**Low Power VLSI Circuits and Systems**

**Prof. Ajit Pal**

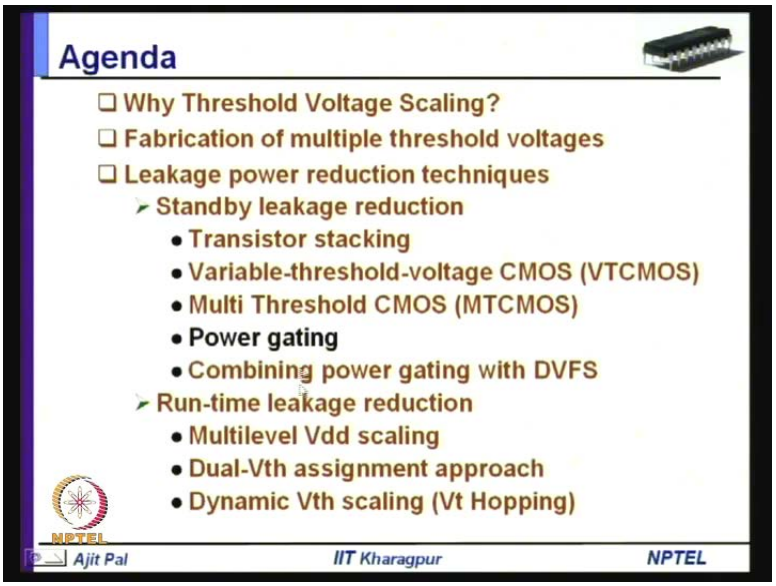
**Department of Computer Science and Engineering**

**Indian Institute of Technology, Kharagpur**

**Lecture No. # 33**

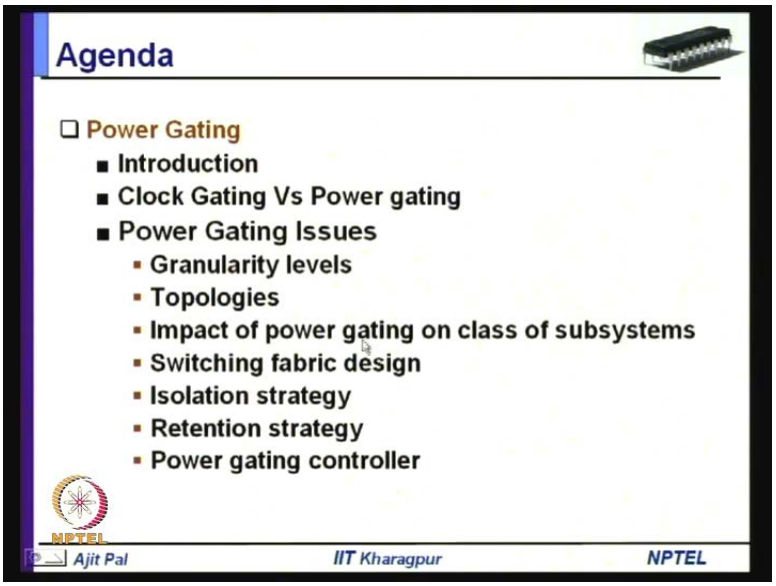
**Minimizing Leakage Power – II**

(Refer Slide Time: 00:30)



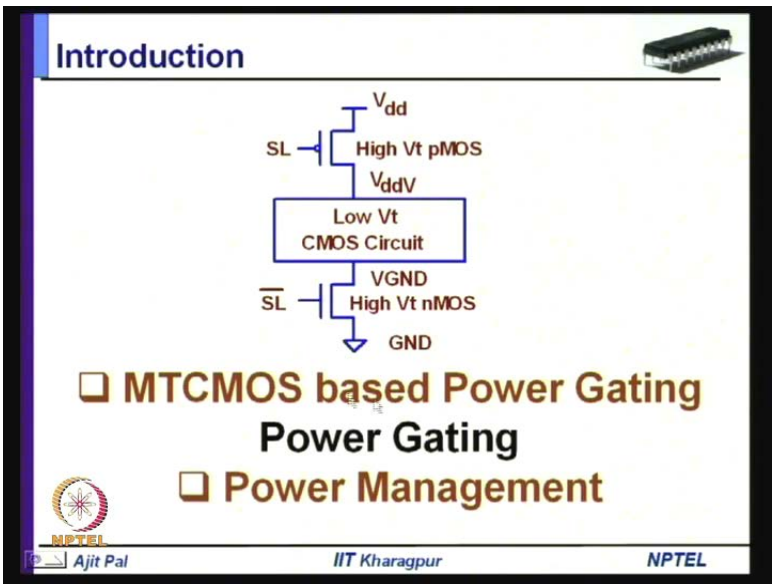

Hello and welcome to today's lecture on Minimizing Leakage Power. In the last lecture we have started our discussion on how leakage power can be minimized, and we have discussed some of the techniques by which you can minimize standby leakage power, such as transistor stacking, variable threshold voltage CMOS known as VTCMOS, and then multi threshold voltage CMOS, MTCMOS. This MTCMOS approach has now become very matured, and now it is it is also called Power Gating. Today I shall discuss about this approach in detail, and in my next lecture we should continue with other techniques for leakage power minimization.

(Refer slide Time: 01:07)




So, here is the agenda of today's lecture after a giving brief introduction about Power Gating. I shall discuss about Clock Gating versus Power Gating, and then various issues related to Power Gating such as granularity levels, topologies impact of Power Gating on different class of subsystems. Then switching fabric design, isolation strategy, retention strategy and then I shall end my lecture with Power gating controller.

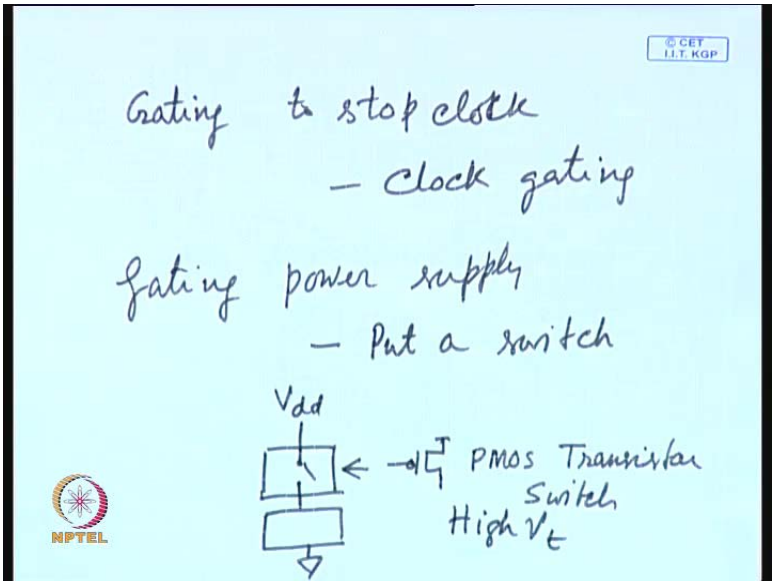
(Refer Slide Time: 01:38)




I have already discussed how the multiple threshold voltage transistor, I mean transistors of multiple threshold voltages are used to minimize leakage power, we have seen we can

add two SLEEP transistors one is this header transistor pMOS, and another is nMOS transistor which is known as footer transistors, and these are of high high higher higher threshold voltage. And the original circuit is realized by using low VDCMOS transistors. So, this is how it is being done this is also known as MTCMOS Power Gating. Why why is it called Power Gating? As you can see essentially what we are doing we are getting the supply voltage with the help of this high threshold voltage transistor.

(Refer Slide Time: 02:33)




Normally, we do we do gating to stop clock known as Clock Gating, and there we have seen the clock is not allowed to reach the circuit either resistors or combinational circuit for… So, that the switching activity is reduced, but in this case what you are doing we are getting power supply.

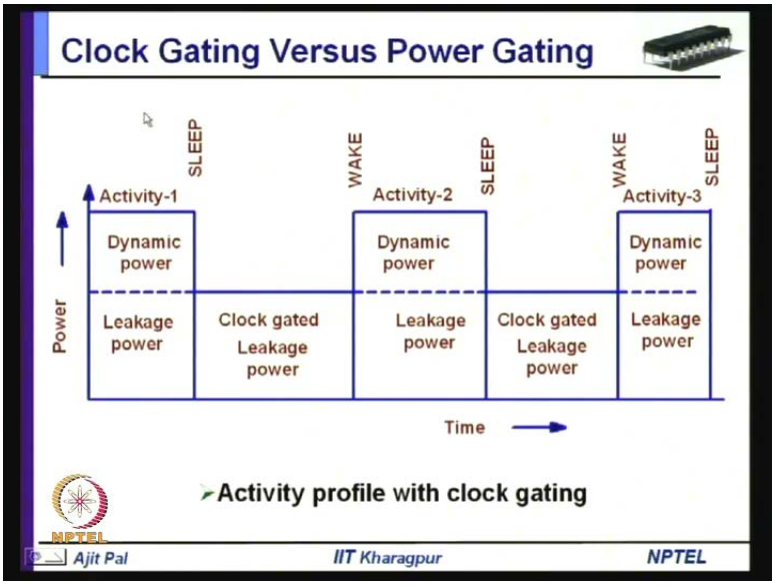
How can it be done? Obviously we have to put a switch in between power supply so we have got your Vdd, and then we will put a switch and the output of the switch will go to the circuit and the other end may be grounded. So, question naturally arises how do we really realize the switch? Actually the switch is realized with the help of a pMOS transistor of high threshold voltage, so this is the pMOS transistor switch that is used in series with the circuit. So, switch is realized with the help of a pMOS transistors; that is exactly what is being done here, you may be asking why do we put a high VTPMOS transistors, the reason for that is as we know our main intention is to reduce leakage

power. The dynamic power can be reduced by using Clock Gating, but here we are we are interested in reducing leakage power, and obviously leakage power can be reduced only when the threshold voltage of the transistor is high. And that is the reason why the transistor that is being used in this in this switch in realizing the switch is a pMOS transistor having high v t, and and that is the reason why this approach is also called MTCMOS based Power Gating, and subsequently this MTCMOS based part have been removed. And it is simply called Power Gating and with the help of this Power Gating what essentially we are doing is power management.

What is the objective of power management? Objective of power management; what do you really mean by management? Management sometimes, I mean you will follow an approach by which the leakage power can be reduced. How you are going to do that? We shall be turning on some parts of the subsystems, and turning on off some parts of the subsystems at appropriate time in appropriate manner; that is your management.

So that is the job of the management and in fact this Power Gating is precisely done in this field; that means you are essentially doing power management by getting the supply voltage, and that is the reason why it is also called power management. So, the same thing is being told I mean same thing is being done, whenever you do MTCMOS based Power Gating or power gating or power management.

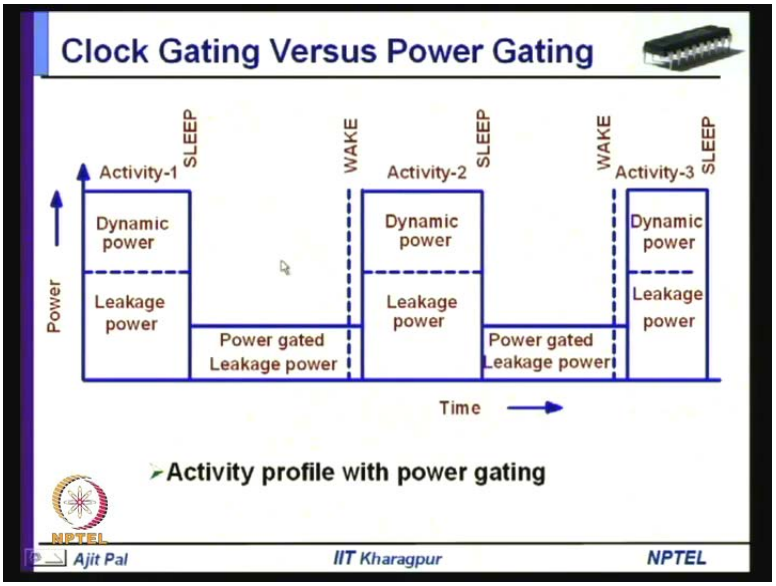
(Refer Slide Time: 06:10)



So any of the three actually means the same thing, now let me start our discussion with this Clock Gating versus Power Gating. So, here as you can see whenever you do Clock Gating in this part, this dynamic power is reduced, but leakage power remains the same, I mean leakage power as as you can see in this case, I mean neither Clock Gating is done nor Power Gating is done here only Clock Gating is done.

So, the leakage power is present, so clock gated leakage power, and so whenever you are in the SLEEP mode; the the you are you are you are essentially reducing the leakage power by getting the clock. So, you can different points of time you can do that in this part, in this part and so on. So, from this diagram we find that the leakage power remains same only dynamic power is reduced when you do Clock Gating.

(Refer Slide Time: 07:08)



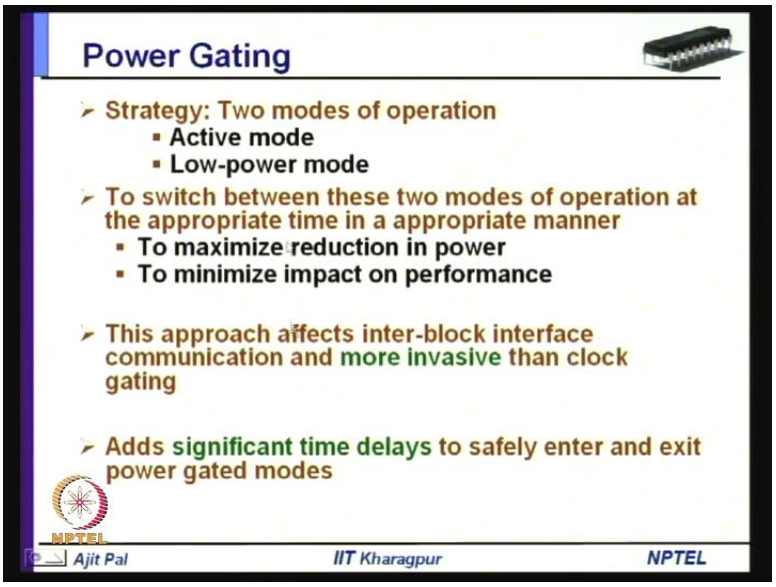

Now, let us see what happens, whenever you do Power Gating as you can see here when you do Power Gating not only the dynamic power is reduced, because whenever you when Power Gating is being done the not only the dynamic power is reduced not only the clock is gated, the supply voltage is also gated and as a consequence the dynamic power is reduced. And also leakage power is significantly reduced, so this is the gate power-gated leakage power, and that can happen at different points of time time depending on the application.

So, this is the activity level 1, and this is the in this part; there is no activity, no circuit can be power-gated. Similarly, here activity level 2, and again power turn is turned on

and having it is having dynamic power as well as leakage power dissipation, and in this part again Power Gating is done. So, you have got so reduction in leakage power. You may be asking that whenever the supply voltage is isolated, why there should be leakage power dissipation? The reason for that is you are doing Power Gating using a PMOS transistor switch, and MOS transistor switch obviously will have some leakage current, it is not an idle switch that means resistance is not infinite. As a consequence the the leakage power will not become 0, but it will definitely be much less than whenever you do not do power gate.

So compared to power gate, if you do not do Power Gating compared to that you can see there is a reduction in leakage power that is essentially possible, because of the use of high v d transistors in series. So, this is the difference between Power Gating and you know this is the difference between Clock Gating and Power Gating.

(Refer Slide Time: 09:21)




Now, let us come to little bit about the Power Gating strategy. So, Power Gating has essentially two modes one is known as Active mode, another is Low-power mode, so this Active mode means whenever the circuit is in full operation clock is applied power supply is applied. And the the standby mode or Low power mode is when power is gated. So, the two switch between these two modes of operation, and appropriate time at the appropriate time in an appropriate manner, as I told Power Gating is nothing but power management, you have to switch these two between these two modes at

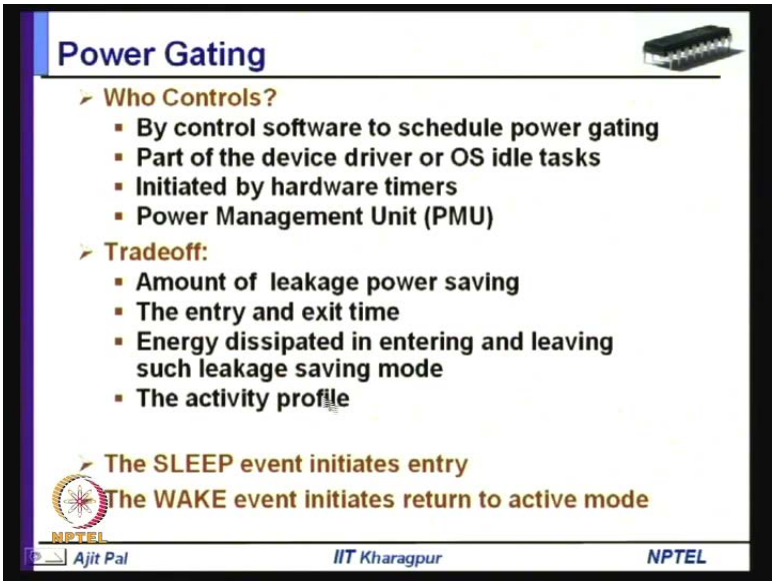
appropriate time in appropriate manner with this goal, what are the goals to optimize reduction in power and to minimize impact on performance.

So, as I told the the challenge of any power reduction technique is to maintain the performance level same, that means without reduction or compromise in performance you have to do the Power Gating, and you have to maximize the reduction in power.

So, obviously as we shall see this approach affects inter-block interface communication and in more invasive than Clock Gating. In case of Clock Gating, we have seen you have to you have to add some additional hardware. However, that additional hardware is not does not make significant difference and as a result it is not that much invasive, but in this particular case as we shall see, It will affect the timing and all the characteristics of the circuit. as a consequence it is more invasive than the Clock Gating, and it will add significant time delays. And we have to see whenever it enters this form I mean Active mode to Low-power mode, you have to safely enter it and exit from exit power-gated mode.

So, there is significant time delays to safely enter and exit Power Gating modes, that means you cannot really instantly switch from one mode to another mode, it will always involve some delays. You have to do it slowly slowly, you have to go from Active mode to Low power mode, and again you have to come back to the Active mode from the Low-power mode, so this will involve significant time delays.

(Refer Slide Time: 12:01)



Now question naturally arises if it is a management technique, who is doing the management, who is controlling? There are several possibilities number one is by control software to schedule Power Gating. So, you have to I mean the there is a software, it is controlled by software there is a control software which will do the Power Gating that means maybe that is part of the operating system, which will find out when to do Power Gating that means when to switch from Active mode to Low-power mode, and again come back to the Active mode, another part is part of the device driver.

So or OS idle tasks that means in this particular case you know that operating system knows when a particular subsystem is in idol condition. So, for example, there is a game controller, so the operating system knows when DMA mode (( )) transfer has be perform, so the then the device driver of that particular particular subsystem can be put to Low power mode when it is not in use.

So, the idle idle situations can be found out, and another possibility is it can be initiated by hardware timers, so inside the processor you will find some hardware timers known as timer counters. For example, you want a particular event to occur after certain time and that time can be programmed, and after that time is over that time out is reached and that event is triggered. For example, as you have as you see whenever some missile is launched and various; for example, even when you are using a using a television you can program it so that it will switch off after certain time.

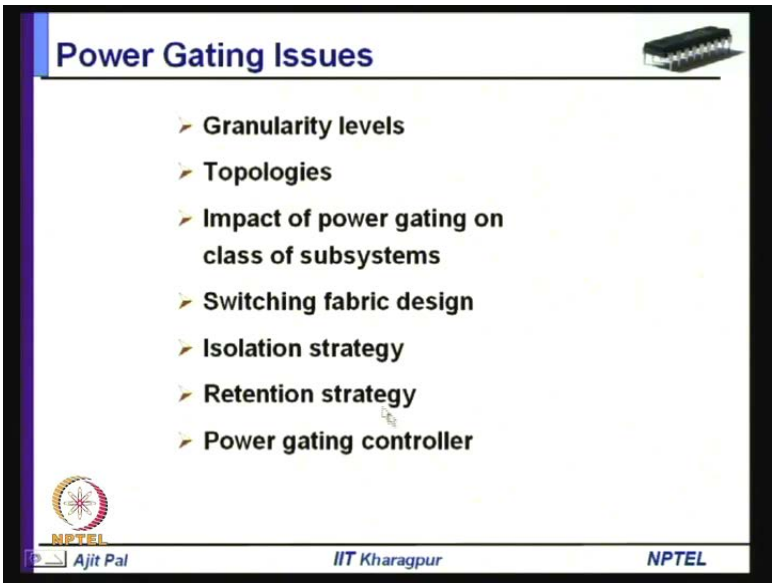
So, this kind of thing can be triggered with the help of hardware timers and most sophisticated one is to have power management unit. Now a days particularly on silicon on chip systems you have got power management unit, so that power management unit will perform this task of switching from one particular mode to another particular mode. And so this can this control is done in many ways and obviously from application to application it will vary.

Then whenever you do Power Gating there are several tradeoffs, the tradeoffs are how much leakage power saving you can do by doing Power Gating, and the entry and exit time. As I told the entry to the power-gated mode will take some finite time, so you have to there is a you have there is a tradeoff how quickly it can be done, and and how much power saving can be achieved then the energy dissipated in entering and leaving. Later on we shall see whenever you enter a Power Gating mode; there is some some overhead.

So, energy dissipated in entering and leaving such leakage power modes that there is a tradeoff of that; that means by not doing Power Gating whatever power dissipation is there that should be larger than energy dissipated in entering, and leaving such leakage saving mode, then it will again be there is a tradeoff between the activity profile based on the activity profile you have to judge whether you will be doing Power Gating or not.

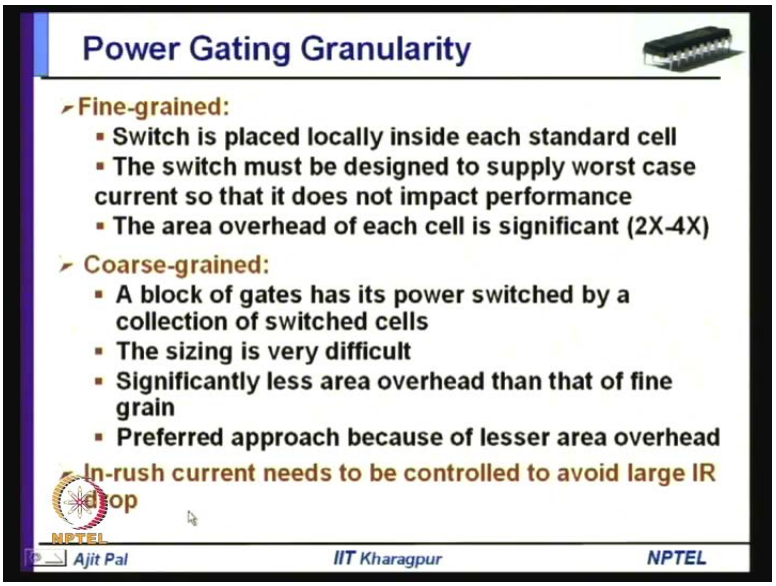
But as as we have already seen the it is it is essentially controlled by two entities, two events, number one is SLEEP event which initiates entry into the power-gated mode, and another is WAKE event which initiates returns to the return to the Active mode. So, these two events will trigger or will switch from one mode to another mode.

(Refer Slide time: 16:17)



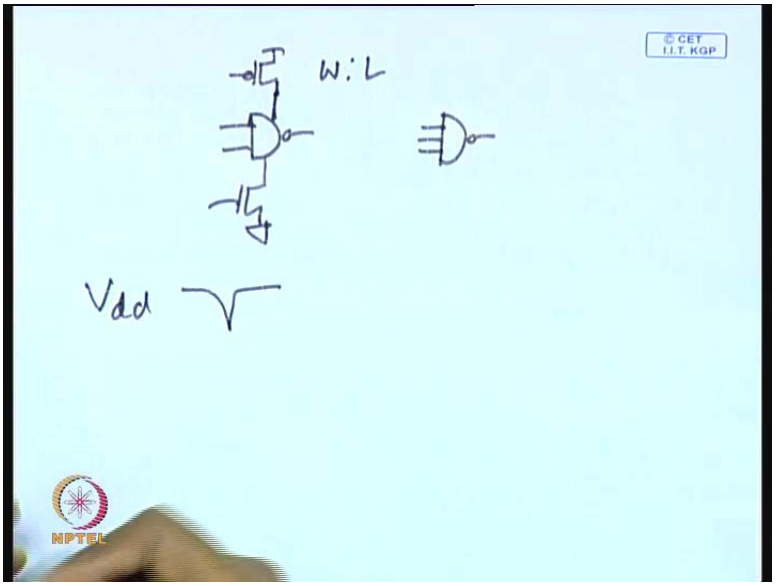

And as I have told there are number of Power Gating issues such as granularity levels, and so on I shall discuss them one after the other.

(Refer Slide Time: 16:29)




First let us focus on granularity levels. The granularity levels can be essentially of three types: first one is fine-grained, so switch is placed locally inside each standard cell and the switch must be designed to supply worst case current. So, that it does not impact performance; that means suppose you have got a standard cell standard cell can be a a two input NAND gate or a three input NAND gate.

(Refer Slide Time: 16:56)




So far each of them you can have a Power Gating circuit, and obviously in this case you have to whenever use Power Gating circuit; it will it will have some area overhead and

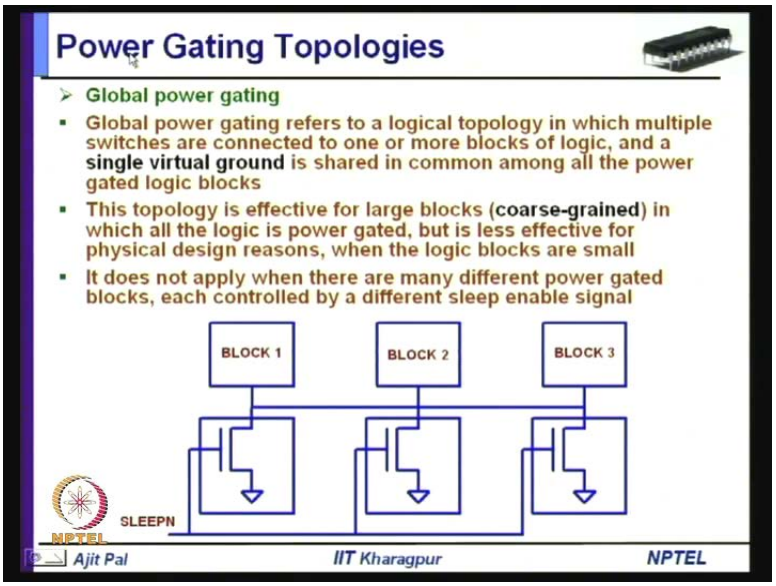
this area overhead of each cell is quiet significant it can be two times to four times and you have to design the switch particularly you have to since you are inserting some transistors in series pMOS transistors in series or nMOS transistors in series, you have to choose the width and length. In such a way that the the it can supply worst case current because larger the width, it will be able to supply more current, but it will occupy larger area.

So this kind of fine-grained technique has large area overhead, and and that is the reason why coarse-grained Power Gating granularity is used, in such a case a block of gates has its power switched by a collection of switched cells. That means you can do kind of clustering several gates can be combined together, and a single switch will take the job will perform the switching from Active mode to Low-power mode and so on.

So, and in this case sizing is very difficult because you do not know exactly which components will be turned off turned on, and also the current requirements of different gates. So, this in this particular case sizing is more difficult, and however it has got significantly less area overhead than the fine-grained approach, and this is preferred this is preferred approach, because of lesser area overhead. And whenever you use whether fine grained or coarse-grained techniques, you have to take care of the in-rush current to be controlled to avoid large IR drops; that means you should not suddenly switch large number of you should not turn on large number of switches, so that there is a power droop, that means Vdd, it does not drop leading to glitches and power line problems.

So, inrush current has to be controlled and we shall see later on how it can be done. So, we have seen two type two types of granularity that is being used.

(Refer Slide Time: 19:30)

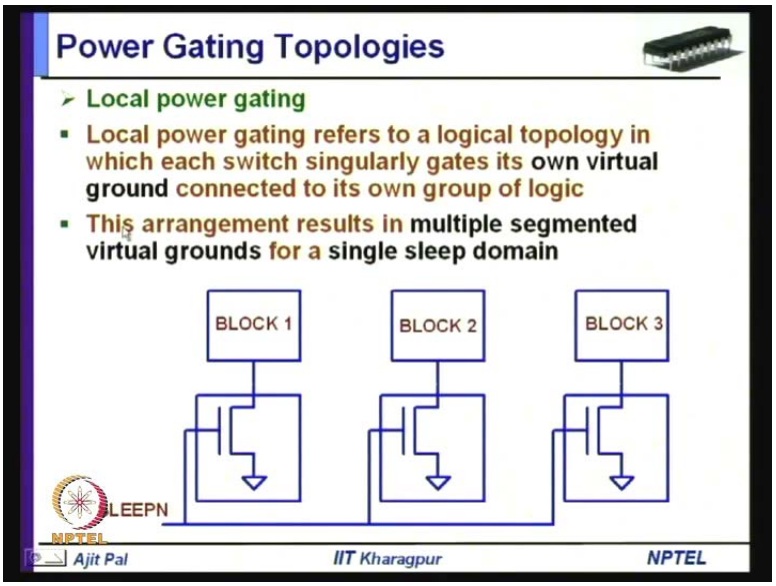



Now, let us consider the various topologies. There are three different topologies number is is Global Power Gating, it is so in this Global Power Gating, this refers to logical topology in which multiple switches are connected to one or more blocks of logic and a single virtual ground is shared in common among all the power-gated logic blocks. So here you find one this is one logic block, this is another logic block and this is the third logic block. So all of them are having a single virtual Vdd line, so here is the rather in this in this particular case it is virtual ground line, so this is a single virtual ground line. Although you have got multiple switches, and these multiple switches are controlled by using a single SLEEP control signal.

So, this topology is effective for large blocks in which all the logic is power-gated, but is less effective for physical design reasons when logic blocks are small, so this particular approach is is effective when you are using large blocks, and particularly we have seen this can be used when you are using coarse-grained technique.

So, these granularity and the topology are kind of interrelated, so this particular technique does not apply when there are many different blocks each controlled by different SLEEP enable signals. That means whenever you have got a large number of blocks which are controlled by a single control signal only, then you can use use this kind of topology.

(Refer Slide time: 21:10)

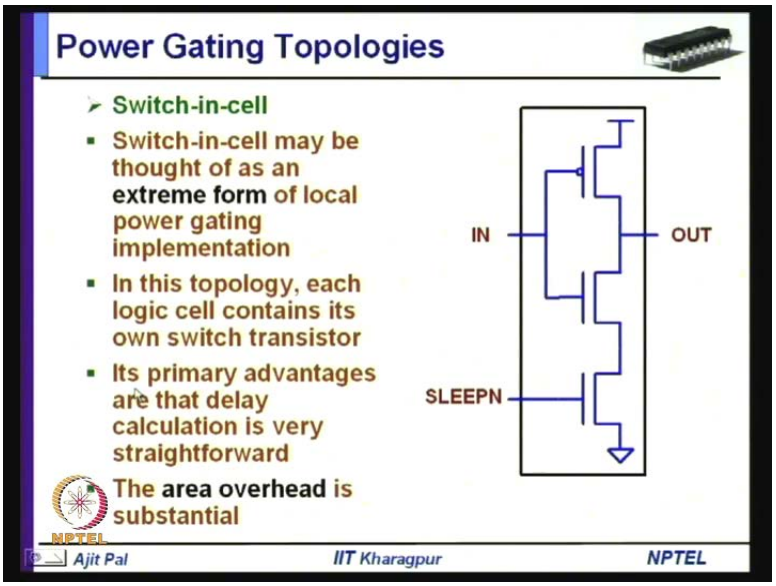



Coming to the second technique, which is known as local Power Gating, in this case the this this particular technique refers to logical topology in which each switch singularly gets its own virtual ground connected to its own group of logic. So, in this case as you can see this is 1 block, and this virtual ground is not connected to the virtual ground of block 2.

So there is a independent virtual grounds for each of the logic blocks, so this arrangement results in multiple segmented virtual grounds for a single SLEEP domain, so you you know that there is only a single SLEEP control signal. Although there is a single SLEEP control signal you have got multiple virtual grounds for different blocks of logic, so this particular technique can be particularly used when you are using this fine grained granularity. For example, you are doing switching is done at gate level.

So, this arrangement results in multiple segmented virtual grounds as I have told with a single SLEEP domain.

(Refer Slide Time: 22:32)



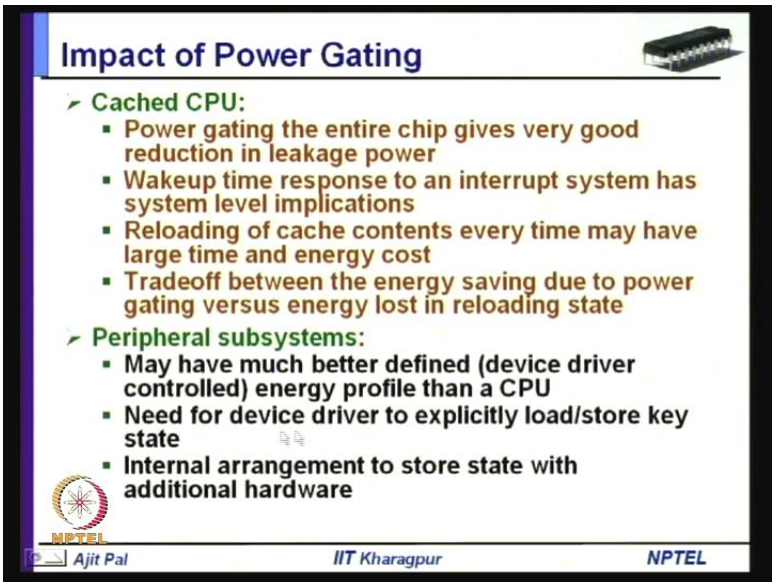

Coming to the third topology, this is known as switch-in-cell. So, switch-in-cell maybe thought of as an extreme form of logical Power Gating implementation. So, in this topology each logic cell contains it own switch transistor as you can see this is an inverter, so this inverter is is provided with another additional transistors obviously this will be a transistor with high threshold voltage which is built-in when this inverter is fabricated. That means you have got kind of standard cells with this kind of built-in gating hardware, and each of these standard cells will be having its own gating hardware built-in as part of this circuit. So, each cell is independently I mean is designed separately, and its primary advantage advantages are that the delay calculation is very straight forward.

So, a delay calculation is very easy, because it can be very easily characterized, so you have got an inverter, and in series with that you have got a switch that means your high threshold voltage transistors. You can very easily calculate the delay of this particular circuit and so the calculation of the delay is quiet easy, so whenever you do static timing analysis - the static timing analysis can be done very conveniently without much difficulty.

However whenever you use this approach the area overhead is substantial. So, there is substantial area overhead, because each of the standard cells will have its own Power Gating hardware built in. So, as a consequence there is no sharing no sharing of Power

Gating hardware is taking place as a consequence, it will have large are overhead however, timing analysis is very easy, and it is easier to implement.

(Refer Slide time: 24:39)



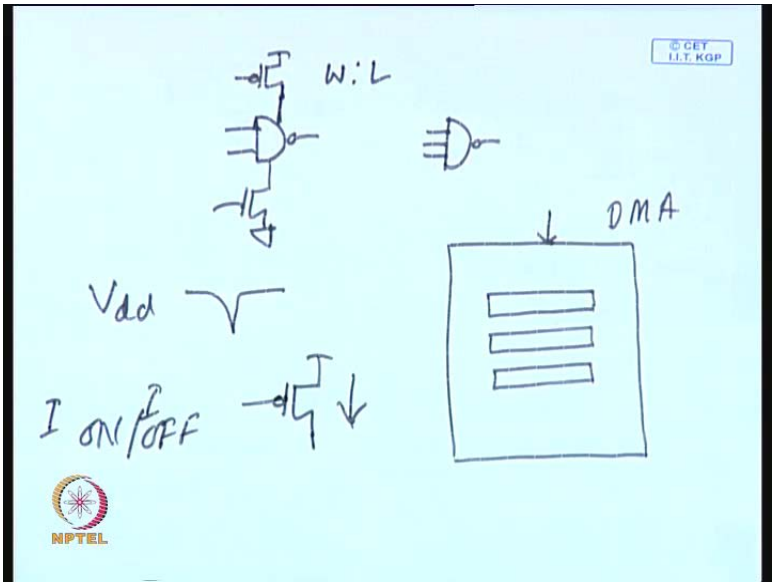

So, this is so much so about the topologies. Now, let us consider impact of Power Gating of for different subsystem. Let us consider a scenario where you have got a Cached CPU; that means you have got a central processing unit with cache memory. As you know normally your program is residing in the main memory which which is usually outside the processor, CPU.

Now, you have got a CPU with on chip cache memory. When the CPU is power-gated obviously the cache memory will will the information stored in the cache memory will be lost, so whenever you do WAKE up; wakeup time response to an interrupt system has system level implications, because you have to do it will lead to cache misses and you have to transfer from the main memory to the cache memory. And reloading of cache contains every time may have large time and energy cost; that means whenever you have got a cache CPU as you go from Active mode to Low-power mode; that means, you do Power Gating of the CPU the information will be lost, and then whenever you do the WAKE up you have to transfer or in other words if you do not transfer there will be cache misses. And those cache misses will lead to large overhead, and so there is a tradeoff between energy saving due to Power Gating versus energy lost in reloading

state; reloading state of the processor I mean that is the which which which is being stored in these in those cache memories.

So, this is one type of environment or one type of scenario, second is peripheral subsystem. So, peripheral subsystem may have much I mean in this particular case energy profile is much much well defined than a CPU. Why it is much more well defined? The reason for that is here a particular peripheral; a peripheral can be a programmable IO like DMA controller, interrupt controller or some other programmable IO, they are parallel programmable IO which are used in microprocessors, and microcomputer based systems those peripheral subsystems have got much well defined energy profile, and in the case you need to need needs for higher.

(Refer Slide Time: 27:18)




I mean need for device driver to explicitly load store key state you will you will you vwll notice; that these programmable IO devices maybe a DMA controller or interrupt controller of whatever it maybe will have some resistors, which are part of those peripherals or subsystems; those resistors are to be initialized whenever you make them active say this is a DMA controller.

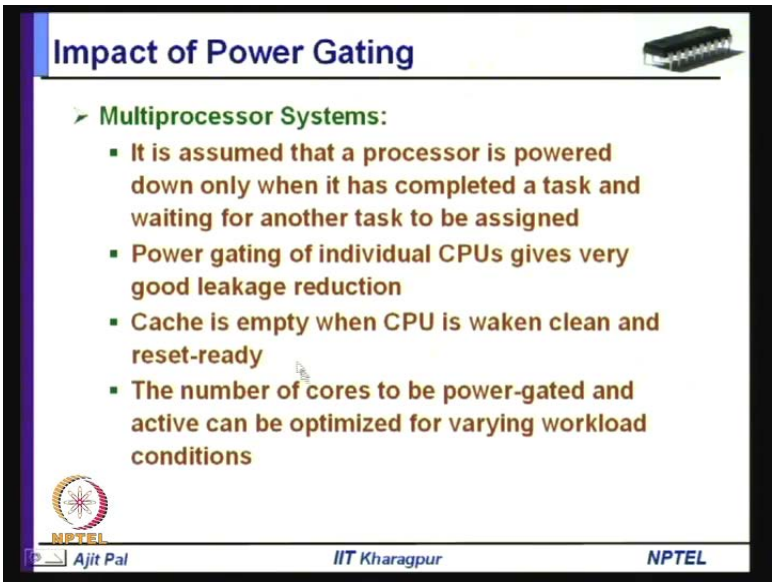
So, DMA controller will have some resistors which will hold the size of data to a transferred between memory, and the IO, and also the from which location to which location I mean which memory location to which memory location data transfer has to be done those loaded in these resistors. That means and the and the you will require device

driver to explicitly load and store key state; that means whenever you are going form active state to Low-power state. That means you are you you are putting a DMA controller or an interrupt controller or some other peripheral subsystem to a Low-power state those resistor contents, if they are having some values they have to be saved in the resistors of the processor, and then as they WAKE up you have to transfer those resistor contents.

Question is who will do that? So, there is a device driver software which has to do that, so the device driver has to explicitly load, and store those key states which are present in these resistors, so the internal arrangement to store state within the additional hardware. So, there is another alternative inside the peripheral subsystem, you can have some built in hardware, where those those those state values - those state conditions; that means content of those resistors can be stored later on I shall come back to this particular topic. And we shall see how inside the subsystems, you can have built-in memory built-in resistor values where you can store these things.

So, there are two alternatives either you do it with the help of a device driver or there is some additional hardware as part of those those those device drivers; that means programmable IO chips.

(Refer Slide Time: 29:48)



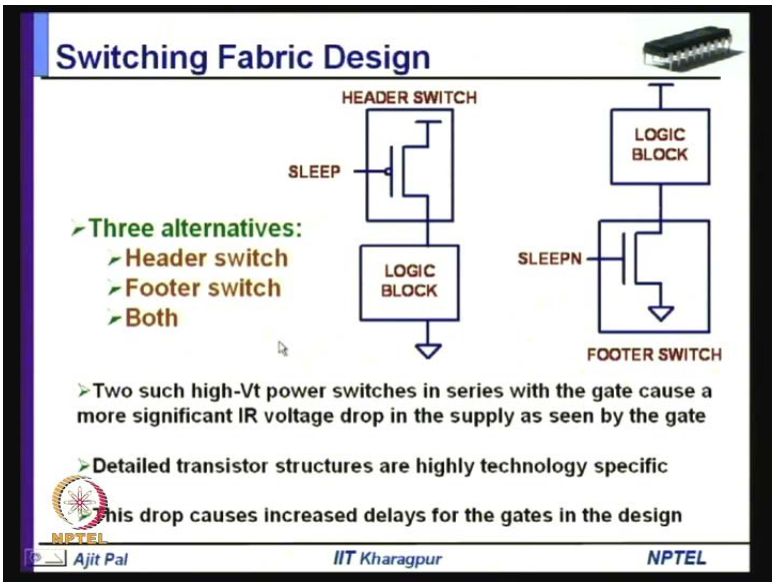

Hard situation is multiprocessor systems. So, these multiple subsystem is becoming very popular in the present context, you can have multiple processors either they have they are

they may not be on same chip; that means two separate chips, two CPU's can be on a single motherboard or other more than two processors can be on a single motherboard that is one possibility. Another possibility is that nowadays you can have multiple processors on a single chip known as cores, multi core processors are becoming increasingly popular two core, four core. And so core whatever it maybe it is assumed that a processor is powered down only when it has completed a task and waited for another task to be assigned, so in a multi-core architecture there is an operating system which will do the task scheduling. And if you want to save power what can be done, the task scheduler I mean scheduled task tasks to a particular processor, and when that task is complete it can be it it it that particular core or processor can be can go can be switched off; that means you can put it to Low-power mode.

So, Power Gating of individual CPU's gives very good leakage power reduction, so the in this particular case the depending on the application your computation or computation requirement may be different, and the whenever it is done with the help of this task scheduler. Then you know a particular task is over; that means cache is empty when the CPU is waken clean, and reset-ready; that means a particular processor was put to Low-power condition when it had completed its tasks, so when it is waken up the cache is empty. And obviously there is no need for no need for loading the its own status automatically it will load the contents of cache memory as a new task is scheduled to it.

So, the number of cores to be power-gated gated and active can be optimized for varying workload conditions, so the task scheduler can identify by monitoring the workload condition when one or more cores can be put to Low-power mode. And when when they have to waken up depending on the workload condition, so these are the three and scenarios I have discussed and their impact on Power Gating I have mentioned.

(Refer Slide Time: 32:36)

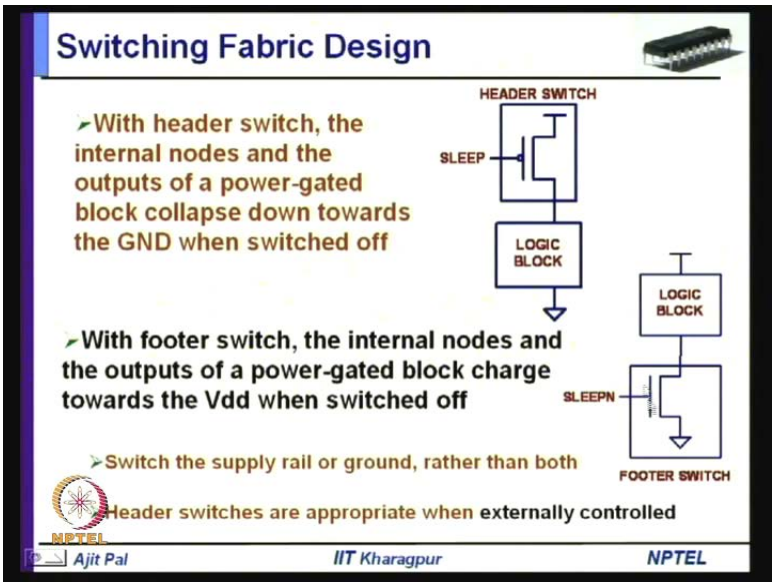



Now, coming to switching fabric design, you can have three alternatives. You can have header switch as it is shown here; that means you will have a pMOS SLEEP transistor or you can have a footer switch which is a nMOS transistors obviously; both these switches will have high threshold voltages or you can have both. So, originally that MTCMOS approach that was proposed by (( )) at all they proposed both the switches to be present header switch as well as footer switch, so there are many authors some of them claim that both the switches are necessary, and most of them claim that one is sufficient.

So, whenever you use both the switches to such high v t power switches in series with the gate causes more significant IR drop in the supply as seen by the gates; that means when you have got two switches quiet large amount I mean significant amount of power drop take place particularly when the supply voltage is getting gradually reduced, and it is now in the in the range of 1 volt. So, detailed transistor structures are structuring has to be done and these are highly technology specific.

So, this drop whenever this IR drop causes increase delays for the gates in the design, so the switching fabric design is very complex, and you have to do it properly.

(Refer Slide Time: 34:10)



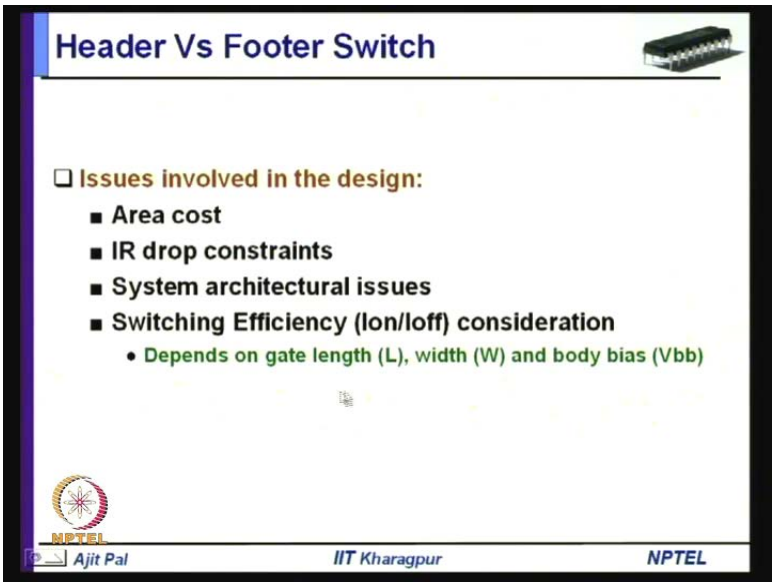

So, that you get the benefit of leakage power reduction technique. So, with the header switch the internal nodes and the outputs were power-gated block collapse down towards ground when switched off, so whenever you are using a header switch that means the switch is turned off. And that means what will happen in that case this this particular line, you know this particular point that virtual ground will will be very close to the ground voltage, that means the the outputs of the power-gated blocks collapse down towards ground when switched off; that means when it is switched off this voltage will become very close to the 0 volt, because this will behave virtual ground voltage.

Similarly when a footer switch you are whenever you are using this is the header sorry whenever you are using a footer switch in that particular case, this voltage will gradually move towards Vdd. That means with the footer switch the internal nodes and outputs of a power-gated charge towards Vdd when switched off; that means whenever you switch it off this voltage will gradually charge towards Vdd, but you know switch the… So what has been found. As I told switch the supply rail or ground rather than both usually one of the two switches are used and particularly header switch is more appropriate when externally controlled.

So, later on you will see this Power Gating is done in two ways. Switch can be external to the processor, external to the chip or it can be inside the chip. So, there can be a tree like structure, so whenever outside the chip this particular approach this header switch is

more appropriate, because in that case this power distribution particularly that power rail can be can be distributed distributed in a very systematic way .

(Refer Slide Time: 36:11)

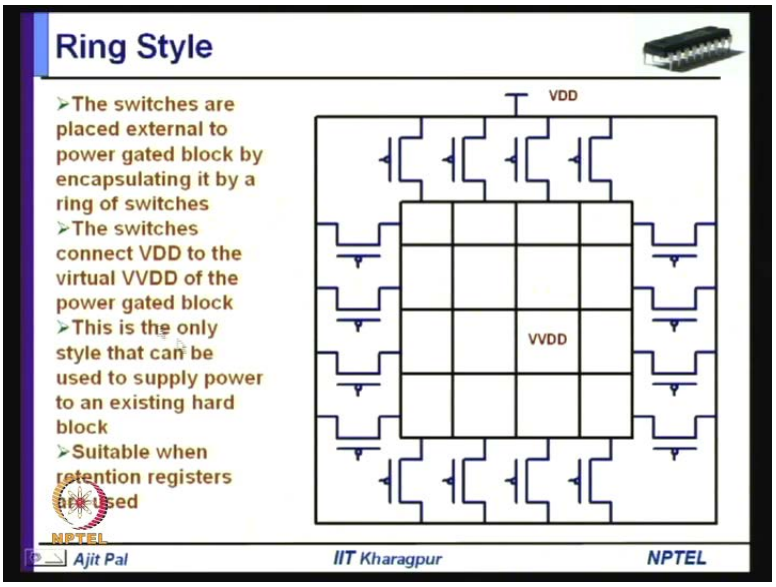



So, whenever you are designing the switches; there are several issues involved in the design number: One is the area cost as I told that that with two length ratio, we will decide the the current that can flow or IR drop, so the there is a IR drop constraints and this there is a tradeoff between the area. And the reduction in leakage power and system architectural issues, which I shall discuss little later it has some impact on various other architectural issues, and another very important thing is switching efficiency consideration.

So, whenever you are putting a switch when it is on it can pass some current so on; current is dependent on the width versus length of these device, and whenever you put them off, the soft current is dependent on the threshold voltage, and various other parameters. This on to off ratio this I on to I off this ratio is very important, and it depends on the gate length width, and body bias which essentially controls the threshold voltage. So, the switching efficiency is known as I on versus I off; that means on on current versus off current, and obviously this on current versus off current will be different different for different types of switches, so far pMOS transistors this on current versus off current will be will be of one type, and for nMOS transistors this on current versus off current will be of different type.

So, the switching efficiency you have to consider whenever you are you will be designing this power switches.

(Refer Slide Time: 38:06)



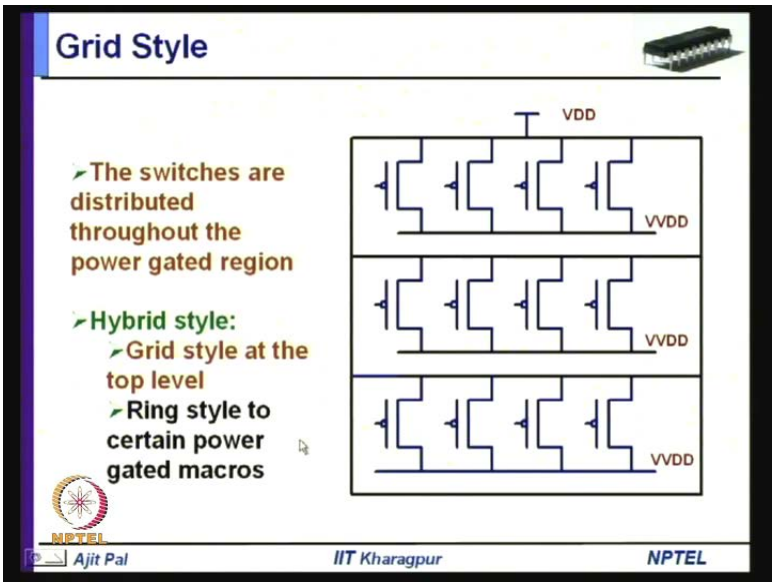

Coming to different styles of power switch, as I as you have already seen whenever we are using a switch essentially a single switch is not used, a single transistor is not used a large more than one transistor is usually used.

Now, how you will put these transistors and there are two basic topologies one is known as ring style, so here you can see the switches are placed external to power-gated block by encapsulating it by a ring ring of switches. So, you have got a ring of switches and here is the power-gated block, so the switches connect Vdd to the virtual Vdd of the power-gated block, and this is only style this is the only style that can be used to supply to supply power to existing hard block.

So normally you may be using hard i p core, so whenever you are using a hard i p core maybe a c p p or some memory or some some special purpose processor whatever it may be whenever you are using a hard i p core, in such a case this is the only technique that can be used and it is suitable when retention resistors are used.

So later on I shall discuss about these retention resistors, which are necessary to shape the state of these power gated block and later on we shall discuss about it.

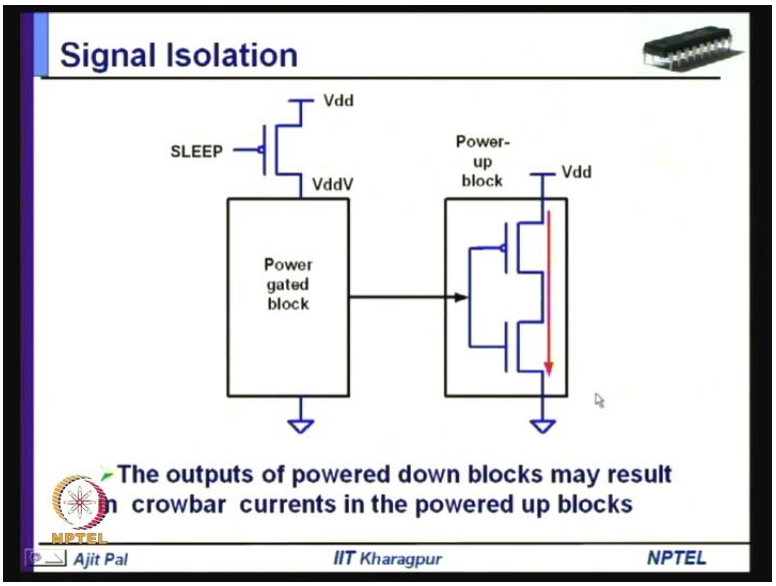
(Refer Slide Time: 39:39)




Then another style is grid style. So, in this case the switches are distributed throughout the power-gated region, so here you can see this is the supply voltage, and this is one virtual Vdd, so this is this will go to one part of the circuit this, will go to another part of the circuit, this will go to another part of the circuit. So, this is how the power rail is distributed throughout the power-gated region. However, you can use a kind of hybrid style where you will be using this ring style as well as this grid style.

So grid style at the top level, so you will use this in the top level; that means you whenever you are using high style then at the top level you will be using this grid style, and then at lower level you will be using the other style, that is your ring style. So, grid style at the top level ring styles to certain power gated macros or hard i p cores. So, this is how we can combine both ring and grid style, the switches in your harder in your circuit.

(Refer Slide Time: 40:49)




Now coming to a very important issue, that is your known as signal isolation. What do you really mean by signal isolation? To illustrate single isolation let us consider a power-gated block. So the Power Gating is done with the help of a SLEEP transistor, header switch which is connected here so here is your virtual Vdd. Now as I mentioned whenever you turn the switch off this voltage is supposed to rise to Vdd sorry this voltage is supposed is supposed to go back to to reach the ground level, so in this particular case it discharges through this path because of higher larger current of this power gated block, so the outputs of power gated down blocks may result in power crowbar current in the power powered down blocks.

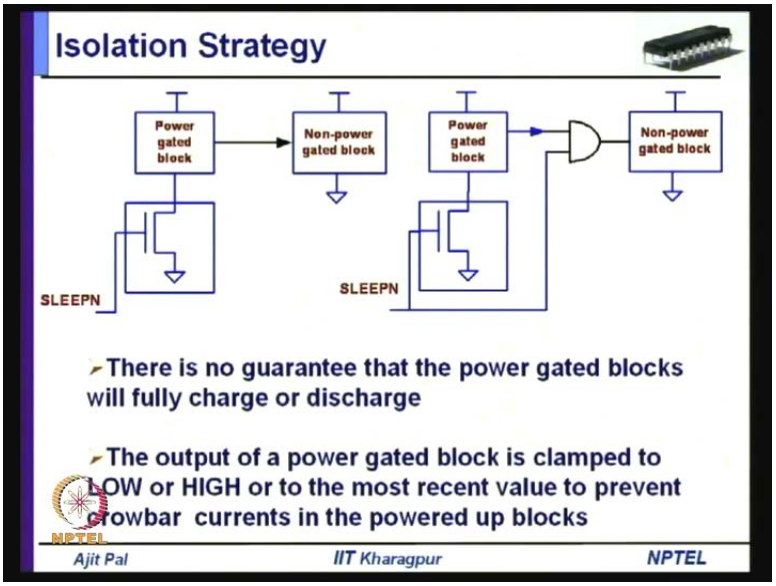
So, what can happen you cannot really guarantee that these voltage will will be discharged to ground level or whenever you are using a footer switch this voltage will charge to Vdd level. This you cannot guarantee, because it will depend on the relative value of the the leakage current of this SLEEP transistor, and the leakage current of this power gated block and also the l by w ratio various other parameters that I have discussed.

So, what may what can happen? This the output of a power gated block can have some intermediate value, and that intermediate value can lie in the range of that it may be greater than v t n; that is the threshold voltage of the nMOS transistor. And less than Vdd by v t p that is your Vdd minus v t p where v t p is the threshold voltage of the pMOS

transistor. In such a case what will happen? As you know in such a case this this voltage will lead to what is known as short circuit current or crowbar current.

So, the input voltage here is such that this transistor as well as this transistor both are on as a result current will flow through this part of the circuit, I mean short circuit current will flow or crowbar current is flow that is another name of short circuit current that can flow through this part of the circuit. Question is how can it be stopped? How can this this short circuit current can be prevented? To prevent this, you have to use what is known as signal isolation so you have to isolate this power gated block with the power of block. You have two types of circuit present in a in a in a in a circuit in operation some part will be power gated some part will be active.

(Refer Slide Time: 43:46)



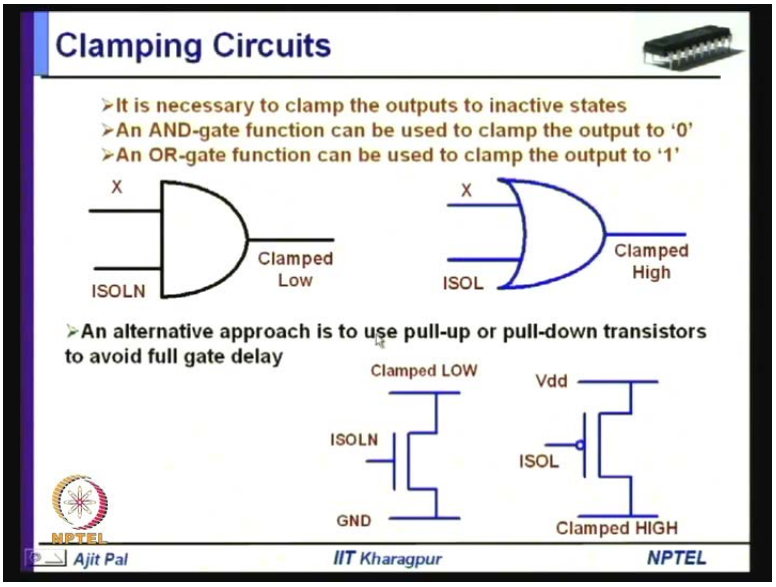

So, there is an interface between power gated block, and a power up block so what you have to do you can see here you can you have got power gated block, and this is diving a non power gated block. So, if this voltage is not appropriate if it is not 0 or Vdd, then this will lead to short circuit power dissipation as I have told, so what you have to do you have to add additional hardware, and this is known as isolation hardware, why do you need it, because as I told there is no guarantee that the power gated blocks will fully charge or discharge.

So, the output of a power gated block is clamped to low or high or to the most recent value to prevent crowbar current in the powered up block, so what can be done you will

add some additional hardware in between this power gated block, and non power gated block and what is the what will be the output? Either this will be clamped to low level or to high level or to the most recent value.

So, whenever you have to output the most recent value; obviously, you will require some kind of memory as part of this hardware isolation hardware.

(Refer Slide Time: 44:58)



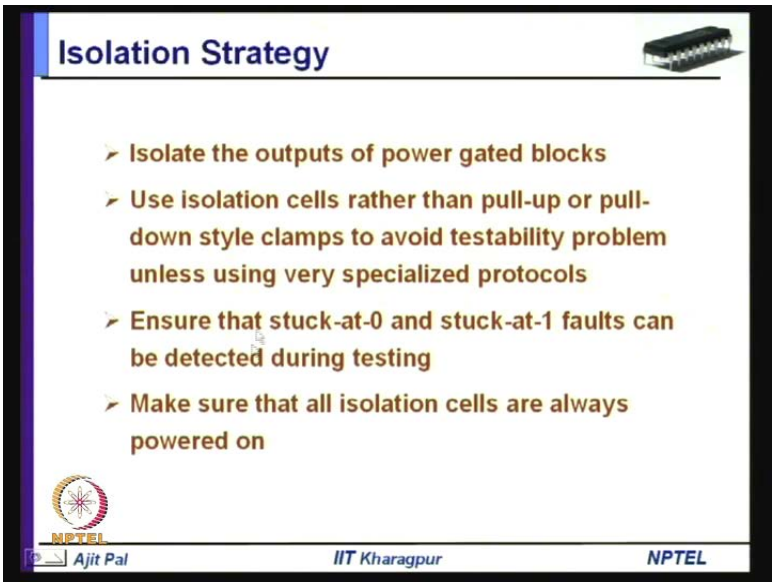

So, this let let us see how it can be done. So, first of all it is necessary to clamp the output to an inactive state and an AND gate function can be used to clamp the output to 0 level so you can use just an AND gate, so this is the output of the non power gated block that this isolation control signal, so here it is coming from that power gated block. So power gated block input is coming here, and this output will clamped low irrespective of the if this input here, if this input is 0 this output will be 0 that means when this particular this signal, I mean this signal is low that means you have done Power Gating then what will happen, this output will be this will be this will be clamped to low level.

Similarly, you can use an OR gate. In this particular gate this will this output will be clamped to high level that means irrespective of the inputs at the at this input, whenever this is this is one this will be clamped to clamped to high level, and as a consequence that uncertainty will not arise here. And alternative approach is to use pull up and pull down transistors so instead of gates, so you know that gates will be having some delay, so if

you want to avoid these delays, you can use simple simpler hardware known as you know pull-up and pull-down transistor.

So here you can see, this is the isolation signal and a transistor is there, so this signal is used here to control this output, so whenever it is high this will be clamped clamped down to low level. Similarly, in this particular case whenever this is low the signal is low this will be clamped down to high level, so you can use this type of this type of simple I mean clamping circuits instead of using gates when delay requirement is large.

(Refer Slide Time: 46:59)

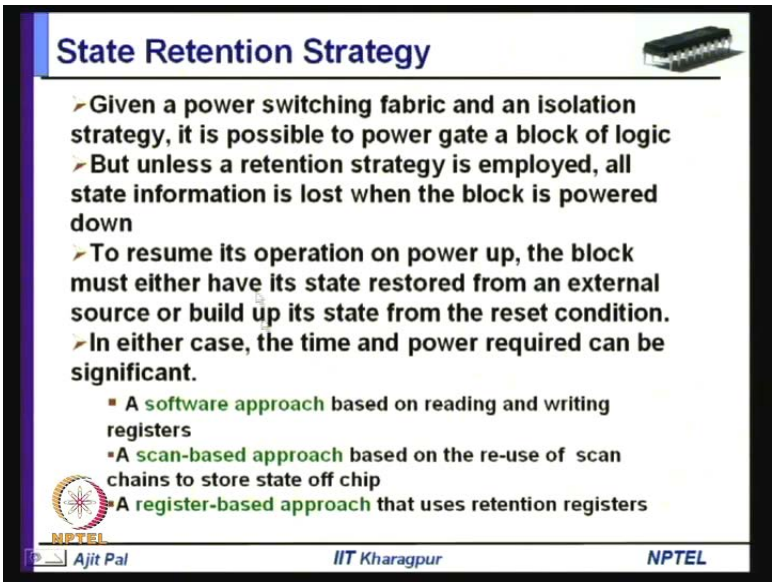



However, there are some strategies which are used to make the circuit full proof. So you have to isolate outputs of power gated blocks use isolation cells rather than pull-up, and pull-down style clamps; that means it is being recommended not to use this type of hardware rather than it is better to use this type of gates to avoid testability problem unless using very specialized protocols. Actually you have to consider the testability of the circuit and you have to ensure that stuck at 0, and stuck at faults can be detected during testing. So, if you use this type of circuit this type of testability can be hampered we are not going into the details, but using this type of circuit that testability can be ensured by applying proper signal here, but which cannot be done in this particular situation.

So, make sure that all isolation cells are always powered up another important thing that you have to remember this additional hardware should be always powered on, that means

this this power power supply of this should not be clock gated it should not be gated. So, it should not be power gated, so they should get continuous power supply.

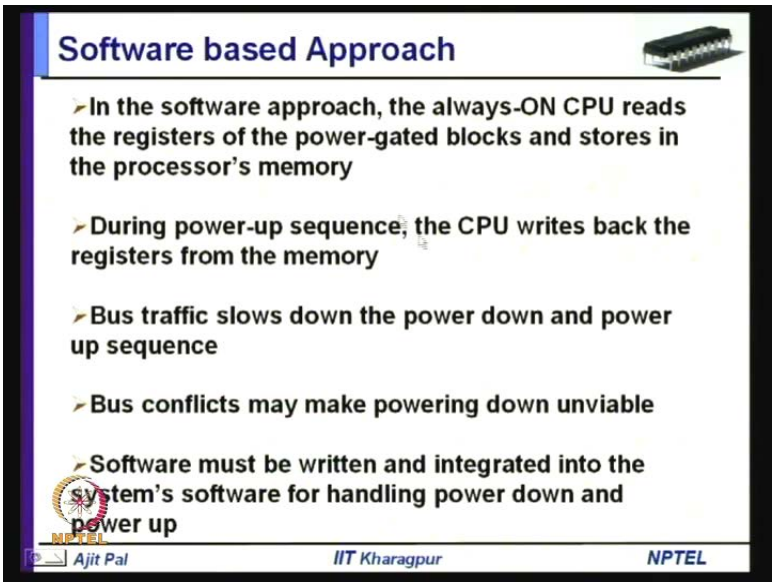
(Refer Slide Time: 48:12)




So, this is the isolation strategy now coming to state retention strategy, so given the power switching fabric, and an isolation strategy it is possible to power gate a block of logic, but unless a retention strategy is employed all state information is lost when block is powered down as I have already discussed. How it happens? And to resume its operation on power up the block must either have its state restored from an external source as I told the the device driver can do it or built-up its own state from the reset condition.

So there maybe some built-in hardware so in either case the time and power required can be significant and there are three basic approaches, first one is known as software based approach on reading and writing registers. Second is scan based approach based on re-use of scan chains to store state off store state off chip, and register based approach that uses retention register.

(Refer Slide Time: 49:18)



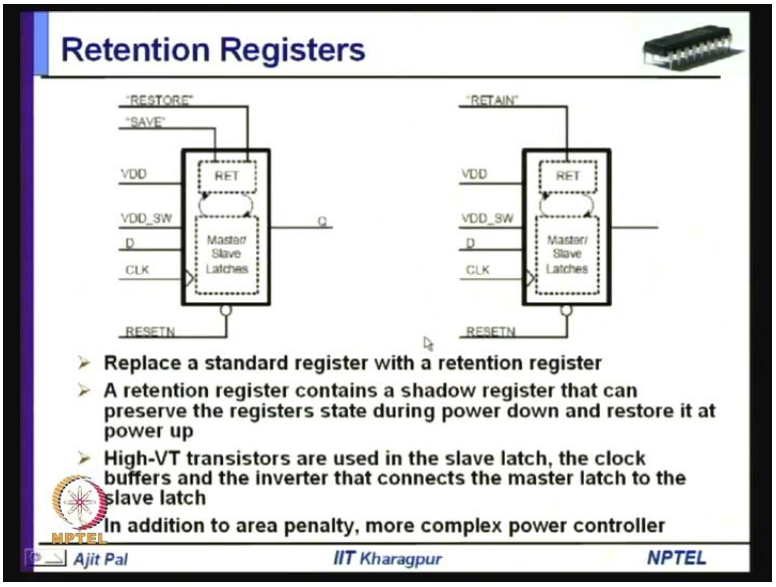

Let me very quickly go through; these three techniques in the software based approach you have got a always on CPU, so what you are doing here a program is running which will whenever you have decided to power down a particular block, there states will be saved will be transferred to some internal resistors of the CPU by programming. And during power-up sequence the CPU writes back the registers from the memory or you can store them in the memory either in the register on in the memory. So, bus traffic slows down the power down and power up sequence.

However, whenever you do that the activity this this storing, and loading will take place through the system bus, and this will definitely influence the traffic on the on this bus, and bus conflicts may take may make power down unviable. So, there may be bus conflicts and that may lead to unviability of this approach, and software must be written and integrated into the systems software for handling power down, and power up so you have to write appropriate software which will take care of this.

Second approach is scan chain based you know nowadays you are using built-in self test, so built-in self test technique require some additional resistors for inputting; the inputting the test sequence and outputting the response. And these are the resistors which are used at the time of you know testing built-in self test. Those resistors can be used to store the states of those power, those blocks which you which you intend to put to SLEEP condition.

So, scan chains used in the BIST can be re-used and during power down sequence the scan register outputs are routed to a on-chip or off-chip memory, and this may lead to significant saving in saving, because you are not using a any additional hardware this scan chain resistors are being used to save the the state of the those blocks which you intend to power intend to do Power Gating.

(Refer Slide Time: 51:40)




Third approach is based on retention registers as I was telling you can have built-in hardware present inside those blocks, so what you what is being done you have got two sets of resistors; one is master, another is slave. So, whenever you have to a save signal is received the information is transferred from the master to the slave.

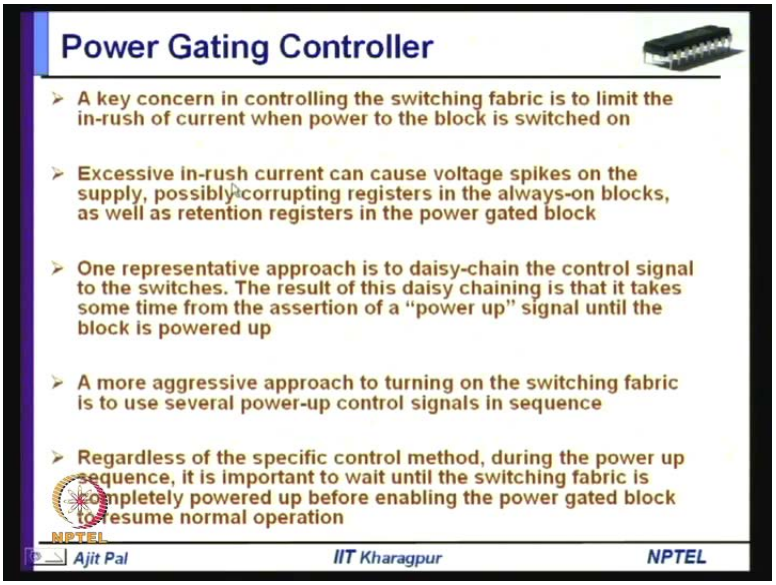
So, and whenever again when restore signal comes the information is transferred back from the slave to the master, you may be asking that means whenever you are in the power; power gated block this part you know is switched off, not this part that means that is the reason why you have two power inputs. So, this is the Vdd switched gated, so the master will receive this switch power, and the other hand this slave resistors will require non gated power, because they have to store the information.

Now, this will require additional control signal coming from the controller, so save and restore, so before before going to low power state you have to save. And before you WAKE up you have to restore the signal are generated. However, you can use a single signal retain; that means when depending on whether it is 1 or 0 you can either retain or

you can transfer to the master, you can do it this way; and these retention resistors are realized by using high threshold voltage transistors, and as a consequence the power dissipation in these resistors is quiet small compared to the leakage power that takes place in these masters.

So, in addition to area penalty you have got more complex power controller is required, because these signals are to be generated from the power controller.

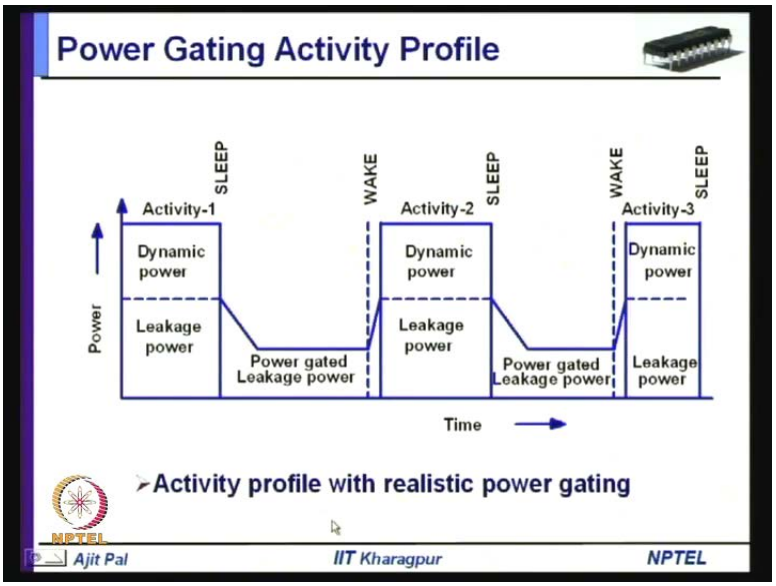
(Refer Slide Time: 53:41)




Let us now come in to the Power Gating controller. So, a key concern in controlling the switching fabric is to limit the in–rush of current when power of the block is switched on, an excessive in-rush can cause voltage spikes on the supply possibly corrupting registers in the always-on blocks as well as retention registers in the power gated block, so one representative approach is to daisy-chain the control signal of the switches, the result of this daisy-chaining is to it takes some time from assertion of a power up signal until the block is powered up.

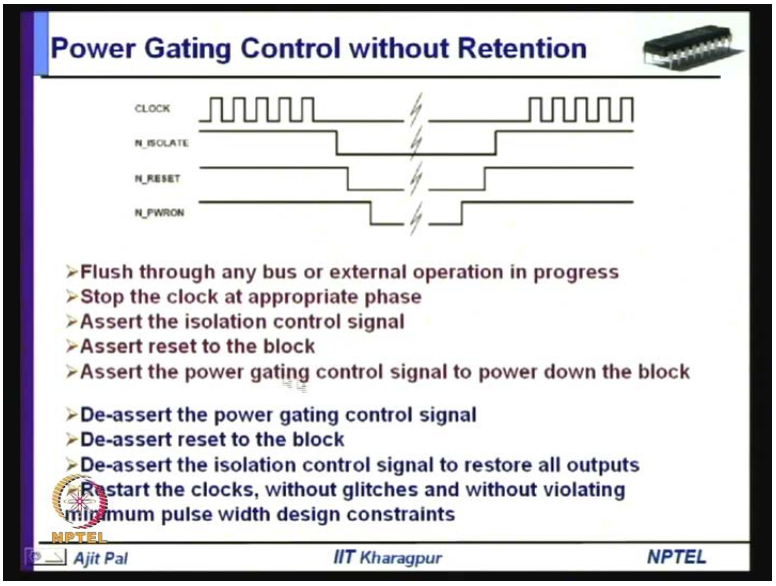
So, a more aggressive approach to turning on the switching fabric is to use several power-up control signal in sequence, so regardless of the specific control method during power up sequence; it is important to wait until the switching fabric is completely powered up before enabling the Power Gating block not to resume normal operation.

(Refer Slide Time: 54:45)




So, this is the requirement of the Power Gating controller. And let us I will look at the activity profile whenever you do the Power Gating, as you can see you are ramping down the supply and ramping up the supply. You are not abruptly changing the supply voltages, and this is the activity profile with realistic Power Gating.

(Refer Slide Time: 55:03)

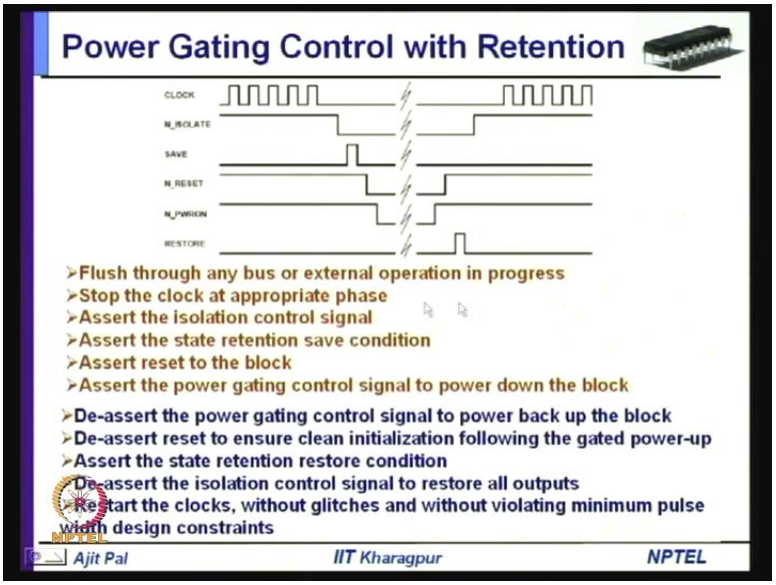



And the Power Gating controller which is will do we generate the sequence when Power Gating has to be done, you can see the signals are generated control signals are generated in sequence. First you are doing isolation signal, then you are doing reset after it has

been isolated, then you are doing Power Gating that means the switches are turned off, and now during this period the circuit is in power gated condition.

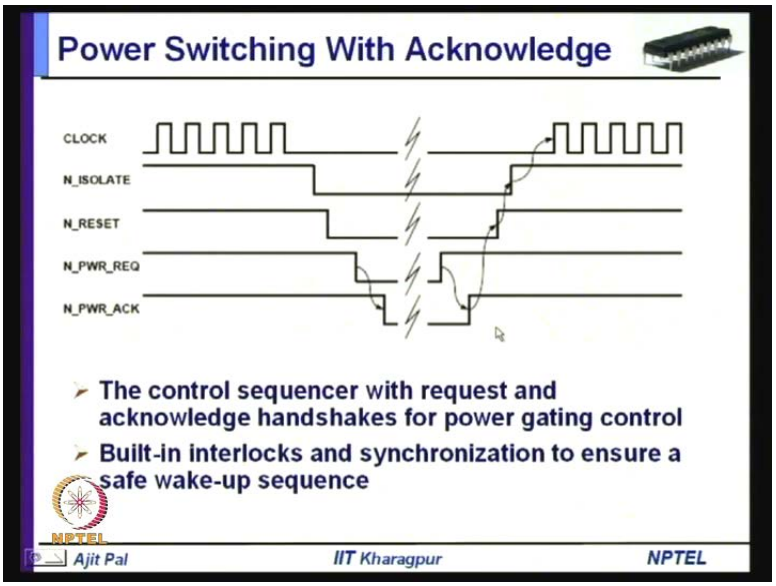
Similarly, you have to de-assert in the reverse sequence that means for the power switch is turned on, then you are doing the reset after it has stabilized then you will isolate and deactivate the isolate signal and start the clocking. So, first thing you have to do is to stop clock and you have to follow the sequence when you do when you go to Low-power mode, and this is the sequence you have to follow when you come back in the when you WAKE up when the circuit wakes up.

(Refer Slide Time: 56:01)




So, this can be done with retention also only difference is that you will requires two additional signals as you can see save and a restore rest of the thing is same. When you go to Low-power mode you have to use this additional signals same before you reset the hardware. And similarly, as you before you WAKE up you have to restore the resistors after turning on the power, then you have to apply the de-asserted the signals in the reverse order. First the power of the hardware then apply reset signal, then isolate the signal then apply clock.

(Refer Slide Time: 56:40)




So, this is the you have to follow the sequence, now whenever you are using this approach you are essentially that power management unit controller is interacting with the switches; the switches can have different delays or nothing depending on the realization actual realization, so they can have variable delays. So to take care of variable delays, you can work in a kind of you know asynchronies mode in which case there is a kind of hand shaking, as you can see after isolating you will generate the reset signal, then the request is send to the switching fabric to to for power down; obviously the switching fabric will take some time to power down it will switch it off, then send an acknowledgement signal.

Similarly, as as the the as you request whenever you have to turn the power on you have to send a signal to the switching fabric, and switching fabric will take some time to turn on and it will send a signal to other parts of the circuit it will send, it will generate the reset signal it will isolate. And then finally, clock is turned on so you have to it has got built-in interlocks, and synchronization to ensure safe WAKE up sequence.

So, we have discussed the the Power Gating technique in detail the various issues involved in Power Gating, we have discussed in detail. In the next lecture, we shall discuss other techniques for minimizing leakage power. Thank you.