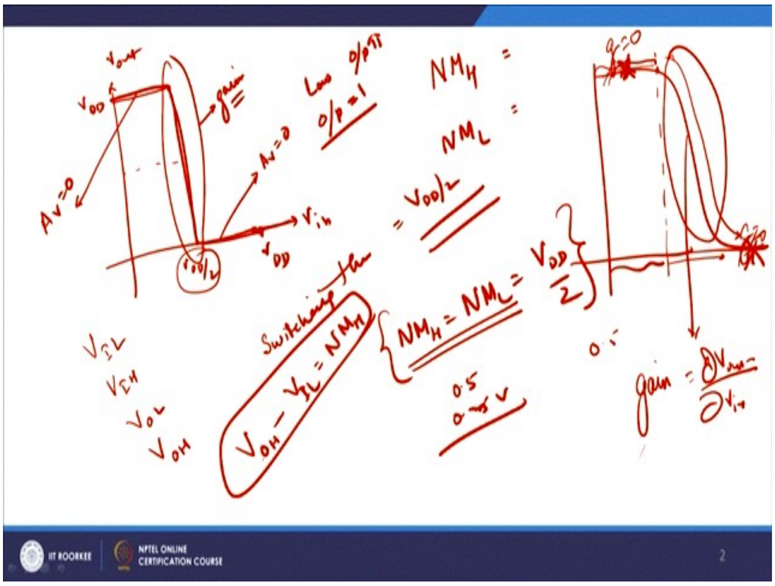
Microelectronics: Devices to Circuits
Professor Sudeb Dasgupta
Department of Electronics & Communication Engineering
Indian Institute of Technology Roorkee
Lecture 19 - CMOS Inverter Basics - II

Welcome to the NPTEL online course on Microelectronics: Devices to Circuits. We start today's lecture titled as CMOS inverter basics part III, in our previous module we have learned how an inverter works, what is the meaning of inverter which is basically a CMOS inverter and what you mean by V_{IL}, V_{IH}, N_{ML} and N_{MH}. So just to refresh your memory let me give you what the idea is.

(Refer Slide Time: 1:00)



So, generally we define two types of noise margin N_{MH} and N_{ML}, so N_{MH} is defined as high noise margin and it gives me an impression that high noise margin primary means that when your input is low right, and your output is high, then at that point of time that means if your output is equals to 1, then how much amount of input right, in the input side how much amount of noise can be given even without changing the output from 0 to 1 right, from 1 to 0.

So let us suppose I have a VTC in this manner right, we have already discussed this point in our previous slide in this manner, then if you look very carefully then typically this much amount of input voltage, even if I give my output voltage will still, so if this is the input voltage I give my output will still be high. But yes, if I across this value and go to this side, the output will fall drastically to this value and output will be equals to 0. So high noise margin is defined as that value or that voltage in the input side, maximum noise voltage which can be given, so that my output does not change from 1 to 0, right and therefore higher the value of N_{MH} or high noise margin better the design is.

The ideal value of your design is something like this, it is like this, it is the ideal value, which you see. This is V_{DD} right, this is your V_{DD} by 2 and this is also V_{DD}, so this is your V out versus V in and this is the profile which you get for all practical purposes, you get the profile something like this, which means that I will expect to see a switching at somewhere around V_{DD} by 2. So we define the switching threshold as or switching threshold as V_{DD} by 2 for ideal case right.

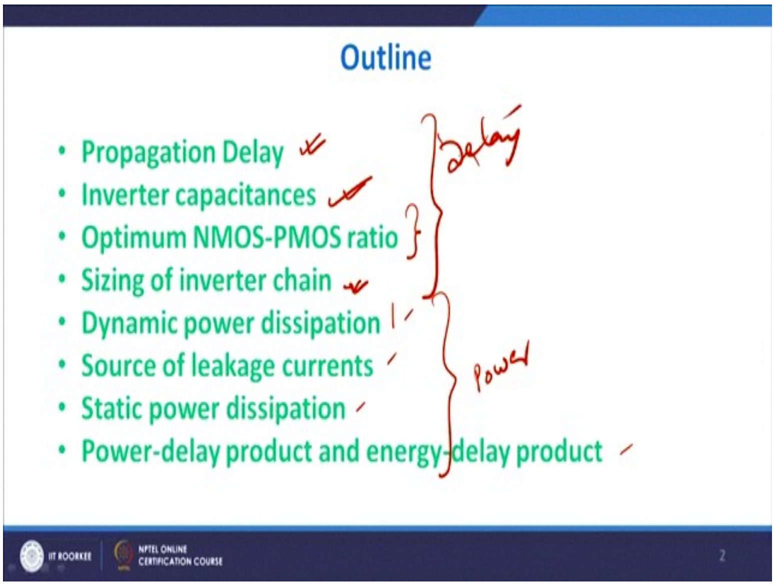
So if you, for ideal case and therefore if you look very closely in for ideal case N_{MH} equals to N_{ML} equals to V_{DD} by 2, which means that the high noise margin and low noise margin are both equals to V_{DD} by 2. Now if you want to lower your V_{DD} for whatever reasons, your noise margin will also be lowering itself automatically and therefore its property of rejection of noise will be compromised, right?

For example if your V_{DD} is 1 volt then N_{ML}, N_{MH} equals to 0.5, if it is now 0.5 then this will be 0.25, 0.25 volt, it primarily means that if a noise comes whose equivalent value of voltage is 0.25, I would expect to see output going from 1 to 0, right, so that is what basically the idea is? We also saw in the previous discussion that for digital applications, we generally put it our we bias your device somewhere here or here right. Whereas for analog applications we need to bias it here and at this place only I will get a gain which is given as $\partial$of V out, ∂ of Vin, why? Because at this stage if you look very carefully at this stage my ∂ V out is equal to 0 here, ∂ V out is also equals to 0 here.

So gain will be 0 here, gain will be 0 here right, so whenever you have this bias right, this and this the voltage gain is always equal to 0 and therefore not used for analog applications, whereas this point is the point where you get a fast change in the output for a small change in the input and this is the point where you define your gain to be there. And if you want to bias, it is an analog device, you need to bias it somewhere in this, this region right. The problem with this region is that it is so unstable that for a small change in the input I will see a large change in the output and that makes it slightly unstable in design, from design or design aspect point of view.

Now, so this is what we have learned, we have also learned how to calculate V_{IL} input low, V_{IH} input high, similarly V_{OL} output low and V_{OH} output high right and similarly we also learned that if you subtract the V_{OH} from $V_{IL,}$ I automatically get N_{MH} and so on and so forth right, we have already learned all these things.

(Refer Slide Time: 5:30)



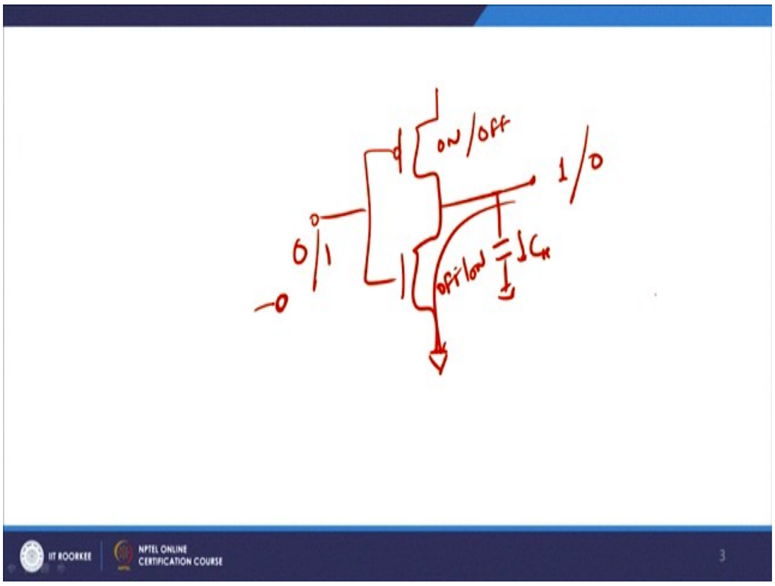
Outline

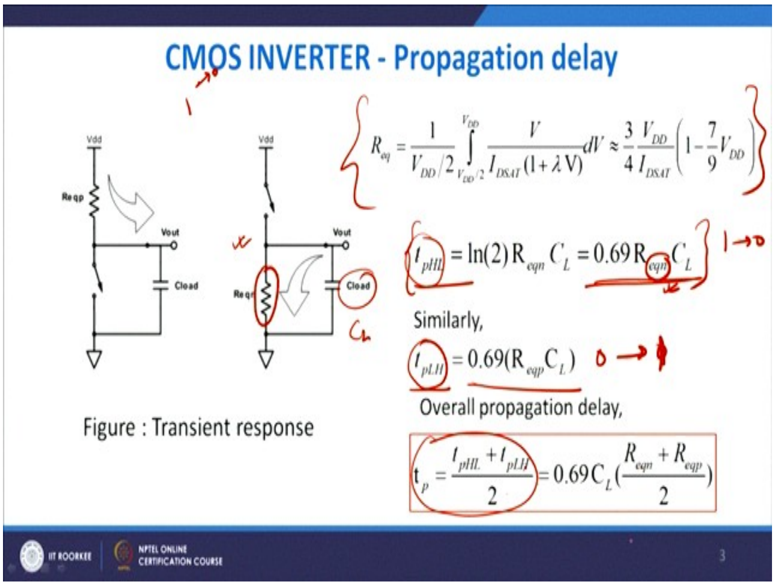
- Propagation Delay
- Inverter capacitances
- Optimum NMOS-PMOS ratio
- Sizing of inverter chain
- Dynamic power dissipation
- Source of leakage currents
- Static power dissipation
- Power-delay product and energy-delay product

So let me explain to you what are the various things which we will doing today. We will be actually looking into what is known as propagation delay, we will look into inverter capacitances, we will have a look at what is the optimised NMOS by PMOS ratio so that I get the minimum delay position available to me. We will look into the sizing of inverter chain and then start with the power dissipation, dynamic power dissipation, leakage power dissipation, static power dissipation and power delay product.

So the 1st half of our lecture will be concentrating on the delay and the 2nd half will be concentrating on power, so this will be power, where this will be delay. So given a inverter, given an inverter can I optimize its power and delay? So that will be the major motivation behind this module and that will be the major motivation for this module or this lecture series.

(Refer Slide Time: 6:40)





CMOS INVERTER - Propagation delay

$$R_{eq} = \frac{1}{V_{DD}/2}\int_{V_{DD}/2}^{V_{DD}} \frac{V}{I_{DSAT}(1+\lambda V)}dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}V_{DD}\right)$$

$$t_{pHL} = \ln(2)R_{eqn}C_L = 0.69R_{eqn}C_L \quad 1\rightarrow 0$$

Similarly,

$$t_{pLH} = 0.69(R_{eqp}C_L) \quad 0\rightarrow 1$$

Overall propagation delay,

$$\boxed{t_p = \frac{t_{pHL}+t_{pLH}}{2} = 0.69C_L\left(\frac{R_{eqn}+R_{eqp}}{2}\right)}$$

Figure : Transient response

Let me come to how, what is inverter delay. Now as I discussed with you in earlier slide or the earlier presentation that when my PMOS is switched on and NMOS is on right, my output actually falls to 0 right. So we have already discussed this in our previous discussion, in our previous slides that suppose I have got PMOS right, I have got a PMOS here and NMOS here right and this is my PMOS and this is my gate. So I give a, let us suppose I initially add 0 here, so when it is 0, this is on, this is off and this capacitance charges to 1 and therefore output goes to 1, right, this C_{HS} charging.

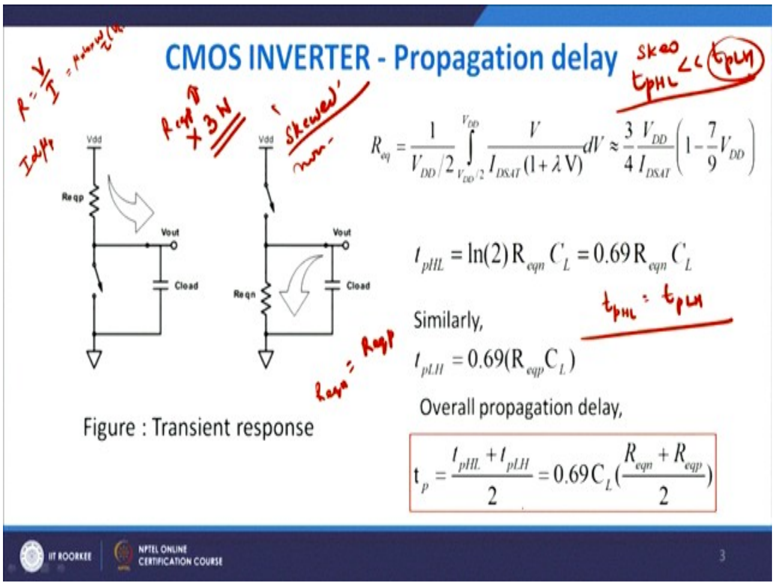
In the 2^{nd} half cycling 0 to 1, this goes to off, right, this goes to on and therefore this charge finds the discharging path to ground and therefore this 1 goes to 0, this was there. We defined now therefore two delays, one is defined as the propagation delay low to high, high to low and we also defined low to high right. Now what is high to low? When the output goes from 90 percent of high to 90 percent of low, we define that to as a propagation delay of high to low, right.

So t_{PHL} is basically defined as t_P is propagation delay and HL primarily meaning high to low. High to low means when the output goes from 1 to 0, so that is what we are defining. We are defining, we are not deriving it here, but primarily this is what the equation is 0.69 R equivalent into C_L, now what is R equivalent? For example in this case, this is the high to low transition, R equivalent is nothing but the resistance offered by the PMOS, NMOS inverter in its on state, in its on state ideally it should be 0.

But in reality it is never 0 because there will be some transit resistance available to you, as a result there will be always, t_{PHL} will be always have some finite value equals to 0.69 R equivalent into C_L, where C_L is basically the load capacitances. So this C load is basically your load capacitance, right and R equivalent is given by this formula which you might have a look into it, but not very critical at this stage of time, more important is this one.

Similarly, when you are discussing output low to high, then we get t_{PLH} to be equal to low to high, again, the formula is 0.69 R equivalent P into C_L. So you see it is R equivalent N here because output is going from 1 to 0 right, 1 to 0 and it is equivalent P, so output is going from 0 to 1 right, t_P high to low. Overall propagation delay in the formula is something like this, that your t_{PHL} + t_{PLH} by 2, so we average of these two we take out and we therefore, so this is common, C_L is common you take it outside and then R equivalent P + R equivalent N by 2.

(Refer Slide Time: 9:14)



CMOS INVERTER - Propagation delay

$$R_{eq} = \frac{1}{V_{DD}/2}\int_{V_{DD}/2}^{V_{DD}} \frac{V}{I_{DSAT}(1+\lambda V)}dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}V_{DD}\right)$$

$$t_{pHL} = \ln(2)R_{eqn}C_L = 0.69R_{eqn}C_L$$

Similarly,

$$t_{pLH} = 0.69(R_{eqp}C_L)$$

Overall propagation delay,

$$\boxed{t_p = \frac{t_{pHL}+t_{pLH}}{2} = 0.69C_L\left(\frac{R_{eqn}+R_{eqp}}{2}\right)}$$

Figure : Transient response

Now the idea is that maybe we can discuss it here itself, or maybe we will see later on, but if you look at the idea here and let us see what the idea is. That, sorry, see if you want that t_{PHL}, if you do not want, if you want this to be true, but t_{PHL} equals to $t_{PLH,}$ which means that high to low propagation delay is exactly equals to low to high propagation delay. Then what should you do? Then you should actually make your R equivalent N exactly equals to R equivalent to P because C_L in any case is equal and that will make you equal propagation delay.

But the problem is that holes, mobility of charge carriers, so what is? It is given as voltage by, so resistance will be voltage by current, applied voltage by current right. Now current in a fat device if you remember by our previous discussion is equals to μn C oxide, W by L $(V_{GS} - V_{TH})^2$ something, which means that the current is directly proportional to the mobility of the charge carriers. In case of therefore PMOS with the majority current carriers are holes, holes have got a much lower mobility as compared to electrons and therefore if you do not do any manipulation in the device structure or the circuit, the current by this............................

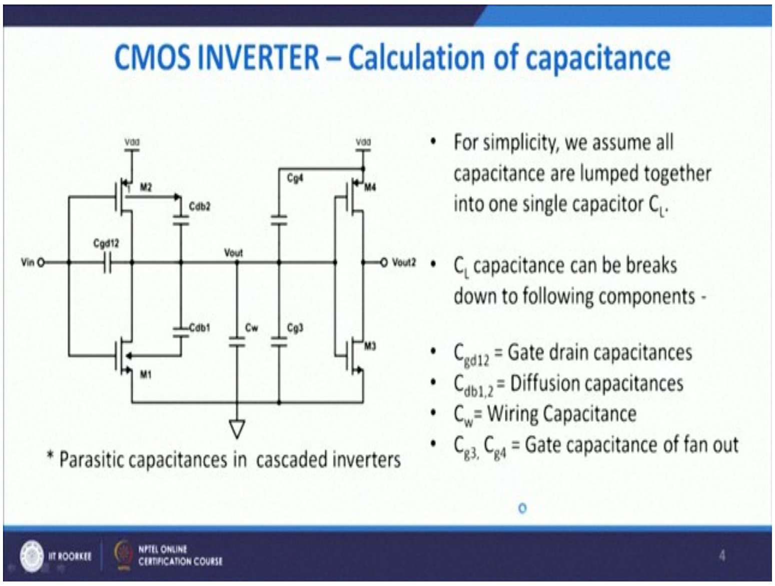
So if the aspect ratio of the device is same NMOS and PMOS is same, W by L ratio is same, voltage also same, then since the mobility of the charge carriers of hole is approximately 2 to 3 times smaller as compared to that of electrons, my current will be also 2 to 3 times smaller and therefore my resistance will be 2 to 3 times larger because it is inversely proportional to the current, right. So this is the problem area which you will face that if you do not do any manipulation, the value of the resistance will still remain the same right, so this is the problem area and the resistance will be different and therefore t_{PHL} will not be equals to t_{PLH}.

If you want to make them equal then the process which you should follow is something like this: that try to make the, since your resistance is large, R equivalent P is large you need to reduce it by how many times? By approximately 3 times, how can you do that? If you make the width of my PMOS 3 times larger right, then the area actually becomes large and the resistance falls to one-third of its value, to the same value as PMOS. So therefore if you want the t_{PHL} it must be equals to t_{PLH}, you have to simply make the PMOS width approximately 3 times larger as compared to NMOS and you automatically get the same values of t_{PHL} equals to t_{PLH}, right.

So that is what I want you to say that, if your propagation delay is 1, then you to make this thing, this is also known as a skewed transistor right, skewed transistor. When you have skewed t_{PHL} equals to t_{PLH}, non-skewed if you have got then t_{PHL} will be smaller as compared to t_{PLH}. So if you do not have any skew, then t_{PHL} high to low will be smaller as compared to

t_{PLH} because this has got a higher resistance and therefore higher current right, so this is just for information sake, a quite important one.

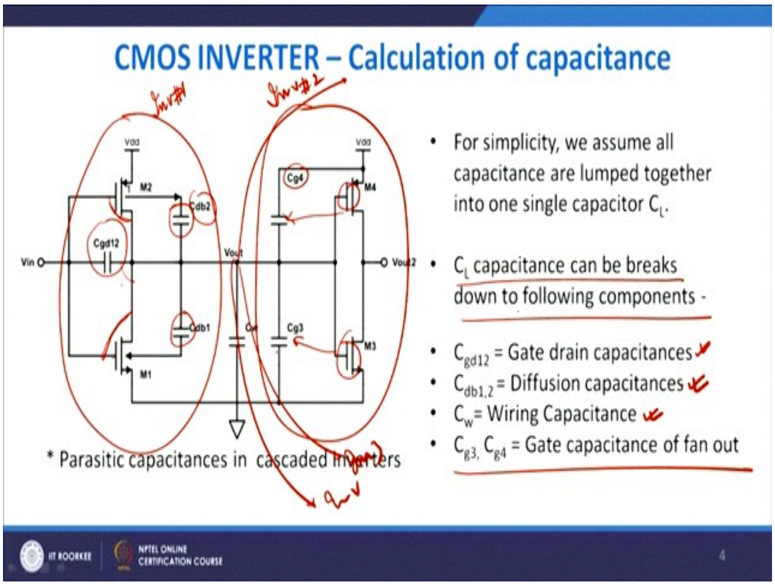
(Refer Slide Time: 12:34)



CMOS INVERTER – Calculation of capacitance

- For simplicity, we assume all capacitance are lumped together into one single capacitor C_L.
- C_L capacitance can be breaks down to following components -
- C_{gd12} = Gate drain capacitances
- $C_{db1,2}$ = Diffusion capacitances
- C_W = Wiring Capacitance
- C_{g3}, C_{g4} = Gate capacitance of fan out

* Parasitic capacitances in cascaded inverters



CMOS INVERTER – Calculation of capacitance

- For simplicity, we assume all capacitance are lumped together into one single capacitor C_L.
- C_L capacitance can be breaks down to following components -
- C_{gd12} = Gate drain capacitances ✓
- $C_{db1,2}$ = Diffusion capacitances ✓
- C_W = Wiring Capacitance ✓
- C_{g3}, C_{g4} = Gate capacitance of fan out

* Parasitic capacitances in cascaded inverters

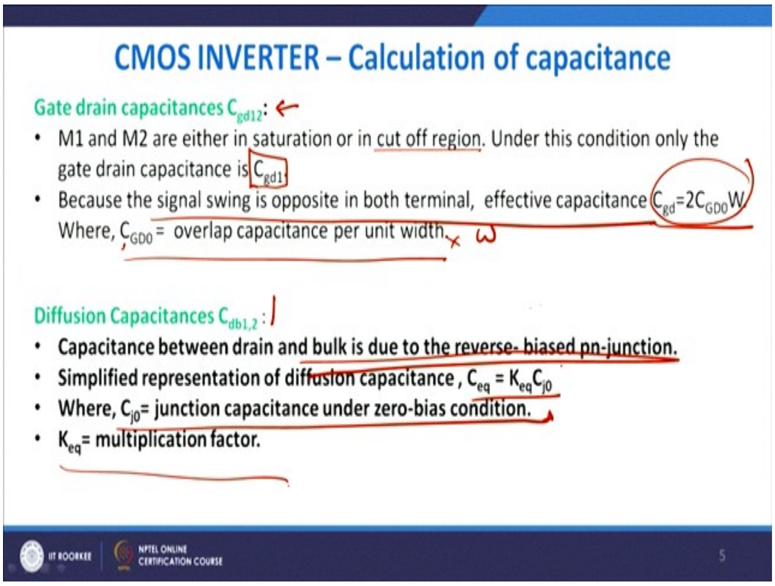
Now with this we are going into another transient analysis, we have done the DC analysis of CMOS inverter, we also wish to have a look into the transient analysis and to understand the transient analysis, we actually have to look into the various capacitive models available within the CMOS inverter. Now, what we, but people generally have done over the years is that all the inverters, all the inverters they have taken all the capacitance together, lumped it together and thrown in the output side and then they have assumed that the inverter itself is capacitance free.

So the inverter capacitance is free, so the inverter is free of any capacitance and all the capacitance will be lumped together and it has been thrown in the output side as C_L or whatever. Now, but if you break it down, then you say this is one inverter, this is one inverter and driving another inverter here, so let us suppose this is inverter 1, this is inverter 2 and it is driving it. So you see this is drain to bulk capacitance, you have gate to drain, you have actually drain to bulk here and then drain to bulk here, drain to bulk for NMOS and drain to bulk for PMOS, this is gate to drain, this capacitance gate to drain right and this is also gate to drain. So these two gate to drain, sorry sorry……, these two gate to drain together form $C_{GD\,12}$.

Similarly, you will have a gate capacitance here, so only gate capacitances right and this is gate capacitance is falling down, so this gate capacitance is appearing at this particular point, this is gate capacitance is appearing at this point, right. So when you lump this, this, this, this, this and this, together, we get the load capacitance, so that is what I was saying that C_L, the load capacitance can be broken down into following components, $C_{GD\,1\,2}$, which is this one, gate to drain capacitance, C_{DB} drain to bulk of 1 2, this is your diffusion capacitance, sorry depletion capacitances and this is your diffusion capacitances. C_W is basically the wiring capacitance, so any wire just like your R_{NC} element can be broken down, there will be a wiring capacitance there.

You will have C_G 3 and 4 the gate capacitance of fan out, so this is converted to inverter 2, this is inverter 3 here, inverter 4 here, so this is inverter 3 and inverter 4, then inverter 3 and 4 will have C_G 5, 6, C_G 7, 8 so on and so forth, all will be added together to form the value of C_L. So C_L consists of the previous stage diffusion and depletion capacitance and the next stage gate capacitance plus the wiring capacitance, that takes care of approximately all the values of the inverter capacitances.

(Refer Slide Time: 15:26)



CMOS INVERTER – Calculation of capacitance

Gate drain capacitances C_{gd12}:

- M1 and M2 are either in saturation or in cut off region. Under this condition only the gate drain capacitance is $\boxed{C_{gd1}}$
- Because the signal swing is opposite in both terminal, effective capacitance $C_{gd}=2C_{GD0}W$
  Where, C_{GD0} = overlap capacitance per unit width. × ω

Diffusion Capacitances $C_{db1,2}$:

- Capacitance between drain and bulk is due to the reverse- biased pn-junction.
- Simplified representation of diffusion capacitance , $C_{eq} = K_{eq}C_{j0}$
- Where, C_{j0}= junction capacitance under zero-bias condition.
- K_{eq}= multiplication factor.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 5

Now how do we calculate therefore this C_{GD}, which is gate to drain capacitance, right? How do we calculate C_{GD}, which is gate to drain capacitance here? Now when M1 and M2 are either in saturation or in cut-off, now then you only have the C_{GD} 1, gate to drain 1, right, why is it true? Because if it is cut-off, then obviously there is no channel formation taking place, when the saturation, the channel is formed, but it screens, it screens of the depletion region from the gate region, right?

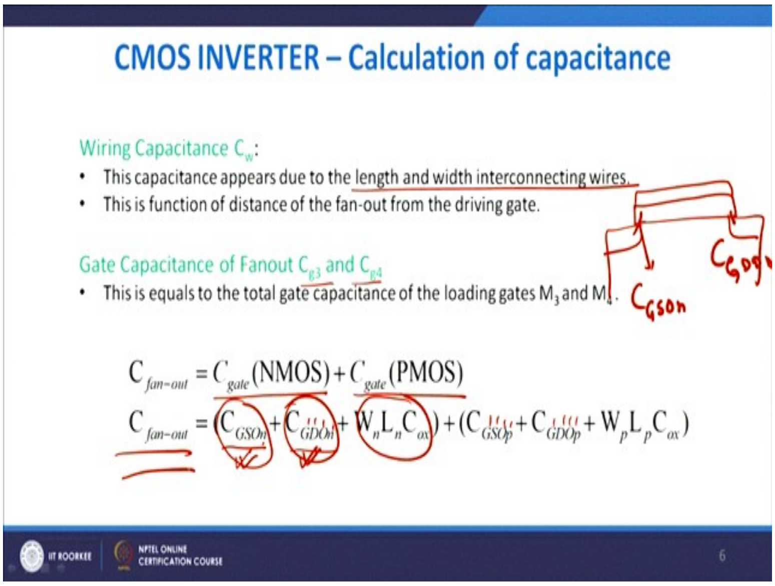
Only, that is the reason I am saying only gate to drain capacitance is C_{GD} 1 right, so C_{GD} 1 is only available to you gate to drain, whenever you want to cut-off on the saturation region. Now, since the signal swing is opposite in both the terminals, C_{GD} is equal to 2 C_{GD} 0 into W, where C_{GD} 0 is overlap capacitance per unit width, so you see why we are multiplying it by 2 W or 2 into W? Because since C_{GD} 0, this C_{GD} 0 is per unit width to multiply with width, you get the total capacitance and since you have 2 devices connected to the same input, we multiply that by 2.

So overall C_{GD} happens to be equal to 2 times C_{GD} 0, right, and the signal swing is opposite, so when this is positive for PMOS, it is negative for NMOS and vice versa, right, and so both will come out, effective capacitances will be just double of that because they will be in parallel to each other.

Now there are 2, so this was basically your drift, your gate to drain capacitance, we also have drain to bulk right, 1 and 2 and this is primarily because as we discussed in our previous term p-n junction reverse bias right because of the p-n junction reverse bias we will always have a capacitance between drain and bulk and it is given as C equivalent equals to K equivalent C_J0, where C_J0 is basically the junction capacitance in zero-bias condition and K is the multiplication factor, K multiplication factor depends upon how much amount of bias you have given in the input side.

So the first one is gate to drain, the second one is diffusion one, the gate to drain is a depletion one, this is a diffusion one right. Diffusion primarily occurs because when p-n junction is reverse biased, the depletion region will be formed, which depletion region will be depending on the type of bias you have given and level of the minority and majority current carriers available with you.

(Refer Slide Time: 17:41)



CMOS INVERTER – Calculation of capacitance

Wiring Capacitance C_w:

- This capacitance appears due to the length and width interconnecting wires.
- This is function of distance of the fan-out from the driving gate.

Gate Capacitance of Fanout C_{g3} and C_{g4}

- This is equals to the total gate capacitance of the loading gates M_3 and M_4.

$$C_{fan-out} = C_{gate}(NMOS) + C_{gate}(PMOS)$$

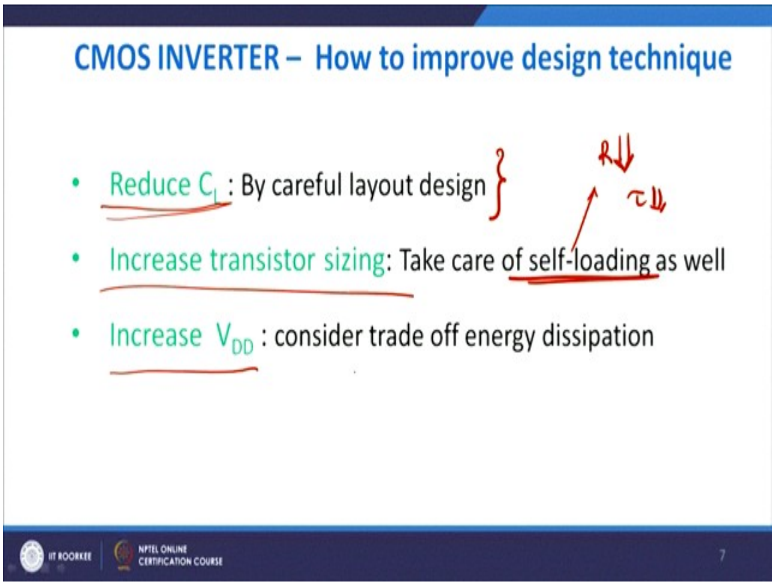
$$C_{fan-out} = (C_{GSOn} + C_{GDOn} + W_n L_n C_{ox}) + (C_{GSOp} + C_{GDOp} + W_p L_p C_{ox})$$

Now to calculate wiring capacitance as I told to you is due to length and width of interconnecting wires and also it depends upon how many fan outs are there. So your fan out is 4, then there will be 4 wires typically which will be emanating from the driver device onto the driven device and that will add up to the wire capacitance there, so larger the length more will be the capacitor.

The gate capacitance C_G 3 and C_G 4 comes out from the gate capacitance of the subsequent stage and therefore fan out is equal to C gate into NMOS into C gate into PMOS, I think, very simple and straightforward way of looking at it, C gate can be again broken down into two parts C_{GS} ON and C_{GD}. So you see the channel can be broken into two parts, one contribution between gate and source, so that is this one in the on state and the overlap capacitances, and then gate to drain overlap capacitance, so this is O means overlap, gate to source overlap capacitance for N channel and gate to drain for overlap for N channel, this is gate to source overlap P channel, gate to drain overlap P channel, right?

So this is basically the overlap channel, this is also the overlap channel which you see and this one is primarily the gate oxide, so when you have plotted, when you have actually drawn this, so let me say you have drawn something like this and you have this thing, then it is something like this that, this overlap here drain side and source side, this is basically your, so this is actually your C_{GS} ON and this is your C_{GD} n right, D n, D0 N, so that plus W n into L n means, width of NMOS into length of NMOS. Since C oxide is the oxide per unit area, multiply with that you get the total oxide, you add those 2 oxides and you get the total fan out oxide which is available with you, and it is quite a large sum which you see.

(Refer Slide Time: 19:36)



CMOS INVERTER – How to improve design technique

- Reduce C_L : By careful layout design
- Increase transistor sizing: Take care of self-loading as well
- Increase V_{DD} : consider trade off energy dissipation

R⇓
τ⇓
IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
7

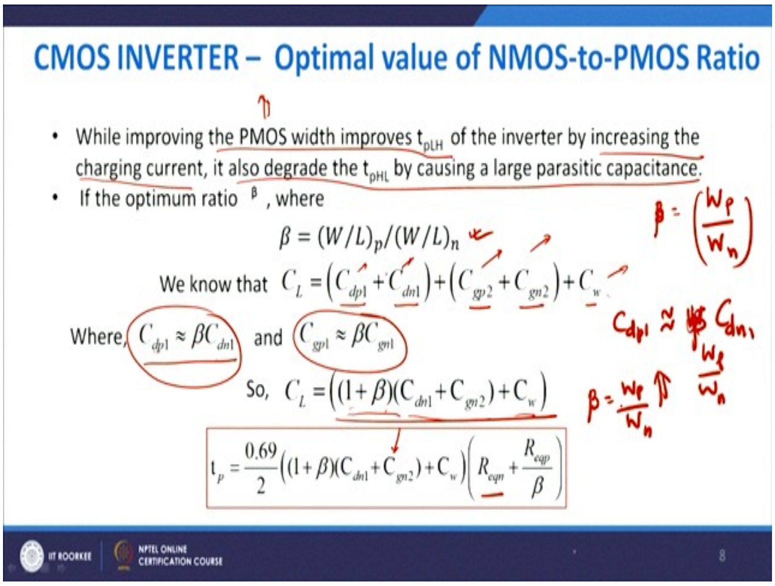
Now, so how to design a very good inverter? If you want to improve the speed, of course the best way to do that is reduce C_L and this can be done by layouts, so either by layout or by critical sizing of the transistors and you can reduce the value of $_{CL}$. Similarly, the idea was is, if you remember t_{PHL} and t_{PLH} was depending on the value of R equivalent N, which means the resistance offered by the transistor. In the 1st case when you are reducing C_L right, this can be the, so if you reduce C_L obviously your τ reduces.

Similarly, if you increase a transistors size, well, that will make your R actually go low right, R go low, but when R goes low obviously your Tau reduces and you can work at much faster pace but then be very careful about what is known as self-loading. Right, what is the meaning of less self-loading is? That you are increasing your W right, fine, why you are increasing your W?

So that the area under the gate goes on increasing and you automatically have a smaller resistance available to you, but then what happens is as we go on increasing the value of R, W you also tend to increase the capacitance of the gate. So gate has to drive in a much better manner in order to invert the channel and therefore your actually the first point is neglected and you will not able to reduce the value of C_L, your C_L value starts to rise again because the W has increased drastically, this is known as self-loading right.

Why should you increase V_{DD}? Very important that if you increase V_{DD} for the same amount of charges the current will be large, you got the point. So if the current is large, I_D is large, then your charging and discharging process can be done much faster right because for the same amount of time more charge will be collected or discharged if I_D is typically very large and τ therefore reduces. So therefore if you want to increase the speed 3 things, reduce capacitive loading, increase the transistor sizing by increasing the W ratio, W width and then increase V_{DD}. But if you increase V_{DD} you will have also higher power dissipation in a much larger manner and so power dissipation will be much larger in that case, right?

(Refer Slide Time: 21:54)



CMOS INVERTER – Optimal value of NMOS-to-PMOS Ratio

- While improving the PMOS width improves t_{pLH} of the inverter by increasing the charging current, it also degrade the t_{pHL} by causing a large parasitic capacitance.
- If the optimum ratio β, where

$$\beta = (W/L)_p/(W/L)_n$$

We know that $C_L = (C_{dp1} + C_{dn1}) + (C_{gp2} + C_{gn2}) + C_w$

Where, $C_{dp1} \approx \beta C_{dn1}$ and $C_{gp1} \approx \beta C_{gn1}$

So, $C_L = ((1+\beta)(C_{dn1} + C_{gn2}) + C_w)$

$$\boxed{t_p = \frac{0.69}{2}((1+\beta)(C_{dn1} + C_{gn2}) + C_w)\left(R_{eqn} + \frac{R_{eqp}}{\beta}\right)}$$

$\beta = \left(\frac{W_p}{W_n}\right)$

$C_{dp1} \approx \beta C_{dn1}$

$\frac{W_p}{W_n}$

$\beta = \frac{W_p}{W_n} \Uparrow$

Ok, let me see the optimal value for NMOS to PMOS ratio and now as I discussed with you, while improving the PMOS width right, improving the PMOS width improves t_{PLH} low to high, if you increase PMOS width t_{PLH} will go on reducing by increasing the charging current as I discussed with you, but it also degrades t_{PHL} by causing a large parasitic capacitances, why? Because as you go on increasing the PMOS width right, the PMOS itself might be doing very good, but then it will start loading your external load capacitance.

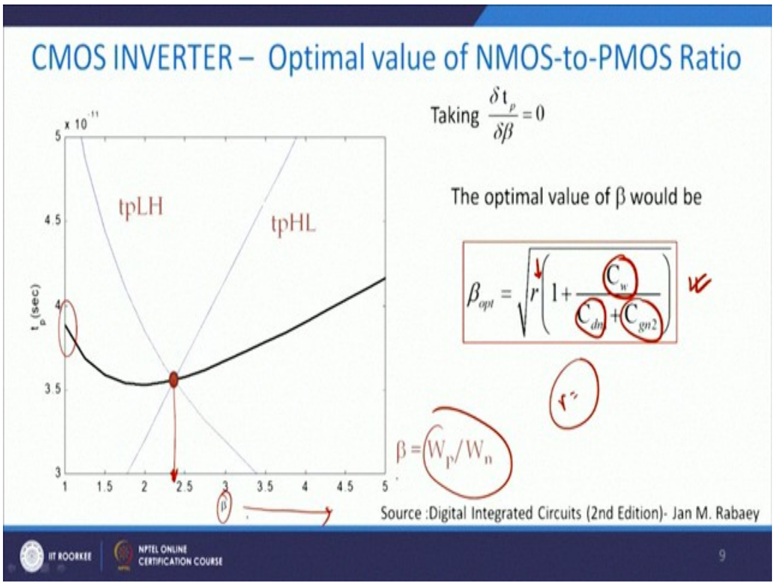
So the load capacitance will be, parasitic capacitance will be larger and larger. So at one hand, you are actually reducing the value of R equivalent P but in the same instance of time your capacitance is going on increasing, so we do not know exactly whether it will be always win-win situation, so you need to optimize it. Now if the optimising ratio β is given W by L of P upon W by L of N, we know that CL is given by dpl 1, dn2 plus gp 2 gn 2 plus wire, so this is your gate of the second fan out transistor and this is your dp, dn of the first transistor and this is wire transistor, where C dp 1 is β times C dn 1.

Why? Because the ratio, if you look at the β ratio, β ratio is given as, so if L is constant, it is nothing but W_P by Wn right, so if you want to find out C dp, C dp1 will be approximately equals to W n times right or maybe β times C dn 1 right. So this is nothing but we develop W_P by Wn right and therefore you can automatically say that this holds good, which means that the output capacitance will be approximately equals to β times C dn 1 and C gp will be approximately equal to C gn1 right.

So this is drain, so this is C dp 1 means this is basically your depletion capacitance of PMOS 1 will be β times depletion capacitance of 1 because width has increased by β times. Similarly gate capacitance of the 2nd case will be approximately β times because you see capacitance is directly proportional to the width and therefore you directly multiply, so if you look at C_L and if you just put this into this formula, I get 1 + β times C dn 1 C $_{gn}$ + C $_{gn}$ 2.

Now if you t_P if you find out 0.69 by 2, 1 + β because remember, it was C_L, so this is C_L, this is C_L, this C_L comes here and into R equivalent plus R equivalent P by β, you understand why it is by β, because you actually multiplied this β ratio by this thing, which means that beta is equals to W_P by Wn, so higher the value of β when you add it, you will divide it by β right, and therefore you get t_p equals to R equivalent n plus R equivalent Q by β.

(Refer Slide Time: 24:46)



CMOS INVERTER – Optimal value of NMOS-to-PMOS Ratio

Taking $\frac{\delta t_p}{\delta \beta} = 0$

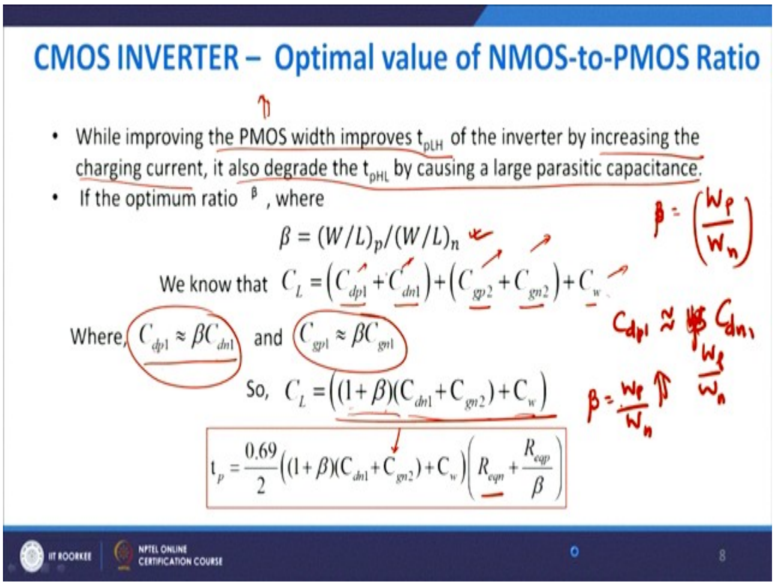
The optimal value of β would be

$$\boxed{\beta_{opt} = \sqrt{r\left(1 + \frac{C_w}{C_{dn1} + C_{gn2}}\right)}}$$

$r =$

$\beta = (W_p/W_n)$

Source :Digital Integrated Circuits (2nd Edition)- Jan M. Rabaey



CMOS INVERTER – Optimal value of NMOS-to-PMOS Ratio

- While improving the PMOS width improves t_{pLH} of the inverter by increasing the charging current, it also degrade the t_{pHL} by causing a large parasitic capacitance.
- If the optimum ratio β , where

$$\beta = (W/L)_p/(W/L)_n$$

We know that $C_L = (C_{dp1} + C_{dn1}) + (C_{gp2} + C_{gn2}) + C_w$

Where, $C_{dp1} \approx \beta C_{dn1}$ and $C_{gp1} \approx \beta C_{gn1}$

So, $C_L = ((1+\beta)(C_{dn1} + C_{gn2}) + C_w)$

$$\boxed{t_p = \frac{0.69}{2}((1+\beta)(C_{dn1} + C_{gn2}) + C_w)\left(R_{eqn} + \frac{R_{eqp}}{\beta}\right)}$$

$\beta = \left(\frac{W_p}{W_n}\right)$

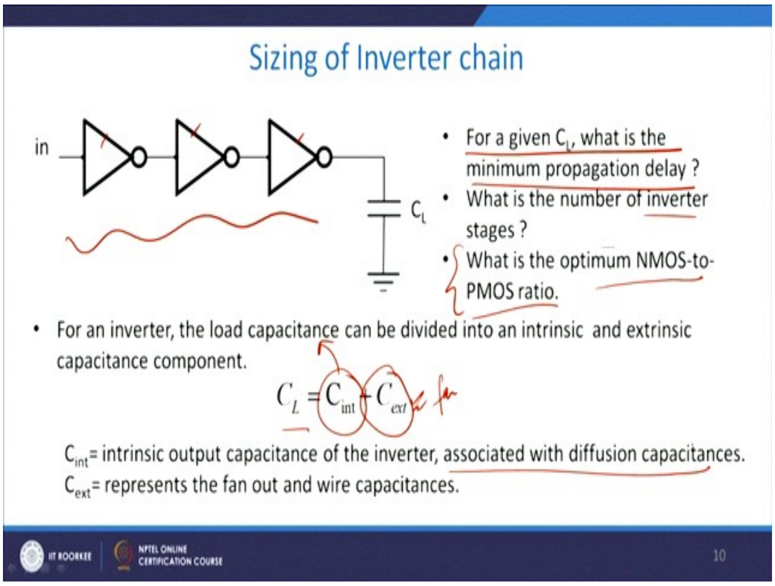
$C_{dp1} \approx \frac{W_p}{W_n} C_{dn1}$

$\beta = \frac{W_p}{W_n} \Uparrow$

Now if you differentiate ∂τP with respect to β, I get equals to 0, the optimal value of β comes out to be this value, where β equals to W_P by Wn and R if you look is basically the ratio which you see, R is basically the ratio which you will get this ratio, R is basically your, R equivalent P by R equivalent n is the ratio of R right. So for a fixed value of R, right, wiring capacitance it depends upon the diffusion capacitance and the gate capacitance of the second one and if you plot the propagation delay with respect to β, we see that somewhere near 2.5 to 2.5 of β I will expect to see t_P equals to 0 and this is also true from my previous understanding that if I am able to sustain the width of my PMOS 3 times more as compared to NMOS I automatically get a very reduced profile.

Therefore why, think about it. Why with therefore larger beta more than 2.5, I actually see a reduction in the value of t_P, for a low value of β is very simple, for low value of β since my width my PMOS is smaller as compared to NMOS, relatively smaller, it is not 3 times, it is less than 3 times, therefore the mobility is so large, so small that the current is very small and therefore resistance is high and that is the reason you get a larger t_p. And you minimize somewhere around 2.5, beyond 2.5 your actually t_{PHL}, your t_{PHL} starts to go grow, high to low starts, goes high, and as a result the overall gain starts or overall delay starts to become larger. So if you are biasing your device, please bias it in a manner such that it is approximately equals to 2.5 to 3 times, right?

(Refer Slide Time: 26:35)



Sizing of Inverter chain

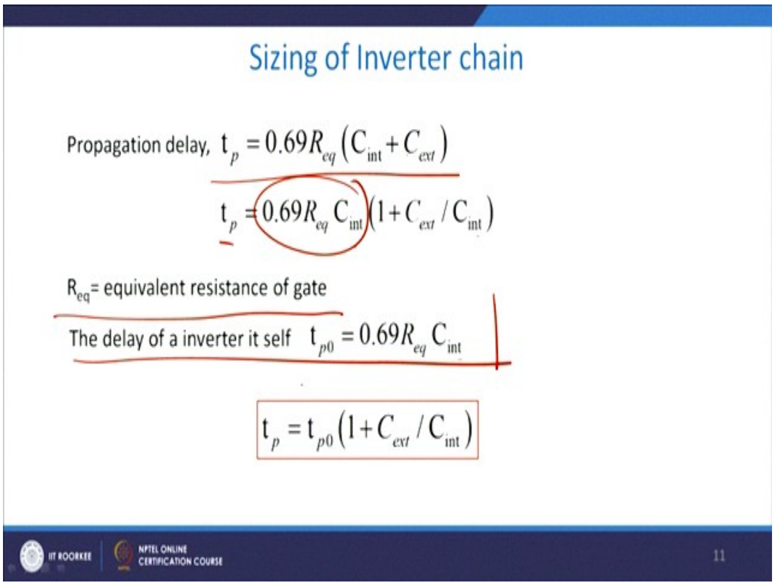
- For a given C_L, what is the minimum propagation delay ?
- What is the number of inverter stages ?
- What is the optimum NMOS-to-PMOS ratio.
- For an inverter, the load capacitance can be divided into an intrinsic and extrinsic capacitance component.

$$C_L = C_{int} + C_{ext}$$

C_{int}= intrinsic output capacitance of the inverter, associated with diffusion capacitances.
C_{ext}= represents the fan out and wire capacitances.

Now, let me come to the sizing of the inverter, the idea is for a given C_L or for a given load capacitance what is the minimum propagation delay? Right, How? If you have a buffer? Means you have a large 1, 2, 3 inverters right and therefore you need to find out what are the optimum number of stages and what is optimum NMOS to PMOS ratio. As we have discussed already that a load capacitance can be distributed by internal capacitance or intrinsic capacitance and external capacitance, so this is primarily because of fan out, this because of fan out and this is because of intrinsic output capacitance of the inverter associated with the diffusion capacitance and so on and so forth, so I get C_L equals to C_{int} plus C_{ext}.

(Refer Slide Time: 27:18)



Sizing of Inverter chain

Propagation delay, $t_p = 0.69R_{eq}\left(C_{int}+C_{ext}\right)$

$t_p = 0.69R_{eq}C_{int}\left(1+C_{ext}/C_{int}\right)$

R_{eq}= equivalent resistance of gate

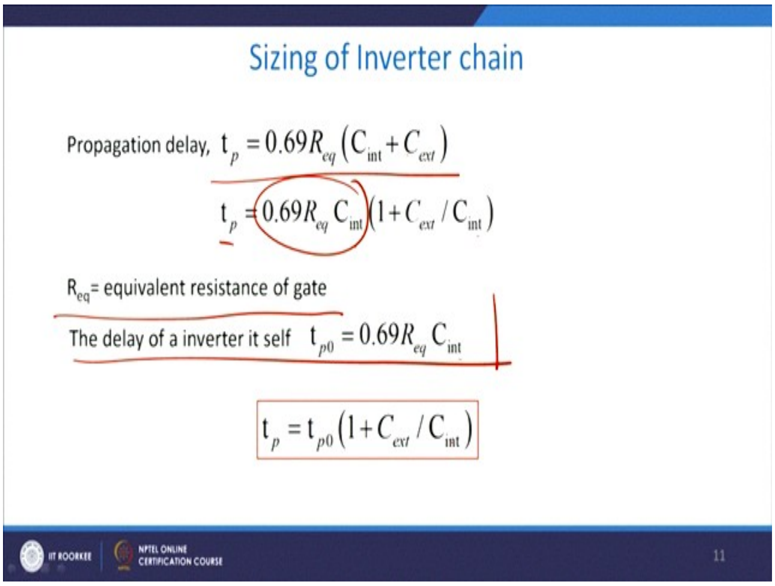
The delay of a inverter it self $t_{p0} = 0.69R_{eq}C_{int}$

$$\boxed{t_p = t_{p0}\left(1+C_{ext}/C_{int}\right)}$$

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
11

So propagation delay tp will be equals to 0.69 R equivalent to C internal plus C external we have seen this point. If C internal as output, then I get tp equals to 1 plus C internal by C external, where R equivalent is basically the equivalent resistance of the gate, whatever gate your trying to use. So the delay of the inverter itself is given by this value that tp0 equals to 0.69 R equivalent to C internal and therefore we replace this by tp0, tp0, you understand that the inverter itself will have some intrinsic delay and that is given by 0.69 R equivalent to C internal and I get tp0 C external by C internal.

(Refer Slide Time: 27:53)



Sizing of Inverter chain

Propagation delay, $t_p = 0.69R_{eq}\left(C_{int}+C_{ext}\right)$

$t_p = 0.69R_{eq}C_{int}\left(1+C_{ext}/C_{int}\right)$

R_{eq}= equivalent resistance of gate

The delay of a inverter it self $t_{p0} = 0.69R_{eq}C_{int}$

$$\boxed{t_p = t_{p0}\left(1+C_{ext}/C_{int}\right)}$$

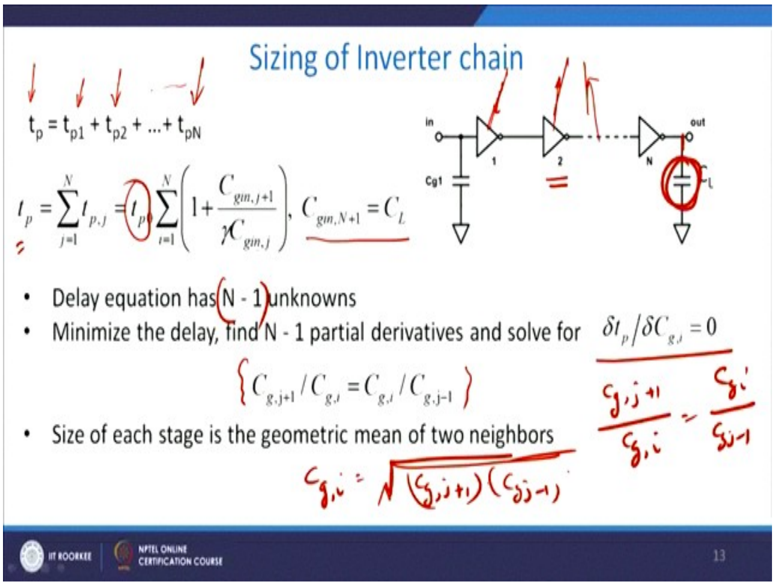
IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
11

Now therefore sizing up an inverter will reduce its delay because its R_S getting reduced but this is same instance of time, its input capacitance also increases, so we are not very sure whether what is happening, so there must be some optimal value of increase or the sizing of the inverter which will ensure me a minimum delay.

Now let us suppose my the input gate capacitance and intrinsic output capacitance obviously are the function of gate size, so therefore I say that C intrinsic is γ times Cg, where γ is basically a proportionality factor depending upon the function of technology, so whatever technology you want to use you can use and therefore I can safely write down tp to be equals to tp0, 1 plus f by γ right, what is f? f is C external by C internal.

So by, this is C external by C internal is f right, f is C external divided by γ if you write down, then I get tp equals to this, this, which means that the delay of the inverter is a function of external load and its internal capacitance, intrinsic capacitance. So if you take an inverter, right, then the delay of the inverter is basically ratio between the external loads C_L, load capacitance and its intrinsic capacitances, what is intrinsic capacitance? Again diffusion capacitances, gate to drain depletion capacitances, overlap capacitances, so on and so forth. So it depends upon the ratio of your output capacitance to input capacitance for a basic inverter.

(Refer Slide Time: 29:29)



Sizing of Inverter chain

$t_p = t_{p1} + t_{p2} + \ldots + t_{pN}$

$$t_p = \sum_{j=1}^{N} t_{p,j} = t_{p0} \sum_{i=1}^{N} \left(1 + \frac{C_{gin,j+1}}{\gamma C_{gin,j}}\right), \quad C_{gin,N+1} = C_L$$

- Delay equation has (N - 1) unknowns
- Minimize the delay, find N - 1 partial derivatives and solve for $\delta t_p / \delta C_{g,i} = 0$

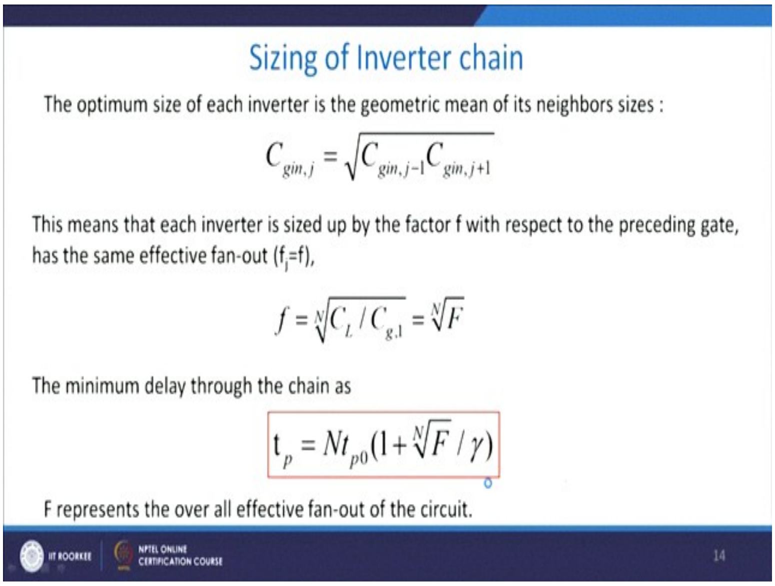
$$\{C_{g,j+1} / C_{g,i} = C_{g,i} / C_{g,j-1}\}$$

$$\frac{C_{g,j+1}}{C_{g,i}} = \frac{C_{g,i}}{C_{g,j-1}}$$

- Size of each stage is the geometric mean of two neighbors

$$C_{g,i} = \sqrt{(C_{g,j+1})(C_{g,j-1})}$$

in
out
Cg1
1
2
N
C_L



Sizing of Inverter chain

The optimum size of each inverter is the geometric mean of its neighbors sizes :

$$C_{gin,j} = \sqrt{C_{gin,j-1}C_{gin,j+1}}$$

This means that each inverter is sized up by the factor f with respect to the preceding gate, has the same effective fan-out (f_j=f),

$$f = \sqrt[N]{C_L / C_{g,1}} = \sqrt[N]{F}$$

The minimum delay through the chain as

$$\boxed{t_p = Nt_{p0}(1+\sqrt[N]{F}/\gamma)}$$

F represents the over all effective fan-out of the circuit.

So let us see how we can size an inverter chain. So what we do is that tp which we say is the total delay is equals to tp1 for the 1st gate, 2nd gate so on and so forth till nth gate, for nth gate you go on adding all the delays together, so tp therefore will be given as this quantity, summation tpJ, J equals to 1 to n, whereas tp0 is the intrinsic delay and summation 1 to n, if you do it we just saw that, if you take for example this inverter, 2nd inverter, then 1 plus $C_{g\ in}$ means input gate capacitance of this one divided by γ times $C_{g\ in}$ of the previous one, this one, right and this we get.

So I get $C_{g\ in}$ N +1 equals to C_L, which means that the last capacitor you are loading with load capacitor C_L, fine. So the delay therefore has got N - 1 unknowns because there are N - 1 transistors available to it. We need to solve therefore ∂tp, ∂$C_{g\ i}$ and equate it to 0 and if you do that, we get something like this into our consideration that C_J, $C_{g\ J}$ + 1 upon $C_{g\ i}$ equals to, so I get $C_{g\ J}$ + 1 divided by $C_{g\ i}$ equals to $C_{g\ i}$ divided by $C_{g\ J}$ - 1. So if you take C_g, if you want to find out $C_{g\ i}$ it is nothing but square root of $C_{g\ J}$ + 1 multiplied by $C_{g\ J}$ - 1, right.

So if you want to find out the input gate capacitance of 2nd one then find the geometric mean of the 1st and 3rd one right, with that geometric means you see square root of this gate capacitances. If you are able to fix the value of second one such that it is the square root of the two subsequent ones, we automatically get a reduced factor. We will take care of the next profile in the next subsequent lecture. Thank you very much!!!!!!!