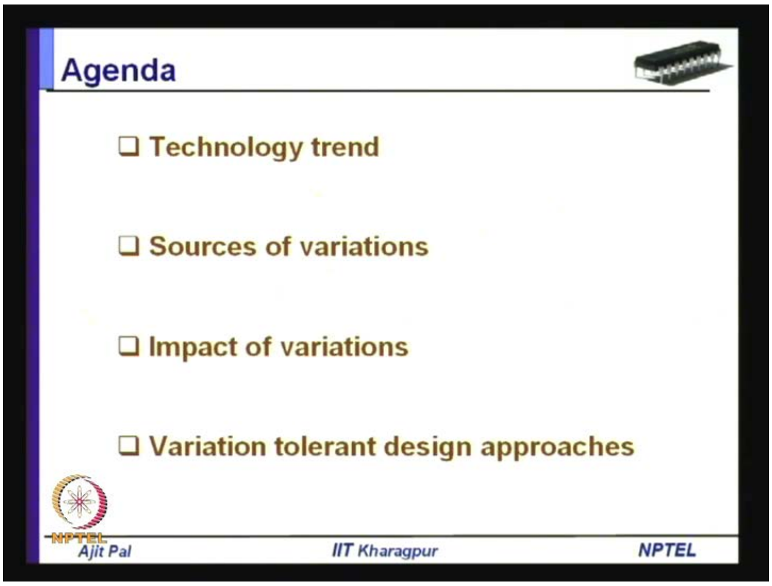
Low Power VLSI Circuits and Systems

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

Lecture No. # 35
Variation Tolerant Design

Hello and welcome to today's lecture on variation tolerant design. In the last couple of lectures, we have discussed minimization of leakage power techniques and before that we have discussed how we can minimize switch capacitance and also, we can minimize the minimize the dynamic power.

(Refer Slide Time: 00:41)



Agenda

❑ Technology trend

❑ Sources of variations

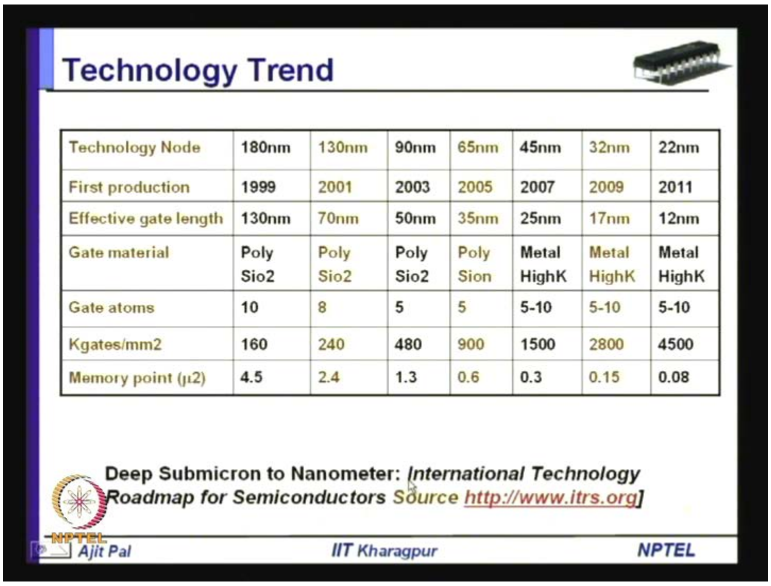
❑ Impact of variations

❑ Variation tolerant design approaches

Ajit Pal IIT Kharagpur NPTEL

Now, we shall consider another very important aspect of the V L S I design that is your variation tolerant design. And here is the agenda of today's lecture first the technology trend and then I shall discuss about the sources of variations then the impact of these variation on this circuit. And finally, we shall discuss about some variation tolerant design approaches.

(Refer Slide Time: 01:11)



Technology Trend

Technology Node	180nm	130nm	90nm	65nm	45nm	32nm	22nm
First production	1999	2001	2003	2005	2007	2009	2011
Effective gate length	130nm	70nm	50nm	35nm	25nm	17nm	12nm
Gate material	Poly Sio2	Poly Sio2	Poly Sio2	Poly Sion	Metal HighK	Metal HighK	Metal HighK
Gate atoms	10	8	5	5	5-10	5-10	5-10
Kgates/mm2	160	240	480	900	1500	2800	4500
Memory point (µ2)	4.5	2.4	1.3	0.6	0.3	0.15	0.08

Deep Submicron to Nanometer: *International Technology Roadmap for Semiconductors Source http://www.itrs.org]*

Ajit Pal IIT Kharagpur NPTEL

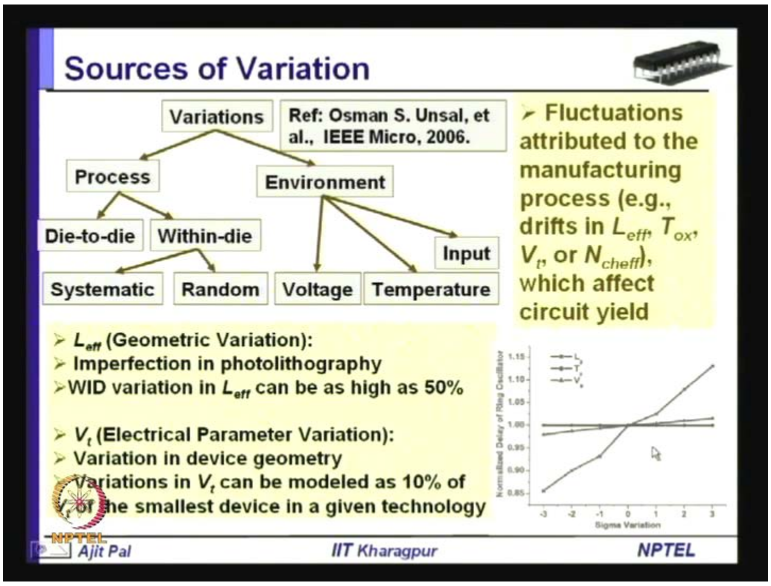
Coming to the technology trend this is the trend which had been predicted by international technology road map of semi conductor in short I T R S we can get it from their website www.itrs.org. And as we can see over the years how the technology node is changing; that means, how the verified size is changing? If we consider 1999 it was it is the 180 nanometer then 2001 130 nanometer, 2003 90 nanometer, 2005 65 nanometer, 2007 45 nanometer 2009 32 nanometer and 2011 22 nanometer. So, this is the trend which has been predicted by I T R S and as you can see as you as the technology changing, these are the effective gate lengths for different technologies for 180 nanometer it is 130 nanometer for 130 nanometer 70 nanometer for 90 nanometer 50 nanometer and so on.

So, if we consider 45 nanometer, which is the mature technology that is been used in the present V L S I circuits the effective channel length is 25 nanometer. And it will come down to 17eventeen nanometer and 12 nanometer as the technology node changes from 32 nanometer and tends to 32 nanometer and 22 nanometer. Now, as we are growing from one technology generation to another technology generation, there is the change in the in the gate material as we can see traditionally poly silicon. And silicon dioxide that is been used so on top of silicon dioxide poly silicon is deposited to form the gate and that is continued till 65 nanometer.

But as you go from go reach the 65 nanometer you can see the number of atoms in the gate region; that means, that silicon that silicon dioxide layer is reduce into 5 atoms; that means, that silicon dioxide comprises only atoms of I mean 5 atoms. So, you can imagine the thickness of that silicon dioxide layer and; obviously, if we want to continue in reducing the size we have to switch to high k dielectric. Obviously, we cannot use silicon dioxide, but we have to move to some of that some of the material which is which has got high k dielectric high k means, that high had dielectric constant. And with that; obviously, you can have a little more thickness, 5 to 10 at 10 atoms and that will continue as we go from, 45 nanometer to 32 nanometer and 20 nanometer.

So, whenever, you change the switch from you can see high k (()) to high k material we cannot use poly silicon, as the gate material we have to use metal on top of that high k dielectric to realize the gate. And there these are the other parameters, like number of gates kilo gates power millimeter square it will it will keep on increasing 160 and 240 then 480 and 900 to 1500 then 2800 and 4500 that is power millimeter square. So, this is another parameter memory point let us not go into the details of this. But what we are we are observing? As the device size is sinking it and with the mean, as we go from one technology generation to another technology generation. The device size is sinking and we can you are able to put more and more transistors per millimeter square or more and more number of gates per millimeter square.

(Refer Slide Time: 05:23)



Sources of Variation

Variations

Ref: Osman S. Unsal, et al., IEEE Micro, 2006.

Process

Environment

Die-to-die

Within-die

Input

Systematic

Random

Voltage

Temperature

- Fluctuations attributed to the manufacturing process (e.g., drifts in L_{eff}, T_{ox}, V_t, or N_{cheff}), which affect circuit yield

- L_{eff} (Geometric Variation):
- Imperfection in photolithography
- WID variation in L_{eff} can be as high as 50%

- V_t (Electrical Parameter Variation):
- Variation in device geometry
- Variations in V_t can be modeled as 10% of V_t of the smallest device in a given technology

Ajit Pal IIT Kharagpur NPTEL

And this will lead to some other problem and the problem we can define as variation process parameter variation. So, the variations can be of several types' process and environment you can borrowed categorize into two types process and environment. So, for as the process variation is concerned again it can be categories into two type die to die. So, we are having different dies for the cated on a silicon wafer. So, this is die to die what is the variation form one die to another die and this is this represents the variation within die. So, there can be variations in within die and again the variations can be of two types systematic variation; that means, from one end of the die to the another end that the there will be variation.

Because of you know, you are fabricating with the help of photolithography techniques, because of a variation and various other problems in the fabrication process there can be inose in systematic variation from one end to the another end. So, that is your systematic variation another possibility is random variation, these variation can be random in nature. And then coming to the environment we can see there are three parameters voltage, voltage may keep on changing because of various regions that supply voltage may change or because of I r drop on the power line voltage may change there may related.

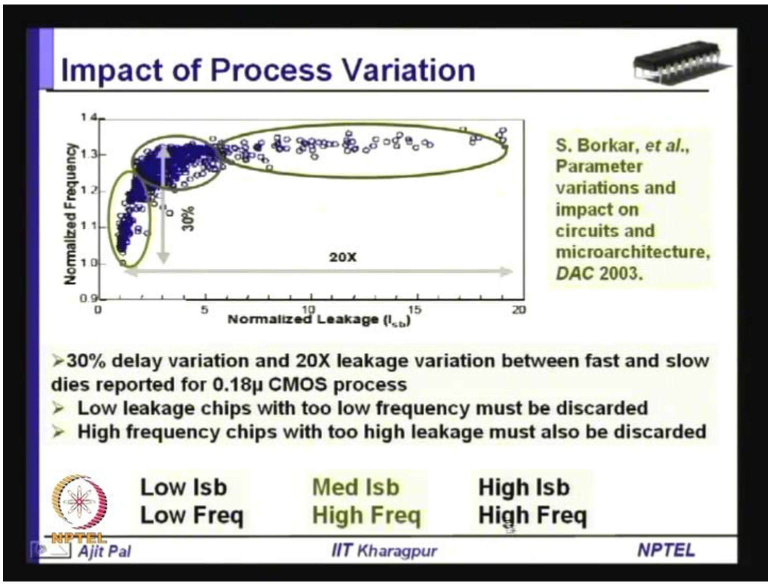
So, this is the; that means, the environment in which the CPU chip is working is changing then temperature can change as we know. The temperature is generated within the chip and if there is no heat sink, I mean proper heat sink the heat cannot be dissipated and as a result temperature will rise. And again another environment is input that you are keeping on change in the input and input has as we know the impact has a lot of impact on the dynamic power; that means the number of transitions that can take place is dependent on the input.

So, environment is input and you know that manufacturing process variations can be activated to the drift in effective channel length. Variation in the that thickness of the silicon dioxide and that will lead to variation in the threshold voltage and also, variation in the doping concentration in the within the channel. So, these changes will lead to this will affect the yield later on we shall the see how the yield is affected. So, yield effective is the effective channel length this variation occurs, because of the imperfection in photolithography as I have told. And within very within the variation in effective channel length can be as high as 50 percent.

And threshold voltage variation that is your electrical parameter variation that occurs, that is also because of variation in device geometry. And variation in V t can be modeled as 10 percent of the V t of the smallest device in a given technology. So, we find that various parameters are changing physical parameters channel length thickness of the silicon dioxide doping concentration and so on. But stimulation result has shown, that the variation of the channel length is the widest we can see this is the; that means, this is the zero point. And you can see the sigma variation on both directions seems quite high on both directions and this is this line corresponds to the effective channel length L g.

So, effective channel length, that variation and the other two are t g and V g; that means your thickness of the silicon dioxide and the threshold voltage. So, we find that, this channel length variation is maximum that is reason why? We have to focus on techniques by which these impacts of this channel length variation can be mitigated. And as we know this channel length variation also leads to variation in the threshold voltage, because the because of V d h roll up you may recall that in deep of micron technology as the channel length varies. Then it leads to I mean variation of the threshold voltage; that means, if the channel length become shorter and shorter the solt voltage becomes smaller and smaller and here is the impact of process variation.

(Refer Slide Time: 10:05)



Impact of Process Variation

S. Borkar, *et al.*, Parameter variations and impact on circuits and microarchitecture, *DAC 2003.*

➢30% delay variation and 20X leakage variation between fast and slow dies reported for 0.18μ CMOS process

- ➢ Low leakage chips with too low frequency must be discarded
- ➢ High frequency chips with too high leakage must also be discarded

Low Isb
Low Freq

Med Isb
High Freq

High Isb
High Freq

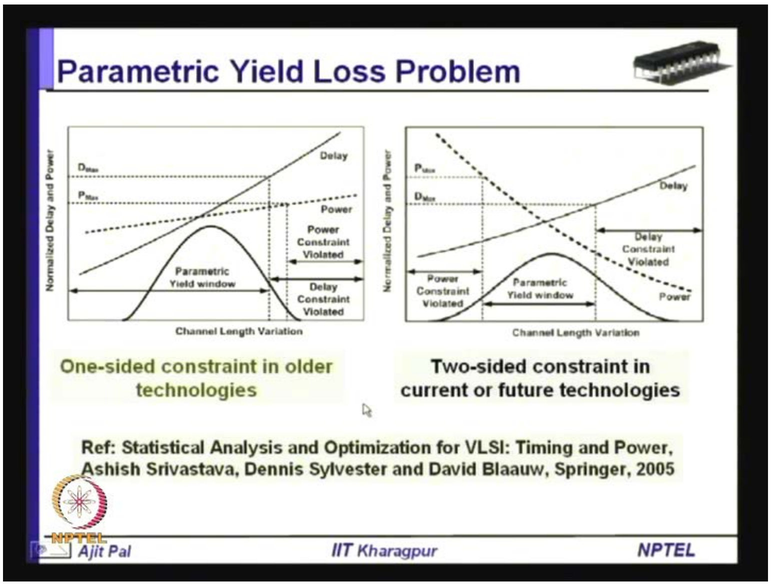
Ajit Pal IIT Kharagpur NPTEL

So, this has been taken from a paper by s borker et al so, parameter variations and impact on circuits and micro architecture. So, as you can see this vary here the variation of the normalized leakage and variation normalized frequency, these two have been plotted. So, you can see that 30 percent delay variation. So, there is a 30 percent of the delay variation in this in the X in the Y axis that is the delay variation. You can see these are the plots of the delay of the different dies that is been fabricated on a on a chip and you can see the variation is 30 percent. So, for as the frequency is concerned or delay frequency delay there they represent the same thing. However, the if we look at the variation in the leakage current we find that that ISB that is the leakage current that variation is 20 X; that means, the range over which the variation is occurring is 20 times.

So, 20 time of this nominal variation so, 0 and this is roughly two and here it is close to 20. So, you can see twenty times twenty times variation in the leakage between the fast and slow dies that has been reported in 0.18 micron C M O S process. So, you can imagine that was the result that was actually, obtained from 0.18 micron technology. So, as the device size is sinking that variation is increasing. So, in the present technology which have realize by using 45 nanometer of 33 nanometer technology the variation will be more. And low frequency chips with low frequency, low leakage chips with low frequency, this is the low leakage chips with low frequency they are to be discarded.

Similarly, high frequency chips with too high leakage; that means this will lead to very large power dissipation they are they also have to be discarded. So, we find that a large number of devices which have fabricated are to be discarded, because of this process parameter variation.

(Refer Slide Time: 12:34)



Parametric Yield Loss Problem

One-sided constraint in older technologies

Two-sided constraint in current or future technologies

Ref: Statistical Analysis and Optimization for VLSI: Timing and Power, Ashish Srivastava, Dennis Sylvester and David Blaauw, Springer, 2005

Ajit Pal IIT Kharagpur NPTEL

So, we have to reduce the impact to these, variation if we want to improve the yield. And you can see the parametric yield loss problem that can be represented by this diagram here is the channel length variation as we as I have already mentioned, the channel length variation is the maximum. So, you can see as the channel length varies the how the delay and power is changing? You can see in the earlier generations the leakage current was insignificant and as a consequence, the power dissipation was primarily the dynamic power. And as a consequence as the channel length increases, these capacitance increases and also, delay increases as you can see in the earlier generation technologies.

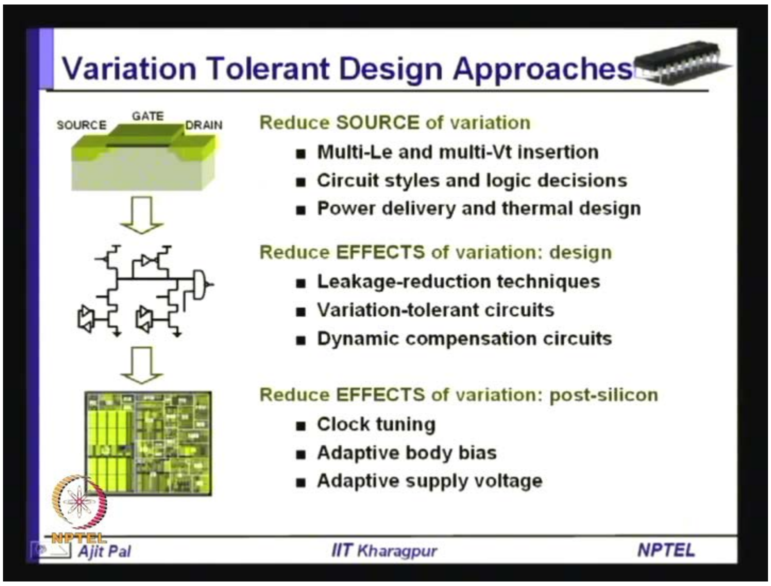
So, delay increases and also as the length increases, the power dissipation increases because of the increase in capacitance so; that means, both delay and power both increases as the channel length increases. So, you have a kind of limit on the delay and also on the power you can see if you said this that d max this is the d max and this is the p max. So, this the this limits the these two limits on these two parameters delay and parameter actually, we will decide we will forced you to discard the chips; that means, those dies which are lying in this range. So, essentially, this is one sided constant in older technology. But unfortunately the situation and change in the present day technology where it has become a two sided constant.

As you can see the power is no longer increasing as the channel length is increasing as the channel length decreases, then the then the you know that threshold voltage decreases and leakage power increases exponentially. And this has lead to increase in leakage power as we reduce the channel length. So, it is different from the earlier plots. So, in this particular case it has now, that the delay is on the right side delay constant; that means, the longer the channel length larger will be the delay. So, this is the limit on the delay. So, the best on performance you have to discard the dies which are falling in this range, but the dies which will fail, because of larger power constant and this is the range over which the power constant is violated.

So, the dies within this range are to be discarded because of larger power dissipation. So, you find that now, you have got a two sided constant and; obviously, larger number of dies are to be discarded either, because of delay or because of power variations. Earlier it was only on one side and; obviously, the yield will decrease significantly. Because of this two sided constant in current or future generation technologies. And this particular thing has been taken from paper by Ashish Shrivastava and Dennis Sylvester and David Blaauw, the paper is statistical analysis and optimization for VLSI timing and power.

So, we find that we the yield is decreasing, because of variation in process parameter. So, what has to be done? We have to go for variation tolerant design approaches; that means, variation will take place we have to design in such a way that circuit is tolerant to the variations how can it be done?

(Refer Slide Time: 16:17)



Variation Tolerant Design Approaches

SOURCE
GATE
DRAIN

Reduce SOURCE of variation

- Multi-Le and multi-Vt insertion
- Circuit styles and logic decisions
- Power delivery and thermal design

Reduce EFFECTS of variation: design

- Leakage-reduction techniques
- Variation-tolerant circuits
- Dynamic compensation circuits

Reduce EFFECTS of variation: post-silicon

- Clock tuning
- Adaptive body bias
- Adaptive supply voltage

Ajit Pal IIT Kharagpur NPTEL

So, there are three basic approaches, first one is reduce source of variation so, it is essentially, prevention is better that cure. So, we would like to prevent the source of variation how can it be done? It can it has to be done by using suitable fabrication of the device. So, device has to be fabricated in such a way that it will reduce the source of variations. So, for that purpose what you have to do? You have to realize circuits with multiple channel length and multiple threshold voltage. So, multi L E and multi V t insertion so; that means, circuit design has to be done; that means, at the time of fabrication you have to realize with multiple channel length and multiple threshold voltage.

Then you have to use suitable circuit styles and logic decisions. So, that the sources of variation is reduced and power delivery and thermal design has to be done; that means, the way the power is delivered; that means, the power network. So, you have to use sophisticated power tree and by which the supply voltage is distributed within the chip. And also, the thermal design has to be done you have to identify you have to do placement of the devices in such a way that the there is no hot spot in a particular chip area; that means, the heat is distributed uniformly, throughout the circuit. So, it means, so, happen that a particular area is more heated.

So, you have to use thermal design; that means, replacement and routing are to be done in such a way such that the hot spot does not occur and that; that means the heat dissipation from different parts is more or less uniform. Second techniques is reduce effects of variation; that means, here you know already chip has been fabricated and in spite of all the techniques that you use to reduce sources of variation there will some variation. And that variation can be effect of these variations can be reduced by suitable circuit design.

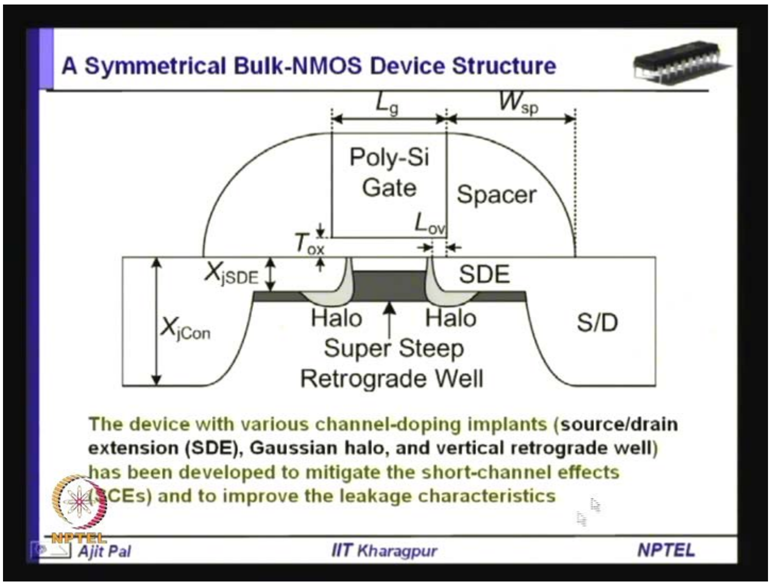
So, I have already discussed about leakage reduction techniques, those leakage reduction techniques various leakage reduction techniques can be used to reduce the effect of variation. And also we have to go for variation tolerant circuits; that means I shall discuss about some of the techniques by which you can realize variation tolerant circuits. Then dynamic compensation of circuits compensation circuits; that means, the circuit will be designed in such a way that the effect of these variations will be automatically compensated. So, that is a kind of dynamic compensation circuit will be incorporated that will build in the part of circuit, which will which will really compensate the dynamic variations.

This is the you know you may say that, these are the pre silicon techniques; that means, before the chip is fabricated you have to do the design in such a way that the impact of these variations is reduce. And another way of doing is based on post silicon so, reduce effects of variation by using post silicon techniques and there are several techniques which have been developed. Number one is clock tuning or this is also, known as frequency breeding after the chips have been fabricated you can divide you know into you can distribute the dies into different frequency beans; that means, some of the device are slow you can put it in lower bean.

Some of the dies will be faster you put them in lower bean and mark them then you sell them accordingly; that means, instead of discarding them. You will divide into different you know based on the frequency of operation you will put them in different categories or different slots of frequencies and or this is known as frequency breeding or clock tuning and; that means, the dies different dies will operate at different frequencies. That after the fabrication you will do measurements and then you will do the distribution of these, dies and two different clocks or different frequencies. Then adaptive body bias after the fabrication has been done you can use different body bias to different dies.

So, depending on the variation the body bias value will be different later on we shall discuss more about it in details. Then another approach is the adaptive supply voltage. So, you instead of using a fix supply voltage you can use different supply voltages to different dies; that means, the if a particular dies dissipating too much of current you can instead of using high supply voltage. You can apply lower supply voltage and similarly, if a if a die is slow you can apply higher supply voltage and so, that the speed of operation increases. So, in this way you can use adoptive supply voltage or you can use adoptive body bias, these are post silicon techniques we shall discuss in detail later on. So, these are the various variation tolerant design approaches first.

(Refer Slide Time: 21:53)



A Symmetrical Bulk-NMOS Device Structure

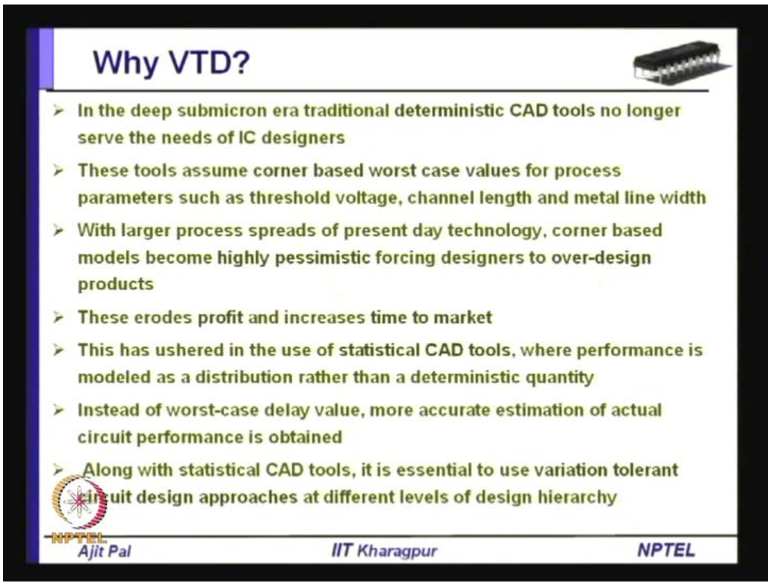
The device with various channel-doping implants (source/drain extension (SDE), Gaussian halo, and vertical retrograde well) has been developed to mitigate the short-channel effects (SCEs) and to improve the leakage characteristics

Ajit Pal IIT Kharagpur NPTEL

Let us focus on the first technique; you can see here this is a single transistor structure. So, this has been design in such a way so, this device with various channel doping implants source drain extension S D. So, you can see source drain extension has been done then Gaussian halo so, here there is Gaussian halo and vertical retrograde well so, you can see vertically there is a variation. So, you can see these things has been done to they have been developed mitigate short channel effects and to improve leakage characteristics.

So, you can see a single transistor has to be fabricated with all these you know different you know different types of features and there will be a spacer on both side. So, whenever, you fabricated with all these features; obviously, the fabrication cost will be enormous; obviously, the most of the short channel effects will be mitigated and as a result impact of variation will be smaller, but the cost of fabrication will be too much. So, this approach is not really very attractive of the from the for the fab houses. So, they will rather pass only balk to the designer so, we have design whatever variation that will occur that problem has to be solve by the designers. So, the fab house usually passes from the balk to the designers and we shall see how they tackle by the design.

(Refer Slide Time: 23:37)



Why VTD?

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

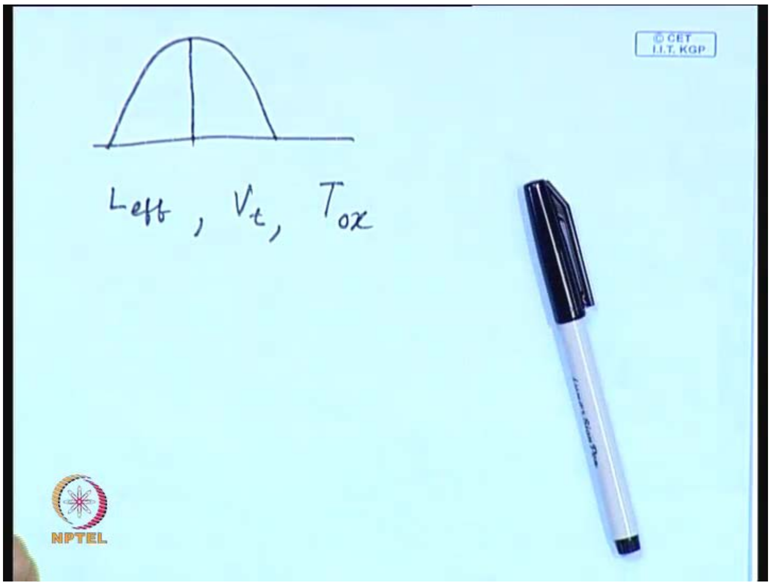
Ajit Pal IIT Kharagpur NPTEL

So, in the deep submicron era traditional deterministic cad tools no longer serve the needs of IC designers. Now a days the cad tools that you are assuming or that you are using are based on deterministic approach by that, I mean they assume that those channel length. The threshold voltage and all other physical parameters they are fixed; that means, for a particular technology when you are fabricating they are fixed; that means, there you assume a fix threshold voltage you assume a fix channel length you assume a fix silicon dioxide thickness and so on. And based on that you do the design using different cad tools and; obviously, these tools assume corner based worst case values; that mean, they will identify the worst possible delays worst possible participation.

And they will find out the worst case corners and for process parameter such as threshold voltage channel length and metal line. And based on that they will do they will do the design and with large larger process spreads of present day technology corner based models become highly pessimistic forcing designers to over design the products. That means you will be designing in such a way that in spite of worst case variation in the power dissipation in the delay the circuit will operate satisfactory. And obviously, you have to design has to be very conservative.

So, whenever you do a very conservative design; obviously, the you are not getting the most benefit out of it; that means, the profit erodes profit erodes means, lot of devices are to be discarded and it will increase the increase the time to market. So; obviously, this is not these techniques are not very attractive for the present day technology. So, this has ushered in the use of statistical cad tools so, you cannot really use those deterministic cad tools you have to go for statistical cad tools.

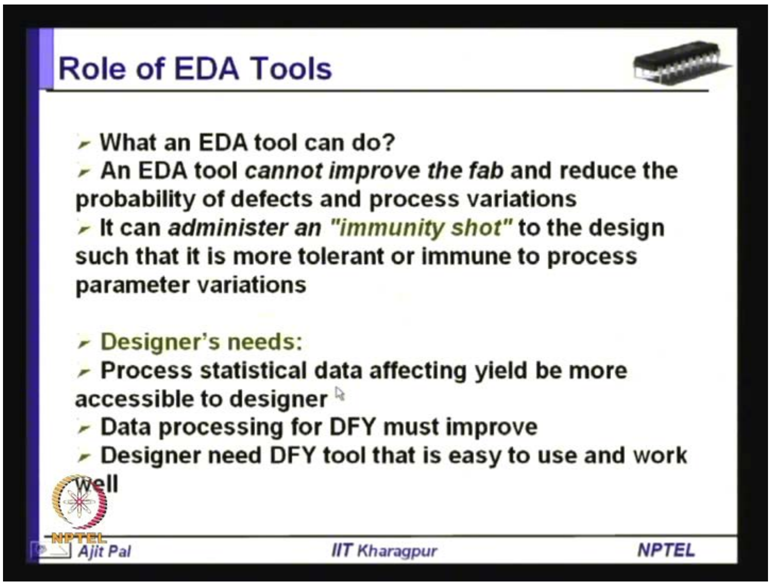
(Refer Slide Time: 25:58)



What do you really mean by that? In this case you have to assume that a channel length L effective is no longer a fix value. So, it will have a kind of Gaussian distribution over the nominal channel length. So, using this kind of distribution of a particular parameter so, it can be effective channel length it can be V t it can be thickness of the silicon dioxide and so on. So, all these parameters are no longer fixed and constant there kind of random variable. So, you have to use statistical approach in synthesizing the circuits.

So, instead of worst case delay value more accurate estimation of actual circuit performance is obtained whenever you use these statistical techniques. And along with statistical cad tools it is essential to use variation tolerant circuit design approaches at different levels of design hierarchy. I shall discuss about some of the variation tolerant circuit design approaches, which can be use at different levels of design hierarchy.

(Refer Slide Time: 27:00)



Role of EDA Tools

- What an EDA tool can do?
- An EDA tool *cannot improve the fab* and reduce the probability of defects and process variations
- It can *administer an "immunity shot"* to the design such that it is more tolerant or immune to process parameter variations

- Designer's needs:
- Process statistical data affecting yield be more accessible to designer
- Data processing for DFY must improve
- Designer need DFY tool that is easy to use and work well

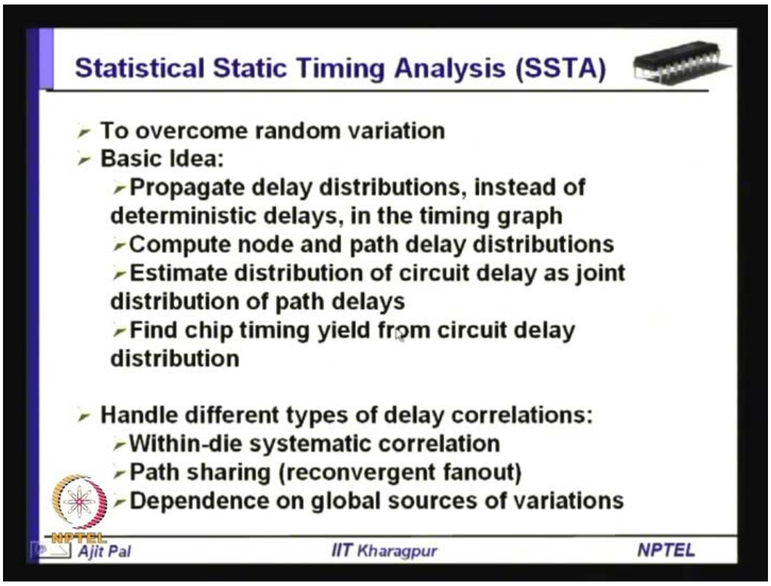
Ajit Pal IIT Kharagpur NPTEL

And obviously, in this case what is the role of the E D A tools? So, an E D A tool cannot improve the fab so, that as I told the fabrication house will keep on fabricating chips and; obviously, they will fab their inherent process parameter variations, that cannot be changed and cad tools cannot really do that. So, only thing the E D A tools the electronic design automation tools can do, they can administer and immunity shot to the design; that means, they can use a kind of immunity shot means they can design in such a way that it is more tolerant or immune to process parameter variations. So, that means, it is somewhat like error free communication through error prone channel.

So, normally you know whenever, you do digital communication you do applied developed techniques so, that when the signal is passing a channel which is error prone. But your objective is to have error free communication and for that purpose, lot of redundancy and various techniques we use. So, this is somewhat similar to that you have to do the design in such a way that it is more tolerant or tolerant or immune to process parameter variations. And for this purpose what the designers need is process statistical data affecting yield and unfortunately, usually there is a gap communication gap.

Between the fab houses they keep some of their fabrication I mean fabrication techniques or fabrication some secret; that means, they do not they do not divert all the information of their fabrication to the designers. And as a consequence a kind of a communication gap exits and if the designer has to do proper design, then the process statistical data affecting yield be more accessible to the designers. So, the fab houses should make these information process statistical data make available to the designers and, but always there is some gap. And data processing for design for yield must improve and designers need, design for yield tools that is easy to use and work well. So, these are the designer needs for developing suitable cad tools for the future generation technologies.

(Refer Slide Time: 29:36)



Statistical Static Timing Analysis (SSTA)

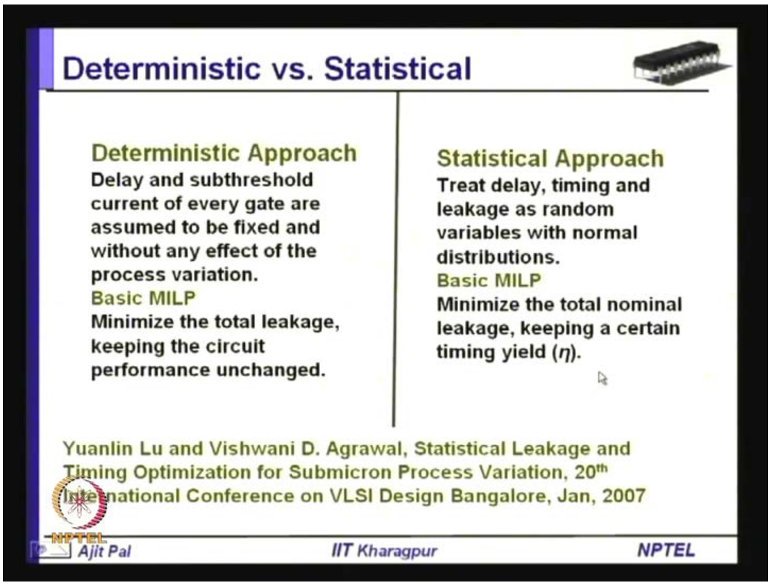
- To overcome random variation
- Basic Idea:
  - Propagate delay distributions, instead of deterministic delays, in the timing graph
  - Compute node and path delay distributions
  - Estimate distribution of circuit delay as joint distribution of path delays
  - Find chip timing yield from circuit delay distribution
- Handle different types of delay correlations:
  - Within-die systematic correlation
  - Path sharing (reconvergent fanout)
  - Dependence on global sources of variations

Ajit Pal IIT Kharagpur NPTEL

Now, one technique is your statistical static timing analysis. So, to overcome random variation the basic idea is propagate delay distribution instead of deterministic delays. I have already told that delay is no longer a fixed value; that means, you cannot assume that a particular gate will have a fix delay. Say three input and gate will have a fix delay of three nanosecond or something like that you cannot tell it that way. The as I told it is a it is a random variable and that delay will be distributed maybe it is a it will be Gaussian distribution. And that distribution has to be propagated instead of these deterministic delays in the timing graph.

And compute node and path delay distributions estimate distribution of circuit delay and joint distribution of path delays. And find cheap timing yield from circuit delay distribution. So, you can see this is the I mean instead of simple static timing analysis you have to statistical static timing analysis and you have to consider different types of delays. So, handled different types of correlation within die systematic correlation as I have already told that can be systematic correlation. So, that means, different parts of the die can have different variations, because of the imperfection in the photolithography process. And path sharing there will be reconvergent fanout and dependence of global source of variations. So, these are the I mean you have to handle different types of delay correlation by the statistical static timing analysis.

(Refer Slide Time: 31:18)



Deterministic vs. Statistical

Deterministic Approach
Delay and subthreshold current of every gate are assumed to be fixed and without any effect of the process variation.
Basic MILP
Minimize the total leakage, keeping the circuit performance unchanged.

Statistical Approach
Treat delay, timing and leakage as random variables with normal distributions.
Basic MILP
Minimize the total nominal leakage, keeping a certain timing yield (η).

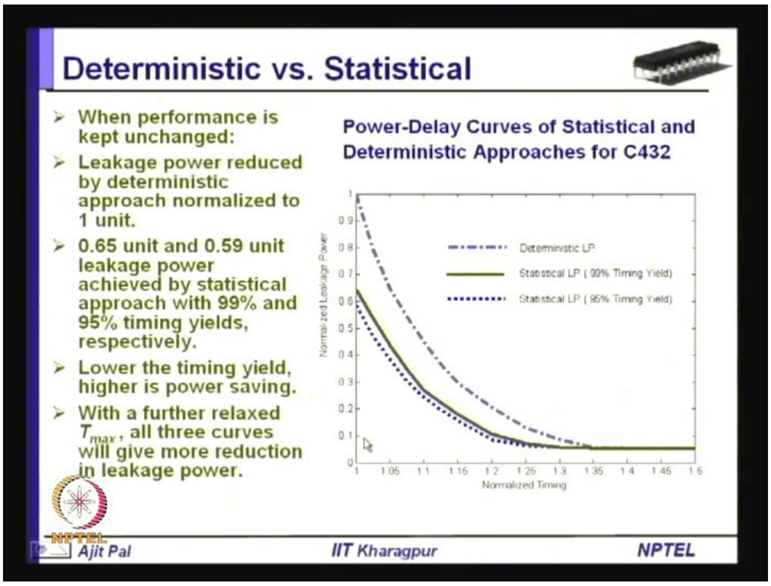
Yuanlin Lu and Vishwani D. Agrawal, Statistical Leakage and Timing Optimization for Submicron Process Variation, 20th International Conference on VLSI Design Bangalore, Jan, 2007

Ajit Pal IIT Kharagpur NPTEL

Now, I shall tell about a paper which has been recently published by Lu and Agrawal, statistical leakage and timing optimization for submicron process variation it has been published by in that V L S I design conference in the year two thousand and seven. So, this is the traditional deterministic approach, where delay and sub threshold current of every gate are assumed to be fixed and without any affect on the effect on the process variation. And basic you know basic M I L P is the is minimize the total leakage keeping the circuit performance unchanged. So, this is how the integral linear programming is done for the deterministic approach.

On the other hand for statistical approach you have to treat delay timing and leakage as random variables, with normal distributions and the basic M I L P for this particular approach is minimize the total nominal leakage keeping a certain timing yield. So, this will be the integer linear minimization minimizing the minimization criteria in integer linear programming, minimize the total nominal leakage keeping a certain timing yield. So, based on this approach in this paper they have reported some result.

(Refer Slide Time: 32:45)



Deterministic vs. Statistical

- When performance is kept unchanged:
- Leakage power reduced by deterministic approach normalized to 1 unit.
- 0.65 unit and 0.59 unit leakage power achieved by statistical approach with 99% and 95% timing yields, respectively.
- Lower the timing yield, higher is power saving.
- With a further relaxed T_{max}, all three curves will give more reduction in leakage power.

Power-Delay Curves of Statistical and Deterministic Approaches for C432

Ajit Pal IIT Kharagpur NPTEL

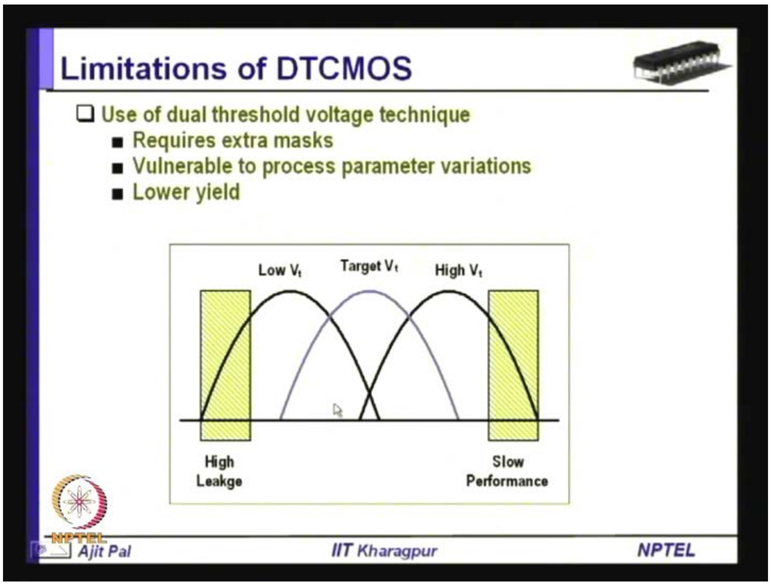
So, where you can see the impact of the statistical approach compared to deterministic approach. So, in this plot if you can see the first curve corresponds to deterministic L P deterministic linear programming. So, leakage power reduces by normalize to one unit. So, this is one and; obviously, as you increases the timing so, then the leakage power reduces; that means, if you are if you are ready to accept small I mean more and more delays. If you are if you can tolerate more delays then; obviously, leakage current can be reduce, which have already demonstrated by a technique called energy efficient energy you know that we have discussed two approaches of leakage power minimization.

One is your delay constant another was energy constant. In energy constant approach you have seen if we can tolerate some delay then leakage can be reduced significantly with by accepting say four to nine percent of delay . So, it is here essentially the same approach is being used. So, if you if you the increase the normalize timing then leakage power can be reduced so, this is the deterministic approach. Now, you can see 0.65 unit or 0.59 unit leakage power is achieve by statistical approach with 99 percent and 95 percent timing yield respectively.

So, here you can see the 0.65 this point corresponds to 0.65 and this point corresponds to 0.59. So, you can see this is the statistical L P this red line with 99 percent timing yield. And this one corresponds to statistical L P with 95 percent 95 percent timing yield. So, you can see lower the timing yield higher is the power saving so, that means, if you are ready to accept larger delay you can have higher yield. So, with a further relaxed t max all three curves will give a more reduce reduction in leakage power.

So, if you are ready to accept larger delay as I told leakage power can be minimize. So, this is the power delay curves of statistical and deterministic approach for a particular circuit C 432 this is a scratch bench mark circuit C 432. And for that particular circuit this is the stimulation result published in this paper I have already mentioned. So, you can see this is one technique I mean this is a statistical approaches, which has been used in the design of a circuit.

(Refer Slide Time: 35:45)



Limitations of DTCMOS

❑ Use of dual threshold voltage technique

- Requires extra masks
- Vulnerable to process parameter variations
- Lower yield

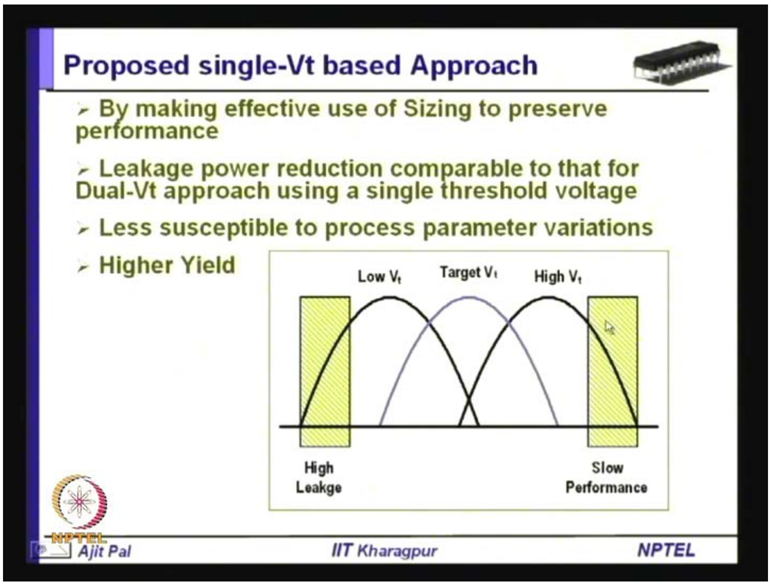
Ajit Pal IIT Kharagpur NPTEL

Let us come to another technique we are we have already introduce the dual threshold CMOS technique. So, whenever you use dual threshold CMOS circuit we know that we are using two types of transistors in the realization. For realizing the gates on the critical path we are using transistors, of lower threshold volt and to realize gates on the non critical path we are using transistors of higher threshold voltage. So, now, let us consider this situation where both the threshold voltages lower threshold voltages and higher threshold voltages are no longer fixed, they are you know they are having kind of normal distribution. So, here you can see this is the low threshold voltage and which varies over this range and this is the high threshold voltage which varies in this range.

Now, you can see the low threshold voltage variation you can see whenever, the threshold voltage becomes in this range; that means, the devices in the critical part will have. So, much of power dissipation that leakage current will be very high and you have to discard those chips, those dies you have to discard. Because they will be having lot of power dissipation on the other hand on this side those dies will have very slow performance, because some of the non critical path delays will exceed the you know timing budget so, as a consequence you have to discard them. So, here you can see this use of this well threshold CMOS technique not only requires extra marks, but they are vulnerable to process parameter variations.

So, and this leads to lower yield; that means, this dual threshold CMOS implementation has inherent limitation of lower yield; that means, the whenever you switch from single threshold implementation to dual threshold implementation there is a possibility of smaller yield. So, we use dual threshold technique to reduce leakage power, but as you do. So, it is became it is pronged to pronged I mean larger process parameter variations so, it is vulnerable to process parameter variation leading to lower yield how can it be overcome.

(Refer Slide Time: 38:23)



Proposed single-Vt based Approach

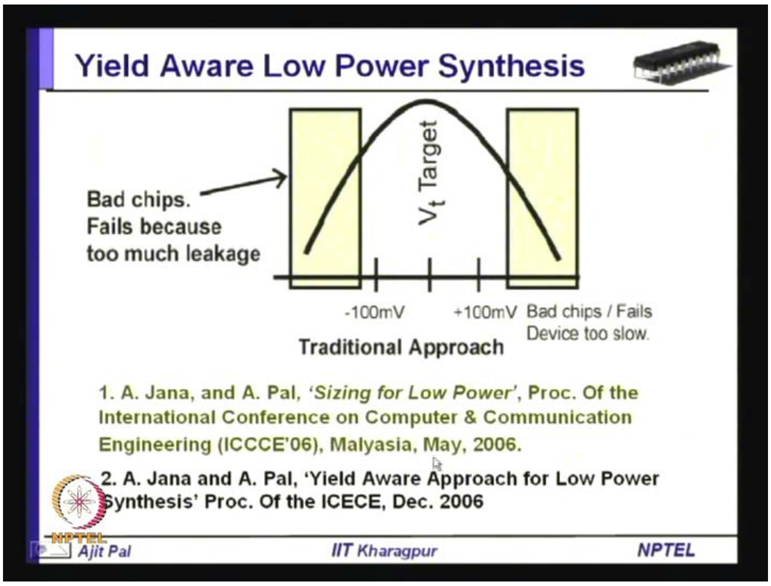
- By making effective use of Sizing to preserve performance
- Leakage power reduction comparable to that for Dual-Vt approach using a single threshold voltage
- Less susceptible to process parameter variations
- Higher Yield

Ajit Pal IIT Kharagpur NPTEL

So, a single V t based approach has been proposed and I shall discuss about that. So, this single V t based approach has been based on making effective use of sizing to preserve performance. So, here this sizing has been used judiciously to preserve performance and leakage power reduction has been achieve by using single threshold voltage. Single threshold voltage and it has been it has been found that the leakage power reduction is comparable to that of dual V t approach using a single threshold voltage. So, that is the beauty of this approach I mean without using dual threshold voltage you are able to achieve leakage power reduction comparable to dual threshold voltage approach.

So, by using a single threshold voltage and; obviously, it is less susceptible to process parameter variation and it gives you higher yield. So, this is the this is the target V t; that means, you are using a single threshold voltage; obviously, this will be more than the lower threshold voltage and less than the higher threshold voltage whenever you use dual V t. So, you have to really identify this target V t and how can it be done? I shall briefly discuss and you can see this is neither violating the leakage larger leakage nor it is violating the larger delay.

(Refer Slide Time: 39:56)



Yield Aware Low Power Synthesis

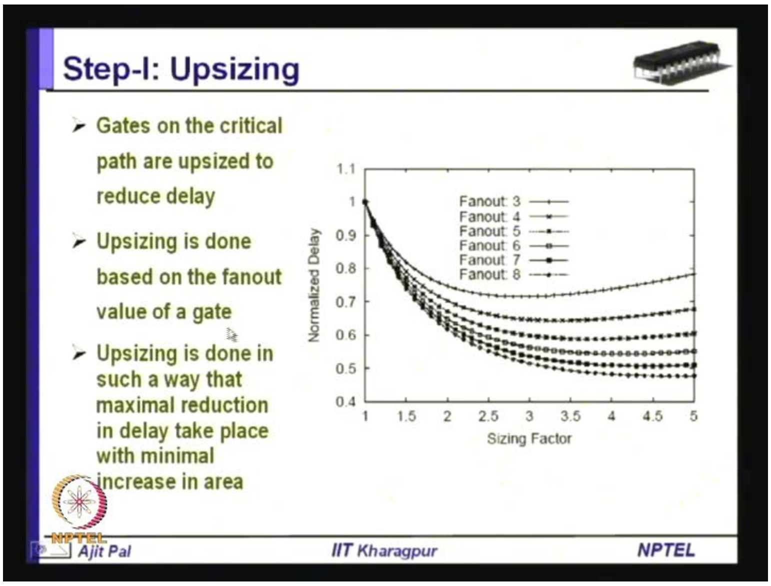
1. A. Jana, and A. Pal, *'Sizing for Low Power'*, Proc. Of the International Conference on Computer & Communication Engineering (ICCCE'06), Malyasia, May, 2006.

2. A. Jana and A. Pal, 'Yield Aware Approach for Low Power Synthesis' Proc. Of the ICECE, Dec. 2006

Ajit Pal IIT Kharagpur NPTEL

So, this will lead to better yield. So, these are the two this is based on the two papers sizing for low power this was published in published in a conference in Malaysia in May, 2006. Another paper yield aware approach for low power synthesis of this team proceeding of I C E C E in December 2006 and subsequently a journal paper has also come out from this so, you can see.

(Refer Slide Time: 40:34)



Step-I: Upsizing

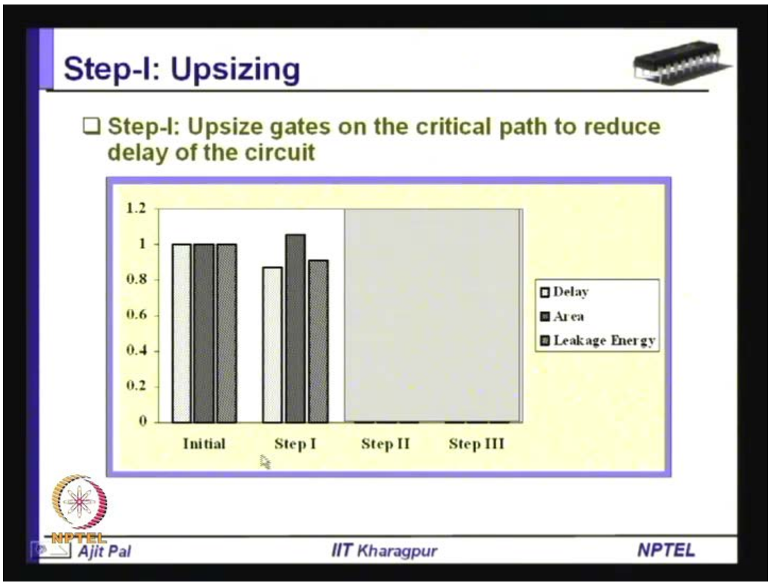
- Gates on the critical path are upsized to reduce delay
- Upsizing is done based on the fanout value of a gate
- Upsizing is done in such a way that maximal reduction in delay take place with minimal increase in area

Ajit Pal IIT Kharagpur NPTEL

So, let us see the basic approach the step of this approach. So, here step one is upsizing first step is upsizing; that means, what is being done gates on the critical path are upsized to reduce delay, but it is not done by upsizing all the gates uniformly. What is being done upsizing is being done based on fanout of the gate? Because some experiment was carried out this is been found that larger the fanout you can see this is the fanout larger the fanout. And this is the this is the plot for different fanout this is for fanout 3 this is for fanout 4 this is for fanout 5 this is for fanout 8. So, we find that the reduction in delay is more for gates of higher fanout; that means, if you do the if you increase the size three times; that means, make the width three times wider then you can see the reduction in delay will be you can see it is 0.5.

Whenever, the fanout is 8 compared to reduction of only 25 percent instead of 50 percent 25 percent when the fanout is three. What does it mean? It means that if you do the upsizing of the gates with higher fanout with smaller increasing area you will be able to achieve larger reduction in delay and that is the reason why the gates with larger fanout are upsized instead of gates with smaller fanout. So, essentially upsizing is done in such a way that maximum reduction in delay takes place with minimum increasing area.

(Refer Slide Time: 42:20)



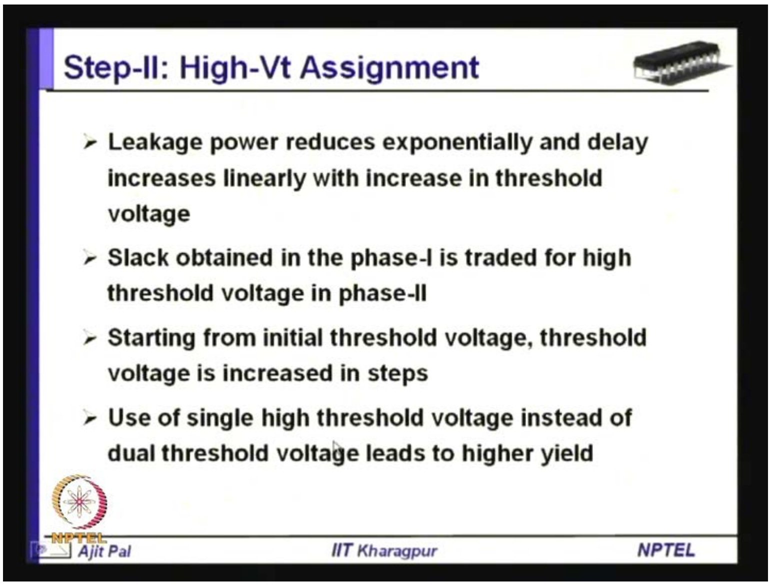
Step-I: Upsizing

❑ Step-I: Upsize gates on the critical path to reduce delay of the circuit

Ajit Pal IIT Kharagpur NPTEL

So, this is the first step and at the end of this first step this is the initial you know delay area and power. And after the upsizing is done; that means, some of the gates on the critical path have been increase size has been increased and; obviously, the area will become more as you can see this area will become more. However, delay will reduce, because the delay of the critical path has been reduced by upsizing. As you know as you increase the size the delay reduces as I have shown in the previous curve you have seen that the delay reduces. So, delay of the critical path will reduce, which is shown by this and; however, the leakage energy will I mean will remain more or less more or less the same value. So, there will be increase in area, but reduction in delay so, this is the outcome of the step one.

(Refer Slide Time: 43:11)



Step-II: High-Vt Assignment

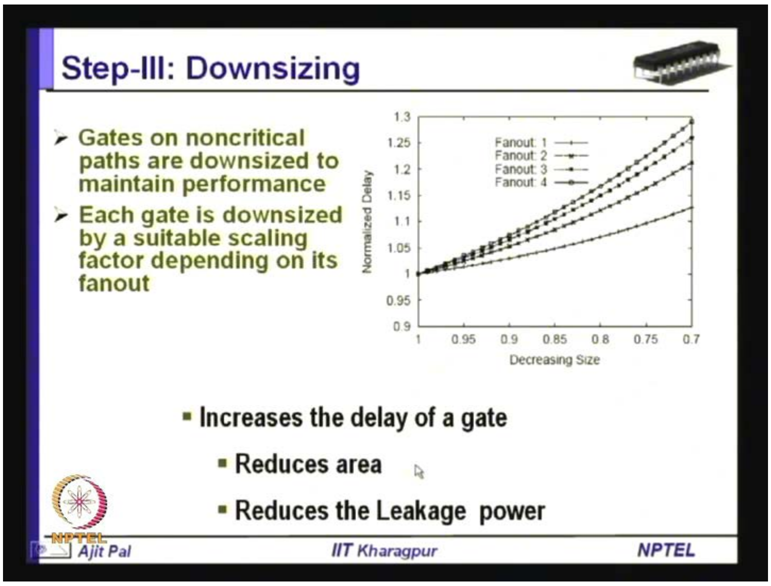
- Leakage power reduces exponentially and delay increases linearly with increase in threshold voltage
- Slack obtained in the phase-I is traded for high threshold voltage in phase-II
- Starting from initial threshold voltage, threshold voltage is increased in steps
- Use of single high threshold voltage instead of dual threshold voltage leads to higher yield

Ajit Pal IIT Kharagpur NPTEL

Now, let us go to step two where leakage power as you know leakage power reduces exponentially and delay increases linearly with increase in threshold voltage. So, slack obtained in phase one is traded for higher threshold voltage in phase two. So, what is being done in this case? You know we have by upsizing we have achieve some slack; that means, reduction in delay now, that reduction in delay will be traded for reduce in the leakage power by increasing the threshold voltage of those gates. That means, not only the gates of the critical path all the gates in the circuit will be now, threshold voltage will be increase from this minimum value of point to V d t to some larger value. What is the larger value? That we shall find out.

So, slack obtained in phase one is traded for higher threshold voltage in phase two. Starting from initial threshold voltage threshold voltage is increased in steps and use of single high threshold voltage instead of dual threshold voltage leads to higher yield. So, this is what is being done? And you can see and at the end of the step two this is what we find? That means, obviously, there is increase in area, but the now the delay is same as the original circuit because of the use if larger threshold voltage delay has increased from this value. So, delay is now more delay is now same as the original circuit, because we are now not compromising in performance. However, there is a significant reduction in leakage power; because of you knows exponential dependence of leakage on the threshold voltage and as the consequence the leakage power has reduced significantly.

(Refer Slide Time: 44:59)



Step-III: Downsizing

- Gates on noncritical paths are downsized to maintain performance
- Each gate is downsized by a suitable scaling factor depending on its fanout

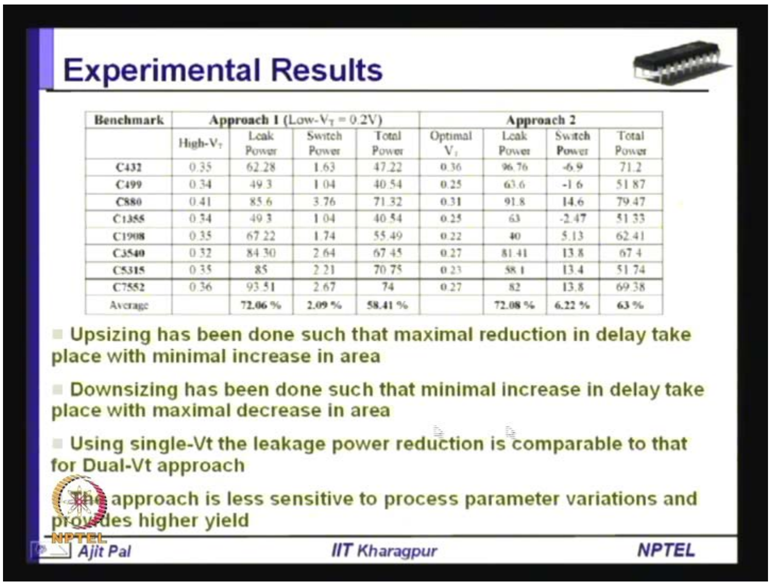
- Increases the delay of a gate
  - Reduces area
  - Reduces the Leakage power

Ajit Pal IIT Kharagpur NPTEL

Now, what is being done in step four? You know that there are the gates in the critical paths are downsized to maintain performance. Now, some of the gates on the non critical path can be downsized. So, what will happen? Their delay will increase, but as long as the delay does not exceed the delay of the critical path there is no problem. And here also, the gates on the non critical paths are not upsized you know not upsized arbitrarily. Again there is a plot; that means, the gates with smaller fan out are upsized rather than gates with larger fan out. Because the gates with smaller fan fan out will have lesser impact on delay; that means, the delay will increase by a smaller amount if you reduce the size of gate with smaller fan out.

So, each gate is downsized by a suitable scaling factor depending upon it is fanout. So, this increase this increases the delay of the gate reduces the area and reduces leakage power. So, ultimately we get a plot like this so, we find that. Now, so far as the delay is concerned it is same as the original circuit so, there is no compromise in performance. But because of downsizing of a large number of gates on the non critical path now, the area has reduced. And this the leakage power has also reduced as you can see by smaller amount from the previous value because of the reduction in area. So, we find that at the end of step three we are getting a circuit which has the same performance of the original circuit, but significantly lesser leakage power dissipation and that too is achieved by using a single threshold voltage.

(Refer Slide Time: 46:46)



Experimental Results

Benchmark	Approach 1 (Low-V_T = 0.2V)				Approach 2			
	High-V_T	Leak Power	Switch Power	Total Power	Optimal V_t	Leak Power	Switch Power	Total Power
C432	0.35	62.28	1.63	47.22	0.36	96.76	-6.9	71.2
C499	0.34	49.3	1.04	40.54	0.25	63.6	-1.6	51.87
C880	0.41	85.6	3.76	71.32	0.31	91.8	14.6	79.47
C1355	0.34	49.3	1.04	40.54	0.25	63	-2.47	51.33
C1908	0.35	67.22	1.74	55.49	0.22	40	5.13	62.41
C3540	0.32	84.30	2.64	67.45	0.27	81.41	13.8	67.4
C5315	0.35	85	2.21	70.75	0.23	58.1	13.4	51.74
C7552	0.36	93.51	2.67	74	0.27	82	13.8	69.38
Average		72.06 %	2.09 %	58.41 %		72.08 %	6.22 %	63 %

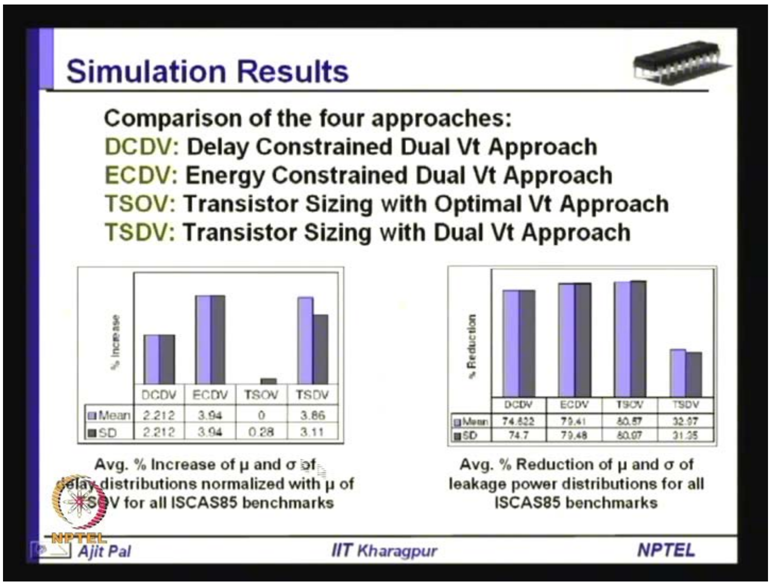
- Upsizing has been done such that maximal reduction in delay take place with minimal increase in area
- Downsizing has been done such that minimal increase in delay take place with maximal decrease in area
- Using single-Vt the leakage power reduction is comparable to that for Dual-Vt approach
- The approach is less sensitive to process parameter variations and provides higher yield

Ajit Pal IIT Kharagpur NPTEL

So, this is the comparison. So, this is the case where you know here you can see this is the result of Dual vth threshold voltage. So, whenever you realize by circuits using Dual vth threshold voltage there is a reduction of 72.6 percent occurs. And on the other hand whenever, used this single threshold voltage by using the approach that have discussed in detail we can see there is 72.08 percent reduction in the leakage power.

So, let us forget about this switching power, that reduction also occurs and total power reduction is 63 percent compared to total power reduction of 58.41 percent in the case of a Dual vth threshold volt approach. This reduction is arising, because of you know downsizing of the gates on the non critical path so, capacitance reduces. So, upsizing has been done such that maximal reduction in delay take place with minimal increase in area downsizing has been done such that minimal increase in delay takes place with maximal decrease in area. So, using single V t the leakage power reduction is comparable to that of dual V t approach and the approach is less sensitive to process parameter variations and provides higher yield.

(Refer Slide Time: 48:07)



Simulation Results

Comparison of the four approaches:
DCDV: Delay Constrained Dual Vt Approach
ECDV: Energy Constrained Dual Vt Approach
TSOV: Transistor Sizing with Optimal Vt Approach
TSDV: Transistor Sizing with Dual Vt Approach

	DCDV	ECDV	TSOV	TSDV
▨ Mean	2.212	3.94	0	3.86
■ SD	2.212	3.94	0.28	3.11

Avg. % Increase of μ and σ of delay distributions normalized with μ of TSOV for all ISCAS85 benchmarks

	DCDV	ECDV	TSOV	TSDV
▨ Mean	74.822	79.41	80.57	32.97
■ SD	74.7	79.48	80.97	31.35

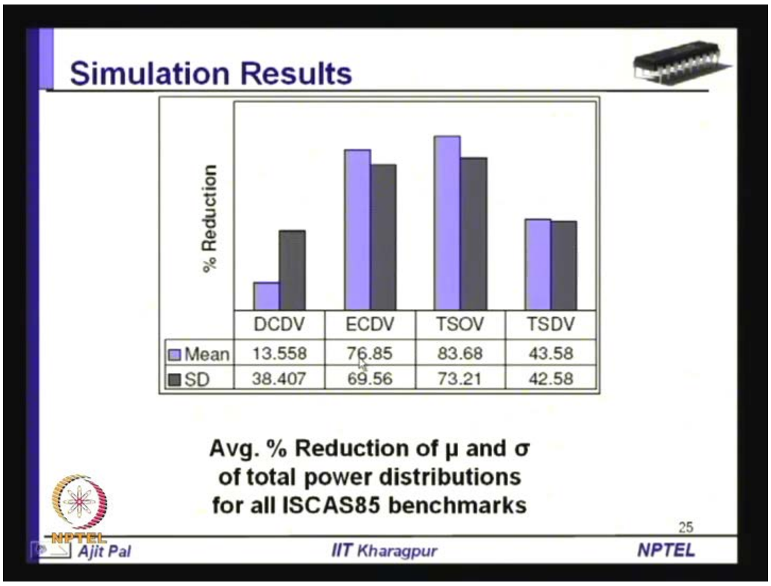
Avg. % Reduction of μ and σ of leakage power distributions for all ISCAS85 benchmarks

Ajit Pal IIT Kharagpur NPTEL

So, this has been verified by so, here is the comparison of four different approaches this is the delay constant Dual vth V t approach, this is the energy constant Dual vth V t approach and this is the approach transistors sizing with optimal V t, which have discussed in detail right now, and another approach is transistor sizing with dual V t So, in this case we find that this approach this transistor sizing with optimal V t you can see, this average percentage increase in mu and sigma of the delay distribution, normalized with mu of the T S of the for all the bend marks.

So, this gives the best result similarly, average reduction in the mu and sigma of the leakage power distribution for all discussed bench mark as shown. So, we find that T S V performs I mean maximum reduction occurs in the case of the third approach which has been proposed.

(Refer Slide Time: 49:04)



Simulation Results

	DCDV	ECDV	TSOV	TSDV
▨ Mean	13.558	76.85	83.68	43.58
■ SD	38.407	69.56	73.21	42.58

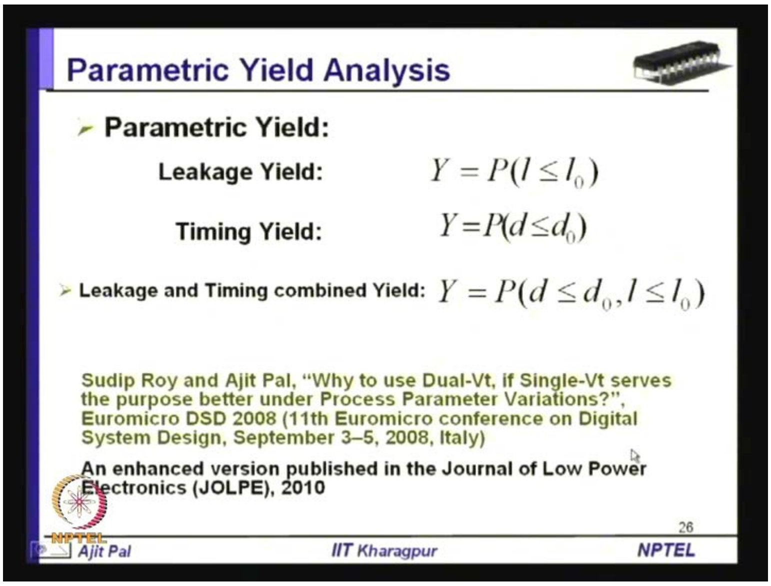
Avg. % Reduction of μ and σ of total power distributions for all ISCAS85 benchmarks

25

Ajit Pal IIT Kharagpur NPTEL

And so, here average percentage reduction mu and sigma for total power distribution and here also maximum reduction occurs for this approach 83.68 and 73.21. So, mu and standard deviation mu is mean and sigma is standard deviation.

(Refer Slide Time: 49:24)



Parametric Yield Analysis

- **Parametric Yield:**
  - **Leakage Yield:** $Y = P(l \leq l_0)$
  - **Timing Yield:** $Y = P(d \leq d_0)$
- **Leakage and Timing combined Yield:** $Y = P(d \leq d_0, l \leq l_0)$

Sudip Roy and Ajit Pal, "Why to use Dual-Vt, if Single-Vt serves the purpose better under Process Parameter Variations?", Euromicro DSD 2008 (11th Euromicro conference on Digital System Design, September 3–5, 2008, Italy)

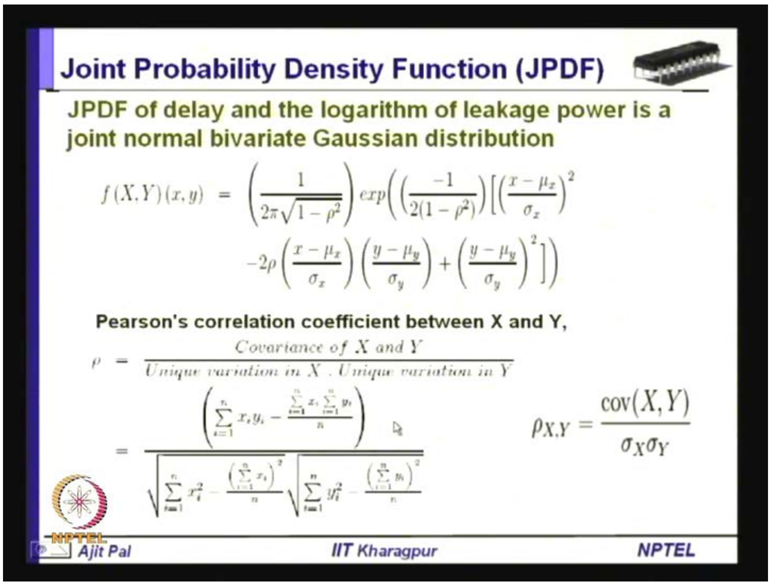
An enhanced version published in the Journal of Low Power Electronics (JOLPE), 2010

26

Ajit Pal IIT Kharagpur NPTEL

So, this approach parametric yield analysis has been we have used it and some statistical approaches has been used to verify the efficacy of this approach.

(Refer Slide Time: 49:37)



Joint Probability Density Function (JPDF)

JPDF of delay and the logarithm of leakage power is a joint normal bivariate Gaussian distribution

$$f(X,Y)(x,y) = \left(\frac{1}{2\pi\sqrt{1-\rho^2}}\right) exp\left(\left(\frac{-1}{2(1-\rho^2)}\right)\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 -2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right)$$

Pearson's correlation coefficient between X and Y,

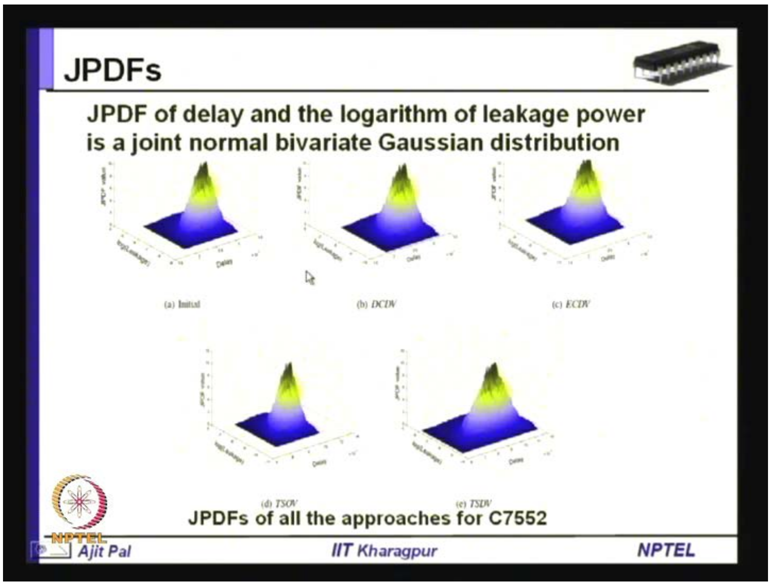
$$\rho = \frac{Covariance\ of\ X\ and\ Y}{Unique\ variation\ in\ X\ .\ Unique\ variation\ in\ Y} = \frac{\left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}}$$

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Ajit Pal IIT Kharagpur NPTEL

And where joint probability distributed function has been computed and here is the plot of the joint probability distribution function of delay and logarithmic, logarithm of leakage power.

(Refer Slide Time: 49:45)



JPDFs

JPDF of delay and the logarithm of leakage power is a joint normal bivariate Gaussian distribution

(a) Initial (b) DCDV (c) ECDV

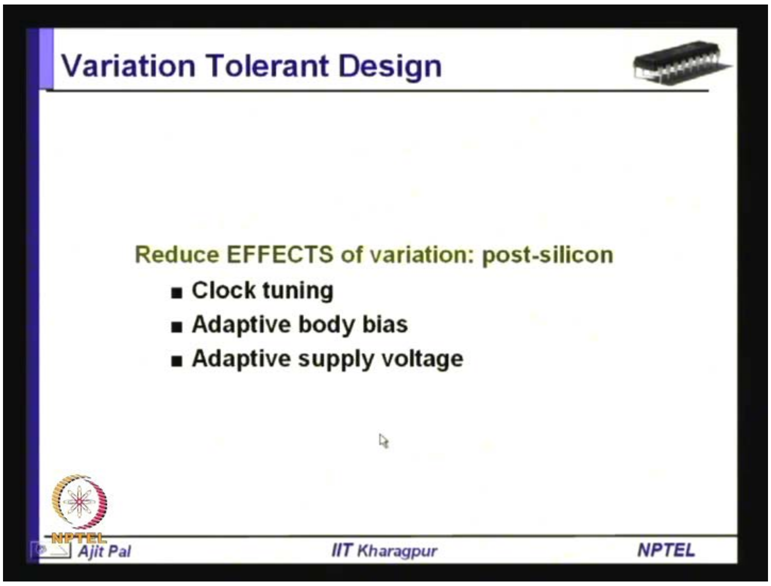
(d) TSOV (e) TSDV

JPDFs of all the approaches for C7552

Ajit Pal IIT Kharagpur NPTEL

That means the J P D F joint probability distribution efficient function is a joint normal bivariate Gaussian distribution for the different approached is given here. So, we find even by looking at the curves this is the narrower so on. This side is the delay on this side is the logarithmic (()) h power for all the approached this is the initial one this is the delay constant dual V t approach this is the energy constant dual V t approach. And this is that T S D V approach and this is the T S O V approach this I have discussed in detail. So, we find that variation is much smaller in such particular approach.

(Refer Slide Time: 50:35)



Variation Tolerant Design

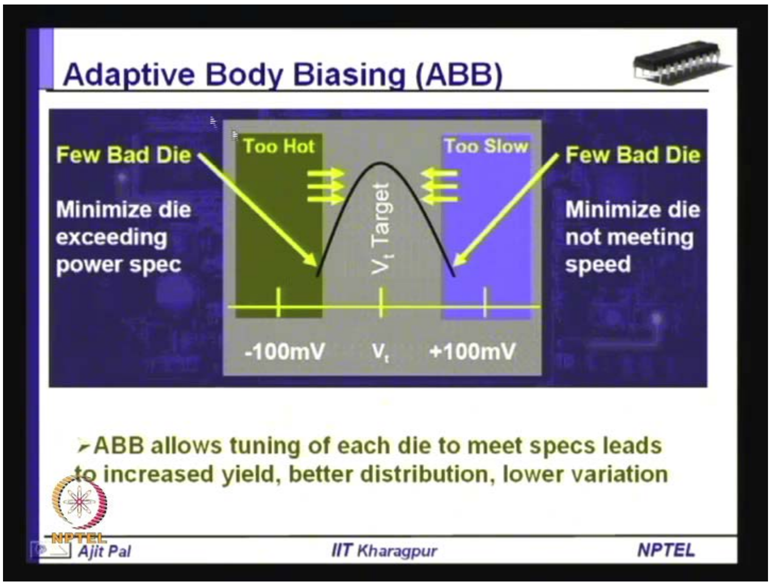
Reduce EFFECTS of variation: post-silicon

- Clock tuning
- Adaptive body bias
- Adaptive supply voltage

Ajit Pal IIT Kharagpur NPTEL

Now, coming to other techniques which I have told clock tuning and adaptive body bias and adaptive supply voltage can be used.

(Refer Slide Time: 50:49)



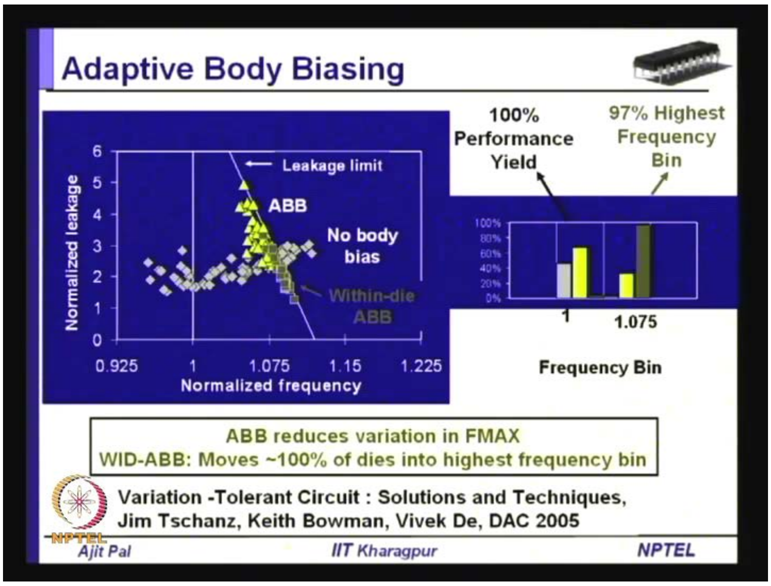
Adaptive Body Biasing (ABB)

➤ABB allows tuning of each die to meet specs leads to increased yield, better distribution, lower variation

Ajit Pal IIT Kharagpur NPTEL

And by here you can use adaptive body biasing to you know you can apply different body bias to different dies. So, if the chip is too hot so, you will increase the body bias you can see and you will push it to closer to the nominal threshold voltage. Similarly, if the device is too slow then you can increase the body bias; that means, you can use forward body bias and you can have you can improve the performance. So, we will find that adaptive body biasing allows tuning of each die to meet specs leads to increased yields. So, what you are doing? After the fabrication has been done each die is provided with different body bias so, this can be done in two ways. So, entire die can be having a single body bias or different parts of the die can have different body bias we shall see.

(Refer Slide Time: 51:49)



Adaptive Body Biasing

ABB reduces variation in FMAX
WID-ABB: Moves ~100% of dies into highest frequency bin

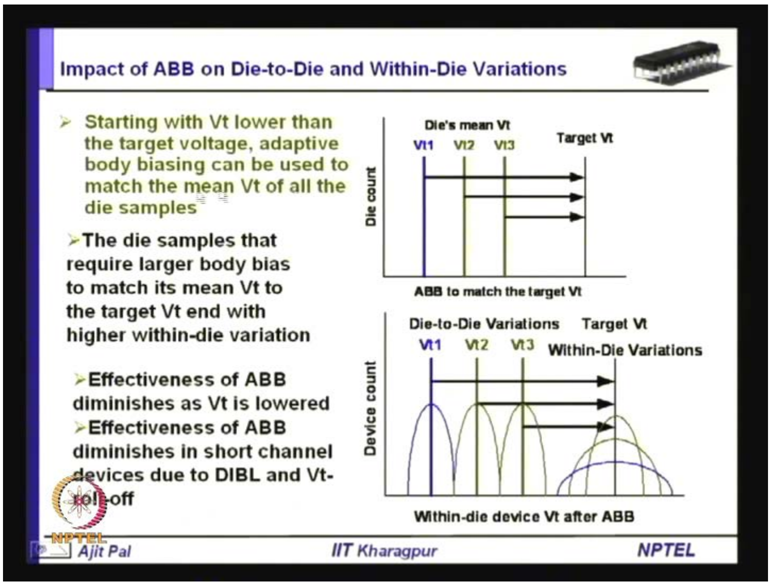
Variation -Tolerant Circuit : Solutions and Techniques,
Jim Tschanz, Keith Bowman, Vivek De, DAC 2005

Ajit Pal IIT Kharagpur NPTEL

So, in this particular plot as you can see these curves are without anybody bias so, we can find wide distribution of the normalized frequency and leakage. So, whenever you do adaptive body biasing those yellow plots corresponds to adaptive body biasing. So, we find that frequency distribution is over a small range; that means the performance is improved significantly. And particularly when we do within die adaptive body biasing; that means, different parts can be can have different body biases. So, this will definitely have better result instead of a single body bias for the entire die.

So, this within die adaptive body biasing leads to still better performance as you can see. Most of the devices those red plots corresponds to within die adaptive body biasing and yellow plots correspond to those I mean a single body bias for a for the entire die. So, this shows, how the adaptive body biasing can be used to reduce the to improve the yield?

(Refer Slide Time: 53:06)



Impact of ABB on Die-to-Die and Within-Die Variations

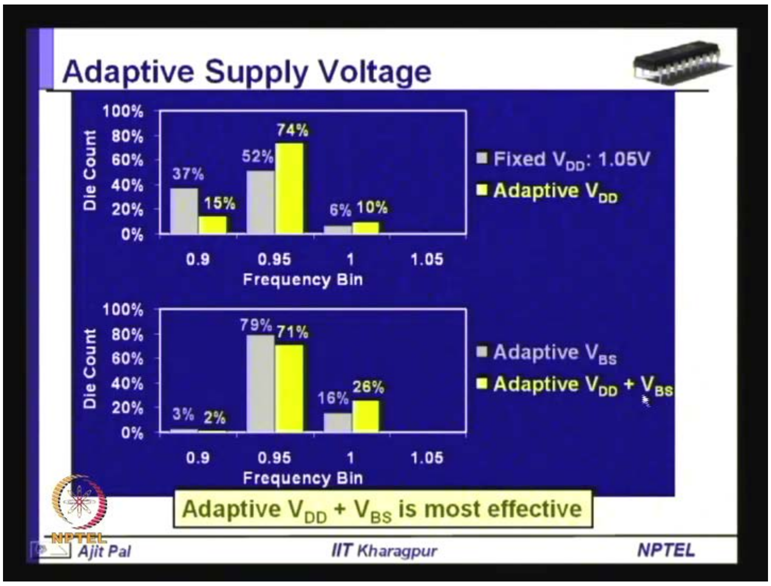
- Starting with Vt lower than the target voltage, adaptive body biasing can be used to match the mean Vt of all the die samples
- The die samples that require larger body bias to match its mean Vt to the target Vt end with higher within-die variation
- Effectiveness of ABB diminishes as Vt is lowered
- Effectiveness of ABB diminishes in short channel devices due to DIBL and Vt-roll-off

Ajit Pal IIT Kharagpur NPTEL

So, this shows the impact of adaptive body biasing on die to die within die and within on within die variations. To a starting with V t lower than the three target voltage adaptive body bias can be used to match the main V t to all the die samples. So, what you can do? You can start with a die with smaller; that means you will fabricate in such a way that, the threshold voltage will be smaller than the original. Then you will be using adaptive body biasing to have the target threshold voltage for all the dies.

So, this is how this die you are using as body bias tube so, that it reaches the target threshold voltage for this die; obviously, different body bias has to be used and so on. So, this is how you can do it with adaptive body biasing to match the target threshold voltage from die to die. And here you can see the for a within die variations a can also be can also use within or within die variations can be reduced by adapt by applying different body biases.

(Refer Slide Time: 54:19)

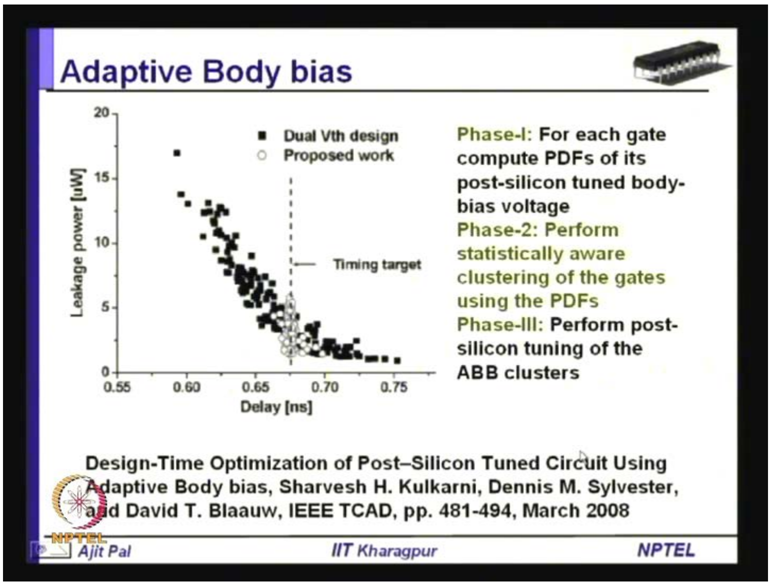


Adaptive Supply Voltage

Adaptive V_{DD} + V_{BS} is most effective

And in addition to that you can use adaptive supply voltage also, you can reduce the impact of these variations; that means, the slower dies will be used in (()) higher threshold voltage higher supply voltage. And faster dies will be using smaller supply voltage and that is how? You can see by using adaptive supply voltage how the frequency of operation is improved; that means, most of the chips are now close to that frequency bean one. So, instead of the original distributions the distribution is improving by using adaptive supply voltage. And here you can use you can combine adaptive supply voltage and adaptive body biasing can be done, you can see by using combined adaptive supply voltage and adaptive body biasing. Most of the device will be will be operating in the normal range; that means, slower devices are not existing.

(Refer Slide Time: 55:25)



Adaptive Body bias

Phase-I: For each gate compute PDFs of its post-silicon tuned body-bias voltage

Phase-2: Perform statistically aware clustering of the gates using the PDFs

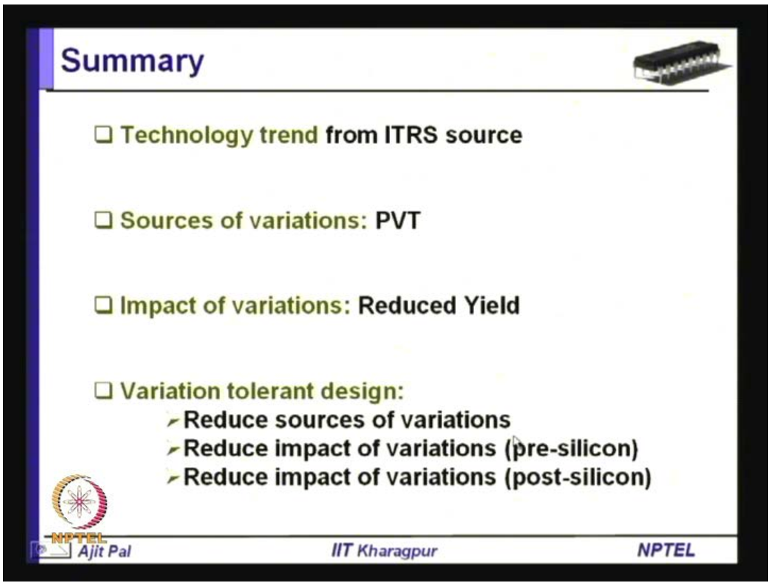
Phase-III: Perform post-silicon tuning of the ABB clusters

Design-Time Optimization of Post–Silicon Tuned Circuit Using Adaptive Body bias, Sharvesh H. Kulkarni, Dennis M. Sylvester, and David T. Blaauw, IEEE TCAD, pp. 481-494, March 2008

Ajit Pal IIT Kharagpur NPTEL

And this is the somewhat similar thing, which has been proposed in this paper by the group of by the Dennis Sylvester group. So, they I have used three phases for each gate compute probability distribution functions of it is post silicon tuned body biased voltage. Then phase two performs statistically aware clustering of the gates using these probability distribution functions. Then in phase three perform post silicon tuning of the adaptive body bias class test and by that you can see you can have very good timing result compared to Dual vth based design.

(Refer Slide Time: 56:10)



Summary

- ❑ Technology trend from ITRS source
- ❑ Sources of variations: PVT
- ❑ Impact of variations: Reduced Yield
- ❑ Variation tolerant design:
  - Reduce sources of variations
  - Reduce impact of variations (pre-silicon)
  - Reduce impact of variations (post-silicon)

Ajit Pal IIT Kharagpur NPTEL

So, with this we have come to the end of today’s lecture. And we have discussed technology trend we have discussed sources of variations process voltage and temperature impact of variations which leads to lower yield. And we have discussed various variations tolerant design reduces to reduce sources of variations. In the three approaches has mentioned at the time of fabrication then the poly silicon design approaches and post silicon approaches I have discussed in detail. So, with this we have come to the end of today’s lecture on variation tolerant design. Thank you.