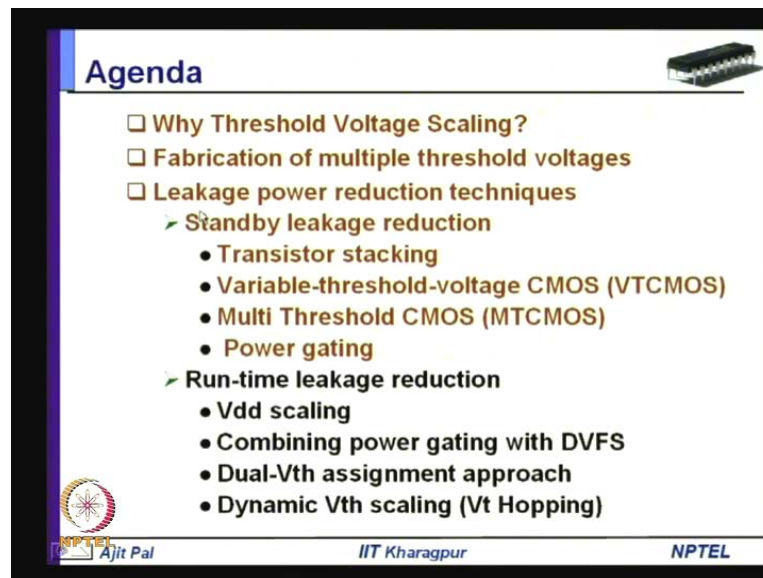


**Low Power VLSI Circuits and Systems**  
**Prof. Ajit Pal**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture No. # 34**  
**Minimizing Leakage Power – III**

Hello, and welcome to today's lecture on Minimizing Leakage Power; this is the 3rd lecture on this topic.

(Refer Slide Time: 00:27)



**Agenda**

- Why Threshold Voltage Scaling?
- Fabrication of multiple threshold voltages
- Leakage power reduction techniques
  - Standby leakage reduction
    - Transistor stacking
    - Variable-threshold-voltage CMOS (VTCMOS)
    - Multi Threshold CMOS (MTCMOS)
    - Power gating
  - Run-time leakage reduction
    - Vdd scaling
    - Combining power gating with DVFS
    - Dual-Vth assignment approach
    - Dynamic Vth scaling (Vt Hopping)

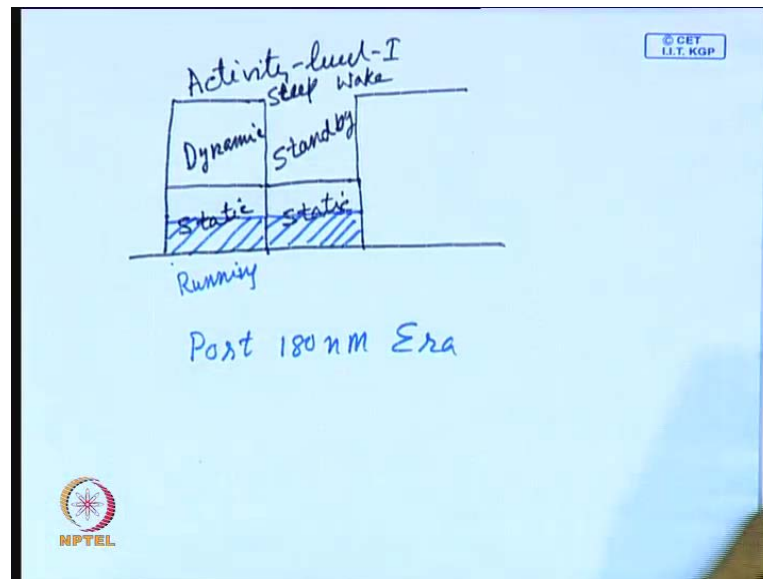
NPTEL Ajit Pal IIT Kharagpur NPTEL

In the earlier 2 lectures, we have discussed various issues related to leakage power reduction, and we have discussed various techniques, such as transistor stacking, variable threshold voltage CMOS, (VTCMOS) in short, multi-threshold voltage CMOS (MTCMOS), and power gating in the last couple of lectures. And these are essentially standby leakage reduction techniques. And today we shall focus on run-time leakage reduction, and discuss about various techniques like Vdd scaling, combining power gating with dynamic voltage and frequency scaling.

Then dual-Vth assignment approach and finally, dynamic Vth scaling and special case known as (Vt Hopping). Before we, proceed to discuss this technique, let me emphasize the important some run-time leakage reduction, and highlight the difference between

standby leakage reduction, and run-time leakage reduction.

(Refer Slide Time: 01:31)



As you know, when a particular circuit is in normal mode of operation, say this is the corresponds to activity level I, in such a case you have got both dynamic and static, which is essentially leakage power reductions. And then, when the circuit goes to sleep mode that means, here **you know** that sleep is activated and it goes to sleep mode, and in the standby condition, you have only the static power dissipation.

Now, **the** then again the circuit may be become active, initiated by wake and here it is, it goes **it goes** to sleep mode, I mean at this point, so in this way it, it goes on; now as you can see, when the circuit is not in operation, we may call it, it is in standby condition. And various techniques that we have discuss tried to reduce this static power dissipation, that means it can be brought down to say from this level to may be this level, by using various techniques, which I have discussed in the previous lectures.

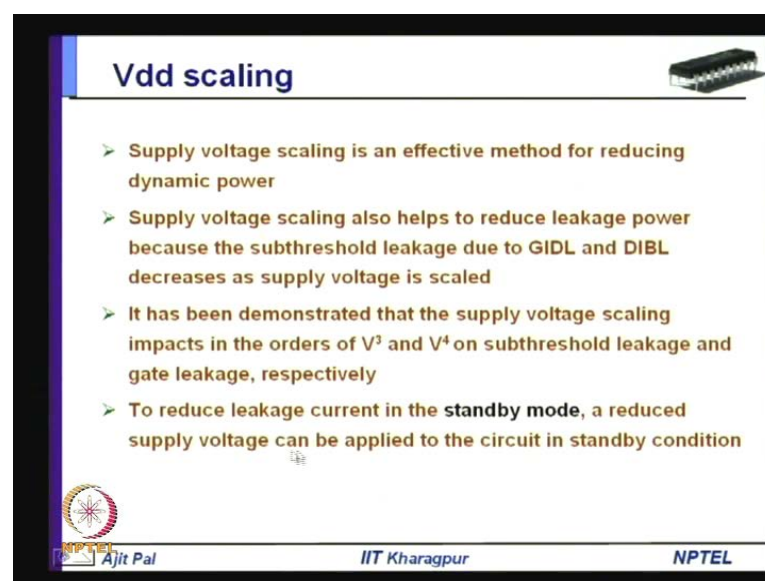
But, now, as we find that in deep sub-micron technology, we are finding that the static power dissipation which also take place when the circuit is in normal mode of operation that means, when the circuit is running, is in operation that time also there is static power dissipation. In the early technology generations that means, before 80 nanometer technology era, this static power dissipation was insignificant, and that is the reason why during normal mode or during run-time the dynamic power dissipation were minimized, and I have already discussed various techniques.

Now, in the **you know** post 180 nanometer eras, what we are observing, that static power dissipation is also significant, so what we have to do, when the circuit is in normal mode of operation or when the circuit is running, we would like to reduce the static power dissipation. As well that means, we would like not only we are interested in reducing the dynamic power, when the circuit is in running by clock gating and various other techniques, we are interested in reducing the leakage power, when the circuit is in run-time condition.

So, this is known as run-time leakage power reduction and today, we shall discuss various techniques, for minimizing this leakage power, when the circuit is in run-time condition. So, I hope you have understood the difference between standby leakage power reduction, and run-time leakage power reduction. Run-time leakage power reduction essentially concerns with the techniques of reducing the leakage power, when the circuit is in operation.





And as I mention, we have various techniques like, Vdd scaling, combining power gating with dynamic voltage and frequency scaling, dual Vt assignment, and dynamic Vth scaling. Let us, start with Vdd scaling, we have already discussed various techniques of supply voltage scaling techniques like, static supply voltage scaling, multi-level voltage scaling then, dynamic voltage and frequency and scaling, which are primarily used to reduce the dynamic power.

(Refer Slide Time: 05:22)



**Vdd scaling**

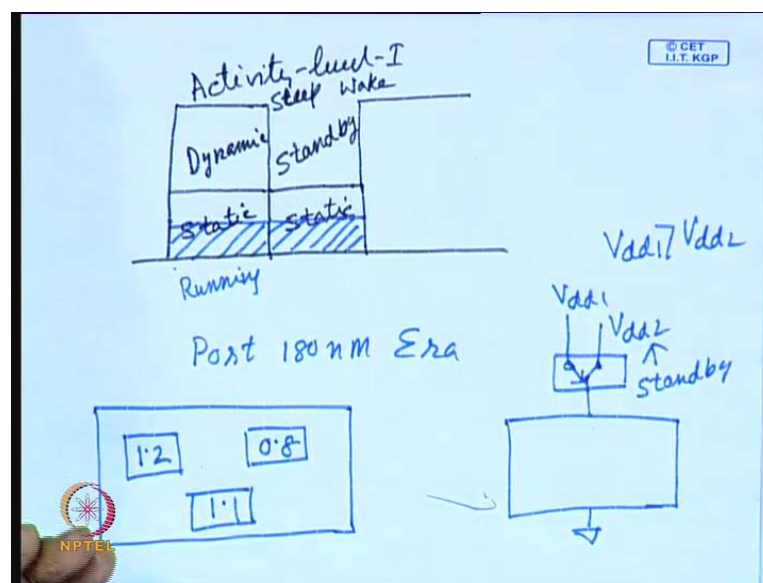
- Supply voltage scaling is an effective method for reducing dynamic power
- Supply voltage scaling also helps to reduce leakage power because the subthreshold leakage due to GIDL and DIBL decreases as supply voltage is scaled
- It has been demonstrated that the supply voltage scaling impacts in the orders of  $V^3$  and  $V^4$  on subthreshold leakage and gate leakage, respectively
- To reduce leakage current in the **standby mode**, a reduced supply voltage can be applied to the circuit in standby condition

  Ajit Pal  IIT Kharagpur  NPTEL

But, as we know that, as we do the supply voltage scaling, then it also helps to reduce leakage power, because of sub threshold leakage due to GIDL, Gate Induce Drain Leakage, and Drain Induce Barrier Lowering, because of these two phenomena. As you reduce the supply voltage, the leakage power reduction also occurs that means, whenever you are doing supply voltage scaling, you are not only reducing the dynamic power, but is also reduces the static power, because of these two phenomena that means, GIDL and DIBL.

So, it has been demonstrated, that the supply voltage scaling impacts in the order of  $V$  cube and  $V$  to the power 4 on sub-threshold leakage and gate leakage respectively. That means, we may consider supply voltage scaling, as a technique to reduce leakage power reduction, and for example, when the to reduce leakage current in the standby mode, a reduce supply voltage can be applied to the circuit in standby condition.

(Refer Slide Time: 06:33)



That means, suppose here, you have got a circuit during normal mode of operation, what you will do, you will pass the signal, through a switch and you can take from two sources, one is your say Vdd 1 and another is your Vdd 2. That means, when the circuit is in normal mode, you will apply Vdd 1 that means, this is Vdd 1 is greater than Vdd 2 that means, the circuit performance will be high.

So, but when the circuit is in standby, that means you have already done clock gating and other you have applied other techniques, so that there is no dynamic power dissipation,

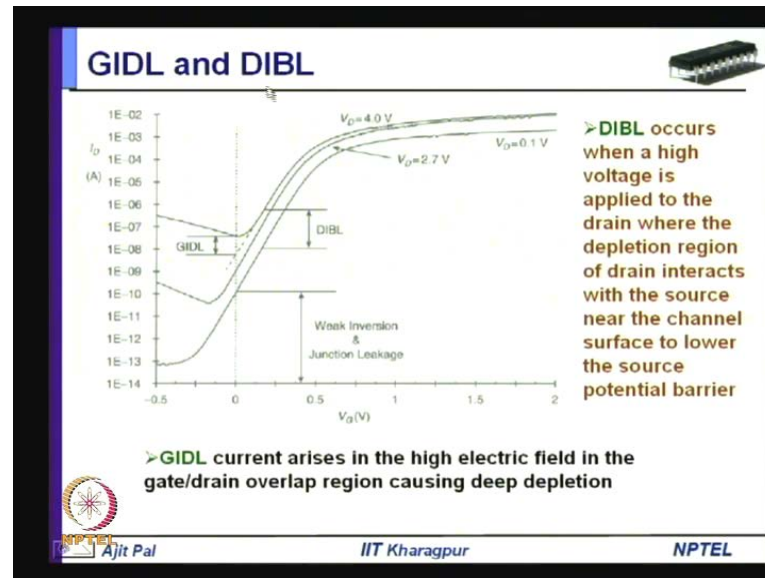
that time to reduce leakage power, what you can do, you can as well reduce the supply voltage; you can apply a lower supply voltage to it, when the circuit is in the standby condition. So, **the** when the circuit is in standby condition, then you apply this lower supply voltage, so this will effectively reduce the leakage power in the standby condition.

However, as I mention today, we are more interested in reducing the leakage power in run-time condition, the run-time leakage power reduction occurs, whenever you are using multi Vdd that means, in a single circuit **you know** as you have discussed earlier, in a single chip, you can have multiple voltage domain; say this may be operating at 1.2, this may be operating at 0.8, this may be operating at 1.1 (Refer Slide Time: 08:04). So, as you do this in the run-time condition leakage power of this particular circuit, which is operating in with voltage domain 0.8, will have lesser leakage power reduction.

So, you may consider it as kinds of you know by product, your main goal was to reduce the dynamic power, but you are also achieving reduction in leakage power. So, in your multi Vdd circuit, that means that is your, you can where you have got multiple voltage domains, there will be reduction in leakage power. Similarly, whenever you are doing dynamic voltage and frequency scaling, that time dynamically you are changing the supply voltage, along with frequency.

So, that time or primary goal is to reduce the dynamic power, but as you reduce the voltage, there will be reduction in leakage power. So, again it is a kind of bonus you are getting that means, primary goal is reduction of dynamic power, but leakage power reduction is also taking place; you are not really applying any specialized technique for that.

(Refer Slide Time: 09:22)



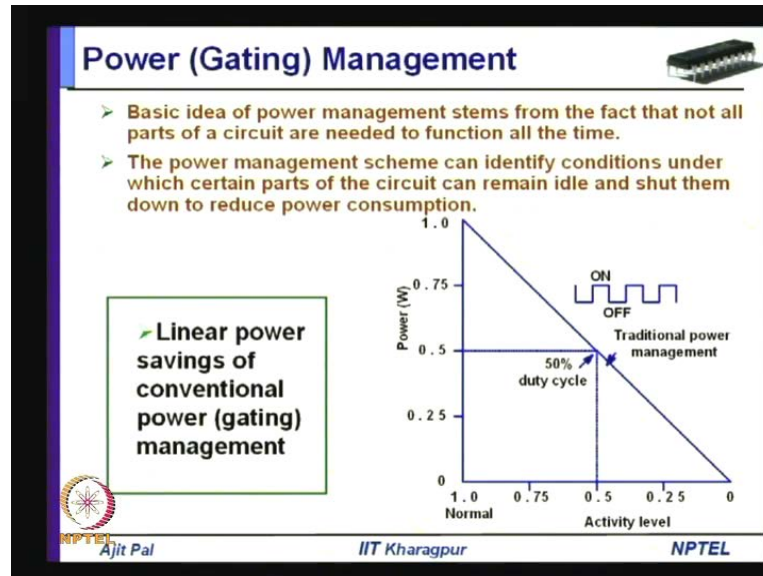
So, this happens, because of this GIDL and DIBL effect, so here these two are again shown, so DIBL occurs, when a high voltage is applied to the drain, where the depletion region of drain interacts with the source near the channel surface to lower the source potential barrier. So, you can see for the same gate voltage, you can have different current, drain current, because of this DIBL effect; that means, because of different supply voltages. So, this supply voltage corresponds to 0.1 volt whereas, this supply voltage corresponds to four volt.

So, as you can see as you reduce this supply voltage for the same gate voltage, there is significant reduction in the leakage current. Similarly, this is the gate induced drain leakage for the same, you know gate voltage you can see there is a reduction in the leakage current, even for the same drain voltage. So, here the drain voltage is same, but for the same gate voltage, I mean for different gate voltage you can have reduction in the leakage current, this will occur as you reduce the supply voltage obviously, at high level you will apply different gate voltage.

So, GIDL current arises in high electric field in the gate drain overlap region, causing deep depletion, I have already discussed these two effects, so we find that this dynamic  $V_{dd}$  scaling can be used to reduce leakage power. Now, let us consider a technique where we can combine power gating with dynamic voltage and frequency scaling, we

have already discussed power gating in detail, in my last lecture, where we have seen, how power gating can be done to reduce leakage power.

(Refer Slide Time: 11:11)

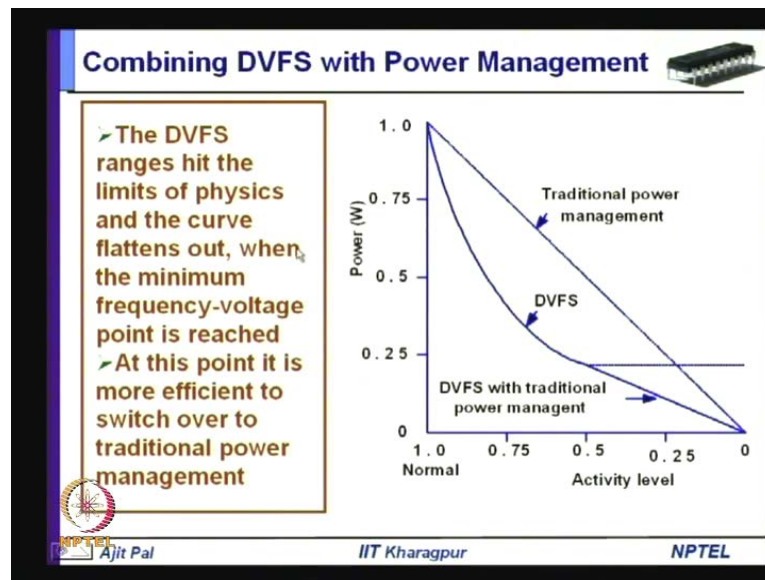


But, unfortunately, as we shall see this **this this this** particular power management or power gating **can be I mean**, we have already seen that this can occur only, when the circuit is in standby mode, that means circuit is in not operation, only then this can be applied. So, power management is essentially applicable, when **the** a particular sub-system or a system is not in operation, so you can do the power gating and reduce the supply, reduce the leakage power, you can achieve leakage power reduction.

But, as you have seen the dynamic voltage and frequency scaling can be applied in run-time condition, so can we not combine them together to achieve significant reduction in leakage power, so this is what is being tried and as you can see, whenever you do this power gating normally, there is a linear saving in power dissipation, depending on the activity level. If the activity level is 0.5 percent, there is a power saving of 50 percent.

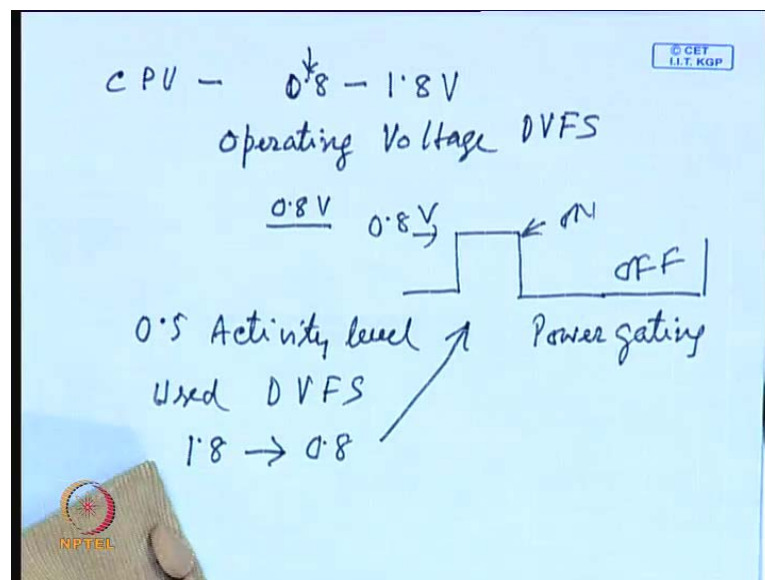
So, the reduction in power dissipation takes place linearly, depending on the activity level. So, depending on the performance required you can do, you can control the duty cycle of power gating and reduce power dissipation.

(Refer Slide Time: 12:41)



On the other hand, we can we have seen that in the dynamic voltage and frequency scaling, you can depending on the activity level; you can reduce the frequency as well as you can reduce the supply voltage, so that you can do dynamically.

(Refer Slide Time: 13:11)



So, what can however, this dynamic voltage and frequency scaling is applicable up to certain voltage why, the reason is we have seen that, a particular processor, any processor will operate as a CPU will operate over a free voltage range may be say 0.8 to may be 1.8 let us assume; this is the 0.8 to 1.8 volt, this is the operating over-rating range



of the processor, so operating voltage range. Now, what can be done depending on the activity level, you can adjust the frequency, and accordingly adjust the voltage to by using **dynamic voltage** dynamic voltage, and frequency scaling to reduce the dynamic power and as you have seen this will also the leakage power. Now, whenever this limit is reached 0.8 volt is the limit **you** is reached, then obviously, you cannot further reduce the supply voltage.

The reason for that is, if the supply voltage goes below this circuit, this processor will not function correctly that means, we have reached a point where the processor will fail to operate. In such a case, what can be done, we cannot really reduce the supply voltage beyond this level, and whatever power dissipation that takes place you have to accept, even when the activity level is much smaller. So, even when the activity level is smaller, you cannot really reduce and as a consequence the **the** for lower activity level, you can see the **the** reduction in power dissipation remains constant.

However, what you can do you will use dynamic voltage and frequency scaling up to this point, as the activity level reduces in this direction, you go keep on reducing the frequency and voltage, and you get a significant reduction compared to this linear reduction, which you can achieve by using this traditional power management (Refer Slide Time: 14:51). However, what you have to do at this point, you will resort to this **this** conventional power management.

So, maybe it will be linear, but still **you can reduce the power, leakage power** as you can reduce the leakage power during this condition that means, when the activity level is say may be in this condition 0.5 and lower, then you can reduce, the you can perform power management that means, you can keep on adjusting the duty cycle. So, for this duration the CPU is on and for this duration CPU is off, and here the voltage level obviously, will be kept at 0.8 volt, because this is the lower limit. So, this voltage will be switched, know up to say 0.5 activity level, you have used dynamic voltage and frequency scaling.

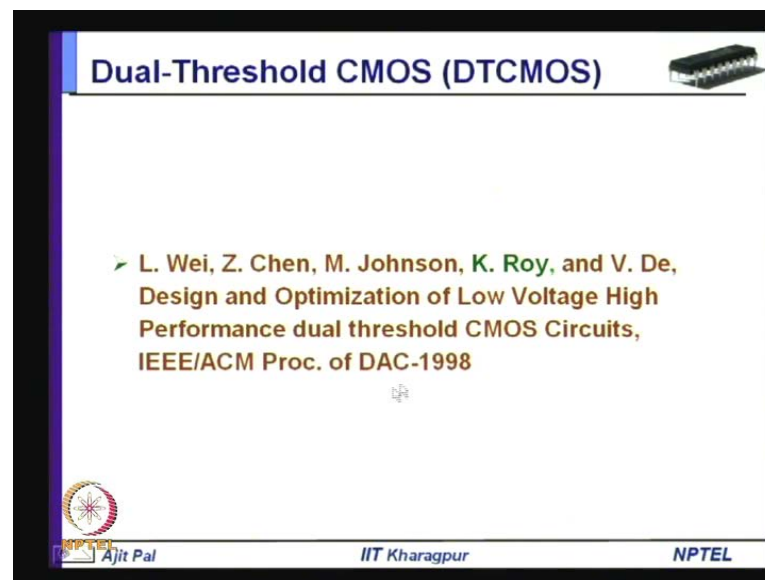
That means, and during this period what has happened, the supply voltage has reduced from may be 1.8 to it has come down to 0.8 since, the supply voltage cannot be further reduce what you do, you now use this power management or power gating. So, that means what you are doing you are combining two techniques, **up to this point** up to this point that means, which corresponds to the point where the voltage level has reached it is

limit that means, which is 0.8 in this particular example. And beyond this, since you cannot reduce the supply voltage beyond, this you will be doing power management, so you will control the duty cycle for power management that means, the on time of the CPU will be controlled depending on the activity level. So, at this point it is more efficient to switch over to traditional power management (Refer Slide Time: 17:11).

So, as a consequence you will find there will be significant reduction not only in dynamic power, but also in leakage power. So, here you have combine two techniques dynamic voltage and frequency scaling, and the complete tradition the power management or power gating to reduce the leakage power **over a over the entire range** over the entire range, by that I mean over the entire activity level.

Even when, the activity level very, is very close to 0, you can make the duty cycle very small, and as a consequence there will be reduction in leakage power. So, **this is** this is a technique where dynamic voltage and frequency scaling and power management have been combined.

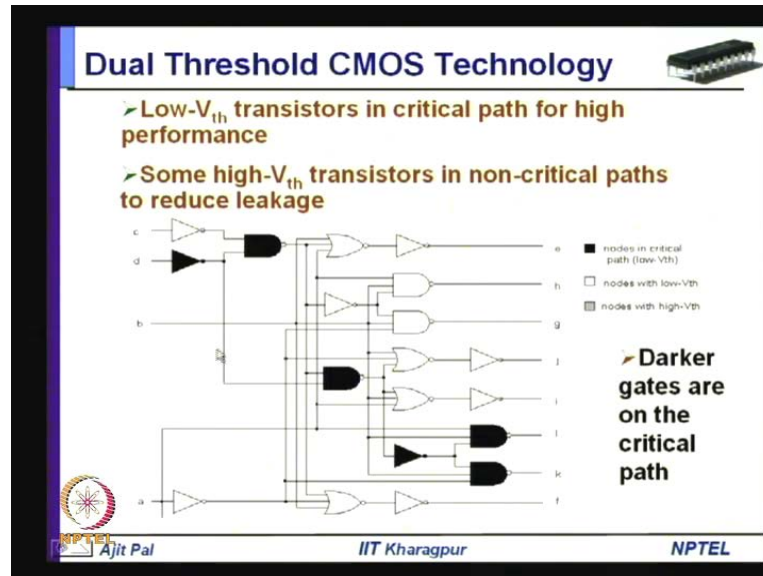
(Refer Slide Time: 18:18)



Now, we shall switch to another technique, which is known as dual-threshold CMOS (DTCMOS), the DTCMOS has been can be used to reduce leakage power in run-time condition, this approach was first proposed by paper, in a paper published in a conference, design auto-machine conference back in 1998, published by Kasha Roy and his group. These are the people working in the team, and the title of the paper was design

and optimization of low voltage high performance dual threshold CMOS circuits subsequently, this paper was also published in transaction on (0).

(Refer Slide Time: 18:56)



So, this was the first time this dual threshold voltage CMOS technique was introduced, and let me explain, the basic approach of these dynamic dual threshold CMOS or DTCMOS technique. So, here is a multi input, multi output net list shown here, and you can see you have got several inputs see this a, b, c, d and several outputs e, f, g, h and so on. So, and you have got n number of gates in this and these inputs may come from registers and this **out these** outputs may go to another set of registers; and in between you have got this multi input, multi output combinational circuit.

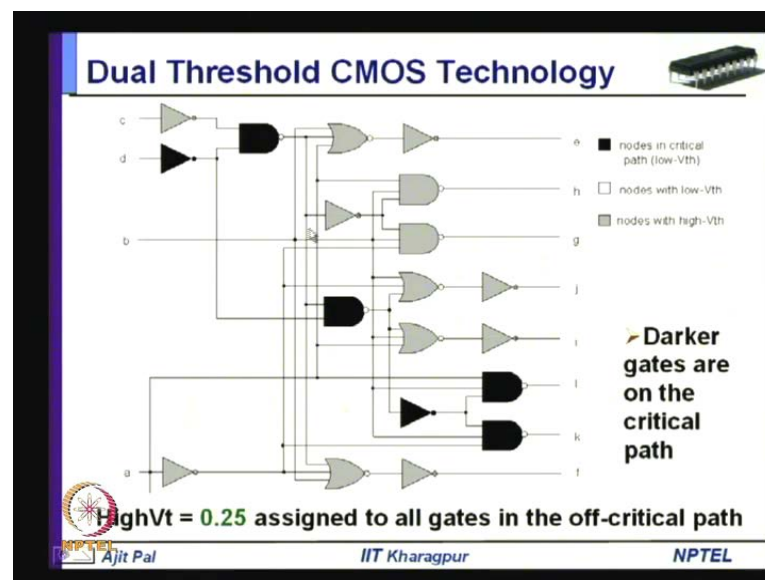
So, this multi input multi output combinational circuit will have a critical path, as you can see the gates, the darker gates are on the critical path, and the gates which are shown in dark form are on the critical path. On the other hand, the other gates which are not dark are on the non-critical path, so what can be done since, **you know** the performance is very important for the gates on the critical path, you can realize these gates, which are darker, I mean which are dark shown in this diagram by using low threshold voltage transistors.

So, as **you know** the low threshold voltage transistors will give you high performance, as **you know low  $V_{th}$**  low  $V_{th}$  corresponds to lower **sorry**, I mean it will be delay will be lower, **and high  $V_{th}$**  and high  $V_{th}$  corresponds to larger delay, **as we have** as we already

know. So, therefore, what you can do, you can assign gates of on the critical path a high  $V_t$  can be assign to them that means, the gates will be realized by low threshold voltage transistors. On the other hand, the gates on the non-critical path **as you know**, there is a slack that means, these gates can be realized by using transistors of high threshold voltage, of course, the delay will increase, but as long as the delay of the critical path is not exceeded, you can use high  $V_t$  transistors in realizing the gates on the non-critical path, so this is the basic approach.

So, low  $V_t$  transistors in the critical path for high performance and high  $V_t$ , some high  $V_{th}$  I mean, what will be the value of high  $V_{th}$  that we shall discuss, transistors in the non-critical path to reduce leakage, so this is the basic approach.

(Refer Slide Time: 21:47)



Now, let us see how the **the the the** number of gates that can be assign high threshold voltage changes by the choice of high  $V_t$ , so low  $V_t$  has been assumed to be 0.2 volt, 0.2  $V_{dd}$ . Here, it has been assumed that the supply voltage is 1 volt, so the low  $V_t$  is 0.2 volt, **as you know** the nominal threshold voltage lies in the range of 0.2  $V_{dd}$  to 0.5  $V_{dd}$ , so that is the range of  $V_{th}$  that you can have.

Now, this is considered to be low  $V_{th}$  now, the high  $V_{th}$  can be in the range of **I mean** which will be more than 0.2  $V_{dd}$  and obviously, it has to be less than 0.2  $V_{dd}$ , 0.5  $V_{dd}$ , so that will be in the range; but, what will be the exact value that we shall discuss (Refer Slide Time: 22:24). Now, the high threshold voltage in this particular example, I have

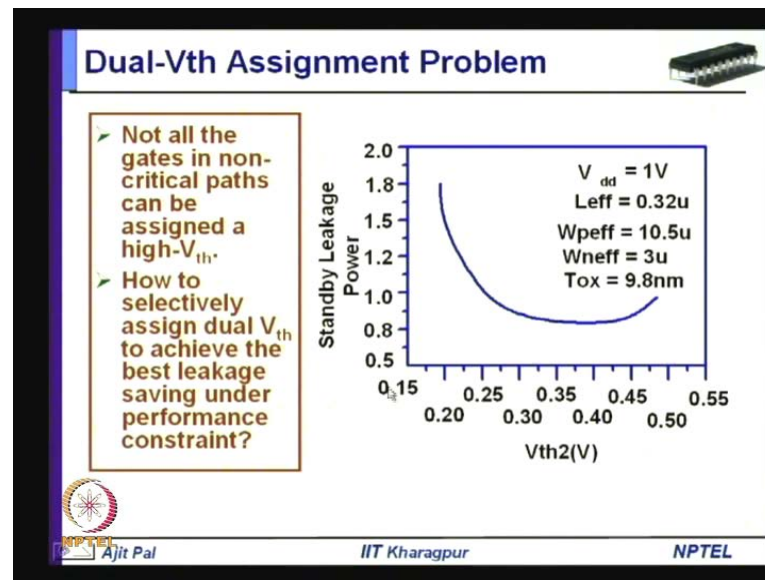
shown is slightly more than this low threshold voltage, so it is  $0.25 V_{dd}$ , so that is  $V_{dd}$  is 1 volt, so high  $V_{th}$  is 0.25. And as a consequence delay of the gates on the non critical path is slightly higher than the delay of the gates on the critical path, and as a consequence all these gates you can assign high  $V_{th}$  by that I mean, you can realize them by using transistors of high  $V_{th}$ . And the that that that the I mean at for none of the non critical paths, the critical path delay will be exceeded that means, as long as the delay of the critical path is not exceeded, you can use high  $V_{th}$ , so you will do that.

And so, all the gates as you can see, you can assign high  $V_{th}$ . Now, let us consider another situation where you have increased the high threshold voltage to 0.396, that means the threshold voltage is substantially higher than the low  $V_{th}$  which is 0.2 0.2 volt. So, but what is the consequence? The consequence is that some gates which are not dark, I mean these are lightly shaded those, lightly shaded gates, you can assign high  $V_{th}$ , but you can see these two gates you cannot assign high  $V_{th}$ , why you cannot use high  $V_{th}$  for them the reason for them.

If you use high  $V_{th}$  to those gates the, for that part the critical path delay will be exceeded, that means if you consider the path of this one say, it is coming from here, then it is going this way, so then it is going here (Refer Slide Time: 24:27). So, this path delay will exceed the critical path delay, if you assign low  $V_{th}$ , I mean high  $V_{th}$  to this particular gate. So, if you realize these two gates by using high  $V_t$  transistors then on these two paths the critical path delay will be exceeded obviously, we cannot really compromise in performance, and that is the reason why, we have to assign high  $V_t$  transistors in realizing these two gates.

So, if we find that the number of gates to which we can assign high threshold voltage is reducing, as we are increasing this high threshold voltage. Now, consider another situation, where high  $V_{th}$  is very close to the upper limit that is 0.5 volt, so in this particular case, we find that more number of gates cannot be assigned high  $V_{th}$  that means, you cannot realize them with transistors of high  $V_{th}$ . So, what will happen only fewer numbers of gates you can assign high  $V_{th}$  obviously, that will lead to reduction in leakage current, but some gates you cannot assign high  $V_{th}$ .

(Refer Slide Time: 25:54)



So, we find the number of gates is reducing I mean, to which you can assign high threshold voltage is reducing as you increase this threshold voltage. So, the you can have a plot like this, I mean low threshold voltage is 0.2 volt, and you can see the supply voltage is 1 volt and these are the various dimensions which are mention for which, I mean these are used in the for the purpose of simulation.

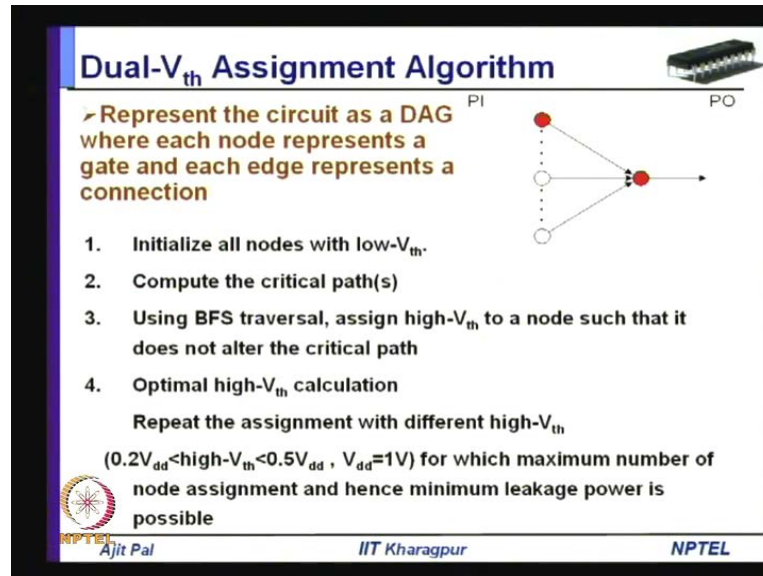
So, you find that as the second threshold voltage, high threshold voltage is increasing, we find that leakage power is initially decreasing because, we are able to assign high threshold voltage to gates on the non-critical path, but it is reaching a lower limit, I mean the leakage the standby leakage power is reaching a lower limit, the leakage power is I mean reaching a lower limit, then again it is increasing. Why it is increasing? Because, the number of gates to which you can assign high threshold voltage is reducing, so not all the gates in a non-critical path can be assigned a high  $V_{th}$ .

And how to selectively assign here, there is a mistake it will not be standby leakage actually, it will be run-time leakage power, so how to selectively assign dual  $V_{th}$  to achieve the best leakage saving under performance constraint.

So, question naturally arises you have a choice of high  $V_{th}$ , it can be in the range of 0.2  $V_{dd}$  to 0.5  $V_{dd}$ , but which one will give you minimum leakage power that you have to find out, and unfortunately that is np complete problem. So, what you can do, you can

use some heuristic base algorithm that was **what was** done by kasha Roy and his group and this is the basic algorithm.

(Refer Slide Time: 27:37)



**Dual- $V_{th}$  Assignment Algorithm**

➤ Represent the circuit as a DAG where each node represents a gate and each edge represents a connection

1. Initialize all nodes with low- $V_{th}$ .

2. Compute the critical path(s)

3. Using BFS traversal, assign high- $V_{th}$  to a node such that it does not alter the critical path

4. Optimal high- $V_{th}$  calculation

Repeat the assignment with different high- $V_{th}$

( $0.2V_{dd} < \text{high-}V_{th} < 0.5V_{dd}$ ,  $V_{dd}=1V$ ) for which maximum number of node assignment and hence minimum leakage power is possible

PI PO

Ajit Pal IIT Kharagpur NPTEL

Initialize all gates with low  $V_{th}$ , then first you have to represent the circuit as a directed acyclic graph that means, each node will represent a gate, and each edge will represent an inter-connection between the gates. So, after this is done, in this intermediate representation he will perform, he will use this algorithm initialize all gates with low  $V_{th}$ , then compute the critical paths, use breath first search traversal, assign high  $V_{th}$  to that to a node such that, it does not alter the critical path.

So, you will assign high  $V_{th}$  to all the nodes, which are not on the critical path and you will keep on doing that, such that the critical path delay is not exceeded. So, in this way you will compute high  $V_{th}$ , so it is essential you will keep on increasing the threshold voltage for gates on the non-critical path; and then you will do the measurement of the leakage power. So, repeat assignment with different high  $V_{th}$  in the range  $0.2 V_{dd}$  to  $0.5 V_{dd}$  with supply voltage of 1 volt, for which minimum number of node assignment, and hence minimum leakage power is possible, so this is how, you will achieve leakage power reduction.

(Refer Slide Time: 29:02)

**Dual- $V_{th}$  Assignment: Another Approach**

□ N.Tripathy, A.Bhosle, D. Samanta and A. Pal, "Optimal Assignment of High Threshold Voltage for Synthesizing Dual Threshold CMOS Circuits", Proc. VLSI Design 2001, pp. 227-232, Bangalore, January 2001

Ajit Pal IIT Kharagpur NPTEL

And subsequently the algorithm that I have discussed has been extended by some of the students N.Tripathy, A.Bhosle, D.Samanta they they were some of my students, and they are they use the different algorithm.

(Refer Slide Time: 29:21)

**Delay-Constrained Dual- $V_T$  Assignment**

➤ Assume the circuit as a DAG where each node represents a gate and each edge represents a connection

Algorithm:

1. Assume  $low-V_T < high-V_T < 0.5V_{DD}$
2. Initialize all nodes with high- $V_T$
3. Compute the critical path(s)
4. Using DFS traversal, assign low- $V_T$  to a node on the critical path
5. Go to Step 3 until all the nodes on the critical path are assigned with low- $V_T$

PI PO

From POs To PIs

Ajit Pal IIT Kharagpur NPTEL

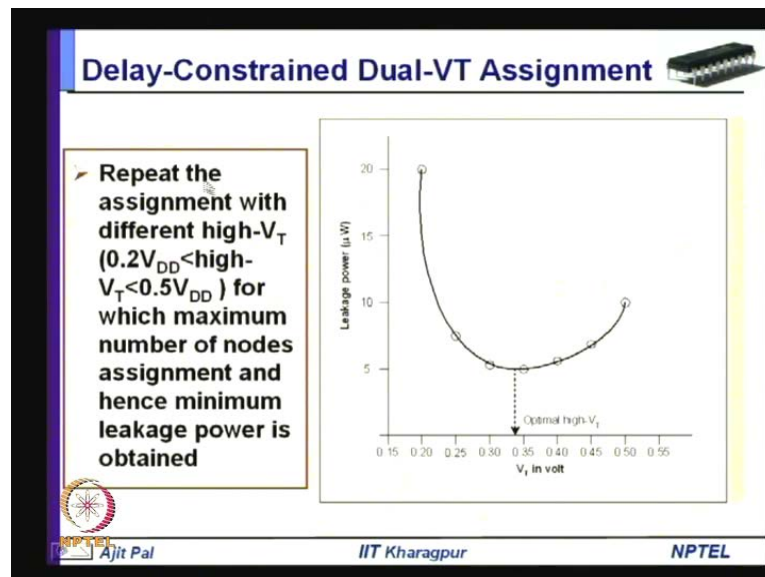
So, here the algorithm is you can see here also, it is assumed that low  $V_t$  and high  $V_t$  is less than  $0.5 V_{dd}$ . So, earlier in the previous algorithm, we have seen initially all the nodes are assigned with low  $V_t$ , but here what is being done all the gates are assigned with high  $V_{th}$ . So, initially all the gates are assigned high  $V_{th}$ , then we are computing the



critical path and then using depth-first search, early they have used breadth first search that means, simultaneously all paths are selected, but here it is depth first search traversal and assign low  $V_t$  to a node on the critical path.

So, the basic difference is earlier initially all the gates are assigned low  $V_t$ , then you are assigning high  $V_t$  to the non **critical** gates only non critical path here, the it is that done the other way, you are initially assigning high  $V_t$  to all the gates, then you are assigning low  $V_t$  to gates on the non on the critical path. So, you will go to step until all the nodes on the critical path are assigned with high low  $V_t$ .

(Refer Slide Time: 30:32)



So, using this approach particularly this **this this** assignment of high  $V_t$  is repeated in this range for which maximum number nodes assignment occurs, and hence minimum leakage power reduction take place; and this is how you find the optimal  $V_t$ , and for a particular example, we find that optimal high  $V_t$  is close to 0.35 volt.

(Refer Slide Time: 30:58)

### Experimental Results

Comparison of our results with [Wei+99]

Benchmark	With approach [Wei+99]				Our approach			
	#Transistor	%Redn in standby leakage power	%Redn in total power	CPU time (s)	#Transistor	%Redn in standby leakage power	%Redn in total power	CPU time (s)
C432	278	59.65	16.52	20	348	87.35	26.17	36
C499	604	51.09	7.18	118	796	64.45	12.68	174
C880	1126	84.87	14.07	55	1208	88.65	19.41	89
C1355	1232	49.36	8.51	198	1346	59.95	13.15	346
C1908	1430	76.21	15.46	225	1684	83.45	21.75	412
C2670	2736	81.24	19.27	269	3092	92.96	24.80	485
C3540	3430	85.60	21.43	301	3698	90.42	32.29	541
C5315	5432	83.12	18.44	342	5516	89.69	31.05	619
C6288	5768	43.38	19.89	564	8950	83.69	45.42	890
C7552	7102	76.41	20.35	387	7786	87.65	22.36	609
		69.01%	16.11%			82.82%	24.92%	

Ajit Pal IIT Kharagpur NPTEL

So, in this way the high  $V_t$  for a large number of **I mean**, it was tried for a large number of benchmark circuits, this was the result of the first approach that was the first paper that I mentioned, where they were able to reduce the run-time, **I mean** leakage power reduction was 69.01 percent. On the other hand, in our case always in a standby leakage power is mention, but this will be all run-time leakage power. So, here we find the second approach the reduction in leakage power is 82.82 percent.

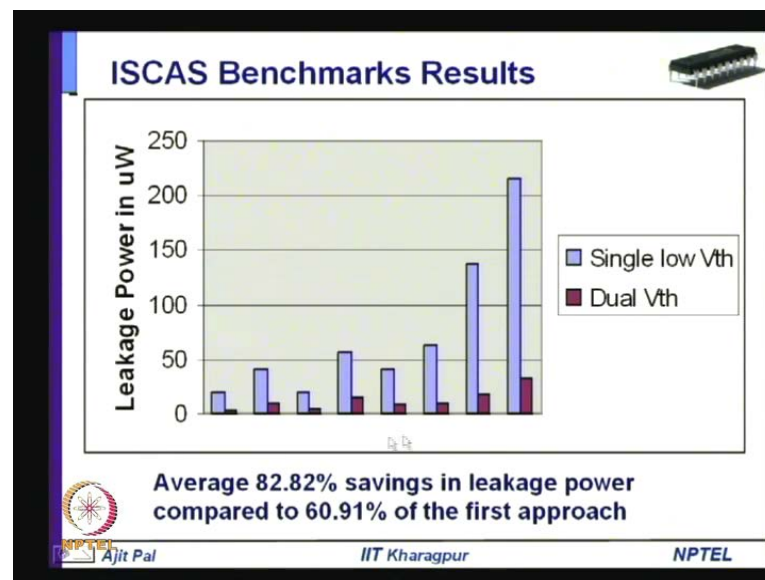
So, much larger reduction is occurring, and primarily due to **you know**, a larger number of transistor assign are assign high  $V_{th}$ . So, you can see here, number of transistors to which the high threshold voltage can be assigned is 278 and here it is 348 and so on (Refer Slide Time: 31:48). So, we find that more and more number of numbers of transistors are assigned with high  $V_t$ , and as a consequence you are able to achieve larger reduction in leakage power.

However, the complexity of the two algorithms are different, the first paper that I mentioned has a linear complexity, because it is doing breadth-first search, and as a consequence the run-time **you know** is smaller, compared to this approach were the complexity of algorithm is  $n$  square. And as a consequence it will have larger run-time, but however we find that, **the** there is definitely increase in run-time, but not significant increase, so about 24.92 percent increase in run-time is occurring.

And as a consequence, **I mean sorry** this is the percentage of leakage reduction, this and these are the run-times 36 instead of 20, 174 instead of 180 second and so on, so there is definitely increase in run-time (Refer Slide Time: 32:46). But, **you know** you will be running the program for each of them, each of these benchmarks only once; and once the circuit is synthesized, and you will keep on getting the benefit throughout the lifecycle of the circuit.

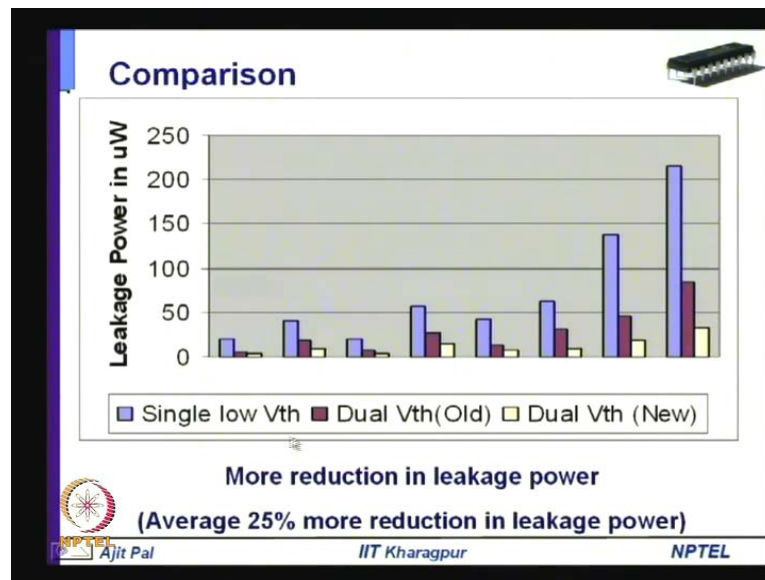
So, run-time, **I mean the** all the competition time is increasing this is really a relevant early event in the present the context, and this approach gives you significant reduction, so from 69 percent to 82.82 percent reduction in leakage power.

(Refer Slide Time: 33:31)



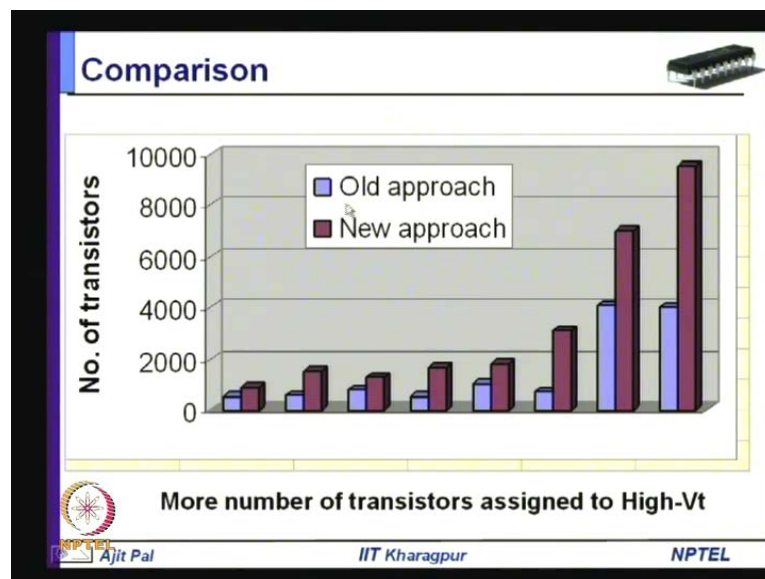
Subsequently, these are the same **same** result shown in bar graph form pi chart form, so here you can see, this is the comparison between single low Vt with dual Vth. So, you can see there is significant reduction in leakage power 82.82 percent.

(Refer Slide Time: 33:54)



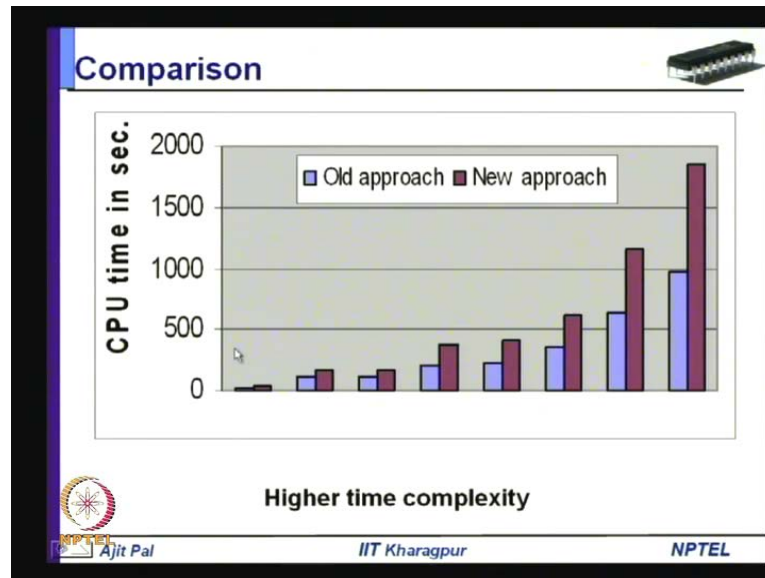
And then here, the reduction has been compared three cases single low  $V_t$ , dual  $V_t$ , the previous approach and the second approach that I have mention. So, this yellow one is you can find significantly lower compared to the single low  $V_t$  realization and also the previous approach. So, **average** on the average 25 percent more reduction in leakage power occurs.

(Refer Slide Time: 34:21)



And here, the numbers of transistors to which you can assign high  $V_t$  have been compared between the two approaches **between the two approaches**.

(Refer Slide Time: 34:31)



And here the CPU time as I told is more for the second approach, because of higher time complexity.

(Refer Slide Time: 34:40)

**Energy-Constrained Dual-VT Assignment**

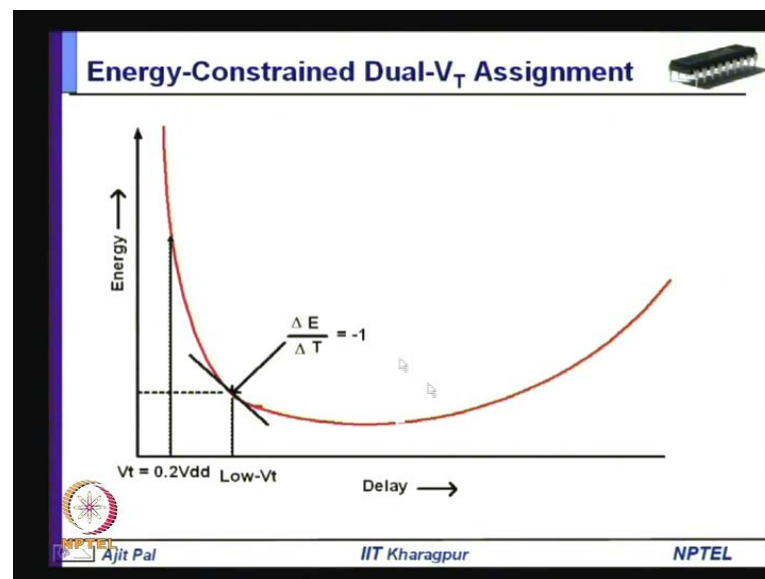
➤ D. Samanta, and A. Pal, *Optimal Dual-VT Assignment for Low-Voltage Energy-Constrained CMOS Circuits*, Proc. 7th ASP-DAC/15th VLSI Design Conference, pp. 193-198, Bangalore, January 2002

Ajit Pal IIT Kharagpur NPTEL

There was another technique, which was proposed where which is known as energy constant CMOS circuit that means, in the previous case we may consider it as a delay constraint, delay constraint by that **I mean**, delay is not allowed to increase, we are keeping the delay of the original circuit same **and by**. And then keeping the same delay we are trying to optimize the number of transistors to which you can assign high

threshold voltage, so that there is reduction in leakage power. Now, what can be done in case of energy constraint approach, we are primarily interested in reduction in energy rather than keeping the delay **I mean**, minimum to the minimum value. So, there are situations where you can have small compromise, little compromise on delay and in such a case, **as we** as we shall see you can have more reduction in leakage power at the cost of **slightly** a slight increase in delay, so that is the basic approaches of this of this second the third paper third approach.

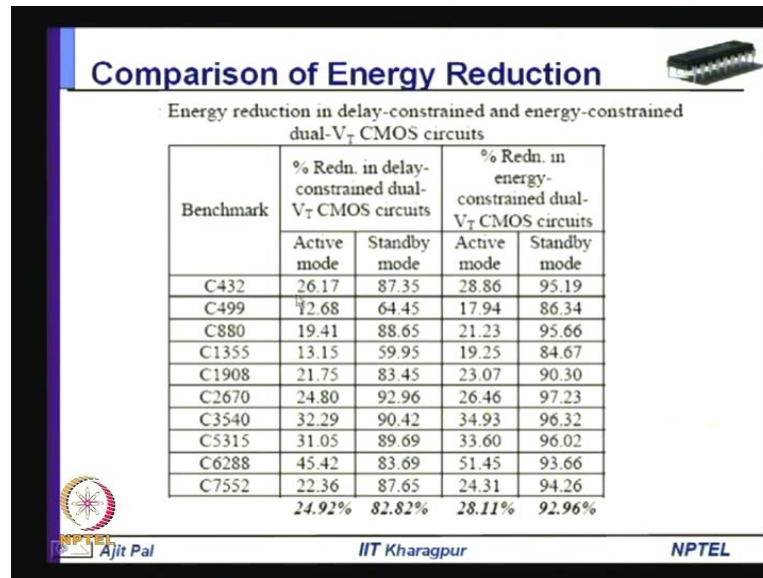
(Refer Slide Time: 35:46)



Here, as you can see the energy has been plotted for different threshold voltage, so whenever you keep the low threshold voltage to **0.2 Vdd** 0.2 Vdd. Here, is the energy requirement of the circuit **I mean**, but if you keep on changing the threshold voltage of the circuit you can see, this is how the delay increases.

Now, you can choose a point where delta E by delta T is equal to minus 1 that means, this point corresponds to delta E by delta T is equal to minus 1, where **you know** you can get reduction in **I mean** larger reduction in delay, but **small increase in sorry** small increase in delay, but larger reduction in energy. So, this point is used is chosen for different circuits is found out for obviously this point will different for different circuits; and this point can be found out, and after this is found out instead of considering low Vth 0.2 Vdd, you can use this voltage as the low Vt that is the basic approach.

(Refer Slide Time: 37:01)



### Comparison of Energy Reduction

Energy reduction in delay-constrained and energy-constrained dual- $V_T$  CMOS circuits

Benchmark	% Redn. in delay-constrained dual- $V_T$ CMOS circuits		% Redn. in energy-constrained dual- $V_T$ CMOS circuits	
	Active mode	Standby mode	Active mode	Standby mode
C432	26.17	87.35	28.86	95.19
C499	12.68	64.45	17.94	86.34
C880	19.41	88.65	21.23	95.66
C1355	13.15	59.95	19.25	84.67
C1908	21.75	83.45	23.07	90.30
C2670	24.80	92.96	26.46	97.23
C3540	32.29	90.42	34.93	96.32
C5315	31.05	89.69	33.60	96.02
C6288	45.42	83.69	51.45	93.66
C7552	22.36	87.65	24.31	94.26
	<b>24.92%</b>	<b>82.82%</b>	<b>28.11%</b>	<b>92.96%</b>

Ajit Pal  
IIT Kharagpur  
NPTEL

And you can see, the reduction in energy that can occur by using this energy constrained dual  $V_{th}$  approach, so the first one as I told is delay constrained approach, where as I mention there is a reduction in leakage power by 82.82 percent. On the other hand, whenever it is, in case of energy constrained approach you can see, there is a reduction in a in leakage power by 92.96 percent. So, you can see, more than 10 percent reduction in leakage power you can achieve, if you can of course, except small increase in delay.


It has been found that, the delay increases by 4 to 9 percent of different circuits that means, **if you consider different benchmarks** we have considered different benchmarks, so the delay increase was found to be 4 percent to 9 percent. So, with this small increase in delay, we are able to achieve significant reduction in energy dissipation due to leakage.


So, obviously, this can be used were you can compromise performance by a small amount 4 to 9 percent, and you can achieve reduction in leakage power, so this is the energy constrained approach. Now, let us come to another approach where we shall use dynamic  $V_{th}$  scaling earlier, we have discussed a technique dynamic  $V_{dd}$  scaling, now we shall consider dynamic  $V_{th}$  scaling.

(Refer Slide Time: 38:40)

### Dynamic Vth scaling

- Just like dynamic the Vdd scaling scheme, a **dynamic Vth scheme (DVTS)** can be used to reduce runtime leakage power in **sub-100-nm** generations, where leakage power is significant portion of the total power at runtime
- When the workload is less than the maximum, the processor is operated at lower clock frequency. Instead of reducing the supply voltage, the DVTS hardware raises the threshold voltage using **reverse body biasing** to reduce runtime leakage power
- Just enough throughput is delivered for the current workload by dynamically adjusting the Vth in an optimal manner to maximize leakage power reduction



 Ajit Pal IIT Kharagpur NPTEL

So, just like dynamic Vdd scaling scheme, a dynamic Vth scheme can be used to reduce run time leakage power in sub 100 nanometer generations, where the leakage power is significant portion of the total power at run time. So that means, this approach can be used only when the leakage power is significant, so leakage power may be more than 50 percent of the total power in such case, what can be done instead of using dynamic Vdd scaling.

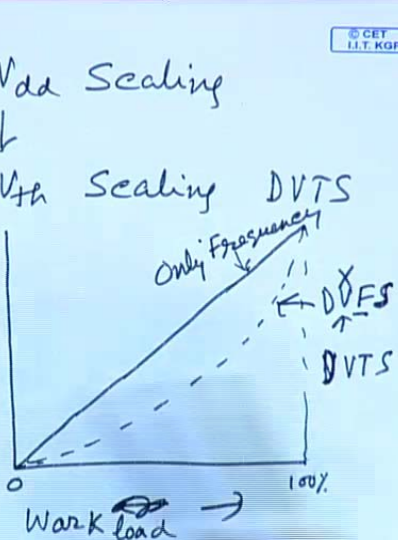
(Refer Slide Time: 39:12)


### Dynamic Vdd Scaling

↓

### Dynamic Vth Scaling DVTS

Lower F  
↓ Vdd → ↑ Delay  
↑ Vth → ↑ Delay



 © CET  
I.I.T. KGP



We shall be using dynamic  $V_{th}$  scaling and it is known as dynamic  $V_{th}$  dynamic threshold voltage scaling. And the basic approach is same, when the workload is less than maximum the processor is operated at lower clock frequency instead of however, instead of reducing the supply voltage the dynamic threshold voltage DVTS hardware, dynamic threshold voltage system hardware raises the threshold voltage using reverse body biasing to reduce runtime leakage power.

So, earlier what we have seen if we plot, so let us assume **this is the** let me plot this way this is the workload, so workload let us assume here it is 0, here it is 100 percent and earlier, we have seen as the workload is reduced **I mean**, if only frequency is scaled the power dissipation can be reduced in this way. So, here by controlling the frequency only frequency scaling, and you were able to reduce the dynamic power by changing the threshold voltage as well as frequency, so this was your DVFS dynamic voltage and frequency scaling.

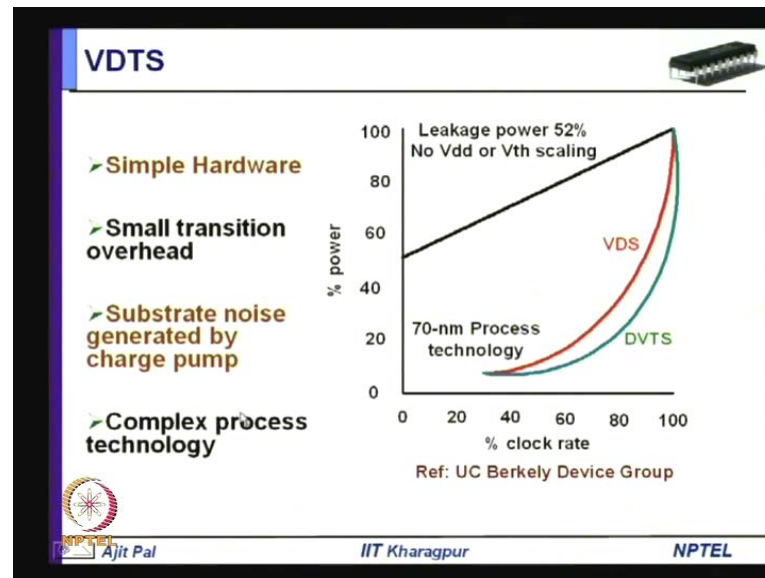
Now, instead of this dynamic voltage and frequency scaling what we will be doing, as you reduce the frequency, you will not reduce the voltage, **instead of** that instead of reducing the voltage you will reduce the threshold voltage  $V_{th}$  dynamic  $V_{th}$ . So, how we will do that, what was the basic purpose of reducing the supply voltage **as you know** as you reduce the supply voltage delay increases, you also know that as you reduce the threshold voltage delay increases that means, the effect is same **higher Vdd sorry** lower Vdd means larger delay and higher  $V_{th}$  means larger delay.

So, they have the same effect that means, as you reduce the frequency, so lower frequency you were using lower Vdd to increase the delay to match the lower frequency. Now, what you will be doing as you lower the frequency, you will increase the threshold voltage to match the lower frequency, so **we** you are achieving the same thing, as you have done in dynamic voltage and frequency scaling, but instead of changing the voltage here you are reducing the supply voltage you are increasing the threshold voltage.

So, what you will be doing just enough through to throughput is delivered for the current workload by dynamically adjusting the  $V_{th}$  in an optimal manner to maximize leakage power reduction. So, what you will be doing as you reduce the frequency, you will increase the threshold voltage such that, the circuit just operates. If you increase the threshold voltage, more **the the** it will the circuit will fail to operate at that frequency. So,

in this way you will maximize the leakage power by increasing the threshold voltage, so that is the basic approach of this dynamic **Vt** Vth scaling.

(Refer Slide Time: 43:13)



And as I have already explained, this plot was done by the device group of U C Berkley, what they did, they assumed **they** this is for a 70 nanometer process technology, in this particular case, it has been found that leakage power is about 52 percent of the total power. So, in such a case, if you do not do threshold voltage, supply voltage scaling or threshold voltage scaling, then **if you simply** if the **if the** as the workload is reduced, you can reduce the clock rate, and you can see by reducing the clock rate, the power dissipation reduces linearly, that dynamic power reduces linearly.

So, **it** you can achieve roughly fifty is it is close to 52 percent, because linear reduction will occur, so 48 percent reduction in dynamic power will take place, but if you use dynamic voltage scaling, along with frequency scaling you can see, you can achieve significant reduction, but you cannot go beyond this thirty percent. Because, **you know** you have reached the limit of the supply voltage that means supply voltage cannot be further lowered beyond this point (Refer Slide Time: 44:29).

So, here the clock rate is 30 percent, and there is a corresponding supply voltage and you can see there is a significant reduction in power dissipation; and obviously, both dynamic power and leakage power is reduced. But, since the leakage power component is more instead of using dynamic voltage scaling, they simulated **they simulated** this for dynamic

threshold voltage scaling and you can see, there is a larger reduction in total power dissipation; if you do dynamic threshold voltage scaling that is DVTS, that is dynamic threshold voltage scaling. And it has been found that this dynamic threshold voltage scaling requires simple hardware compared to dynamic voltage scaling, we have already discussed the hardware that you require for implementing dynamic voltage scaling, you will require a that DC to DC converter you will require phase locked loop for frequency generation and so on.

But, in this particular case **you will require to** you will require much simpler hardware only thing that you have to do, you have to change the threshold voltage by controlling the body biased of the circuit, and body biased can be changed by using **you know** that means, it can be increased by applying reverse body biased using a technique known as charge pump. So, using a technique known as charge pump, the threshold voltage, I mean reverse body biasing can be increased to reduce leakage power; and it has been found that the hardware that you require is simpler compared to DVS approach.

And the transition overheads as you switch from **you know** one particular frequency threshold voltage pair to another frequency voltage **you know** you will, in this particular case what you will be doing, you will be operating at a particular frequency  $F_1$  and a **and**  $V_{th1}$ , then you will switch to depending on the performance required you will switch to another frequency  $F_2$  with a threshold voltage  $V_2$  (Refer Slide Time: 46:26). So, as you switch from one frequency threshold voltage pair to another frequency threshold voltage pair, **you know** it has got some delay overhead, we have seen there is delay overhead whenever you use dynamic supply voltage scaling.

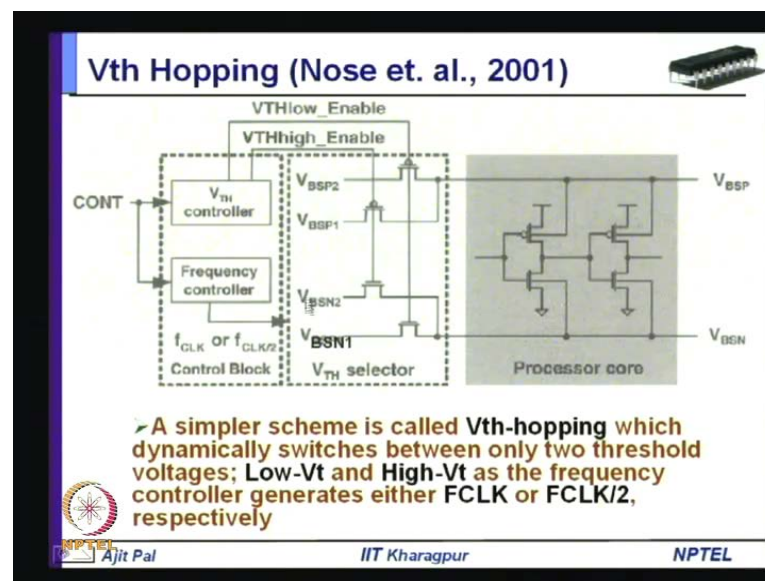
And because, the power supply has **you know has I mean** require quite large time to settle down because **you know**, it has got linear circuits capacitors and so on, so and as a consequence you can switch from one voltage frequency pair to another voltage frequency pair. And that particular delay is quite large compared to this change, **where the** where you are switching from one volt frequency, and threshold voltage pair to another voltage and frequency threshold voltage pair.

So, there is small transition overhead that has been found however, in this world nothing is one sided it has some drawbacks, number one drawback is substrate noise generated by charge pump as I told, the threshold voltage will be changed by using a technique

known as charge pump by which you will be changing the reverse body biased. And that actually leads to some kind of noise in the circuit known as substrate noise, because you are changing the substrate body biasing and this will lead to substrate noise. And so, it this noise will be present whenever you use dynamic volt threshold voltage scaling.

And also the process technology that will be used here is complex, as we have already seen, whenever you are using multiple threshold voltages, then the not only the substrates and the wells are to be separated to apply the different supply voltages, you have to you will require additional hardware to apply the different body biased voltages, and you have to use triple well technology. So, triple well technology is definitely a very complex process technology, and that is the price you have to pay to reduce the leakage power by using dynamic threshold voltage scaling.

(Refer Slide Time: 48:55)



Now, a simpler alternative **is to** is to have a simpler scheme called Vth hopping which dynamically switches between only two threshold voltages, **you know** we have seen that dynamic threshold voltage scaling. There you can have a large number of frequency threshold voltage pairs, discrete voltage and threshold frequency and threshold voltage pair instead of having large number of voltage and **sorry** frequency, and threshold voltage pairs; you will be using only two frequency and threshold voltage pair.

So, in this case, this is known as Vth hopping, it has been proposed by nose at all sometime in 2001 which dynamically switches between only two threshold voltages; low

$V_t$  and high  $V_t$  as the frequency controller generates, either  $F$  clock and  $F$  clock by 2 (Refer Slide Time: 49:32). So, you can see, you will be using only two clock frequencies,  $F$  clock or half of it  $F$  clock by 2 and accordingly, you will be using low  $V_t$  obviously, for higher frequency you will require, lower threshold voltage and whenever you reduce the clock frequency by 2, you can use higher threshold voltage.

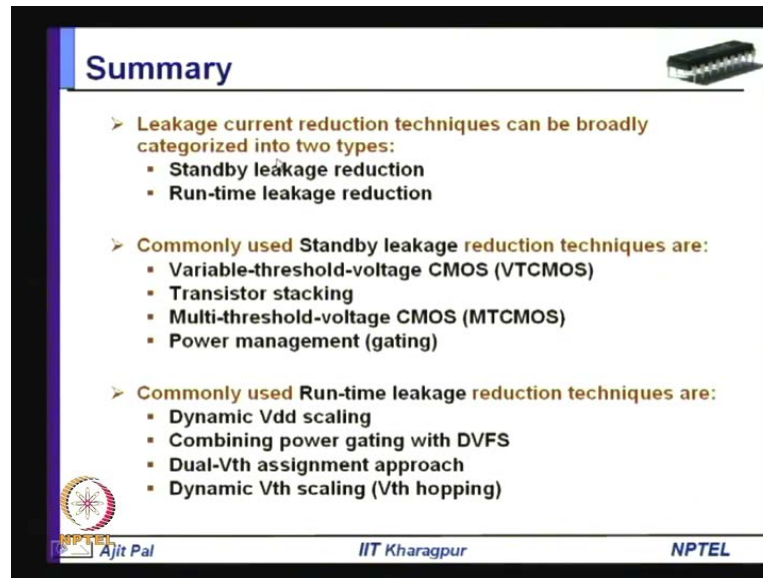
And that means, the switch depending on the activity level, required by the application you will be switching to  $F$  clock from between these two  $F$  clock low  $V_t$  pair to  $F$  clock by 2 high  $V_t$  pair. So, between these two will be switching and here is the corresponding hardware that you will require, so here you can see the control signal, may be coming from the operating system.

So, operating system will send a control signal and  $V_{th}$  controller will select, either a high threshold voltage for the p mos transistor, so high  $B_{SN2}$  and that time, we will be using high  $B_{SN2}$  that means, high threshold volt that means, the body biased can be say, will may correspond to 0 body biased. When the application require high performance and you can use reverse body biased that is, your  $V_{SP1}$  and  $V_{SN1}$  that means, here it will be more positive and here it will be more negative, and that will be done with the help of this  $V_{th}$  selector.

So, you are selecting these voltages with the help of switches, and you are applying to the substrate that means, in the substrate or body, so this is the processor core, where this is the  $V_{sp}$  point, all the p mos transistors are having one body biased point and all the n mos transistors  $B_{VBSN}$  are having one body biased point.

So, by changing these two voltages only **you know**, you can have only two frequency mode of operation and two threshold voltages corresponding to two frequencies. So, this is how you can hop between two threshold voltages and two frequencies that is the reason, why it is known as  $V_{th}$  hopping.

(Refer Slide Time: 52:06)



**Summary**

- Leakage current reduction techniques can be broadly categorized into two types:
  - Standby leakage reduction
  - Run-time leakage reduction
- Commonly used Standby leakage reduction techniques are:
  - Variable-threshold-voltage CMOS (VTCMOS)
  - Transistor stacking
  - Multi-threshold-voltage CMOS (MTCMOS)
  - Power management (gating)
- Commonly used Run-time leakage reduction techniques are:
  - Dynamic Vdd scaling
  - Combining power gating with DVFS
  - Dual-Vth assignment approach
  - Dynamic Vth scaling (Vth hopping)

NPTEL  
Ajit Pal  
IIT Kharagpur  
NPTEL

Let us, now summarize what we have discussed so far, we have discussed two leakage reduction techniques, as we have seen the leakage reduction techniques can be broadly divided into two categories; standby leakage reduction and run-time leakage reduction. We have discussed several standby leakage reduction techniques like, variable threshold voltage CMOS (VTCMOS), transistor stacking, multi-threshold voltage CMOS, and power management or power gating, which are commonly used to reduce standby leakage power.

Then, today we have discussed several run-time leakage power reduction techniques like, Vth dynamic, Vdd scaling, we have seen, how we can achieve leakage power reduction by using dynamic Vt scaling and also, by combining power gating with dynamic voltage and frequency scaling. And we have also discussed dual Vt assignment approach, which reduces run-time leakage power significantly.

And also, we have discussed that dynamic Vth scaling, where you will change the threshold voltage dynamically, as the workload requirement changes, and special case is Vth hopping where the threshold voltage hops between to threshold voltages and frequency pairs. So, with this we have come to the end of our discussion on reduction of leakage power, in the next lecture, we shall discuss a technique known as variation tolerant design Vtd, thank you.