Advanced VLSI Design
Prof. A. N. Chandorkar
Department of Electrical Engineering
Indian Institute of Technology- Bombay

Lecture - 03
Logical Effort-A Way of Designing Fast CMOS Circuits

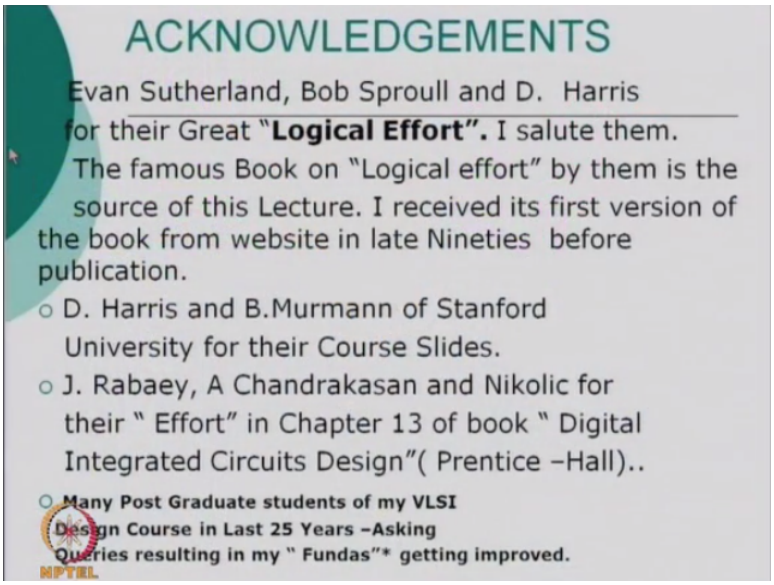
Last time we discussed about the perspective about historical perspective and trained in VLSI design. In that part of that 2 or 3 hours, I discussed with you one of the major requirement of a digital system these days is called high performance circuits and high performance circuit and the other one we said low power circuit and then low standby power circuits. This particular lecture of today which I am going to now go on for next hour or plus.

I am going to discuss something how to design this so called high performance circuit which essentially means fast SIMO circuits. We still believe that SIMO circuit will be going to be continued to have some kind of vehicle for any digital system for many more years to come. Now, this topic particular which I choose today way of designing fast SIMO circuit called logical effort.

Now, why this kind of method I am suggesting, as an analytical person many of you probably are not, most of you are right now designing things on computer using CAD tools but when you are designing a chip on a CAD tool using a CAD tool. Please be remember that there has to be some initial some kind of a back of the envelop calculations done by you to get typically which kind of design one should use to get the required performance.

And the starting point essentially is to be done mostly analytical way, so this part of the lecture essentially talks about how to design or start designing circuits for high performance and do it little analytically and the effort which is shown here is called logical effort.

(Refer Slide Time: 02:08)



Before I start I must tell you that this, I wish to acknowledge many of the people who has helped me to get to this topic. There are 3 people whom I want to actually thank rather salute them they are Evan Sutherland, Bob Sproull and D. Harris for their great logical effort. Actually, I salute them simply because they wrote a book on logical effort way back in 1999-2000 and that is the source of today's lecture.

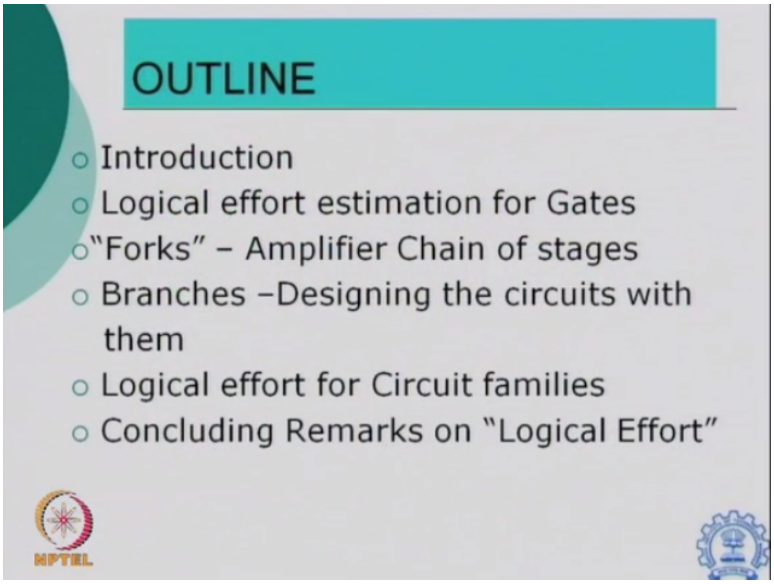
I received this, maybe it is interesting to know that I received its first version of the book from the web site of late 90s before its publication itself one of my colleague at Standford actually forwarded me the link and I could read this book, read this part of the book then and was extremely impressed how to do a design analytically for high speed circuits. I also wish to thank the one of the same author of the book which is D. Harris. he along with Bob Murmann.

Actually gave the first such course on this logical effort in their course at Stanford and I have been happy, I have been lucky I could get some of their slides. The third and the important book which I normally refer for my digital design book is by Jone Rabaey of UC Berkeley. He along with Dr. Chandrakasan from MIT and Nikolic from Berkeley.

They have the written the book on digital integrate circuit design publish by Prentice Hall and in their chapter 13, they also have talked about some part of logical effort and they just gave the word effort there. Of course, there is very concise way they wrote there. This particular lecture

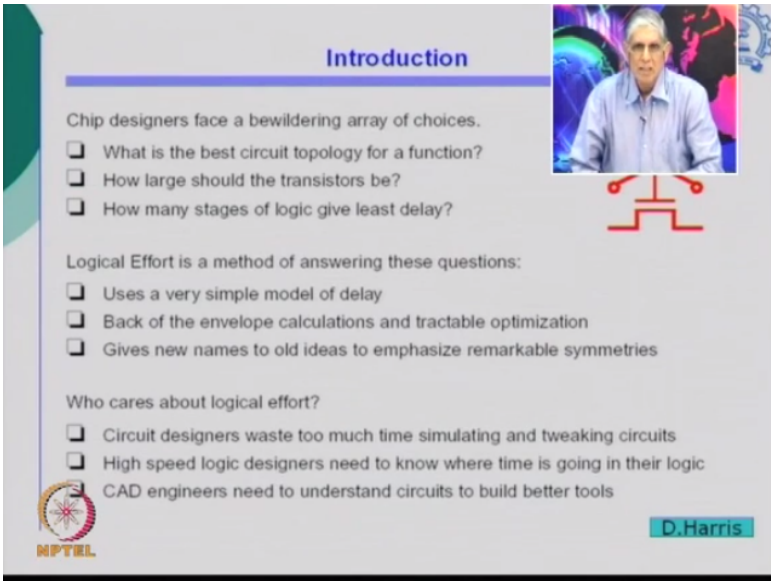
probably expands it much more what is written in this book. I am also thankful to many of my post graduate student of my various design courses in last 25 years. They keep asking me many quarries when I taught this courses and probably that has improved my fundamental to great extent. To start this talk, I have outline for my, this lecture.

(Refer Slide Time: 04:19)



I will start with introduction to some extent, I already gave you. The logical effort estimation of gates will follow, then I will talk about forks, amplifiers chain of stages, branches and logical effort for other circuit families and finally and I give some concluding remarks.

(Refer Slide Time: 04:38)



Now let us say you are a chip designer, if you want to design a chip for a given system you have a bewildering choices to make, this right side shows your difficult situation in which you are put to before you decide what to do, so what are the queries one gets into, so what is the best circuit topology for the function that we wish to implement. How large should be the transistors to actually give the performance?

And how many stages of logic will give the least of the delay. Now these queries are not very easy to answer in many cases. Many of us actually go on the using CAD tools, you have schematic created and start putting initial values of some width to length ratio of each transistors and start doing circuit simulations. You can see when you do such kind of this, tricking a value, is never accurate way of doing things.

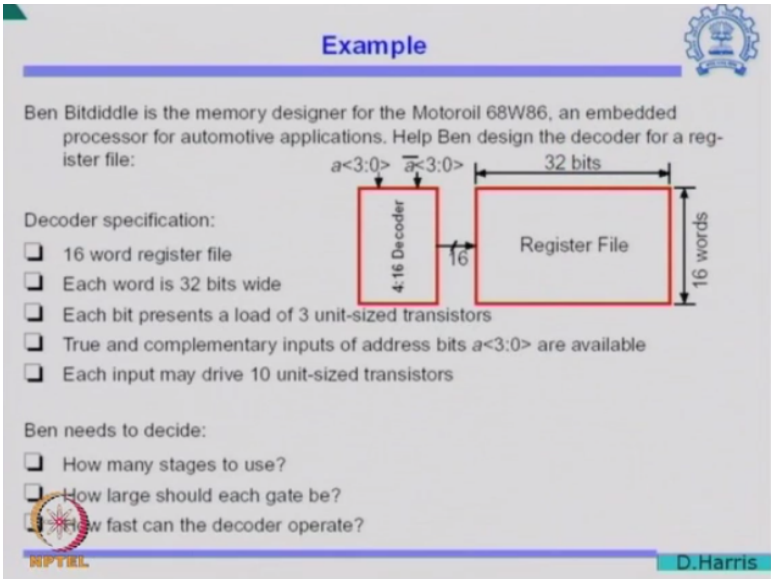
Because it may take larger time or it may never actually converge to a correct solution. So where to start, even that is very crucial even if you are using CAD tool. This logical effort is a method of answering many of these questions, what it does, it uses a very simple model of delay. Then it allows it to do back of envelope calculations and therefore it is very much easier to tractable optimization.

You can always do some kind of optimization because it is much, it is an analytical function at your hand and then gives new names to old ideas to emphasize remarkable symmetry. So, it is not that this idea of generating design method using logical effort is very different from what we are doing on the time but is slightly a modified and better way of doing the same thing. So who cares about logical effort in my opinion.

The people who are circuit designers and many a time they waste too much of time simulation and tricking circuits and therefore they should first start with logical effort work so that they do not have to waste too much time. Then, the people of work on high speed logic design, need to know where time is going and in their logical, which part of the logic has largest dealing and therefore to get the critical paths in the system.

This method may help quickly to know which is the critical path or even in a particular path, which block is giving the slowest propagation and hence contributing to larger delay and then finally of course the CAD engineers need to under circuits to build better tools.

(Refer Slide Time: 07:14)

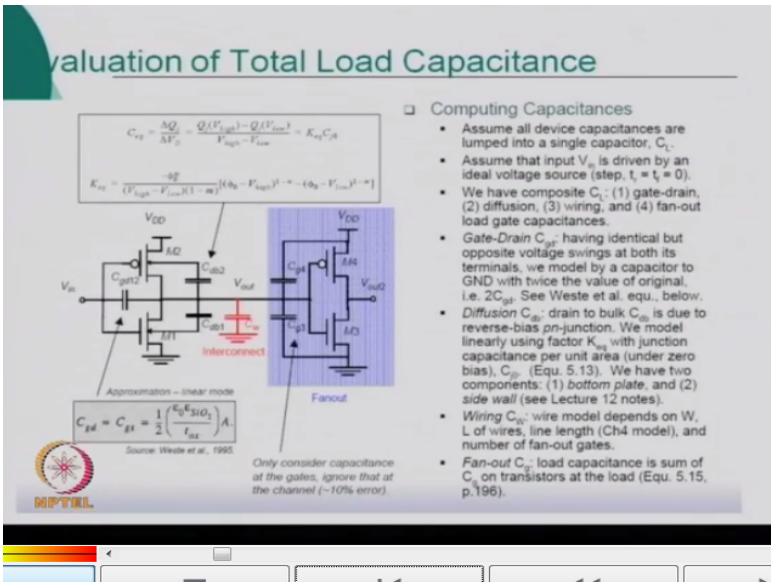


I am always of the opinion that a CAD tool designer should be a good circuit person because only then he can appreciate the problems in circuit and then can create a tool which is much more stronger in actually application of any circuit designer. Here is an example, of course, it was take from Harris lecture, Ben Bitdiddle is the memory designer for a Motoroil 68W86, he is designing a chip 68W86 which is an embedded processors for automotive applications.

Now typical some part of the circuit as shown below and now what is query is help Ben design the decoder for the circuit shown here. Now if you see a typical circuit to have 4 to 16 decoder to a register, which actually drives the register file, which is 32-bit wide and 16 feet depth, so which has essentially meaning 16 word registered file, each has 32-bit width, each bit represents a load of 3-unit size transistors to a complement input of address bits are available.

Each input may drive 10 unit sites transistors. So, Ben need to decide what can this requirement for design is, how many stages to use, how large should each gate and how fast can the decoder operate.

(Refer Slide Time: 08:35)



Now, basically when I talk about the speed, you must appreciate that the speed decided by the propagation delay that is the signal going from input to the output, how much time it takes that decides essentially the speed. Speed means essentially one upon the daily is equal to the frequency of operation and since we are putting a data at a given rate this maximum rate with which data can enter is decided by this clock frequency or is decided by the highest frequency available.

Now here is a simple inverter shows, so what if the are invertor shown, you have a transistor, two transistor, one P-channel transistor as a upper transistor and you have M-channel transistor, M1, which is connected in a complementary mode. Now five given input at VN here which is common to both PN channel as a complementary input and this is my output node.

Now output is connected to its load, may be a similar inverter and maybe different inverter more than one such inverters which is called and the fan out or the load of the circuit through a wire which brings this output signal from the first inverter to the load part, so what essentially are now saying that when the input changes how fast this output will change because the change in this node voltage can be decided by the capacity charge stored from, at this node.

So, when we look into, actually, what is the net capacitance this node is actually seeing when input is changing or input is actually given to this input A. Now, how do we calculate the

capacitance, so we go back to theory of transistors and a look into this fact very carefully, we say okay since there the mass transistor and in a mass transistor you have a gate capacitance associated with the outside capacitance.

And also you have a parasitic capacitance between gate and source, gate entry and similarly for N channel there is a capacitance between gate entry and gate and source. There is also a capacitance which is PN junction capacitance associated with source and drain and there is a capacitance which essentially we call drain to substrate or bulk capacitance and drain to source to this capacitance.

Since there are so many capacitances associated at this node, when asked to then find the net capacitance, so there is a capacitance CDB2 and CDB1 coming from these two transistors at this node. Then, there is a capacitance associated with the wire, wire means any interconnect line between this node to the next node, will have some capacitance and this capacitance is we call it interconnect capacitance, CW.

Finally, since you are some kind of a load whose input will be capacitive in case of mass technology base or systems, so there will be some input capacitance at this stage which essentially in this particular case it to gate to source capacitance of both N-channel and P-channel and essentially one is talking about the capacitance of the gate oxides.

Now, if this net capacitance here which is the load for you plus wiring capacitance, plus the CDB2 plus CDB1 at this the net capacitance is some of all these capacitances and that is the load which this inverter is seeing. Now unless you compute all of them, one cannot see whether the capacitance will charge how fast or how slow it will discharge or how out fast it will discharge, so change in this node voltage potential at the input changes can only be known what is the value of this capacitance.

And what is currents these circuits are able to provide for charging or discharging. Now, I will just give the idea here how do I calculate each capacitance before we proceed further. Let us assume, right now, all device capacitance are lumped into a single capacitor CN, so point when I

am saying, this plus, this plus, this all of them together I call is a load capacitance. Assume that Vin is driven by an ideal voltage source.

That is at this point I can give a step input voltages 0 to 1 and 1 to 0 or 0 to VDD and VDD to 0. Now, I am assuming that this is a step source. Now, we have a composite CL, you know, this plus this plus this, so we want to know the net capacitance coming out of it okay. The first capacitance you see from here this is called CGS or CGD, similarly this is for input site, obviously this is also relevant.

But for output side calculation these capacitances do not come into picture simply because they are not connected at this node but they may connected, if these are driven by some other source, for this source they may act like a load for them. So we proceed ahead, we say okay here inverter driving this capacity loads CL, in which there is a CGS and CGD and two capacitances. These are called gate to drain capacitance and the other is, the idea behind is if there is a channel existing in the transistors.

Basically what we say the gate to channel capacitance can be divided into two parts. Half of it we will say go into the drain side and half of going to the source side by similar logic if this is the N channel transistor, half of the capacitance is lumped at the drain and half of that lumped to the source side and we calculate this CGS and the CGD, both CGS values. Please remember this is the source of the P-channel transistor, this is source of an N-channel transistor.

The drain capacitances, actually drain to bulk or drain to this are not taken care because they do not appear in this node okay. Now this CGS and CGS for P- channel, N-channel will act as net some which is appearing here. There is a diffusion capacitance due to drain to bulk capacitances, you know after is a diode, PN junction diode sitting between source and substrate and drain and the substrate and since these are reversed by it junctions.

They will contribute to a junction capacitance. So that is called diffusion capacitance, CDB and CDB1 for N-channel and P-channel and finally there is a capacitance and of course this you can think of this, they are essentially decided by something we call a model which is linearity model,

which is a factor K equivalent, which decide how much is the junction capacitance, how much is the junction capacitance associated.

Because we know at different biases the junction capacitance will vary because the depletion layer will vary and because therefore we must evaluate the capacitances at given biases during the transition here. Of course, this is being done already in earlier notes of yours, but if you were really want to know more about please look for (()) (16:12) book or any other book on digital CMOS VLSI design and find out how to calculate CDB or any device book for that matter.

Then, there is of course as I said if the wiring capacitance CW which depends on width. Please remember any width and length of the wire epsilon A by D if you see a capacitance the width of the wire and the length of the wire will decide essentially, because the thickness is fixed which is the oxide capacitance thickness of the field oxide and therefore it was not much different from the other side only width and length of the wire may decide the wiring capacitance.

Longer the wire interconnect are flowing, larger will be capacitances associated. So please remember there is a wiring capacitance, there is a capacitance due to the input of the stage which is your load stage and there is an output capacitance coming from your driver's stage. This is the inverter which is driving this load and is connected by wire. Please remember this is called fan out, this is only one inverter shown or one gate shown.

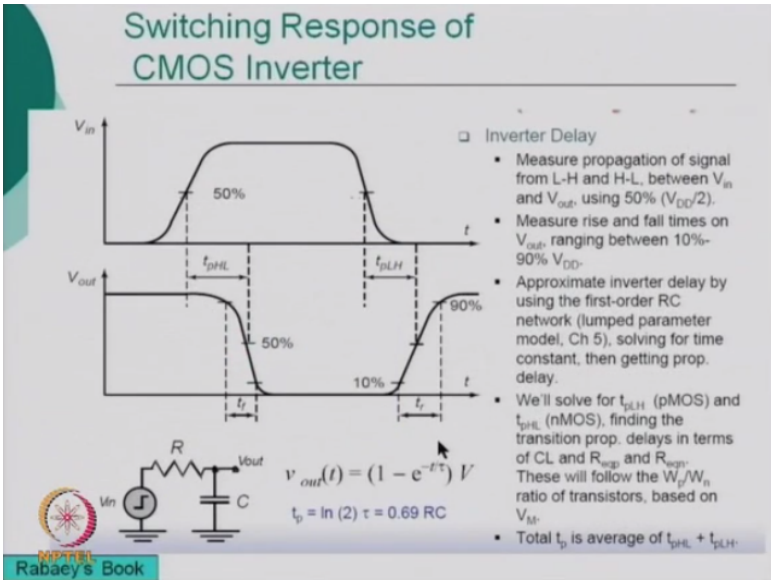
There can be more than one such gate where is this notes is connected down up or many more and is called fan out, so larger the fan out, obviously the load capacitance will also increase. Now, how do we calculate the different capacitances, we know capacitance is delta Q by delta V so e calculate the junction capacitances using simple PN junction theory.

We may assume the PN junction is to be abrupt or you can say stepped the other kind of junction which is linearly grade or otherwise and from that we can calculate some constants and therefore we can evaluate some great extent the capacitances of the junction and different voltages. So I can calculate CJ of the junction at different voltages and there from I can get CDB2 and by similar logic, I can get this.

Please remember the capacitances here which is the capacitance between gate and source and gate and drain if you see very carefully, these capacitances are essentially given by as I say they are lumped capacitances half here and half here and one can say it is nothing but the oxide capacitances of each, so K is epsilon 0, rather epsilon 0, epsilon SI 2 by Tox into half, we are half on this drain side and half on this drain side.

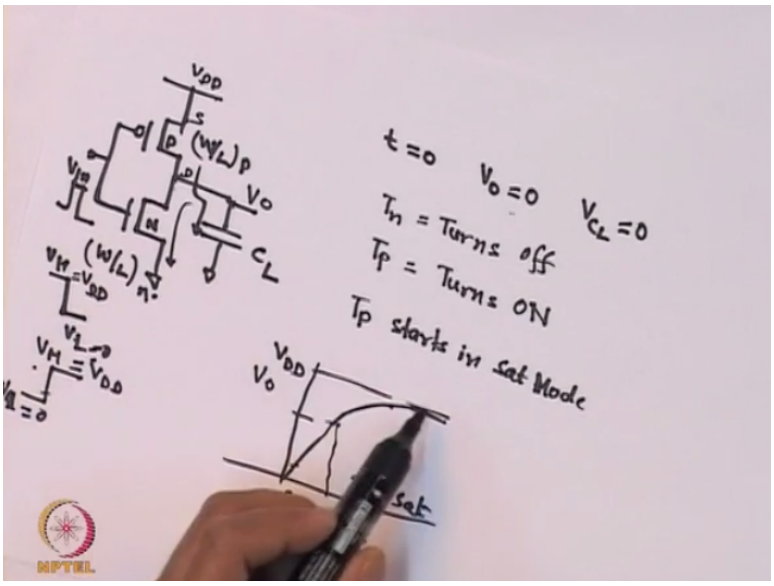
The net capacitance is called CGD which is also equal to the source capacitance, this is half of epsilon 0 epsilon I2 by Tox into area of course and using this ORI is W into L that is the width, obviously these capacitances and all of them are proportioned to the width of the transistor.

(Refer Slide Time: 19:03)



Now switching response, you know, if you want to find out the switching response, what I actually start talking about switching response shown here there is an input pulse given to the circuit shown already by me and based on that input circuit, I am going to actually try to tell you how the currents will flow.

(Refer Slide Time: 19:23)



For example, if this is my inverter, this is my P-channel transistor. This is my N-channel transistor, and this is my VN, okay, input VN and this is my output, this is my V0 and it is essentially I say all of it, all the capacitance lumped at the output is CL. So this is the P-channel, this is N-channel, let us say this is W by L of P, this is W by L of N, is available to known to us and because of that we can evaluate the currents.

Now, if you see the response initially, let us say when you say at T is equal to 0, V0 is 0, therefore VCL is 0, so what does that mean, it essentially means that this transistor, N channel transistor will be fully turned on and unless this is fully turned on, this node voltage cannot go to the ground. So, obviously this voltage is going to the ground because of the CMOS action, when N channel was fully conducting, the P channel must not conduct, is that correct?

It must not conduct. So obviously initially P channel was not conduction, N-channel was fully so essentially you are at this point when VN was high, making N-channel fully conducting and P-channel switched off. When I go the step down, from V high to V low, as it is, let us say, V low is 0 our case, in case of CMOS and VH is normally VDD, so what we say that when VN goes to high to low, initial condition is T is equal to 0, V0 is 0, therefore no charge across the capacitor.

When the input switches down, obviously at the gate of P-channel, you see a voltage zero and whereas you see at the N-channel, you can see a voltage 0, so obviously transistor TN turns off,

whereas TP turns on, call this TP perchance turns on. On means it is conducting, so obviously during the transition this input voltage is changing from high value to a lower value and at the start of that value what is VDS for P-channel transistor.

The VDS for this channel is essentially V0 minus VDD because this is the drain and this is the source of a P-channel, so V0 minus VDD and V0 was 0, so obviously this voltage is very high, any decrease in VN towards 0, the VN minus VS which is the VGS, so VGS minus VT starts actually increasing but at that time initially when you are going down this is the smaller value, you subtract it out of VT, so this is even smaller.

This voltage across drain to source is very high, the device enters TP starts in saturation, in sat mode, transistor is saturated okay. Whereas, if it is in saturation, we know the current of a P channel device in saturation but as the current starts moving coming from here, it start, since N-channel is switched of, we believe of course not all of it initially, we may say all of it but in reality some current may through N-channel, so this power supply through P channel now provides you a current to charge this capacitance to a higher value of voltage.

Essentially means, if I plot V0 versus T, initially voltage starts rising as the current start charging the capacitor. There will be a time when V0 become sufficiently high, VN is already zero, you can see the VGS now is very high. VGS minus VT is high, whereas V0 minus VDD is coming down because V0 is increasing, at that time after certain values, let us say here, the device then, this was sat partial, then we say ahead of this the next part is due to non-saturated P device.

So device enters non-saturation and a then you have a nonsaturating current which function is V0 and which starts charging and finally of course it reaches to VDD and time taken to reach to VDD we called as the rise time 10% to 90% as we shall show you later and we say the switch has occurred, input has gone from high low, output has gone frown low to high.

Now this is called charging transient, by similar argument, one can see from here when you have an input low, V low is 0 and you go to VH, which is essentially VVD, so initially V0 was at VDD but at that time this transistor was fully on as soon as the output, input rises to VDD, P

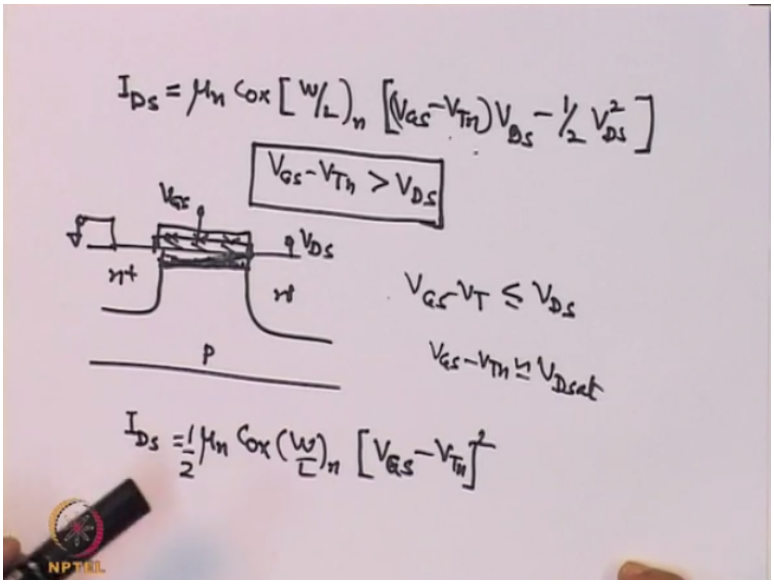
channel switches, N channel turns on and the charge on this capacitor which is CL, VDD now starts discharging through this N channel which is turning on.

Now what is the status in which TN transistor turns on because initially it was a 0, so transistor was off, then VN exceeds VT and channel turns on, but VGS minus VT is still very small, VDS is very large, VDD minus 0 or slightly less than VDD minus V0, so obviously transistor again N channel also wakes up in saturation and once it starts saturating, as this capacitor start discharging, the V0 keep falling and VDS of N channel start reducing and since VN has already reached to VDD.

There will be a point when VDD minus 0 minus VTN is larger than V0 minus 0, which is the VDS and the transistor will finally enter non-saturation. So in both cases the initial rise or initial fall is because of the transistors are in saturation and the next when it is actually one third way down, both devices, both in the case of charging the P channel enters non-saturation and in the case of discharge transition, N channel transistor goes into non-saturation after a while.

Now, this is called transient response and the time taken essentially reflects. So now if you see a current of any N channel transistor or a P channel transistor, one can see from here what is the current one can get.

(Refer Slide Time: 26:57)



Typically, current in a MOS transistor can be given by let us say N channel first, Mu N, Cox, W by L of N channel, VGS minus VTN into V 0 or VDS, minus half VDS square, this is when VGS minus VTN, VNT is greater than VDS. For this condition, we know transistor is said to be in non-saturation.

Channel exist in a MOS transistor, let us say this is N channel shown here, this is a piece substrate, this is your gate oxide and this is your gate, this is your VDS and this is your VS, source is normally grounded okay and you are applying VGS here, so obviously if the channel exist throughout, okay, VGS minus VDS must be larger than VT, which essentially means VGS minus VT must be larger than VDS channel exist throughout.

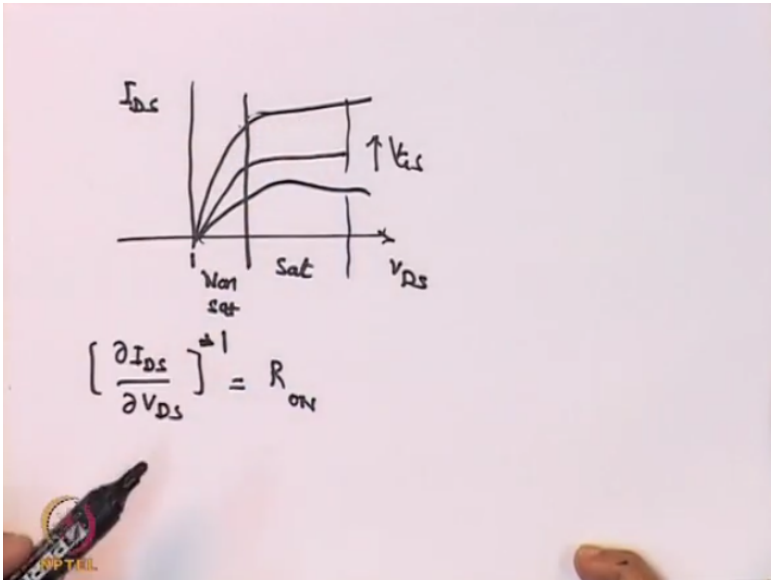
And we say device is in non-saturation. However, if I keep increasing VDS, for a given VGS as well then it may occur that at some point of the time VGS minus VT will be less than VDS and if that occurs, the channel at this end will get pinched off and if you further increase it, channel point will move on the left side to a source and you say you are in saturated region of the transistor.

So since the currents in the case of nonsaturation is shown here, I am showing right now for N channel alone, Mu N Cox, W by L for N channel and then we say it is VGS minus VT whole square. Because we say VGS minus VT is much smaller than VDS, VDS is much larger and at some point for a given VGS minus VTN is we called VD sat, a fixed value at which transistor will enter saturation and then VD sat square minus half VD sat square which is essentially reconverted back to VGS minus VTN square.

Now the point I am trying to make, if I want capacitor to charge or discharge faster from the current point of view, if I had to charge faster I must provider you larger current and if you want to discharge, I must also provide larger discharge current and the current sat proportion to size okay and of course if it is N channel, it is mobility is larger than P channel device, because mobility ratio is larger.

So we see now that if I want to improve the charging transient or the discharge transient, why method of doing is to increase the sizes. Whenever transistor has a size and transistor normally, if you see I-V characteristic of a transistor carefully.

(Refer Slide Time: 30:31)



If you see transistor characteristic, IDS versus output characteristic of a transistor, IDS versus VDS, so you can see for a given VGS, it is something like this, for a different VGS, it is different kind of thing, VGS increasing, so what we are trying to say you that up to this for example device is in non-sat mode and beyond this somewhere here is sat mode, if you see the resistance on this side.

And if you see resistance on this side of the transistor, delta IDS by delta VDS to the minus one is the resistance of transistor, R ON. So obviously if you see R ON, this in nonsaturation, the device has much smaller resistance whereas in the case of saturation, the resistance is very, very high, extremely high in fact, so correspondingly R ON if you see in a transistor, and since the transistor acts like a resistor, we can see from the figure now we go back to the slide here, showing here

I am now talking as if both charging and discharging transition acts like you have a voltage source, you have a transistor resistor equivalent of transistor replaced by resistor both for discharge and charge, charge it will be P channel transistor resistor and discharge it will be N

channel equivalent resistor and this is the load. So how do I charge or discharge a capacitor, particularly if I am charging it, V out is one minus E to the power minus T by Tau into V finally.

So, we now figure it out the charging transient essentially decided by the TU, which one can say typically the delay time is associated is 0.69 RC, propagation delay. So, what essentially is trying to say, Tau is time constant RC, so and if you solve this one can say the appropriate time take to transmit signal from here to here is LN2 into this, this is approximation because currents are not linear but assuming that so in this case is true, so one can say.

If I want to improve the speed, which means the propagation delay I want to reduce, RNC should go down, but C is increasing because that load is not know to me. So, if C is larger or not known to me only where I can change the speed is by changing the R. If I want better this, R should reduce and I just now said, the on resistance where transistor is smaller, if the size of the transistor is larger.

So, one can now get from these simple calculations that all that we are now trying to tell you is that if I want to improve the speed of the inverter, both resistances of N channel and P channel should be correspondingly reduced at a given input voltage for a given load, such that the propagation delay time is minimized.

The figure here, I think having shown you, the figure here shows now a transient response of an inverter, or switching response, this is my input versus time, I am giving a input pulse and I figure out that even after I give my input starts rising, it take finite time before the output start falling down by simple reason that there is a input capacitance also needs to be charged before this will start responding.

Now once N channel start turning on when the input is rising, initially input low essentially means N channel is off and P channel is on, the output is high. Now input is rising, as input rises, the N channel turns on, P channel turns off and because of that I can say the voltage at the output as we discuss just now keeps falling.

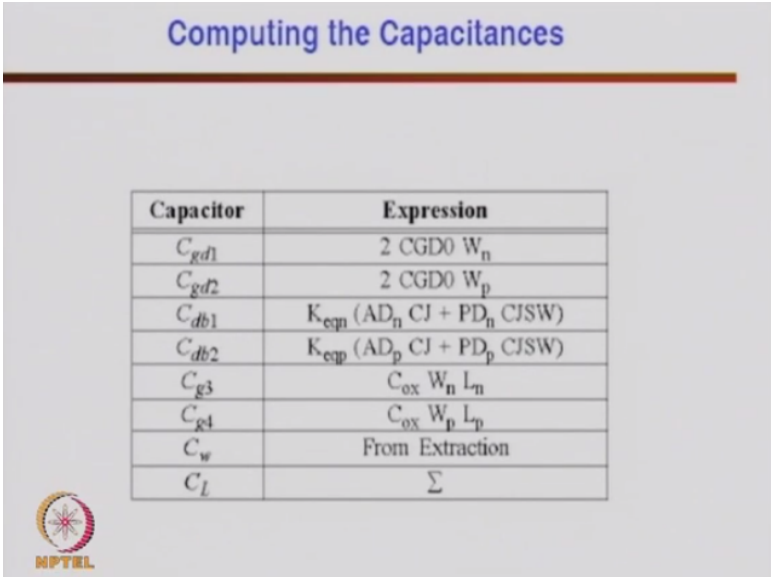
Now, definitions of this is essentially like this. The time taken from 90% to 10% of output from higher to lower is called the fall time by similar logic when you are going from high input to lower input, we say when the input rises from 10% to 90% the charging transition that is called rise time, but we define two other important terms, which we say TPHL and TPLH.

We say 50% point of the input to the 50% of the output the time delay is called TPHL here and for the charging transition 50% of the input to 50% of the output we called TPLH. This we already discussed, all of you know very well, this is very well done in the (()) (35:27) book as well as it is given in Rabaey's book, so you do not have to, but since it was a again a video course, I thought I will repeat what I did early in the case of this.

The net delay one can see from here, average delay one can say, not net. If the total signal has to cross the inverter, you have to go through an average delay of TPHL plus TPLH, average means TPHL plus TPLH by 2, so we say on an average a signal pass, one bit of signal can pass in a delay time which we call propagation time TP, which is nothing but average of TPHL and TPLH.

So, if we reduce both TPHL and TPLH obviously my TP will decrease, and therefore the circuit will become faster. Now, how to compute capacitances, here is something interesting things I shown you.

(Refer Slide Time: 36:25)



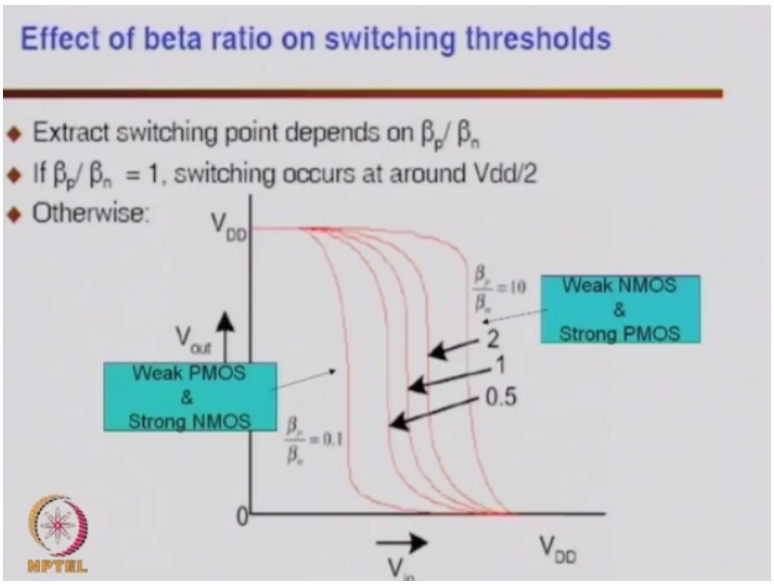
CGD1 is 2, CD0 essentially stand for capacitance per unit length, so if we multiply it by width then it become CGD1, if I multiply by P channel transistor width then it become CGD2. Then these are the junction capacitance, this is junction constant as I derived in my last slide. A is the area, P is the perimeter of this transistor area, on the area and one can then evaluate both junction capacitance CDB1 and CDB2.

If you want to know what is the capacitance at the input, CG3 and CG4, essentially it is net oxide capacitances, so COX is capacitance of oxide per unit area multiplied by areas which is WN into LN, by same argument for a P channel, it is COX WP into LP and of course wiring capacitance decided by technology used the length of the wire you are using, so you need to extract from the circuit layout what is the CW value and finally the CL, the net capacitance is sum of all of these.

Now, here is another issue in the transition one looks for, so one can let me go back again, so if you see this capacitance once again, so obviously if I reduce the capacitance in RC time constant than the speed will improve. So to improve speed, one must look into in the area term, area essentially comes from WN.

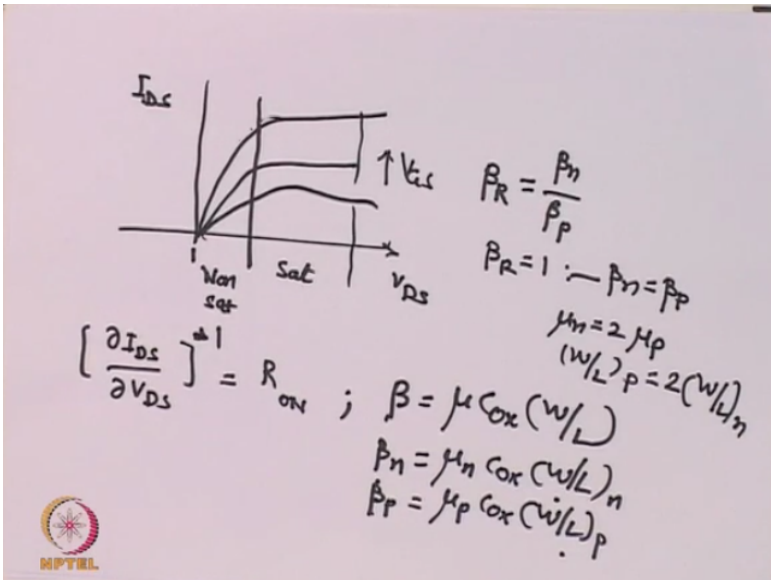
So essentially, the dimension of the transistor WNL, they actually decide what is the kind of speed the circuit is going to have and apply, do we start with the minimum area transistor W equal to L equal to the feature size or do we start with some numbers which is closer to the expect value of TP and therefore can make the guess where to start. Our today's effort is to show you how to get these initial values of W and L.

(Refer Slide Time: 38:33)



Now another issue, which many of us are worried about in the case of CMOS circuit, is what we call as switching thresholds, now why are we worried about switching threshold is the following and typical inverter character output characteristics or transfer characteristics as shown here input against output for different value of beta P and beta N. Beta of course as I say in my calculations, if I have not shown you somewhere, I may repeat.

(Refer Slide Time: 39:05)



In my all of my calculations beta is Mu Cox W by L. Obviously if it is N channel, I will called beta N is Mu N Cox W by L, beta P is Mu P Cox W by L of P and W by N. Obviously, if I define my beta ratio as beta N by beta P, from these two expressions, one can see that if Cox is same

and the mobility ratio of N channel to P channel in the surface mobility ratio is 2, typically, to make beta N is equal to beta P.

The size of the P channel, must be doubled that of size of N channel and that is very important. This is called beta R of one means beta R of one means beta N is beta P, assuming Mu N is 2 Mu P, we say W by L or P is twice that of W by L of N. So, one can see from here that if I change the beta ratio. I am now seeing the transfer characteristics is moving from end to the other end and what is importance of this transfer characteristic.

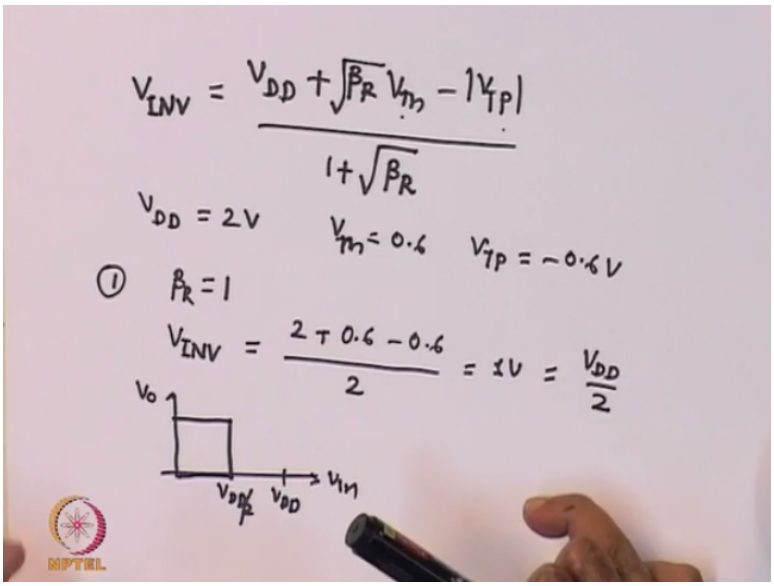
Once can see from here the point this characteristic start falling at this point when beta R is 0.1, this characteristic is say beta R is 0.5, so what we see as I increase beta R, the transfer characteristic starts moving. Now what is the importance of this region where it falls from high to low, this is called a point at which transistor goes from high to low and is the most important characteristics of any inverter. This is called the point at which it actually falls down, is essentially called inverter threshold.

Now, one can see inverter threshold is say for one it is somewhere here, closer to 50% of the EDD whereas if I reduce beta ratio further. I mean beta R ratio higher which is beta P, this is beta R ratio of 10 and this is beta R ratio of 0.5 this, so one can see if I increase the N channel device width to length ratio compared to the P channel larger, my threshold point moved towards left and we increase the P channel device larger, compared to N channel, width wise, then I move towards my right, essentially the points keep moving away from its original.

Let us say at beta R = 1 that is N channel equal P channel beta, you are here and then move either this side or you move on this side. One can see beta R is 10 or something or higher, larger N channel size means strong NMOS, smaller P channel size means weak PMOS by same argument if beta R ratio is 0.1 or lower, then we say you are in weak NMOS and strong PMOS. So, depending on what kind of sizing you do, this inverter threshold actually changes.

What is the important in inverter threshold one can see from here, the inverter threshold typically for a CMOS shows very interesting value.

(Refer Slide Time: 42:48)



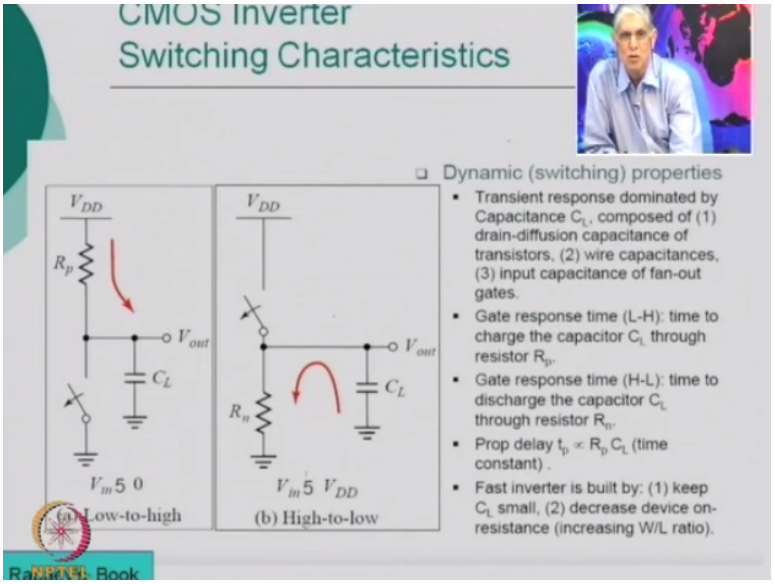
It shows V inverter is equal to VDD plus or other, okay beta R VTN, minus VTP divided by 1 plus root beta R, where beta R is this, VTN is N channel threshold and VTP is P channel which is negative. However, let's take an example, VDD is 2 volt, VTN is 0.6 volt, VTP is also -0.6 volt and let us say initially beta R is 1, case 1. Then, I get V inverter is equal to 2 plus 0.6 minus 0.6 divided by 2 or 1 volt, which is essentially VDD by 2.

So if you have a ratio of beta R equal to 1 and VT are same magnitude wise, then you get 50% point as inverter threshold and that essentially means you have a good switch. This is V0, this is your VN, this is your VDD by 2 and this is your VDD, so obviously 50% of input side you are high and 50% is low. It looks more like an ideal switch and that is what switch circuits are expecting.

So, beta R equal to 1 is very good if switching inverter threshold is 50%, however, many a time these sizes will be decided by the speed since we already said the speed decided by the capacitance. The optimum value of beta R will never come, equal to 1 in case f speed requirement, however for a good inverter threshold to get you expect beta R to be one and therefore there is a design issue which we have to worry about.

Because in whenever your inverter threshold moves from left to right or right to left, obviously the noise margin are comparatively different for rising and discharge transients and because of that you may have a noise problem at the end of the day.

(Refer Slide Time: 45:18)



So, if I go ahead, this is what I think I already said so, maybe I repeat again just because the figure is here is taken from Rabaey's book, when the input goes from, initially input was high and suddenly went to 0, N channel device trans off so as the switches opened, then the P channel is fully on, initially in saturation and then in RP is varying but initially RP is larger and then become smaller as you enter from saturation to nonsaturation.

The current starts moving from VDD towards capacitance, net capacitance would start charging, so input is from low to high, as output is from low to high, initially it was discharged, now it is charging towards the high, so it is called charging transient. In the case of discharge transient, the input goes from high to low suddenly.

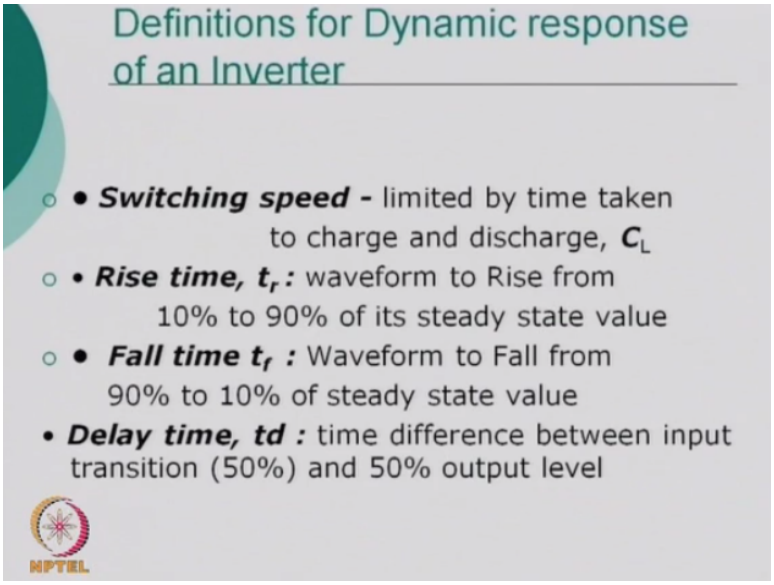
So one can see initially this was fully charged to VDD, low to high and therefore as N channel fully turns on, P channel switches off. Now, this capacitor will be fully charged to VDD voltage actually discharges through this and depend on this value of V out or VC, the status of transistor changes, initially transistors in saturation, higher resistance, so initial transient is slower but as the voltage falls, the transistor enters non-saturation, RN becomes smaller.

And the capacitor then discharges faster. Now, to get an equal kind of charging and discharging one can expect RPCL must be equal to RNCL net, RPs are non-constant but equivalently saying RPCL must be equal to RNCL. If you can achieve this delay, equal to each other, then you say propagation is universal or symmetric. So, if I want to design a propagation delay which is average of TPLH and TPHL, one has to keep CL small which is of course is the ideal.

If I reduce the speed, if I want to improve the speed. I must reduce the capacitance, but that is not in our hand too much, however, some porosities can by minimized by layouts. We must decrease the resistance says RN, RN and RP and to do this we must increase the size of transistor W wire. This is a very simple first order calculation one shows that all I had to do is to increase the size N channel and P channel.

However, to maintain a good inverter threshold I may have to adjust the ratio of beta R as well as and to get a good noise margin, so there is a trade of between noise margin and the speed to some extent.

(Refer Slide Time: 47:57)

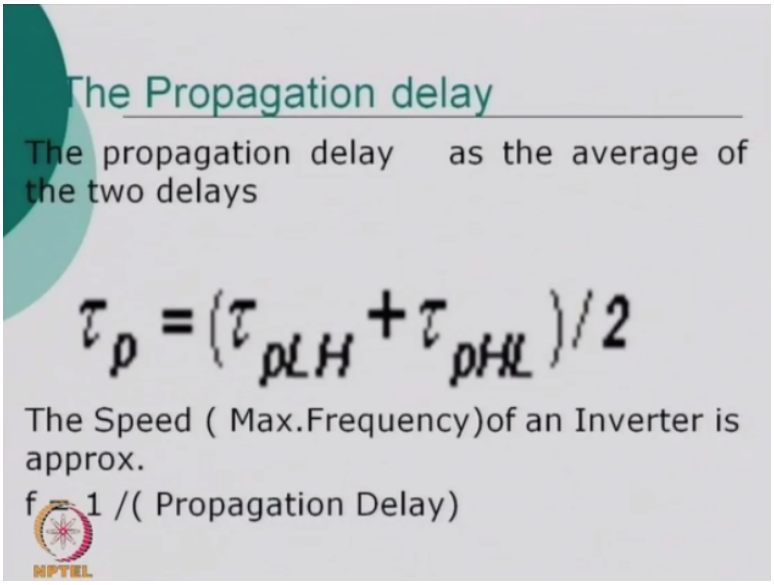


So, before I move further, let me give what we had learned, so far we have not entered the area of logical effort. Before we start the logical effort part, I just want to recapitulate whatever we did earlier or we know about in the case of dynamic response. Four terms we keep talking about,

one is switching speed, which is limited by time taken to charge and discharge CL, then we say rise time.

A waveform to rise from 10% to output, at the output please remember output waveform to rise from 10% to 90% of its steady state value is called rise time and waveform to fall from 90% to 10% of steady state value we call it a fall time and then we say delay time TD, time difference between input transition and output transition is TD, average delay.

(Refer Slide Time: 48:51)



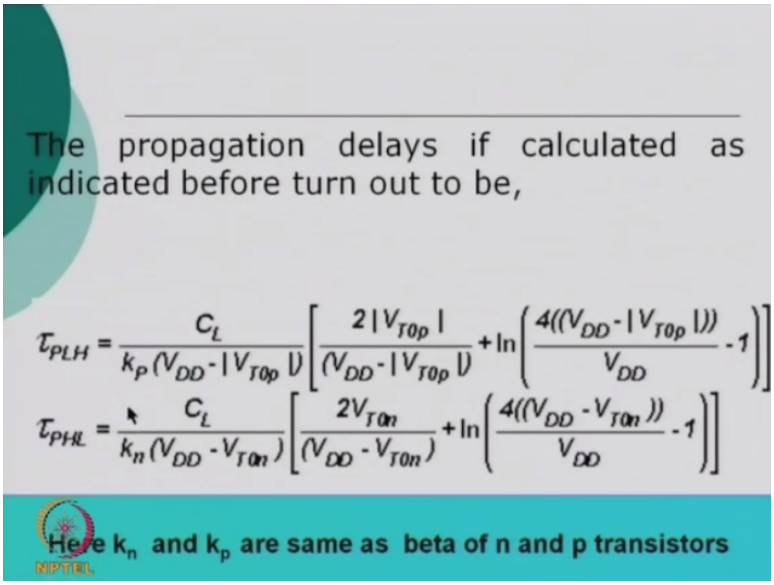
And we define therefore that average delay as a propagation delay which is sum of average of TPLH + TPHL and the maximum frequency which is allowed typically by any inverter for propagation talker is one upon propagation delay. Essentially, many a times a designer does not chose upon toupee as the maximum speed, he assumes then in this calculation of TPLH and TPHL which are directly function of capacitance.

He has underestimated the value of capacitances and the worse case all porosities and underestimation may add to let us say the maximum value may be 4 times the capacitance which was actually there but I estimated only one type and therefore the worse frequency which an inverter can always work is one we say one upon 4 Tau and that is the safest system speed you can work on.

But when you are scaling down the circuits now okay the dimensions of every transistors, actually we are reducing the sizes so much that the propagation delay should become smaller and smaller. So we start taking now this one upon 4 Tau as over universal this, this will be beating the something which we are trying otherwise, we are trying to improve toupee itself without actually going for C.

But now you say okay thing that we have not done well and make it four times Tu, which is not fare and therefore evaluation of toupee now is becoming very, very crucial and essentially. Therefore, it is becoming very important to know what is the capacitance value one has to drive and what is the current each in charge and discharge time I can provide, so that I can minimize TPLH, TPHL with the available capacitances which I may get and therefore what speed a particular circuit can attain.

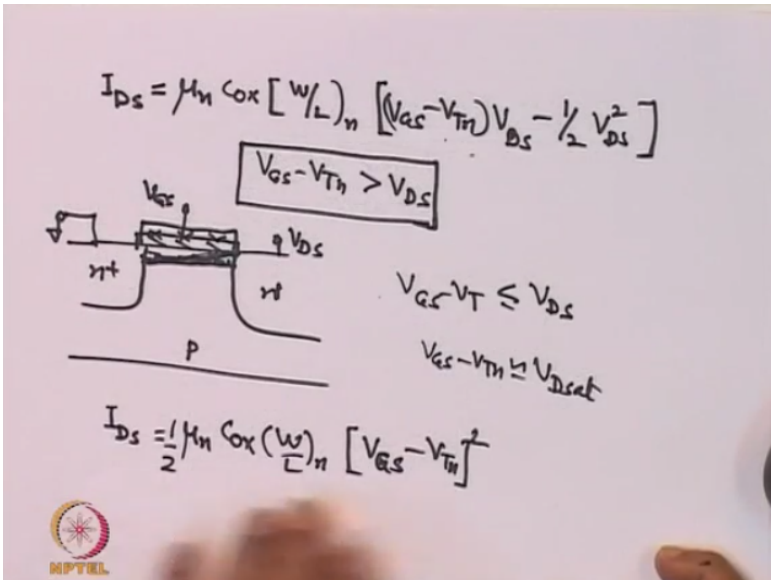
(Refer Slide Time: 50:44)



If I do this analysis which transient analysis I said, typically I can derive this expression, TPLH is CL, net capacitance, KP essentially is half, KP by 2 is beta P but right now we assume 2 has been taken care, beta by 2 is K, beta Mu C of W by L. This is spies version, they say, instead of beta P by 2 or beta N by 2, the write KP and KN, so you can see from here this is beta P by 2 in case and this is beta N by 2, CL upon this which is P channel, this is P channel threshold.

This part of the charging transient, low to high, you can see from here if you see these many term, this essentially looking a linear kind of thing due to the larger resistance provided in the saturation. Whereas, if you see the second part, which shows log term, one can see from here why this log term is appearing. Log term is appearing simply because in the case of non saturation, we are seeing the current which I may show you again.

(Refer Slide Time: 51:59)



In the case of saturation, this is like square log, whereas in the case of nonsaturation, there is this nonlinearity or some kind of a transient system is appearing, VDS is here and VDS square, some quadratic term appear in calculation of VDS, so therefore, if you solve this using these current values, you will essentially get into a logarithmic solution or exponential E to the power T by Tau kind and therefore this will give you log term.

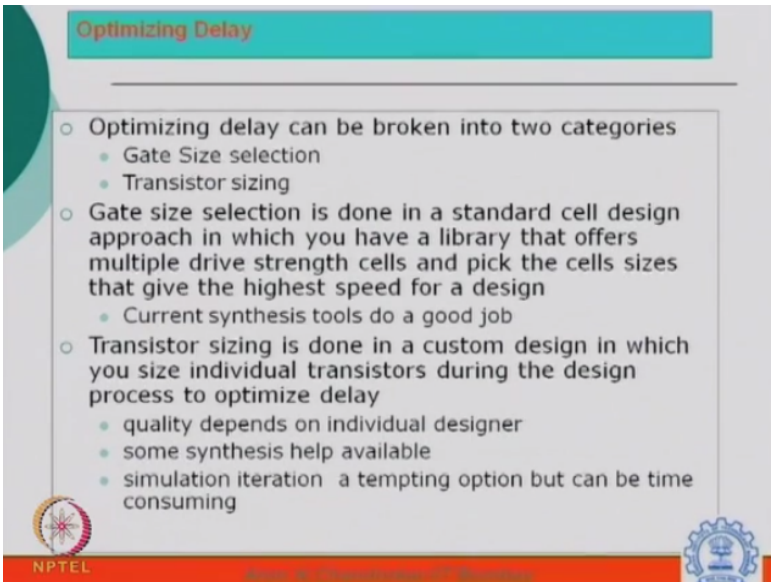
So the long term appears when the transistor is in nonsaturation and linear terms occur when there are devices in saturation. So the first part of the TPLH and TPHL. Both are essentially into saturated transistor mode and the second part that is CLK and VDD, this into this, is essentially into nonsaturated N channel, this into this multiplied is essentially be charging due to these nonsaturated P channel and one can see from here this value is smaller than this, so discharge transition once you reach some value of lower than VDD device will enter nonsaturation and it would discharge faster.

By same logic, initially it will take time to charge but then will charge fast using exponential functions. Again, looking at the expressions or TPHL is, one best method of reducing the TP which is average of the 2 is to reduce CL. You can also say if you increase VDD but you increase VDD the power is going to be larger and the whole effort in (()) (53:40) to reduce power and increase speed together and for a given node VDD also will be given by LTRS so it may be 2.1, 1.5, 1.2, 1 volt, 0.8 volt, 0.6 volt.

So this is not so much in my hand also. VT of course had also technology dependent, depending on the node I use, VDD, I have VT is fixed. So, if you look at very carefully, for a given load, CL cannot be say minimized. The only thing, I can improve the speed is this KP and KP, which is essentially Mu Cox W by L half of course but essentially Mu Cox, Cox is fixed for a technology, mobility of N channel or P channel is also fixed for a technology.

All that I can change to improve the speech is the size of the transistor W by L. Now, therefore, the question arises how do I optimize the delay and that is very important because at the end of the delay, the delay essentially means the time taken for input signal to go to the output, is the propagation delay. So I want to know, how do I optimize this. Now obviously I have different way of, two methods of doing that.

(Refer Slide Time: 54:57)



I can have different gates in the circuit, a chain of gates, so I must do properly what I say gate size selection which kind of gates I should use and what sizing I should do for them, that is how much current they should be able to drive, gate size essentially is proportional to currents required and corresponding the current will provide from the transistors, N channel or P channel for charging and discharge.

I must size the transistor. Please remember future decided size for your current requirements you must believe that since WN is also directly connected to capacitances, it looks sometimes very funny that you improve in RC time constant by scaling, you may reduce capacitance per proportion you may increase the resistance, so RC should remain constant, but that is not so. So, that means at given technology.

No better speed can be attained, so to beat that system I must now somehow find how to reduce the delay for a given logic to implement. Now in real circuit, these days, for example most of the circuit design have done through IPs, available IPs, which earlier we use to call standard cell design approach, in which you have a library that offers multiple dry strength of cells that is it can dry 1 milliamp current, 2 milliamp current, 10 milliamp current or half a milliamp current.

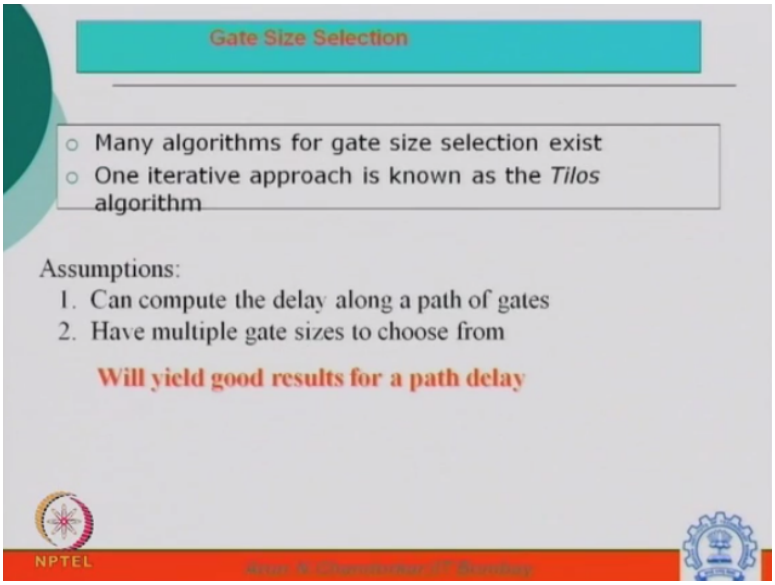
These are predesigned, prefabricated, pretested and their schematics are provided with this data. Of course, the do not tell you what is the technology in and what is the value but they will give what is the kind of schematic value, I mean at the end, what is performance values and current synthesis tool do a very great job for great size selection, so that is not a big issue.

The second part is the sizing of the transistor itself. Now, it is done in a custom design in which you size, individual transistors during the design process to optimize delay, now how does it depend on, the quality depends on individual designer, some synthesis may help or that is available to you synthesizing tools are very greatly available these days and you keep simulating with iteration, attempting option but can be time consuming.

For example, I can have hundreds of W by L ratios and values which I can keep substituting in a circuit and keep simulating, of course, as word go there is no 0 probability system, so some day

you are going to attain a speed or times you are looking for and you say you have achieved (()) (57:33) for that speed but doing so the window for marketing the product would already lost or maybe you are 2 years behind, so no one is going to buy your system anyway and because of that the speed had cracks, where to start simulation therefore is most crucial in starting the design.

(Refer Slide Time: 57:52)

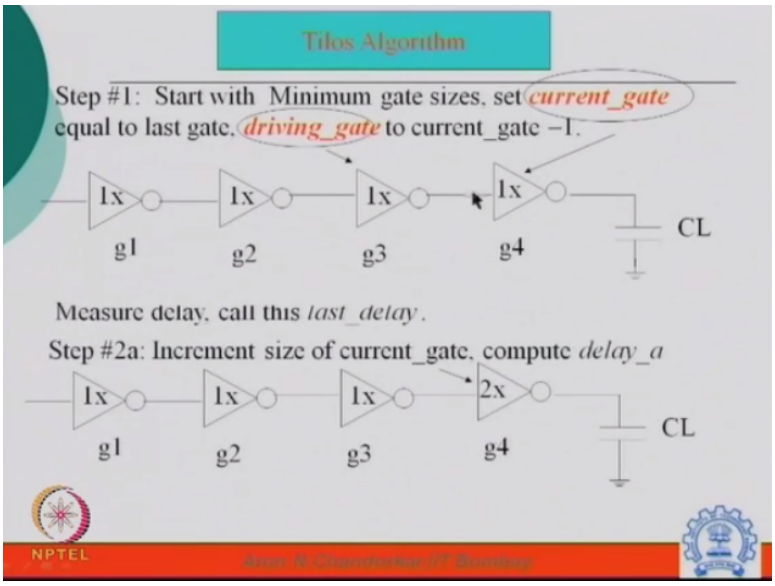


There are many algorithms for gate size selection already exist, so that is not a big issue, and one such approach is very famous, it is called Tilos algorithm. What I am going to talk about is Tilos algorithm is nut shell, can compute a delay along the path of gates, how multiple gate sizes to choose from. These are two assumptions, I can get any size of that is any driving capabilities, 1X, 2X, 4X, 1 milliamp, 4 milliamp, 2 milliamp.

I can choose the size of the gate which can allow me that currents and I can decide the connection so that the path delay can be minimized and this will choose that two of these so that you get the best delay. Here is a very simple algorithm, which I am going to show you which is essentially Tilos algorithm. Why I am showing all these precursors to my logical effort, because after all these have been already done and logical effort actually summarizes in a best way.

All those theorems, methods, algorithms, which we know into a nut shell which gives you a faster design. So, at the end, I am looking for a faster high speed design.

(Refer Slide Time: 59:07)



So here is the step 1, the kind of circuit you are shown here is you have a chain of inverters, each has a gate size such that, is called minimum gate size, which can minimum driving possibility, 1x, 1x, 1x, 1x. No, I give an input and I measure or monitor when the V0 receives the input and the delay I calculate and that delay I call last delay okay, that means all of them are 1x, this gate I saw in the final gate.

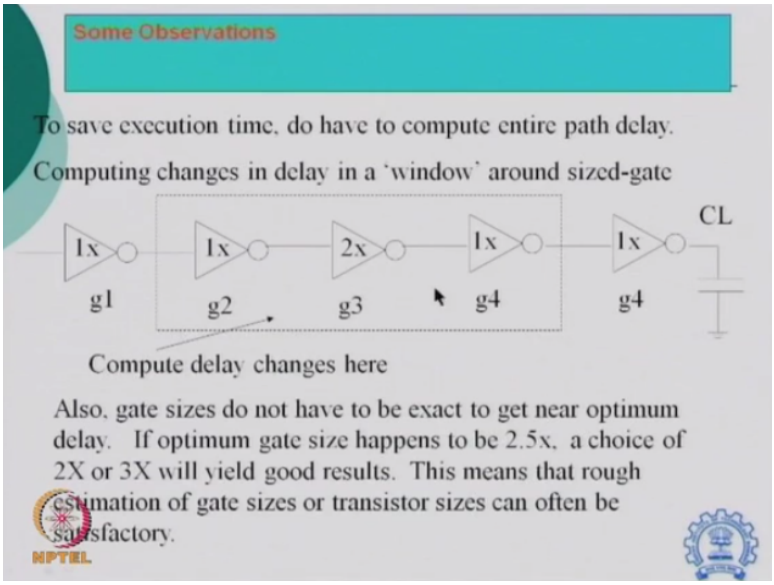
We say is the current gate and this earlier gate which is driving this gate is called driving gate. Now, this driving gate and this current gate and these are right now fixed, so input starts has come here and from here to here for this load of CL, I evaluate the time. The time from here to here is fixed for anything, any size here as well but time to here to here will change depending on what are gate sizes on this.

Then, I take another gate, I say okay, double the size of current gate which means it has the twice the current driving capability. This gate can now provide double the current compared to this and it is still driven by the same size old one, which is 1x. Obviously, if I have larger gate sizes, larger currents, if this has to prove a larger double the current the size of this transistors here or gates should have been sufficiently large enough at least double of that.

So that they can probably double the current. If the size on this increases, obviously the capacitance will increase, which is the output of this has now drive larger capacitance, so then it

will get delayed here, so we calculate the delay from gate G3 to G4 in the 2A step, with 1x as the driving gate and 2x as the driven gate or current gate.

(Refer Slide Time: 01:01:03)



Now, I repeat the same experiment, but now I say my driving gate has double to X capacity of driving but my current gate which is my final gate to drive the load is had the old value and please remember these are not being changed so delay from input coming from anywhere input until this time is same for either case, now you have 2x here and 1x here in the earlier case 2A case, 1x here and 2x here.

So I again monitor the delay coming from here to here in both cases. The first I case A and this one I called 2x, 1x I call case B, then I compare the delays between case A and case B and compare these new delays 1x, 2x and 2x, 1x with all 1x, 1x, 1x case which is called the last delay. Not let us say, the 1x, 1x case has certain delay, and making 1x, 2x I get a delay which is larger than the last delay or lower than last delay and by same argument, if I make 2x, 1x, this kind of statement.

I also have a gate delay which I will figure out whether it is more or less than last delay. So, whichever circuit, A or B, gives me lower delay than the last delay, I would say, I have improved, and this algorithm then tells okay instead of putting, let us say this is better B, then I say at least

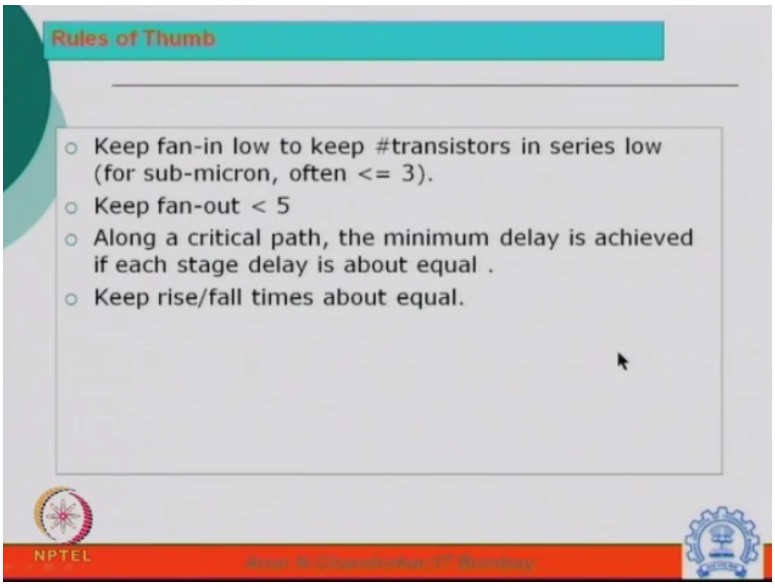
last 2 stages are affixed, you make 2x, 1x as calculations. Now call this, as one current gate, using this current as a gate, I actually now work with this as a driving gate.

So I say now this becomes current gate and this become driving gate and I repeat this performance once again with these two. Then, next time I figure out, now I have standardized this, this, this for the minimum delay, then I use this as my current gate and this as my driving gate and keep doing, but I do not tweak with the first gate.

Because this is essentially coming from the first, which is a fixed size input buffer, so which is fixed, so based on the first one, I now know the size of this, because of this and these 2, I know the size of this, because of this I know the size of this, so I now made all possible gate sizing for the minimal delay. To save executing time may times, you figure out this is 1x, 2x, this is called window around, see how 1 sized gate.

You can now very well know such circuits will give minimum delay here, so you fix this and never redesign it again.

(Refer Slide Time: 01:04:01)

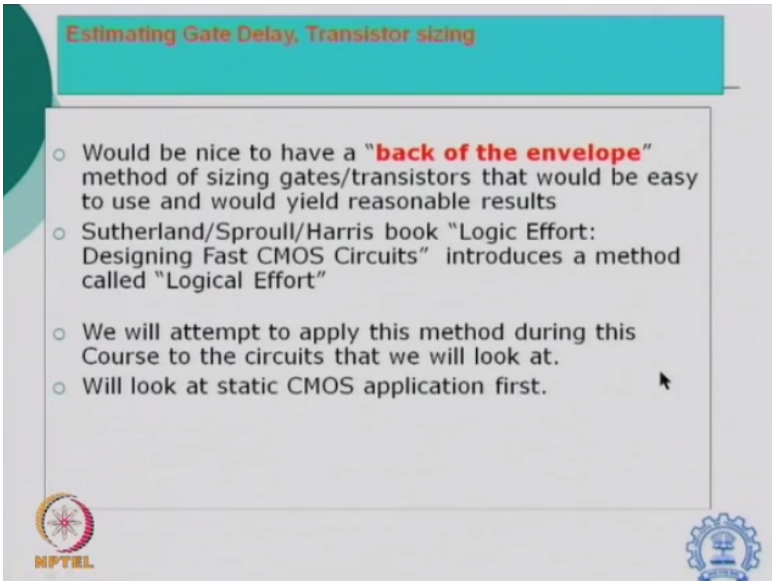


Now, one rule of thumb for delay essentially is larger output capacitance will be decided by how many fan out the output has, typically it is suggested that at no time the output should drive more than 5, preferably, the highest fan out normally can be used is 4 and therefore it is call a case of

FO4, fan out 4, so all circuits are actually for a given standard load which his called FO4, so the maximum fan out be restricted to 4 and no more, less than 5, I would say 4.

Then also since the input capacitance changes with the fan-ins, so also you must say number of fan-ins should be lower and even for more submicron kind of technologies, it should be less than 3. Then, along a critical path, the minimum delays achieved in each stage by delays about equal, keep rise fall time about equal.

(Refer Slide Time: 01:04:56)



Then, before I quit today for the day, I must say before I decided how to go for transistor sizing, sizing I did, but the next part I say is transistor sizing, I would like to do, it will be great or it will be nice if I do some back of envelope calculations. Now, if I can do transistor sizing and get sizing by very simple calculations which will give faster response to me. Then I would say I will be saving my effort.

And we say Sutherland, Sproull, Harris who actually first time talk about the new simple technique based on whatever we were asking in our earlier part, that was called as logic effort and they wrote such a book which is logic effort, design fast CMOS circuit which introduces which method of logical effort. We will attempt to apply this method during this course to circuits that we will look into and we look at static CMOS applications first.

Thank you for the day and we will come back tomorrow and continue with this actual working on logical effort, thanks for the day.