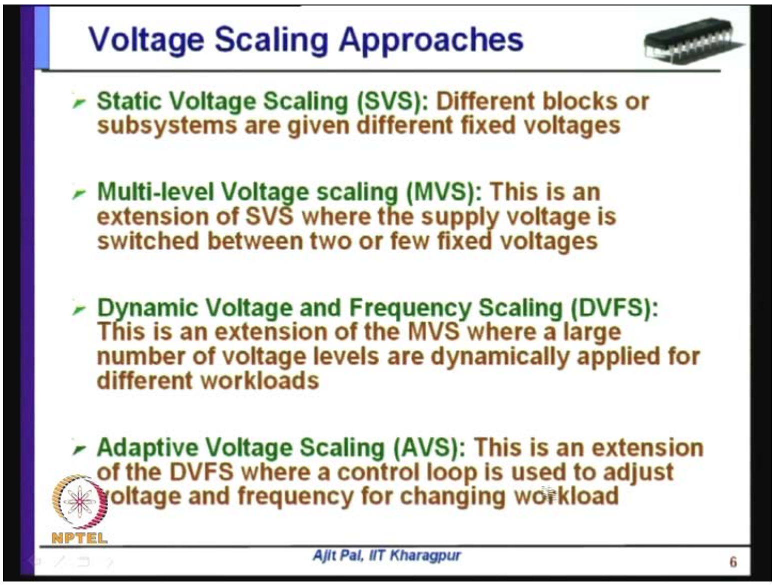
Now, we shall consider various voltage scaling approaches; they can be categorized in to four different types. Now, first one is known as static voltage scaling. So in case of static voltage scaling, different blocks of subsystems are given different fixed voltages;

(Refer Slide Time: 13:39)



That means, what is being done? The entire circuit is divided into different voltage domains say three different voltage domains, and this is applied one voltage may be 1.5; this is given say 1.1; this may be given 0.8. So, these voltages are given to different parts of the circuits. So that means, this is one voltage domain; this is another voltage domain; this is another voltage domain. So, these three voltage domains are identified which will operate in such a way that, the performance will not be degraded, but they will operate in different voltages and which are fixed and this is done at the time of design.

(Refer Slide Time: 14:21)



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

Ajit Pal, IIT Kharagpur 6

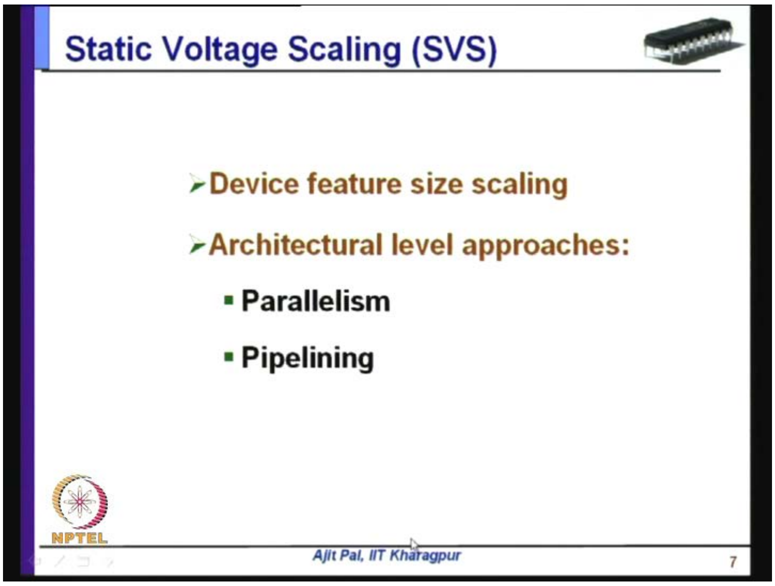
So, this is this is known as static voltage scaling SVS. Then, your multilevel voltage scaling MVS; this is an extension of SVS that is static voltage scaling, where the supply voltage is switched between two or few fixed voltages. Here in your static voltage scaling, we have assumed that these voltages are fixed, but what can be done in MVS this can be made 1.5 volt at some point of time at some other point of time, it can be made 0.8 volt or these voltage can be made 0.8 volt at some other point of time; that means, at different time instances the different blocks of the circuits are given different supply voltages that is your multi-level voltage scaling. So, switched between two or few two or few; note this the number of such voltages is very few may be two or three or four not more than that.

Then, comes the dynamic voltage and frequency scaling where I mean, which is an extension of the multi-level voltage scaling where a large number of voltage levels are dynamically applied for different workloads. So, this particular technique is based on the monitoring of the workload of the processor, and as the workload changes, the voltage and frequency so, we can say V i and f i a pair; this pair is kept on changing depending on the workload; why both are changed because as we know as we reduce the supply voltage, you have to reduce the frequency of operation as well because for as you reduce the supply voltage the delay increases so it will operate lower voltage. So, that is a reason why as you go from one workload to another workload, you will change from V i f i to may be V j f j so, this is another pair. So, this is this is this is known known as dynamic voltage frequency scaling and in this particular case, you have got a large number of voltage frequency pair; not few as was done in case of multi-level voltage scaling.

Then finally, you have got adaptive voltage scaling AVS. This is an extension of the dynamic voltage and frequency scaling, where a control loop is used to adjust voltage and frequency to changing workload. So, the other three approaches are are identified at designed time; that means, based on the analysis of the static timing analysis of the circuit, it is identified at what voltage it should operate; at one for frequency and so on. And that is how depending on the workload, the voltage and frequency pairs are identified. Now, whenever the circuit is running in working condition that time the environment is not known at designed time; that means, it can operate at different temperature and you know that because of that, the the voltage and frequency requirement can be different, and that is the reason why in adapted voltage scaling the the the voltage actually, the it is a close loop technique whatever voltage is required that is identified and frequency of operation that is required for that voltage to sustain that workload is is identified at runtime. So, you have an additional hardware which finds out the voltage and frequency to be used in the circuit.

So, adapted voltage scaling techniques require additional hardware to identify, to find out the voltage and frequency and; however, it gives you much better performance, much reduction in power dissipation because it is done at runtime.

(Refer Slide Time: 18:43)



Static Voltage Scaling (SVS)

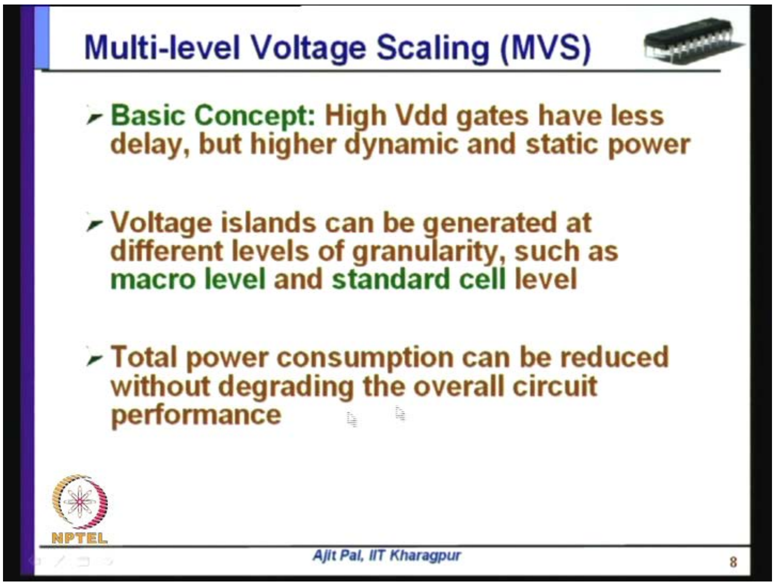
- Device feature size scaling
- Architectural level approaches:
  - Parallelism
  - Pipelining

Now, so far as the static voltage scaling techniques are used, we have discussed two basic approaches; now first one is device feature size scaling. As we know, because of the advancement of technology, the feature size of the devices are decreasing; as the feature size of the devices are decreasing, the supply voltage is correspondingly reduced to achieve lower power dissipation; that is what is being done in your device feature size scaling. So and as you go from one technology generation to next technology generation as the device size is reduced, the supply voltage is also reduced that which leads to reduced dynamic power dissipation. Of course, as you reduce the feature size below 90 nanometer then of course, you will see that the leakage power increases. So although, dynamic power dissipation decreases as you reduce the supply voltage the leakage power increases.

So, we have to use suitable technique to reduce the leakage power, and which can be done by using by controlling the threshold voltage which I shall discuss, which you have discussed in detail. So, that is about device feature size scaling. Now, coming to architectural level approaches we have discussed two techniques; parallelism and pipe lining. Normally, as we know parallelism is used to increase the performance; that means, instead of one processor if we use two processor then performance is doubled, but instead; that means, here in this case what we are doing, we are using parallelism for higher performance, but when our objective is not performance, but power lowering the power dissipation then what can be done? We can use parallelism for low power; that means, we use say more than multiple processors, but which will give you the same performance as the single processor and then, what can be done? The the the voltage frequency of the processors can be reduced, and that will lead to reduced power dissipation; that means, this is this is essentially parallelism for low power. Similarly, as we know pipe lining is used again for improving the performance. Most of the modern processors are pipe line and pipe lining has been used to improve the performance. So as we know, pipe lining essentially dividing a task in to a number of sub tasks then, each subtask is performed in overlap manner; that is the basic idea of pipe lining as we have seen in details.

Now, whenever we use pipe lining then, there is a… Since you are performing subtasks the clock frequency is increased clock frequency is increased increased in pipe line system; that means, if the if you have if you have got 5 stages the clock frequency can be 5 times. So, 5 stages leads to clock frequency 5 into f. So, clock frequency is increase and as a consequence, we get the throughput increases by 5 time performance increases, but instead of this, instead of increasing the clock frequency if we do not increase the clock instead of making it 5 f, if we keep the clock frequency same then, what can be done? For each of the stages, the supply voltage can be reduced because clock frequency is reduced; delay is we can afford larger delay and as a consequence, the clock frequency is reduced; this will give you and also the frequency is reduced. So, reduced voltage and frequency can be used whenever we go for pipe lining for low power. So, these are the static voltage scaling approach and these are decided at design time as I have mentioned.

(Refer Slide Time: 23:37)



Multi-level Voltage Scaling (MVS)

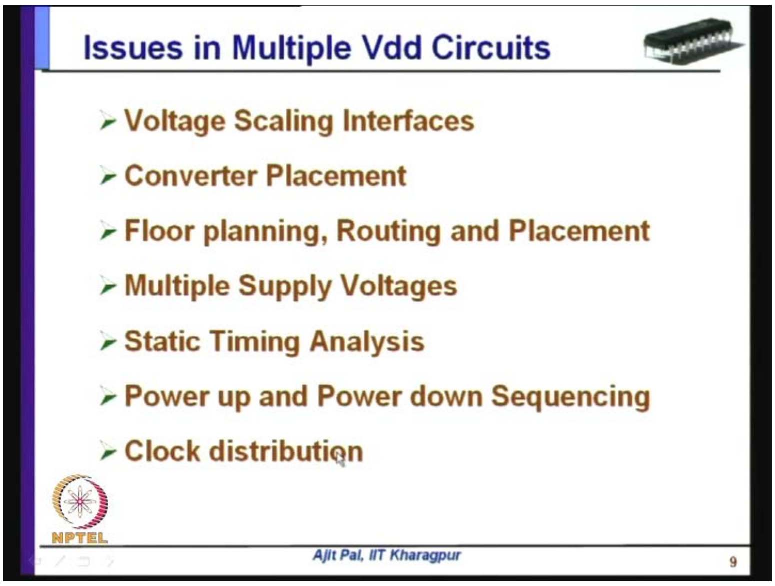
- Basic Concept: High Vdd gates have less delay, but higher dynamic and static power
- Voltage islands can be generated at different levels of granularity, such as macro level and standard cell level
- Total power consumption can be reduced without degrading the overall circuit performance

Ajit Pal, IIT Kharagpur

8

Then comes to multilevel voltage scaling, and here again the basic concept is high V dd gates have less delay, but higher dynamic and static power, what can be done? Voltage islands can be generated at different levels of granularity such as macro level or standard cell level. So, what can be done? We can create islands of circuits which will have different levels of granularity may be macro level or standard cell level, and each of these islands can operate at different voltages and at different voltages at different instances of time; I mean different voltages that can be switched if necessary whenever the workload requirement is different, and total power consumption can be reduced without degrading the overall circuit performance.

(Refer Slide Time: 24:35)



Issues in Multiple Vdd Circuits

- Voltage Scaling Interfaces
- Converter Placement
- Floor planning, Routing and Placement
- Multiple Supply Voltages
- Static Timing Analysis
- Power up and Power down Sequencing
- Clock distribution

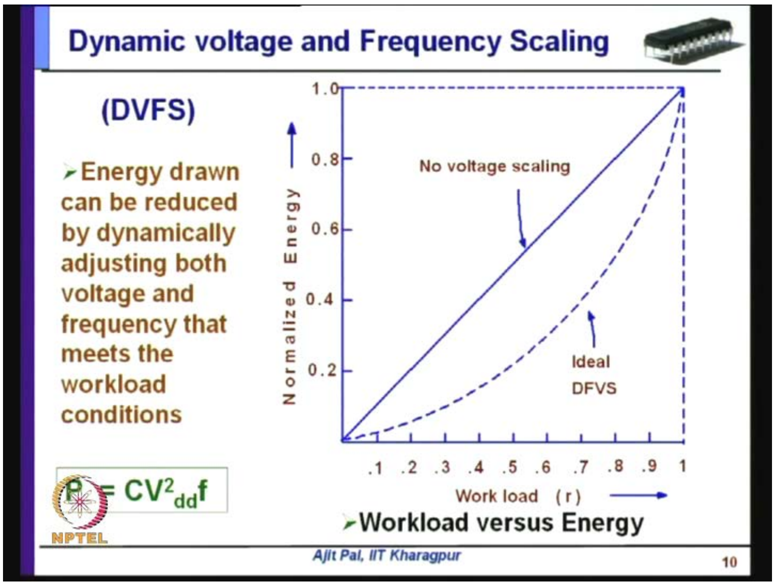
NPTEL
Ajit Pal, IIT Kharagpur
9

And whenever we go for multiple V dd circuits, there are a number of issues to be considered and we have discussed all these issues in detail one after the other. Number one is voltage scaling interfaces; that means, whenever we are using say two such macros or modules and signal is going from one to another in such a case, we have to take appropriate action such that this output does not affect the performance; I mean operation of the circuit this way or the other way. So, sometimes we have to put additional hardware in the form of level converter. So, level converter needs to be used and converter placement is one of the important issues whenever we go for multiple V dd Circuits. The so interfacing in the interface part, we have to consider this converter placement then floor planning routing and placement because different parts of the circuits will be operating in different voltages. S

o they have to be placed properly; we have to make suitable floor planning; you have to do the routing not only of the signals, but also the supply voltage and you have to place them properly then, you have to also switch multiple supply voltages say take different supply voltages and also along with the switches; for example, if this module is operating in multiple voltage switching then, you will require two voltages say 1.1 volt or 0.8 volt and you have to switch these voltage and obviously, this will require additional complexity, because you have to place the switch; you have to also multiple supply voltages are take to be taken to different parts of this hardware, and you have to and in this case, the static timing analysis becomes very complex. So, when the entire circuit is operating at a single voltage, the starting timing analysis that is that is; that means, the delay of the circuit can be very easily modeled and computed. However, whenever we are using multiple voltage levels and that too is changing with time then, static timing analysis becomes very difficult, and this static timing analysis is very important for reliable operation of the circuit because what will be the delay for different supply voltage that has to be identified.

Then, power up and power down sequencing is also very important because the order in which powering up and powering down occurs because sometimes as I have mentioned you will be switching from say 0.8 to 1.1 or you will be switching down from 1.1 to 0.8. So, the how we will do that, in which order we will do and whenever it it is a complex circuit, we have got multiple voltage domains. This powering and powering down sequencing becomes a very important issue and that has to be the sequencing has to be done properly for reliable and correct operation of the circuit. Then, comes the question of power up distribution. You have seen that the circuits are operating at different voltages and frequencies so the clock has to be distributed and if multiple clocks I mean, the different frequencies have to be I mean, applied that also has to be distributed properly. So, these are the different issues to be considered whenever we go for a multiple V dd circuits.

(Refer Slide Time: 28:46)



Dynamic voltage and Frequency Scaling

(DVFS)

- Energy drawn can be reduced by dynamically adjusting both voltage and frequency that meets the workload conditions

$P = CV^2_{dd}f$

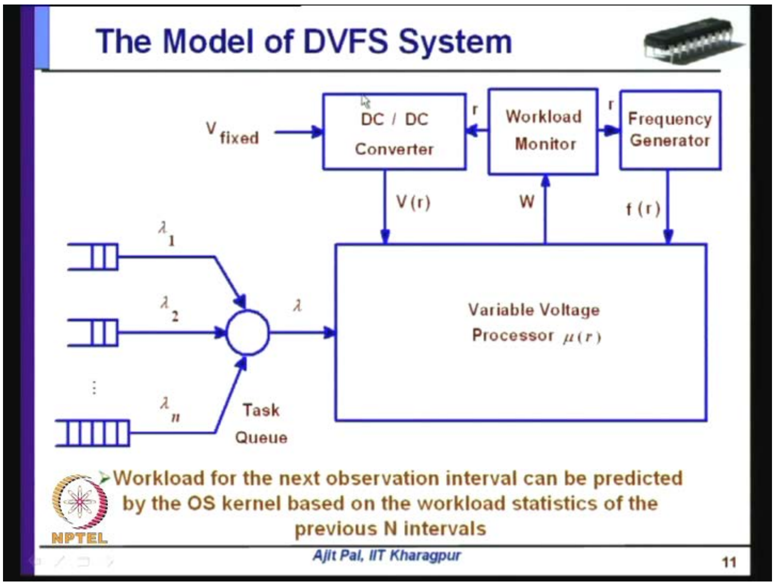
- Workload versus Energy

Ajit Pal, IIT Kharagpur

10

Then another voltage scaling technique is your dynamic voltage and frequency scaling which as I have already told as the workload changes as you can see. Normally, if we do not do voltage scaling only the frequency changes then, linearly the energy reduces. However, whenever we go for reduction of voltage and frequency both then this is how the I mean, this will be the ideal dynamic voltage and frequency scaling curve and as we can see there is significant reduction compared to no voltage scaling; that means, only frequency is reduced for different workload in this. In this particular case, but here both voltage and frequency are reduced leading to significant reduction in power dissipation. So in this particular case, energy drawn can be reduced by dynamically adjusting both voltage and frequency that meets the workload condition, and we have already discussed about this formula that power dissipation P is equal to CV dd square f and I mean, this this this part gives you the idea about how the power dissipation is reduced or energy normalized energy reduction take place.

(Refer Slide Time: 30:12)



The Model of DVFS System

➢Workload for the next observation interval can be predicted by the OS kernel based on the workload statistics of the previous N intervals

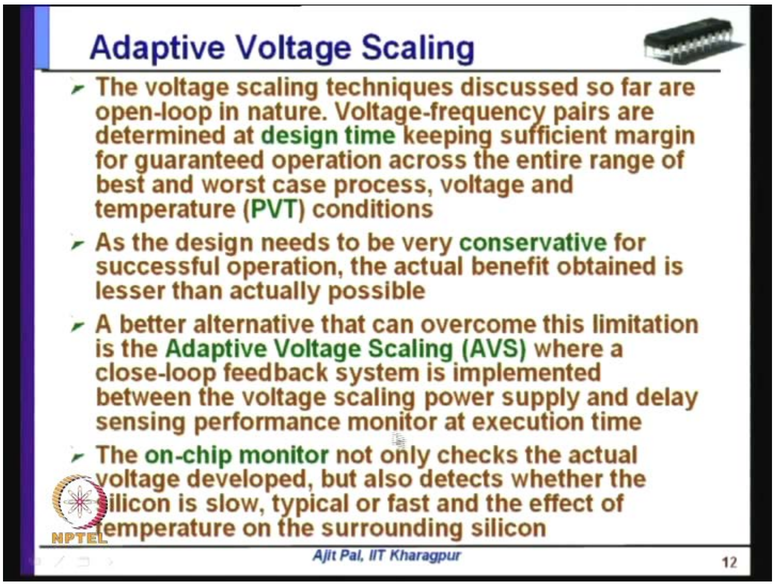
Ajit Pal, IIT Kharagpur

11

And, this is the model of the dynamic voltage and frequency scaling system. As we can see here, this is the processor, variable voltage processor which can operate at different voltage and frequencies for the processor has been designed that way transmitters, Crusoe processor or many other modern processors have been designed which can operate at different voltages and frequencies and here you can see, there is a workload monitor. So the workload monitor based on the workload the it controls the supply voltage; that means, the DC to DC voltage converter is there which is receives a fix voltage and generates different voltages. Similarly, you have got a frequency generator that controls the frequency. So, depending on the workload voltage and frequency can be controlled and you can see, the inputs are coming from different sources and there is a task queue and operating system seduce the tasks to the processor and accordingly the workload changes.

So the based on that workload the voltage and frequency are controlled. So, workload for the next observation interval can be predicted by the operating system kernel based on workload statistics of the previous n intervals. So, it works in this way; that means, based on the previous history, it identifies it predicts the future workload and based on that on voltage and frequency are identified and that controls the circuit. So as it is shown in this particular model.

(Refer Slide Time: 31:59)



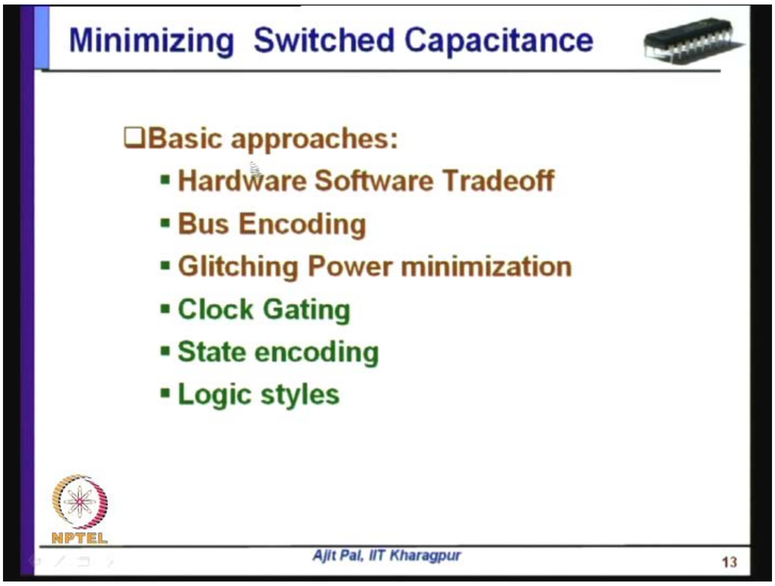
Adaptive Voltage Scaling

- The voltage scaling techniques discussed so far are open-loop in nature. Voltage-frequency pairs are determined at design time keeping sufficient margin for guaranteed operation across the entire range of best and worst case process, voltage and temperature (PVT) conditions
- As the design needs to be very conservative for successful operation, the actual benefit obtained is lesser than actually possible
- A better alternative that can overcome this limitation is the Adaptive Voltage Scaling (AVS) where a close-loop feedback system is implemented between the voltage scaling power supply and delay sensing performance monitor at execution time
- The on-chip monitor not only checks the actual voltage developed, but also detects whether the silicon is slow, typical or fast and the effect of temperature on the surrounding silicon

Ajit Pal, IIT Kharagpur 12

Now, coming to the adaptive voltage scaling as I have told, the voltage scaling techniques discussed so far; that means, the static voltage scaling then your multilevel voltage scaling or dynamic voltage scaling; here, those were open loop in nature; that means, voltage frequency pairs are determined at design time keeping sufficient margin for guaranteed operation across the entire range of best and worst case process voltage and temperature; that means, best and worst variations of you know PVT process voltage and temperature it should operate and as a consequence, the design is very conservative and whenever the design is very conservative then, we do not get I mean I mean, the reduction that we get is not very high. So, as the design needs to be conservative for successful operation the actual benefit obtained is lesser than actually possible. So, this is what happens in in in those conventional designs; that is the reason why in adaptive voltage scaling, a close loop feedback system is implemented where the voltage scaling power supply and delay sensing performance are monitored at execution time. So at execution time, the delay is sensed; what is the delay for a given applied voltage and accordingly I mean, this on-chip monitor not only checks the actual voltage developed, but also detects whether the silicon is slow, typical or fast and the affect the effect of temperature on the surrounding silicon. So, at runtime this process voltage and temperature conditions I mean in that condition, it does the monitoring and it can identify; it it it controls the voltage and frequency to optimize the reduction in power dissipation and that is the reason why, among the supply voltage scaling techniques; this adaptive voltage scaling gives you the best reduction in energy. However, it is much more complex and you require on-chip monitor some extra hardware within the chip.

(Refer Slide Time: 34:28)



Minimizing Switched Capacitance

❑Basic approaches:

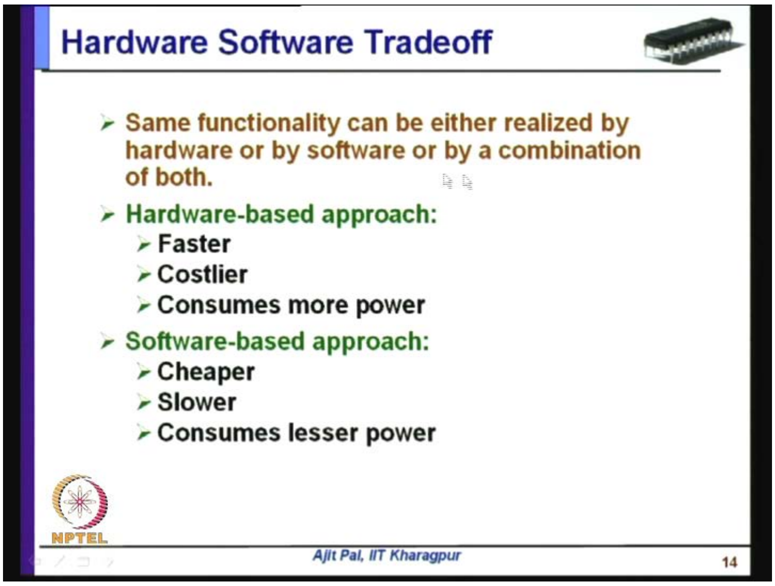
- Hardware Software Tradeoff
- Bus Encoding
- Glitching Power minimization
- Clock Gating
- State encoding
- Logic styles

Ajit Pal, IIT Kharagpur

13

Now, coming to minimizing switched capacitance, we have several approaches and these approaches are mentioned is listed here; first one that we discussed is hardware software tradeoff. As we know same functionality can be realized either by using hardware or by purely by software or by judicious combination of hardware and software. Now, which combination gives you best reduction in power dissipation; that is what is being done in software hardware trade-off.

(Refer Slide Time: 35:14)



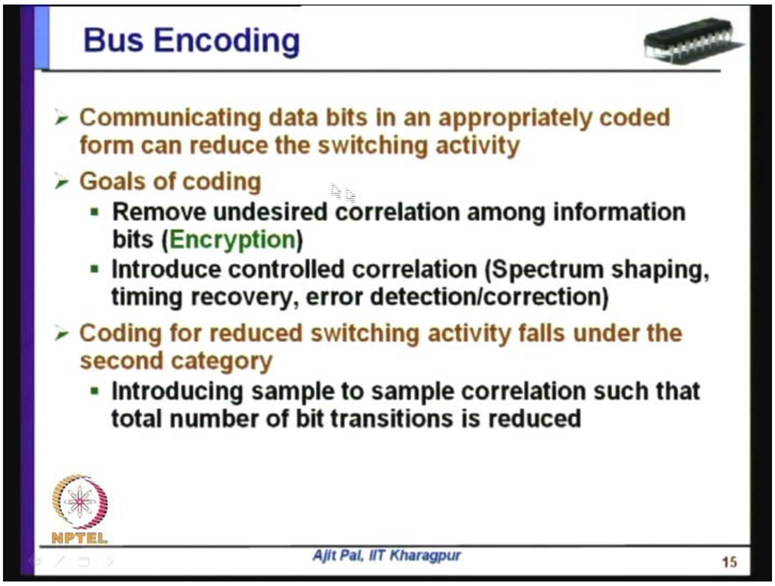
Hardware Software Tradeoff

- Same functionality can be either realized by hardware or by software or by a combination of both.
- Hardware-based approach:
  - Faster
  - Costlier
  - Consumes more power
- Software-based approach:
  - Cheaper
  - Slower
  - Consumes lesser power

Ajit Pal, IIT Kharagpur 14

So same functionality can be either realized by hardware or by software or by combination of both and as we know, hardware based approach is faster, costlier and consumes more power. On the other hand, software based approach is cheaper, slower and consumes lesser power. So, you have to a partition; you have to identify which part you will realize by hardware and which part you will realize by software and then, you know you get best reduction in power dissipation. So, that is the hardware software trade-off.

(Refer Slide Time: 35:59)



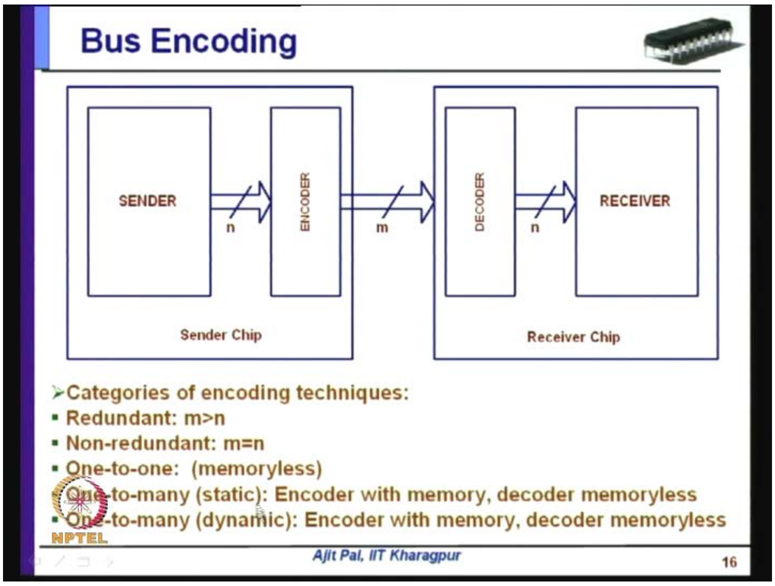
Bus Encoding

- Communicating data bits in an appropriately coded form can reduce the switching activity
- Goals of coding
  - Remove undesired correlation among information bits (Encryption)
  - Introduce controlled correlation (Spectrum shaping, timing recovery, error detection/correction)
- Coding for reduced switching activity falls under the second category
  - Introducing sample to sample correlation such that total number of bit transitions is reduced

Ajit Pal, IIT Kharagpur 15

Then, second technique that is used for reducing I mean, minimizing switched capacitance is bus encoding. We know that whenever we have got we have got say more than one chip or processor or may be within a single chip you have got multiple cores or multiple subsistence; they are communicating with each other. Now, communicating data bits in and appropriately coded form can reduce switching activity. Now, the as the signal is sent through the bus communicated over the bus, you can do the coding and there can be two I mean, can be done for two basic purposes. One purpose is remove undesired correlation among information bits as we do in case of encryption. So, this is this is one goal. Another goal is introduce controlled correlation by spectrum shaping timing recovery I mean, which are used in spectrum shaping, timing recovery and error detection, correction where instead of removing undesired correlation, we are introducing correlation and this bus encoding is a technique which belongs to this category, and which we have discussed in detail encoding for reduced switching activity falls under the second category which is which is being done, and we by introducing sample to sample correlation such that the total number of bit transitions is reduced. So, our objective is to reduce switching activity by suitable bus encoding.

(Refer Slide Time: 37:43)



Bus Encoding

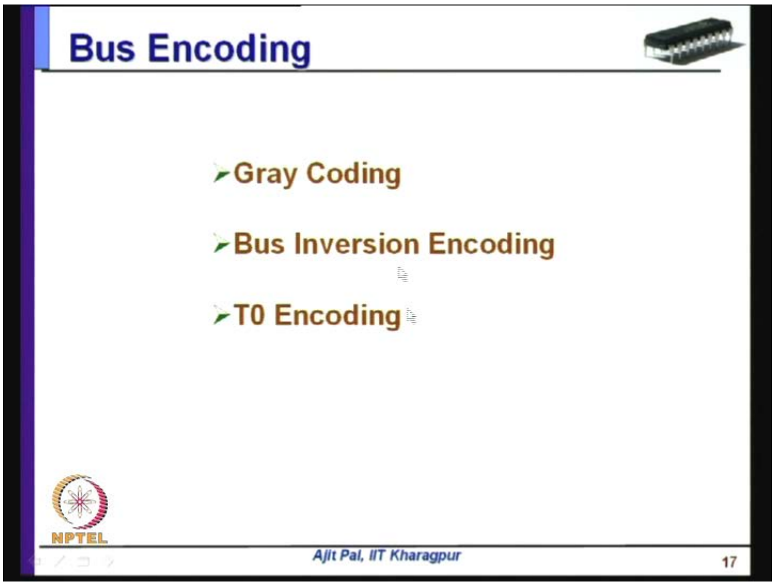
➢Categories of encoding techniques:

- Redundant: $m>n$
- Non-redundant: $m=n$
- One-to-one: (memoryless)
- One-to-many (static): Encoder with memory, decoder memoryless
- One-to-many (dynamic): Encoder with memory, decoder memoryless

Ajit Pal, IIT Kharagpur 16

And there are various bus encoding techniques and this is the basic scheme which is shown here. So, you have got a sender which does the encoding. So, n bit is converted in to m bit and there is a decoder at the other end which again converts m bit to n bit and here is the receiver.

(Refer Slide Time: 38:23)



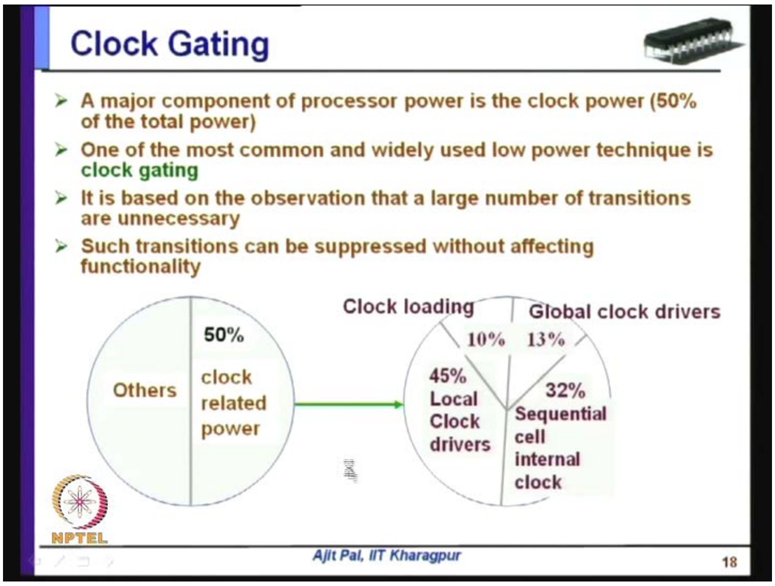
Bus Encoding

- Gray Coding
- Bus Inversion Encoding
- T0 Encoding

Ajit Pal, IIT Kharagpur 17

So, it can it can they can be categorized into redundant, non-redundant, on-to-one, one-to-many and one-to I mean, where your encoder with memory and decoder memory less. On the other hand, one-to-many can be encoder with memory and decoder memory less memory less. There are several techniques like gray coding, which I have discussed in detail and then bus inversion encoding and T0 encoding.

(Refer Slide Time: 39:04)



Clock Gating

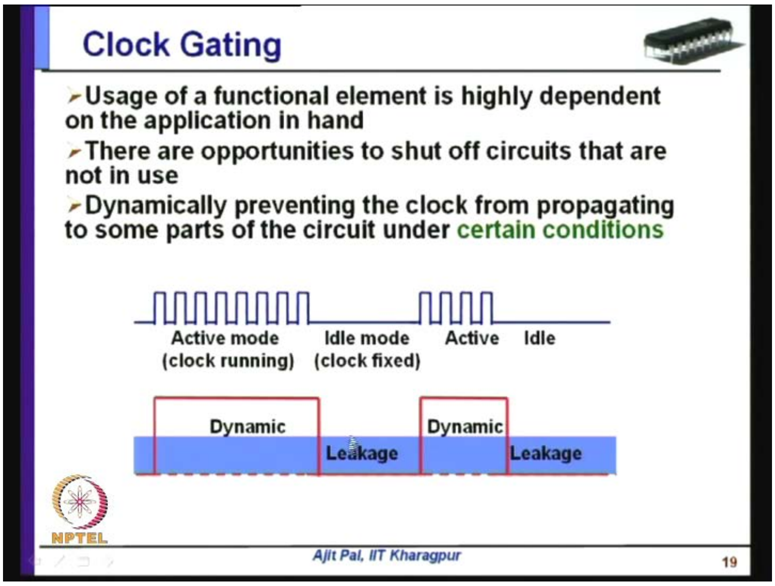
- A major component of processor power is the clock power (50% of the total power)
- One of the most common and widely used low power technique is clock gating
- It is based on the observation that a large number of transitions are unnecessary
- Such transitions can be suppressed without affecting functionality

Ajit Pal, IIT Kharagpur 18

So in gray coding we have seen, it is it is a one-to-one coding and it does not require any memory. On the other hand, bus inversion coding is again one-to-one coding, but it requires memory in the circuit and then T0 encoding, zero transition zero transition encoding there also memory is required and these three encoding techniques, we have discussed in detail and which can be used for bus encoding.

So apart from bus encoding, we can use another technique to reduce the power dissipation is clock gating. So, this clock major component of processor power is the clock power. You know clock has to be distributed throughout the chip, and clock will control the transitions of different gates. So, clock gating can be done to reduce the dynamic power and one of the most common and widely used low power technique is clock gating. So, it is based on the observation that a large number of transitions are not necessary. So, such transitions can be suppressed without affecting the functionality. So here, you know many transitions are taking place within the circuit because you have applied the clock to it and those transitions are unnecessary and when the transitions are occurring power dissipation is occurring.

(Refer Slide Time: 40:19)



Clock Gating

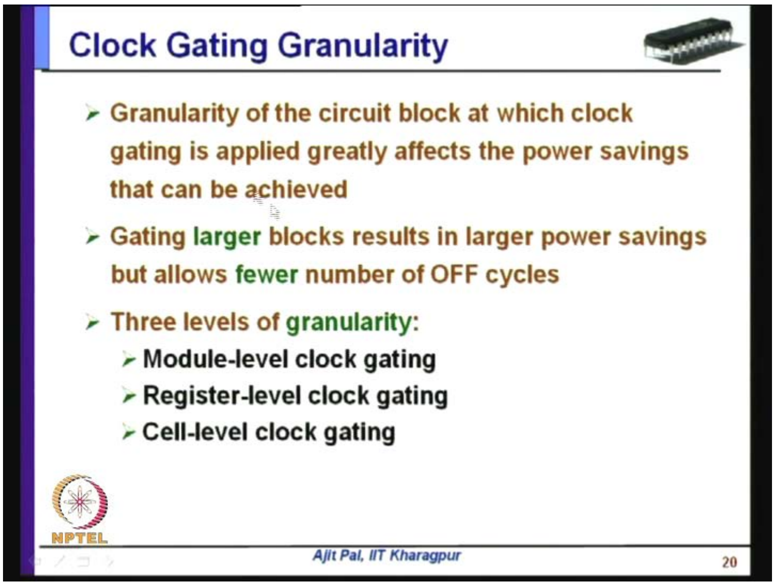
- Usage of a functional element is highly dependent on the application in hand
- There are opportunities to shut off circuits that are not in use
- Dynamically preventing the clock from propagating to some parts of the circuit under certain conditions

Ajit Pal, IIT Kharagpur

19

So, if transitions can be suppressed by clock gating, there will be reduction in dynamic power dissipation; that is the basic idea of clock gating and which we have which we have discussed in detail, and here is the basic idea. Use of a functional element is highly dependent on the application in hand so there are opportunities to shut off circuits that are not in use. So, when it is not in used clock can be gated. So dynamically, preventing the clock from propagating to some parts of the circuit under certain conditions; so, you have to identify the conditions and when the conditions are satisfied then, the clock is gated. As you can see, when clock is not gated then, you have got dynamic power dissipation as well as leakage power dissipation. However, when clock is gated you have got only leakage power dissipation not the dynamic power dissipation. So therefore, there is there is significant reduction in power dissipation that will take place whenever clock gating is done and we have discussed, how you can really achieve clock gating in different parts of the circuit.

(Refer Slide Time: 41:19)



Clock Gating Granularity

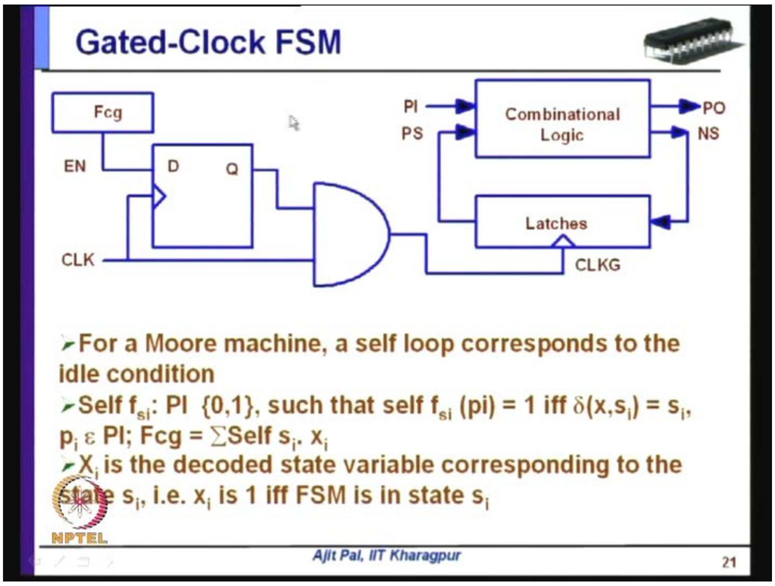
- Granularity of the circuit block at which clock gating is applied greatly affects the power savings that can be achieved
- Gating larger blocks results in larger power savings but allows fewer number of OFF cycles
- Three levels of granularity:
  - Module-level clock gating
  - Register-level clock gating
  - Cell-level clock gating

Ajit Pal, IIT Kharagpur

20

And the clock gating can be done with different levels of granularity of the circuit block at which clock gating is applied greatly affects the power savings that can be achieved. So, if the if the gating I mean, if the if you get larger block then, there is a larger power saving that will take place, but you will get fewer opportunities to do so. So, that is the reason why there is three levels of granularity; module-level clock gating, register-level clock gating and cell-level clock gating. So, whenever you go for module-level clock gating; so for whenever you are able to do the gating, there is a significant saving, but this will this cannot take place very frequently. On the other hand, whenever we go for cell-level clock gating this gating can be done more frequently. However, whenever you do the gating saving in power is not much small saving in power. So, depending on the application and depending on the nature of the circuit, you can go for these three types of clock gating.

(Refer Slide Time: 42:32)



Gated-Clock FSM

- For a Moore machine, a self loop corresponds to the idle condition
- Self f_{si}: PI {0,1}, such that self f_{si} (pi) = 1 iff $\delta(x,s_i) = s_i$, $p_i \varepsilon$ PI; Fcg = $\sum$Self s_i. x_i
- X_i is the decoded state variable corresponding to the state s_i, i.e. x_i is 1 iff FSM is in state s_i

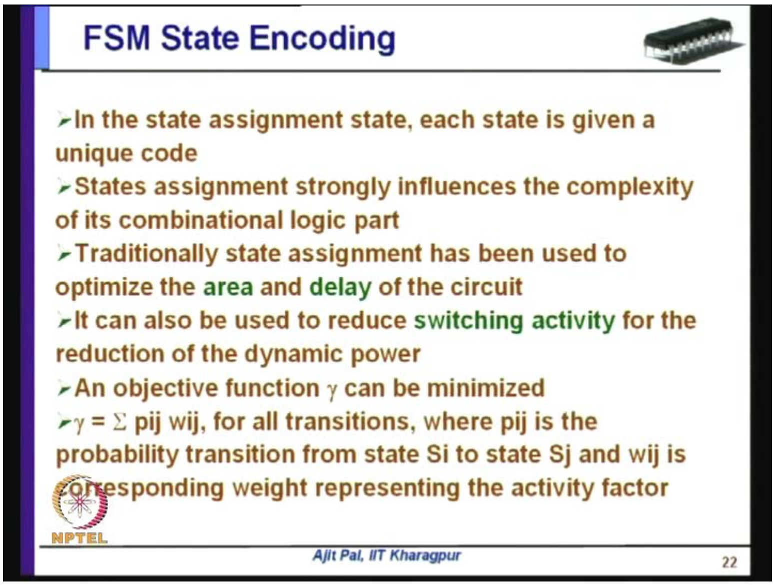
Ajit Pal, IIT Kharagpur

21

Not only clock gating can be done in combinational circuits, the finite state machines can also be gated; so can be clock gated. So here, this is known as gated-clock finite state machine. So here, as you can see normally this is the schematic diagram of your finite state machine and directly clock is applied here and the this the sequential circuits operates finite state machine operates. However, what has been done here; here you have got a circuit Fcg which identifies when clock can be gated for this finite state machine and then it does not allow this this this clock to be applied to this circuit whenever gating is I mean transitions are not necessary.

So for Moore machine, a self-loop corresponds to the idle condition idle condition. So in this idle condition clock can be stopped; that means, in this case you know the output will not change for the input combination state will change. So, those conditions can be identified by Fcg and accordingly the clock gating is being done in gated clock finite state machine, and I have discussed it in detail in this lecture series, how this how finite state machine can be clock gated.

(Refer Slide Time: 44:05)



FSM State Encoding

- In the state assignment state, each state is given a unique code
- States assignment strongly influences the complexity of its combinational logic part
- Traditionally state assignment has been used to optimize the area and delay of the circuit
- It can also be used to reduce switching activity for the reduction of the dynamic power
- An objective function γ can be minimized
- $\gamma = \Sigma$ pij wij, for all transitions, where pij is the probability transition from state Si to state Sj and wij is corresponding weight representing the activity factor

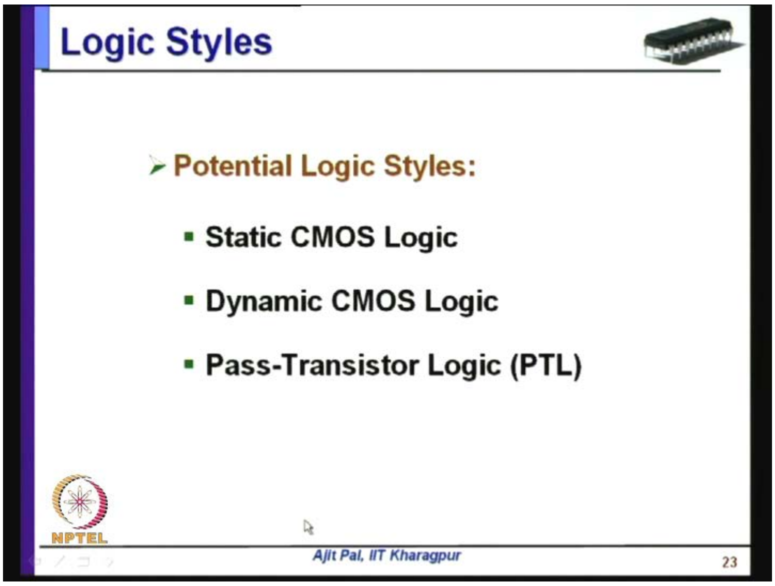
Ajit Pal, IIT Kharagpur

22

Apart from you know clock gating the finite state machine, you can do finite state encoding. As you know, whenever you realize a finite state machine which is which has got different states and each of the states are to be encoded. So, state encoding is usually is done for to reduce the area of the circuit, convention traditional that is what is being done, but you know the that whenever you do the state assignment, state assignment strongly influences the complexity of the combinational logic part of the circuits which affects the area and delay. However, now what you are trying to do? The state encoding is being done which will reduce the switching activity for the reduction of the dynamic power.

So, an objective function gamma is minimized where gamma is equal to summation of pij into wij for all transitions, where pij is the probability transition from state Si to state Sj and wij is corresponds to weight corresponding to the activity factor. Here, the basic idea is as we go from one state to another state, there is a transition I mean, the there is a bit transition. Now, the we have to minimize the bit transitions such that when the finite state machine is operating the power dissipation is reduced. So, that is what is being represented by the subjective function and state encoding can be done to reduce the switching activity apart from I mean area and delay.

(Refer Slide Time: 45:54)



Logic Styles

- Potential Logic Styles:
  - Static CMOS Logic
  - Dynamic CMOS Logic
  - Pass-Transistor Logic (PTL)

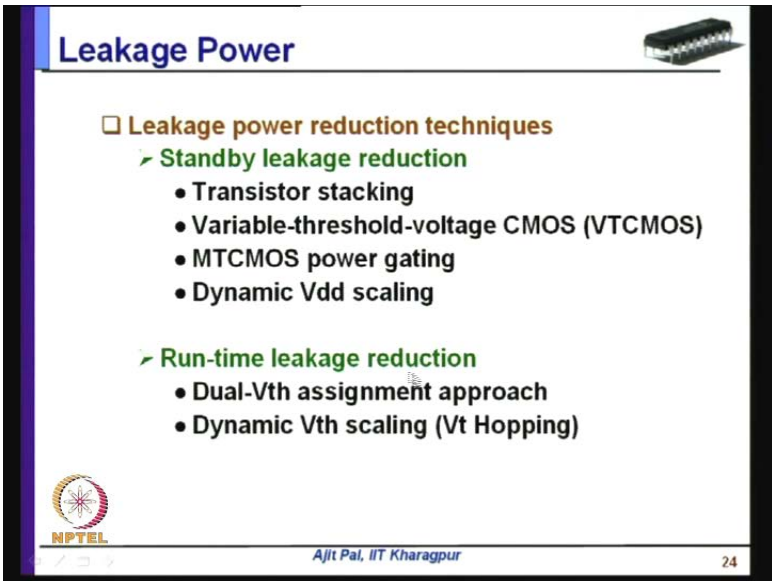
Ajit Pal, IIT Kharagpur

23

Now, we have discussed about different types of logic styles that can be used to realize CMOS circuits. So essentially, you know as I have already mentioned, we our intercourse is oriented towards CMOS circuit which is the technology of the day and you can use I mean, different types of CMOS circuits; one is your static CMOS circuit which is conventionally used then, another is dynamic CMOS circuit; another is pass-transistor logic circuits. So although, static CMOS circuit is a circuit logic is the most popular, but they require I mean, much higher logic circuit; I mean, the number of transistors is much larger in static CMOS circuits and the number of transistors can be reduced in dynamic CMOS circuits and also in pass-transistor logic circuits.

So, instead of realizing circuits by using static CMOS circuits if other alternatives are used then, there is a possibility of reducing the power dissipation and save energy in realizing the same functionality. So, or another alternative is you can make a judicious combination of static CMOS circuit and dynamic CMOS circuit or static CMOS circuit and pass-transistor logic circuit to realize different functionalities leading to overall reduction in power dissipation. So logic styles is one of the important area.

(Refer Slide Time: 47:37)



Leakage Power

- ❑ Leakage power reduction techniques
  - ➢ Standby leakage reduction
    - Transistor stacking
    - Variable-threshold-voltage CMOS (VTCMOS)
    - MTCMOS power gating
    - Dynamic Vdd scaling
  - ➢ Run-time leakage reduction
    - Dual-Vth assignment approach
    - Dynamic Vth scaling (Vt Hopping)

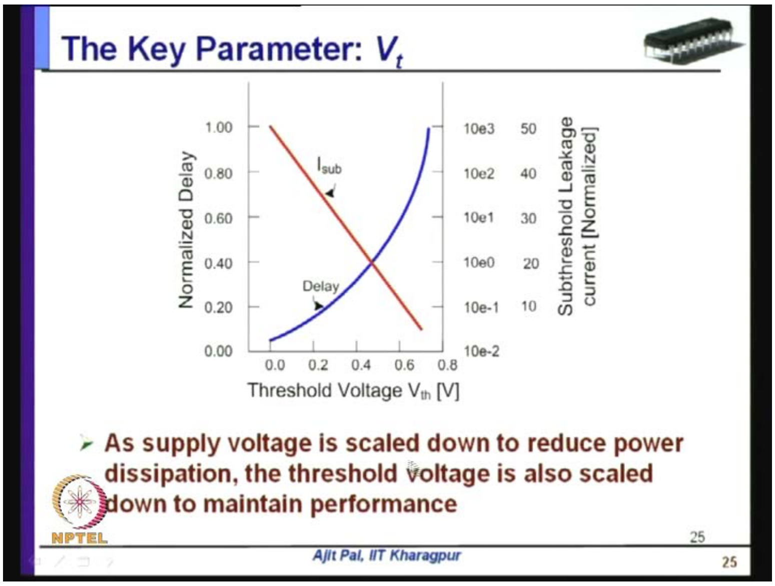
Ajit Pal, IIT Kharagpur

24

Now, coming to the leakage power we have so far we have discussed techniques about the reduction of dynamic power. Now, the leakage power reduction techniques can be broadly divided into two categories. Number one is standby leakage reduction techniques and runtime leakage reduction techniques. So, under the standby leakage reduction techniques, we have discussed various approaches like transistor stacking, variable-threshold-voltage CMOS, VTCMOS then multi-threshold-voltage CMOS power gating and dynamic Vdd scaling. So, these are coming under standby leakage reduction techniques then, you can go for runtime leakage reduction techniques; that means, standby reduction techniques occur can be used when the circuit is in the standby condition; that means, the circuit is not in use. On the other hand, this runtime leakage reduction techniques can be used even when the circuit is in operation.

Earlier, since the dynamic power was a significant component much larger compared to the static power standby; I mean, the leakage power reduction was mainly focused for I mean for standby reduction; that means, these techniques were used standby leakage reduction. However, nowadays in deep submicron technology, the leakage component is comparable has become comparable to dynamic power when the circuit is in operation. So, runtime leakage reduction has become important in dynamic in the deep submicron technology or present day technology. So and for that purpose you can use different techniques like dynamic Vth assignment approach or dynamic Vth scaling, which is also known as Vth hoping. So, these techniques can be used.

(Refer Slide Time: 49:46)



The Key Parameter: V_t

- As supply voltage is scaled down to reduce power dissipation, the threshold voltage is also scaled down to maintain performance

NPTEL

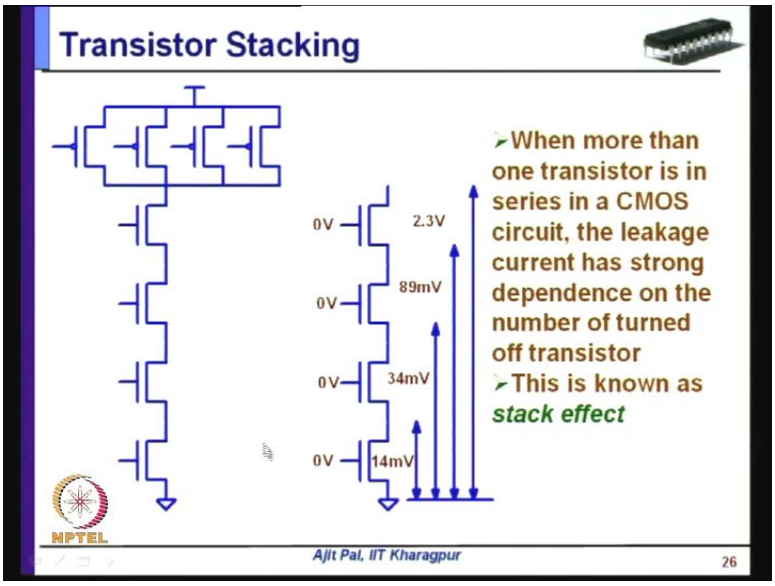
25

Ajit Pal, IIT Kharagpur

25

Let me very quickly give an overview of these techniques. As I mentioned the key parameter that is used for leakage power reduction is the threshold voltage V t and here as you can see as the supply voltage scaling, voltage scaled down to reduce power dissipation; the threshold voltage is also scaled down to maintain performance. So that means, the threshold voltage has strong dependence on the leakage power as well as on the delay. So, we can see that normalized that threshold voltage if it is reduced delay reduces, but the sub-threshold leakage the sub-threshold leakage increases. So, we have to optimize the realize a design in such a way that threshold voltage is properly adjusted or controlled such that the we get overall reduction in leakage power.

(Refer Slide Time: 50:43)



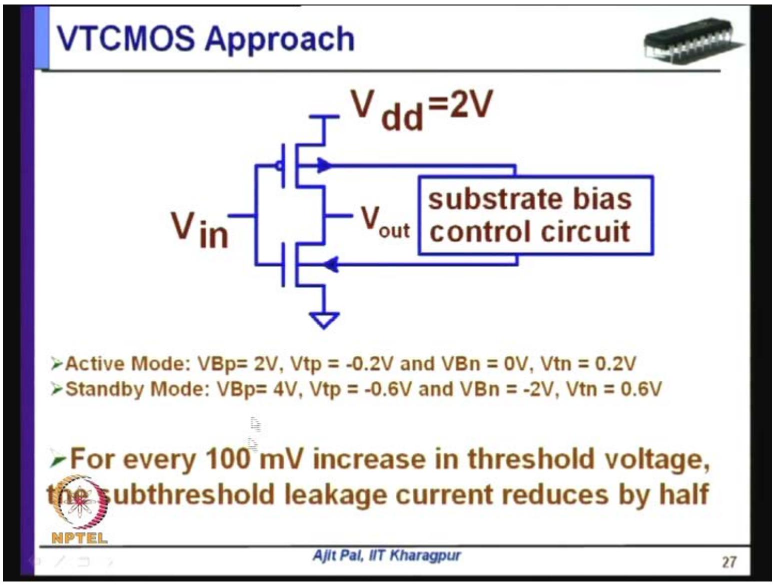
Transistor Stacking

- When more than one transistor is in series in a CMOS circuit, the leakage current has strong dependence on the number of turned off transistor
- This is known as *stack effect*

Ajit Pal, IIT Kharagpur 26

As I told this is one of the standby leakage power reduction techniques that is your transistor stacking. So here, it has been found that… So, this is a four input nand gates static CMOS nand gate; here, it had been found that if the number of transistors that is in series if I mean, if we have got more number of transistors of which are is in series or which are in stack then, the then the leakage power is reduced which is known as stack effect. So, this is known as transistor stacking and this we can use the stack effect to reduce the power dissipation in this way.

(Refer Slide Time: 51:34)



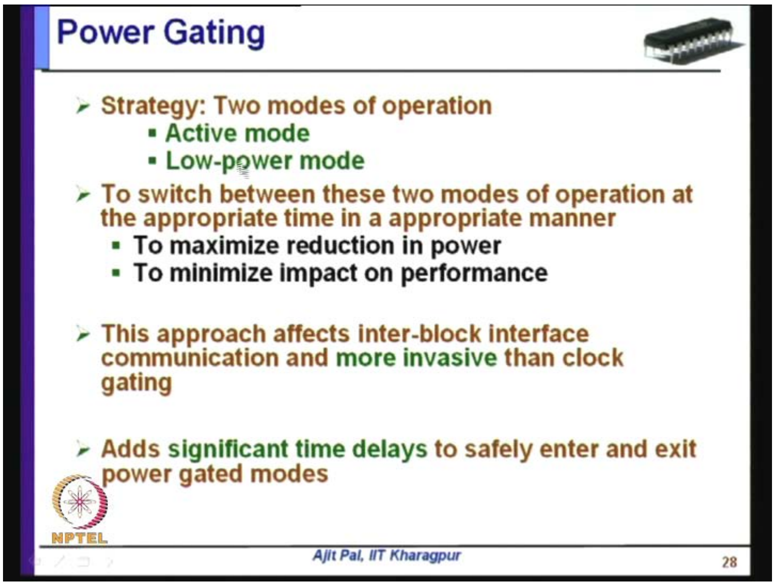
VTCMOS Approach

- Active Mode: VBp= 2V, Vtp = -0.2V and VBn = 0V, Vtn = 0.2V
- Standby Mode: VBp= 4V, Vtp = -0.6V and VBn = -2V, Vtn = 0.6V
- For every 100 mV increase in threshold voltage, the subthreshold leakage current reduces by half

Ajit Pal, IIT Kharagpur 27

Then, we can use the V t CMOS approach where as I have already mentioned, we control the threshold voltage dynamically; I mean, when the circuit is in operation, threshold voltage is lower by adjusting the body bias, and when the circuit is in standby, the threshold voltage is made higher by adjusting the body bias and as you can see for every hundred mill volt increase in threshold voltage, the sub threshold leakage current reduces by half. So, V t CMOS approach is used in this way by for reducing the leakage power by adjusting the body bias to control the threshold voltage inactive and standby mode.

(Refer Slide Time: 52:24)



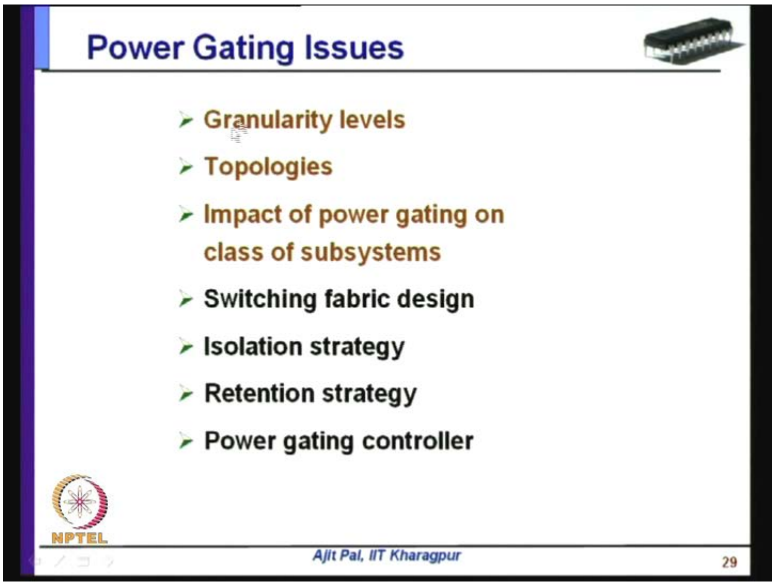
Power Gating

- Strategy: Two modes of operation
  - Active mode
  - Low-power mode
- To switch between these two modes of operation at the appropriate time in a appropriate manner
  - To maximize reduction in power
  - To minimize impact on performance
- This approach affects inter-block interface communication and more invasive than clock gating
- Adds significant time delays to safely enter and exit power gated modes

Ajit Pal, IIT Kharagpur 28

And, power gating can be done which is known as m t CMOS; again you have two modes active mode and low power mode and in active mode, your performance is greater and in low power mode, the there is a significant reduction in power dissipation and this approach power gating approach is more invasive than clock gating because that inter-block interface communication, it affects the inter-block interface communication; it is more invasive and it adds significant time delays to safely enter and exit power gated modes.

(Refer Slide Time: 53:04)



Power Gating Issues

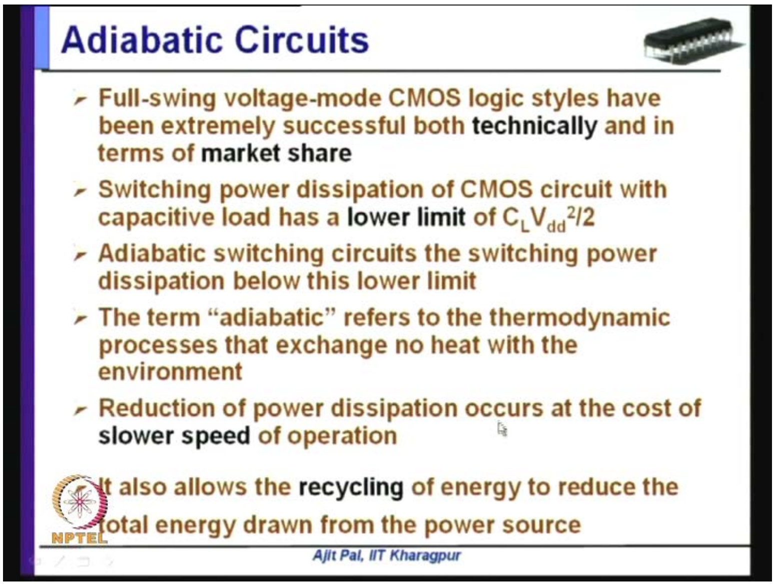
- Granularity levels
- Topologies
- Impact of power gating on class of subsystems
- Switching fabric design
- Isolation strategy
- Retention strategy
- Power gating controller

Ajit Pal, IIT Kharagpur

29

So, power gating is very complex and it involves a number of issues like, at what granularity level it is being done; what is the topology of the circuit; what is the impact of power gating on the classes of subsystems then you have to design the switching fabric; you have to use isolation strategy; we have to use retention strategy when the circuit is in standby; you have to retain some of the states and that retention strategy has to be identified then power gating controller circuit has to be designed. So, these are the techniques which I have discussed in detail in this lecture series.

(Refer Slide Time: 53:45)



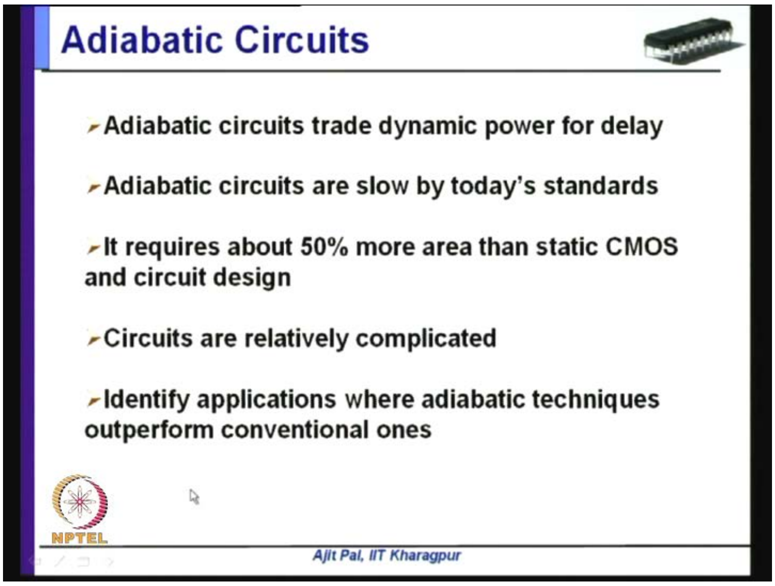
Adiabatic Circuits

- Full-swing voltage-mode CMOS logic styles have been extremely successful both technically and in terms of market share
- Switching power dissipation of CMOS circuit with capacitive load has a lower limit of $C_LV_{dd}^2/2$
- Adiabatic switching circuits the switching power dissipation below this lower limit
- The term "adiabatic" refers to the thermodynamic processes that exchange no heat with the environment
- Reduction of power dissipation occurs at the cost of slower speed of operation
- It also allows the recycling of energy to reduce the total energy drawn from the power source

Ajit Pal, IIT Kharagpur

Then, coming to the adiabatic circuits; so far we have seen that circuit is operating at full swing voltage; that means, circuit operates at full supply voltage V dd; it does not operate at any other intermediate voltage. Now, in (( )) whenever this is being done the lower limit of you know that the switching power dissipation lower limit is C L V dd square power 2 by 2 power transitions. Can we go can we still lower this power dissipation and that can be done by in adiabatic circuits. The switching power dissipation is achieved which is below this limit and actually, this the term adiabatic is referred to thermodynamic processes that exchange no heat with the environment, and this reduction of power dissipation occurs at the cost of slower speed of operation and also in adiabatic circuits, the recycling of energy can be achieved which I have discussed in detail in my in this lecture series.

(Refer Slide Time: 54:54)



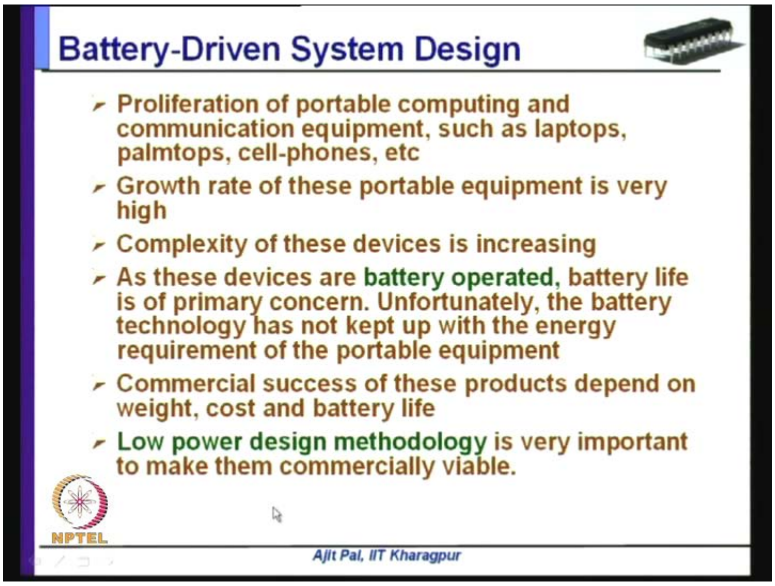
Adiabatic Circuits

- Adiabatic circuits trade dynamic power for delay
- Adiabatic circuits are slow by today's standards
- It requires about 50% more area than static CMOS and circuit design
- Circuits are relatively complicated
- Identify applications where adiabatic techniques outperform conventional ones

Ajit Pal, IIT Kharagpur

So here, as we shall see the adiabatic circuits can operate at a I mean, when the delay is large. So, adiabatic circuits are slow in today's standards; it requires about 50 percent more area than the static CMOS circuit design, and circuits are relatively complicated and; however, you can identify applications where adiabatic techniques can outperform conventional once. So, adiabatic circuits are used not very powerful on nowadays, but it may become popular in future. So, we have to identify the applications where they can be used.

(Refer Slide Time: 55:35)



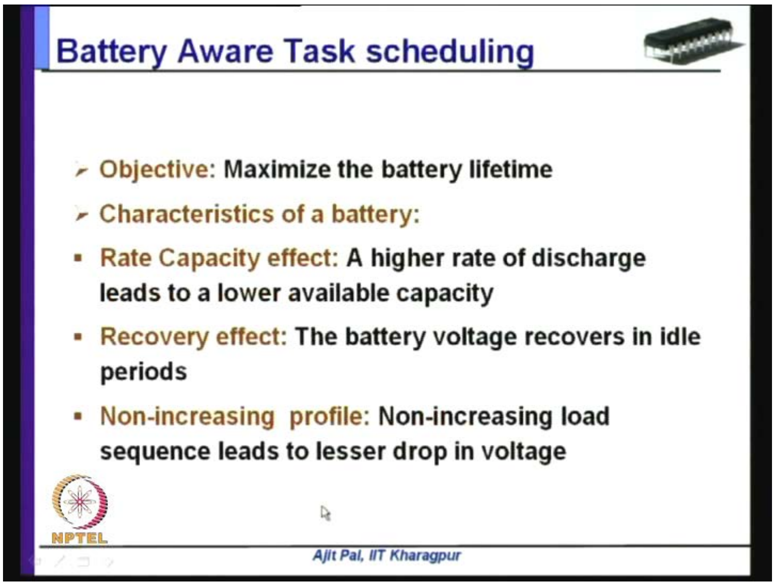
Battery-Driven System Design

- Proliferation of portable computing and communication equipment, such as laptops, palmtops, cell-phones, etc
- Growth rate of these portable equipment is very high
- Complexity of these devices is increasing
- As these devices are battery operated, battery life is of primary concern. Unfortunately, the battery technology has not kept up with the energy requirement of the portable equipment
- Commercial success of these products depend on weight, cost and battery life
- Low power design methodology is very important to make them commercially viable.

Ajit Pal, IIT Kharagpur

Coming to the last topic that is your battery-driven system design, we have as I have already mentioned, there is proliferation of portable, computing and communication equipment such as laptops, palmtop and cell phones has is taking place and they are battery operated, and growth rate of these portable equipment is very high compared to servers, desktops, work stations and complexity of these devices is also increasing. So, as these devices are battery operated, battery life is our primary concern. Unfortunately, the battery technology has not kept up with the energy requirement of the portable equipment. So, we have discussed techniques for battery-driven system design, and we can use low power design methodology to make the products battery operated systems more viable, more commercially viable and you can use different techniques like frequency scaling, supply voltage scaling, dynamic power management, battery-awake task scheduling, in battery-driven system design.

(Refer Slide Time: 56:56)



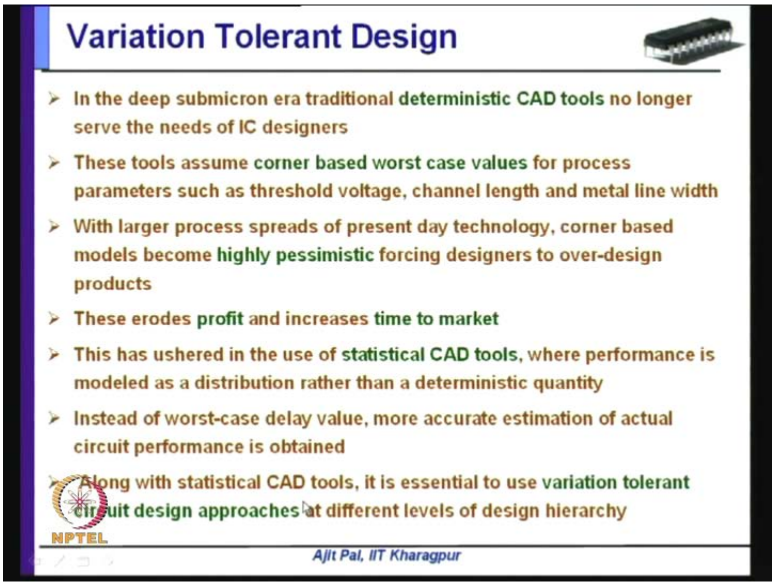
Battery Aware Task scheduling

- Objective: Maximize the battery lifetime
- Characteristics of a battery:
- Rate Capacity effect: A higher rate of discharge leads to a lower available capacity
- Recovery effect: The battery voltage recovers in idle periods
- Non-increasing profile: Non-increasing load sequence leads to lesser drop in voltage

Ajit Pal, IIT Kharagpur

And particularly, what is being done to basic objective is to maximize the battery lifetime by by by exploiting the characteristics of the battery, and these are the three important characteristics which are exploited. Number one is rate capacity effect; a higher rate of discharge leads to a lower available capacity; recovery effect the battery whenever it is kept in you know idle condition the battery charge recovers. So, battery voltage recovery occurs and non-increasing profile non-increasing load sequence leads to lesser drop in voltage. So, by exploiting the important characteristics of the batteries, we are using battery aware task scheduling to reduce power dissipation.

(Refer Slide Time: 57:48)



Variation Tolerant Design

- In the deep submicron era traditional deterministic CAD tools no longer serve the needs of IC designers
- These tools assume corner based worst case values for process parameters such as threshold voltage, channel length and metal line width
- With larger process spreads of present day technology, corner based models become highly pessimistic forcing designers to over-design products
- These erodes profit and increases time to market
- This has ushered in the use of statistical CAD tools, where performance is modeled as a distribution rather than a deterministic quantity
- Instead of worst-case delay value, more accurate estimation of actual circuit performance is obtained
- Along with statistical CAD tools, it is essential to use variation tolerant circuit design approaches at different levels of design hierarchy

Ajit Pal, IIT Kharagpur

Finally, we have discussed about variation tolerant design. Basic idea is this you know in earlier designs, the variations were not much, but in present day technology the the the important parameters like threshold voltage which varies over a range; it is you cannot consider as a fixed so, it is a statistical parameter. So, it is necessary to design, use statistical cad tool so that the circuit can be variation tolerant; that means, circuit will operate even when the threshold voltage and other important parameters vary over certain range. So, this we have discussed in detail in this lecture. So with this, we can come to we have come to the conclusion of this lecture, and I am summarized and given an overview of the entire lecture series thank you.