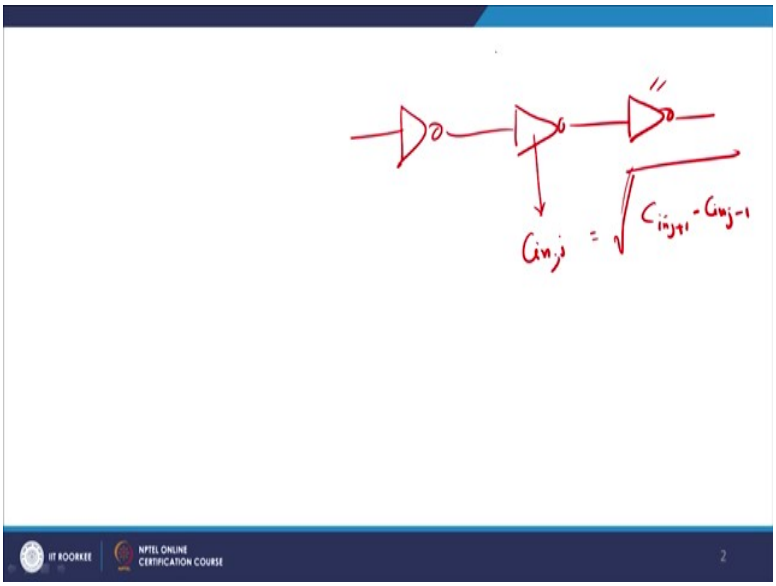
Microelectronics: Devices to Circuits
Professor Sudeb Dasgupta
Department of Electronics & Communication Engineering
Indian Institute of Technology Roorkee
Lecture – 20
CMOS Inverter Basics - III

Hello everybody and welcome to the NPTEL online certification course on Microelectronics: Devices to Circuit, we start today with CMOS inverter basics part 4, what we will be doing in this case is we will be looking into the from the time when we left in the previous case, we saw that in the previous interaction that for an optimized design in terms of reduced delay reduced delay between primary input and output.

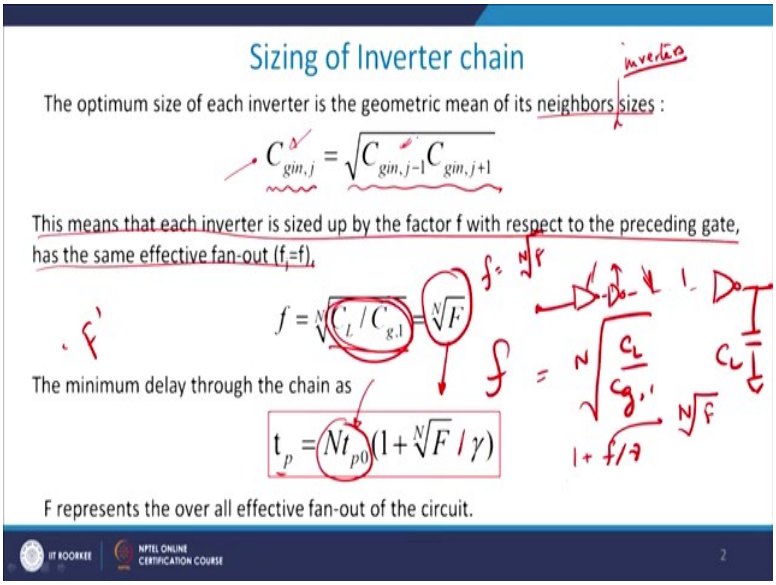
The gate capacitance of the middle transistor or a middle inverter in a chain of inverter should be in a geometric progression of the subsequent and the president inverter.

(Refer Slide Time: 01:18)



So if we have three inverters, so what I wanted to say was that if we have got two inverters so you have got one inverter, two and then three here, right, then if you want that the delay should be minimized, then try to keep the input capacitance of this inverter, right, J let us suppose, it is J to be equals to square root of Cin J plus 1, this 1 multiplied by C in J minus 1 which is this one, right. So we have seen that and therefore, we… how we got it? We have derived in the previous lecture this basic concept that if you want to do this you have to do these basic concepts.

(Refer Slide Time: 02:02)



So let me start from there and show to you that therefore, the optimum size of each inverter J^{th} inverter is the geometric mean of its neighbouring inverters or neighbouring sizes, right, it should be neighbouring inverters as well, right. So what we are telling you is that J^{th} inverter will be basically if you are able to fix its value of input capacitance to be equals to this, then possibly you will get a reduced delay.

Now, this means that each inverter is sized up by a factor of widths F with respect to the preceding gate, has the same effective manner which means that you see, so what I was saying was that you had inverters like this chain of inverters and it ended up at n^{th} inverter, then the last one you had a load capacitance C_L and this was V out.

So, if each of the inverters gives you a, this should be geometric mean of this thing. Similarly, this should be geometric mean of this and this, then what we finally get is that if you take an overall chain do a cross multiplication, then you get C_L by c_g1 square root of n^{th} square root n^{th} root, this will be the function f, fine? I think it is clear to all of you, why?

Because as I discussed with you that as you move from the first inverter to the last inverter in a chain of inverters then you have to progressively size it up as you go from lowest to highest value, but how you will is… the rate at which you will be resizing it up will determine, whether you are optimizing a design or not?

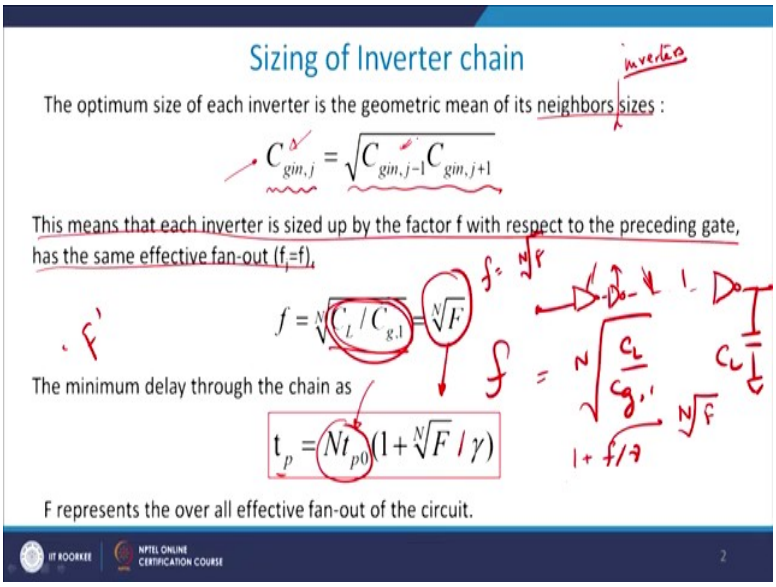
Now, if your last inverter was terminating into a load capacitance whose values equals to CL and the first inverter was basically equals to c_{gi} input capacitance, then C_L by c_{gi} or c_g1 n^{th} root of that will be your f factor which is basically your sizing of factor f. So if this is capital F

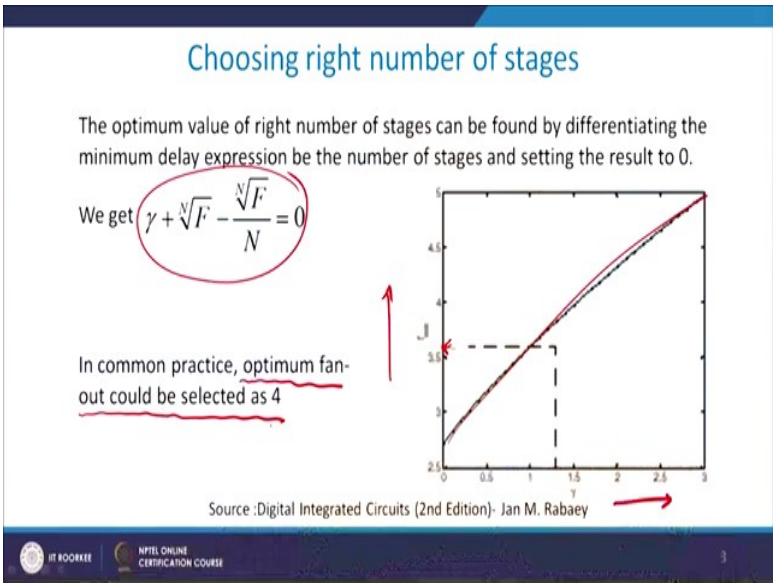
which we term, let us suppose the ratio is termed as capital F also referred to as fan out, electrical fan out, then we refer to small f to be equals to n^{th} root of capital F, which you see in front of it, right.

Once you have known this you just have to feed it into your original equation of chain, so you get tp equals to Ntp0 because see if each inverter has an intrinsic gain an intrinsic delay of tp0 and if you have n such chain, obviously that delay will be obviously larger than Ntp0. And therefore, you see Ntp0 coming here as a quantity.

You also have therefore, this quantity therefore coming up here, divided by γ from where they got this? From the previous, our previous understanding we got this, that if this is so I was saying it is F by γ if you remember. Now this f is nothing but a n^{th} root of f, where capital F is given by this quantity and therefore, I get this by γ to be equals to tp, where tp is the overall delay on this thing. So F represents the effective fan-out, right and we get effective fan out.

(Refer Slide Time: 05:12)





Now what you do is that you need to differentiate the previous equation which equation this equation tp with respect to N and make it equal to 0 and then minimize it and then if it say equal to 0 I get this equation into consideration that γ plus something minus n^{th} root of F by N equals to 0.

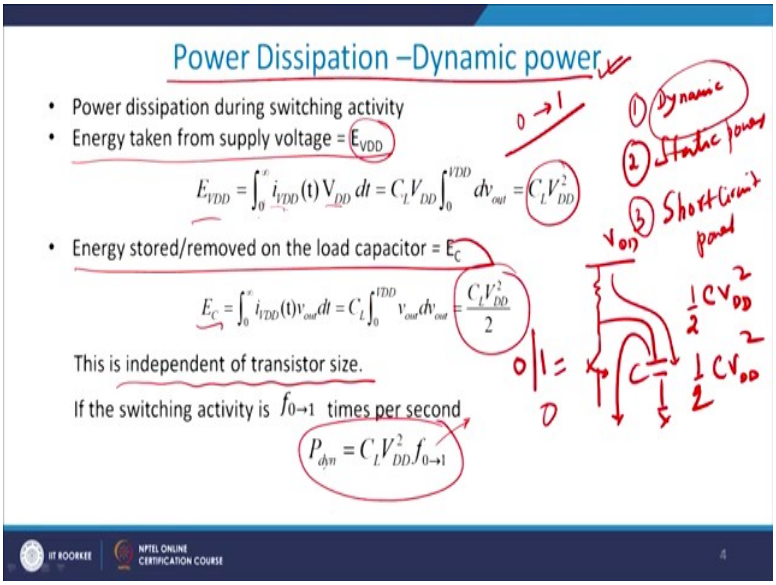
Now this typically means that it has been shown that therefore, if you plot F or if you plot γ on the x-axis and you plot the delay on the y-axis typically you get a curve something like this you will get a curve something like this, somewhere around if you fix γ equals to 1.5 you will get a delay of approximately 3.5 maximum delay which you sent approximately 4.

So in common practice it is always advisable to keep the optimal fan out as equals to 4, , right because this is what you get so approximately 3.5 to 4 you get if your γ is goes to 1.5, gamma is basically your factor which gives you the output ratio of your capacitances.

Now with this knowledge we therefore, with this knowledge we therefore tell that the optimal fan out should be approximately equals to 4 which means that if a single transistor is there and you want to optimize the output you just have to have fan out of 4 available with you and that will give you the best results as far as this design is concerned, right.

Let me now come to an important topic, so we have understood what the delay is all about, how to size a transistor to get the optimal delay. Now a very important property of inverter is that it is basically switching from ON to OFF state or 0 to 1 state in the output side. So since it is switching back and forth from a high value to a low value and vice versa, there is always an energy which is being dissipated or a power which is being dissipated due to the switching action and that power is basically known as dynamic power, right.

(Refer Slide Time: 07:28)



Typically a transistor or an inverter will have three types of power, one is known as dynamic power, right, the other is known as basically a static power, this is static power and third one is actually referred to as a short circuit power, we will explain each one of them individually, but let us first understand dynamic power, right.

Dynamic as the name suggest is basically the power which the CMOS dissipates or draws from V_{DD} rail when you do have a input which is varying from 0 to 1 and 1 to 0, so you do have switching characteristics which is available with you. If you see then as all of you are aware of that energy taken from a supply, suppose EV_{DD} is the energy taken from the supply V_{DD} must be equals to 0 to infinity, I current multiplied by voltage, V into I is basically the power which you get and if you integrate from 0 to T in time domain, so what will happen is this if you if you break down I and V, I get C_L times V_{DD} into integral 0 to V_{DD} dV out, right, if you solve it, I get $C_L V_{DD}$ square, right.

So if you remember from our basic inverter, so I had this, right and I had this, and then this. So every time you are actually charging this capacitor, you are taking it half CV_{DD} square where C is this capacitance, power from the V_{DD} rail, right. In the next half cycle when this was closed the same used to go here and you just remove this half CV_{DD} square, so you take half.

So in one cycle you take half CV_{DD} square and you throw it in the next cycle to the output side, so when you have input equals to 0 your capacitor charges to half CV_{DD} square and

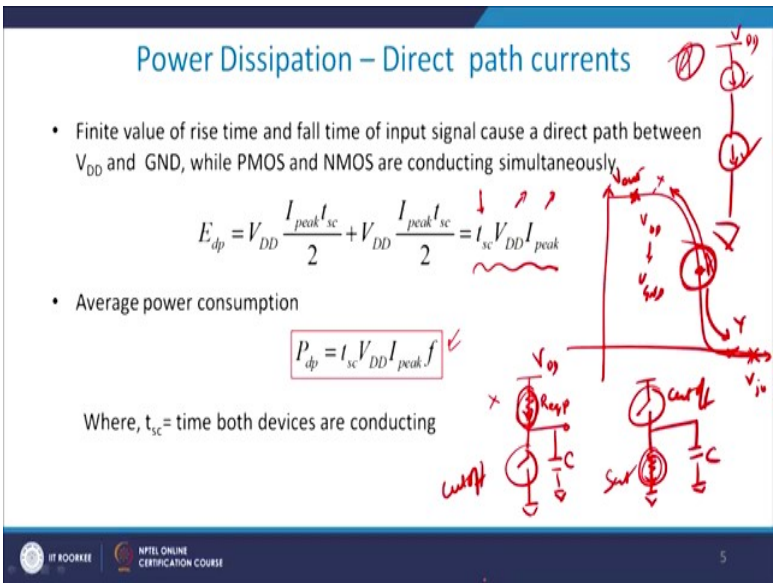
energy is stored is half CV_{DD} square, from where does it come? It comes from the V_{DD} rail, right which is the power deal.

In the next half cycle when input is equals to 1 and output goes to 0 then the charge accumulated on to this is dissipated across the ground and therefore, it goes to 0 and therefore, I get the total power dissipated is equal to half plus half is basically $C_L V_{DD}$ square over two cycle of ranges.

Now energy stored, removed in a capacitor is given by as I discussed with you is $C_L V_{DD}$ square by 2 because half it is there, this is independent of the transistor size. So please understand that your effectively the charging or discharging of the capacitor is independent of the size of the transistor, so it can be any size which you can choose and you can actually have a size which is available with you.

Now, if switching activity is 0 to 1, so you see an important point is that only when your output goes from 0 to 1, right that is the time when you are accepting power from the VDD rail. So please understand the dynamic power, which if you are not dissipating you are getting it from the V_{DD} rail is only happening when your output is going from 0 to 1 because that is the point when your capacitor is getting charged through PMOS and you are drawing power from the V_{DD} rail and that is the reason we say P dynamic to be equals to $C_L V_{DD}$ square multiplied by the frequency of 0 to 1, right what is the frequency of 0 to 1 which is there with me.

(Refer Slide Time: 10:50)



Now that is known as that is as the dynamic power dissipation. Now what happens is that we have discussed this point earlier also, that when you draw the voltage transfer characteristics, right this is what you get, this is your V in, right, and this is your V out, at this stage NMOS is cut off and PMOS is switched on, and at this stage PMOS is cut off and NMOS is switched on and that is the reason grounded.

And therefore, you do not have a direct path between V_{DD} and ground, so if you at this point if you, let us suppose this point X at point X if you want to find out what is the output characteristics it looks something like this, it looks like this and then this is open and this is ch, right, so this NMOS is opened and this is basically by R equivalent P, this is V_{DD} and this is my C capacitance available with me.

So which means that this is saturated and this is cut off, what happens at Y which is this point? Y will be this will be open and then I will represent it by a resistance here, and there will be capacitance here C, so this will be then saturated and this will be cut off, fine, but these are the two extremes where you are seeing it, somewhere in the middle, somewhere here I possibly will be seeing that both will be acting as a current source, both will be saturated, somewhere at this point say A at point A.

As you can see since two current sources in series is basically intrinsically an unstable situation, obviously you do not stay here for a quite long time, a small change in the input will result this A shifting to this point or to this point, but then if you are working at A then you do have a direct path between the V_{DD} and the ground rail, please understand this properly that whenever you, this is somewhere when both the devices are in saturated state at point A, at this point A when you in saturated state both are in on state, on state primarily meaning is that it is basically behaving like a current source in a saturated state and as a result you will have large amount of current flow.

So even I can represent this by a current source both of them right those output impedance is typically very large. Now this will, then so there is a short circuit path between V_{DD} and ground, right, and as a result you will see a large shift current flowing that current,,,, flow of current is given by this formula. So V_{DD} is the applied voltage I peak is the peak current multiplied by how much amount of time this current is flowing is tsc a short-circuit time and therefore, it is given by energy Edp, right.

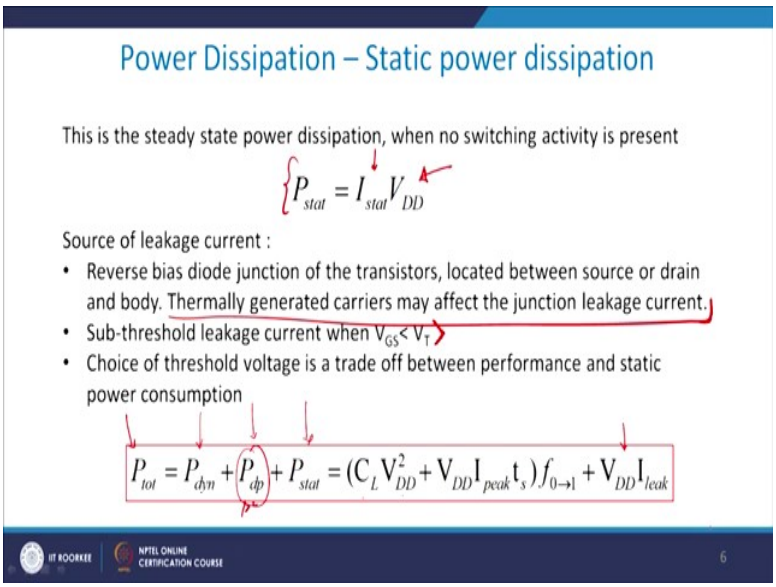
So Pdp will be multiplied by frequency tsc V_{DD} I peak into f so if you go once from a high to low and then you go to low to high, then you are actually traversing through short circuit path to twice and therefore, it is depending on the value of this thing tsc is the time when both the devices are conducting, so these are the two time when they are conducting.

So we have discussed basically the dynamic power we have also understood what is the short circuit power, electrical static power, static power or the steady-state power is defined as that power when you are not in the dynamic position which means that your input is not varying but it is fixed or even your device may be in the cut off state and you expect that there should not be any current flow through the device and I would expect to see that the power dissipation because of should be equals to zero, but it is not and the reason is something like this.

The reason is that whenever you assume that the device is off by saying that the gate voltage is falling below threshold, but if you go to literature available literatures, you will see that not necessarily the device is actually off when your gate voltage falls below threshold. So if your threshold voltage is say 1 volt and you are at 0.9 volt, then as per our understanding it should be off, but in reality there is some sub threshold current still available with there.

And so though you think your device has been switched off, in reality the device is still on and that gives you so small current internally, right that is what is known as a sub threshold leakage when your V_{GS} is less than V_{Th}, right.

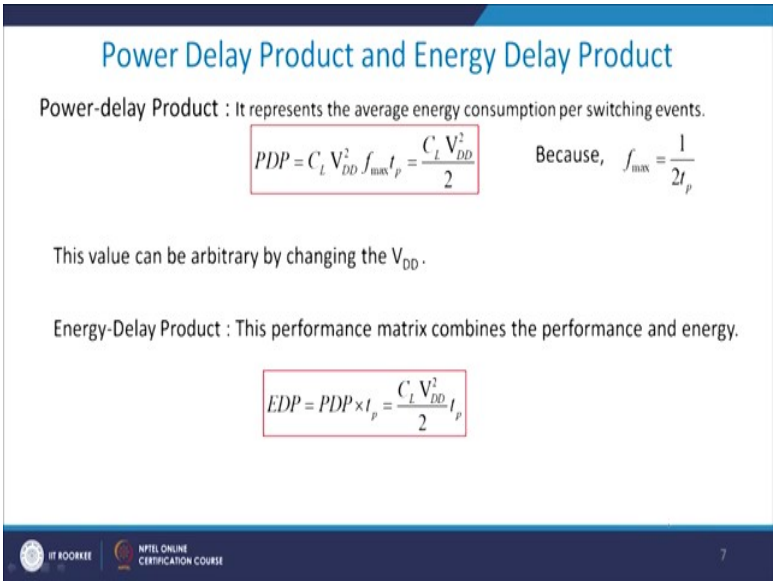
(Refer Slide Time: 15:34)

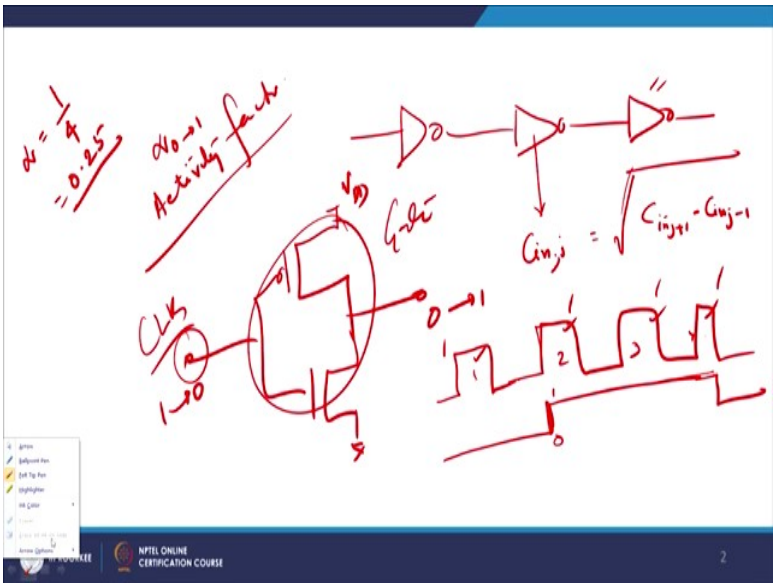


Similarly you will have thermally generated carriers which will affect your junction leakage, remember your base to the source and drain is reverse bias junction you remember and therefore, the depletion thickness is quite large there, but for minority current carriers it is a basically not a hill, but a slope so if the temperature increases by even 10 degree I would expect to see almost doubling of the value of your minority current carriers which might result in a large current.

So these currents which is basically the virtue of reverse by a saturation current is primarily because of these reasons, so I define P_{stat} as the static power to be equals to I_{stat} the static current multiplied by V_{DD}. So we define P total the total power which is with us as equals to dynamic power plus your static power plus your this is short-circuit power, right, it should be SC short-circuit power. So I refer to as $C_L V_{DD}$ square plus V_{DD} into Ise into tsc into f 0 to 1 and this is V_{DD} into I_{leak}.

(Refer Slide Time: 16:52)



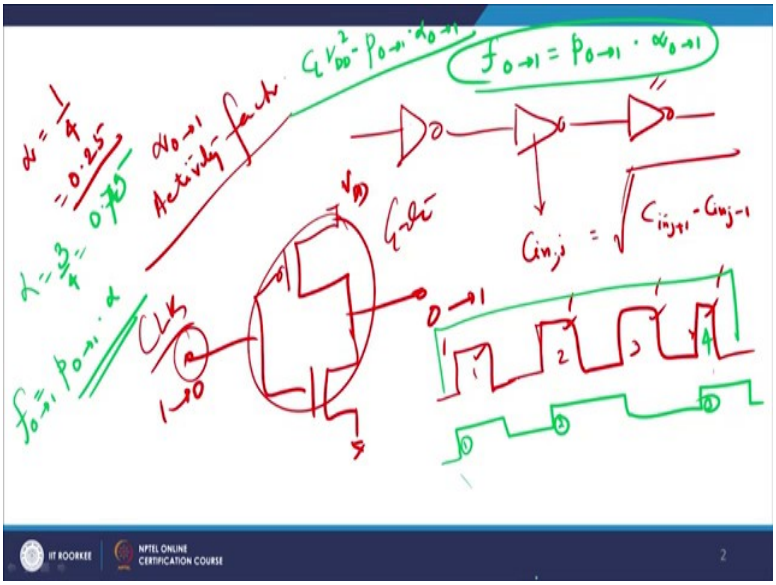


I will explain one important term which is well it is not here, but I will explain it later on maybe I will explain it here. I define a term known as activity factor α, α 0 to 1 also referred to as activity factor. Let us suppose I have an inverter, the inverter will be easy to explain and I have an inverter here and my output is here, input is here, so I just check out how many 0 to 1 transitions are available.

So let me let my input be a clock, so I have four clock cycles 1 2 3 4, right. Now, I have a structure which is not an inverter, right which is not an inverter which is something, some gate, some r bit gate and that r bit gate if you look very carefully has got this output which means that for every 4 clock cycle this is the clock I am giving let us suppose, and this is some gate, very complex gate, my output is showing a 0 to 1 transition only in the second clock cycle, and it is then 1 to 0 in the fourth clock cycle.

So there is only, so for every 4 clock cycle there is only one 0 to 1 transition, then we define α to be equals to 1 by 4 that is equals to 0.25, fine have you understood?

(Refer Slide Time: 18:28)

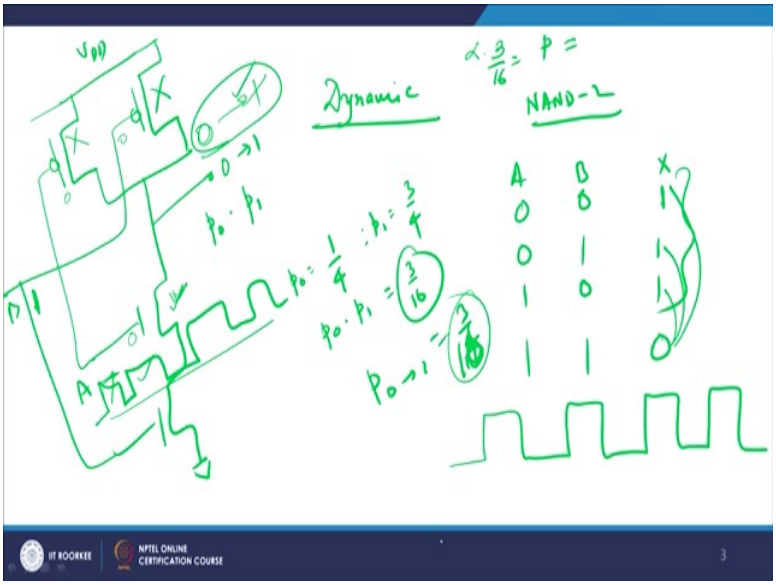


Similarly, if let us suppose, let me just erase this one, right and let me just put it like this that you do have the same clock cycle there are 4 clock cycles, but now what has happened is that rather than one 0 to 1, let us suppose you have one something like this, so there are 4 clock cycles here 4, 1 2 3 4, now you have one 0 to 1 here, you have two 0 to 1 here, you have a third 0 to 1 here, then your activity factor alpha is basically 3 by 4 and it is equals to 0.75.

You will ask me where it is required? Well the frequency which you see f 0 to 1 can be written as probability of 0 to 1 multiplied by α, α is the probability vector. So f 0 to 1 is the frequency of 0 to 1 transitions that is written as probability of 0 to 1 transition multiplied by α 0 to 1 transition, so this is a typical formula which you use, so I can have therefore, the dynamic power dissipation to be equals to $C_L V_{DD}$ square multiplied by probability 0 to 1 multiplied by α 0 to 1, right.

So this is defined as the α is defined is my activity factor for all practical purposes and this gives me an idea if your design is having a… so the gate structure is such that you have a large value of α, then you would expect to see a larger power dissipation, whereas if your gate value has got a lower value of alpha you would expect to see a lower dynamic power dissipation for this case, right.

(Refer Slide Time: 20:25)



One thing which you should be therefore careful about is that whenever you are planning to draw a design or dynamic power dissipation it is not only important that how your design is but what type of gate the design is. For example, I will give you a brief idea, let us take you take NAND-2 logic let us suppose you take. So I have got A, B and NAND-2 logic let us suppose X is the NAND-2 logic, so 0 0 1 1 0 1 0 1, so NAND gate means 0 0 will give you 0 output and 1 1 will get it, you will get 1 1 0, right which means that I will get this, this, this to be a power consuming cycle, whereas 1 to 0 will not be a power consuming cycle, right.

So if you want to find the probability of 0 to 1 you need to find out what is the probability of 0 multiplied by probability of 1, so probability of 0 here is basically P of 0 is 1 by 4 because there are 4 and 1 and probability of 1 is basically 3 by 4 here, 3 by 4. So simply multiply P 0 by P 1 and I get 3 by 16 as the probability of 0 to 1, so probability of 0 to 1 is basically 3 by 16, right.

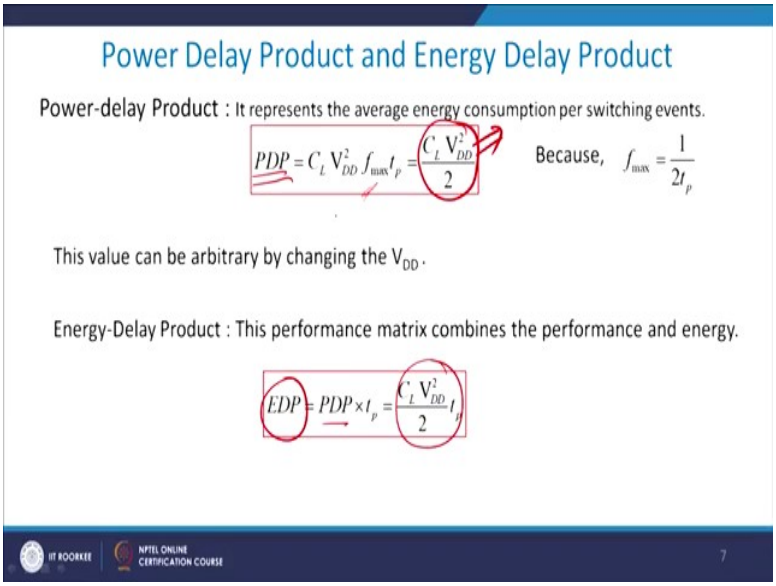
Now in a NAND-2 gate remember therefore, if I give a clock cycle let me say I have a NAND-2 gate, NAND-2 look something like this, right, I have a this is NAND-2 logic, so this is A and let us suppose this is B, so I have got this NAND-2 logic and this is my V_{DD} here and this is a NAND-2 logic. So when it is 0 0 output is 1, 0 0 when it is 0 where 1 output is 1, 1 0 output is 1, 1 1 output is 0 and I get the NAND-2 logic here.

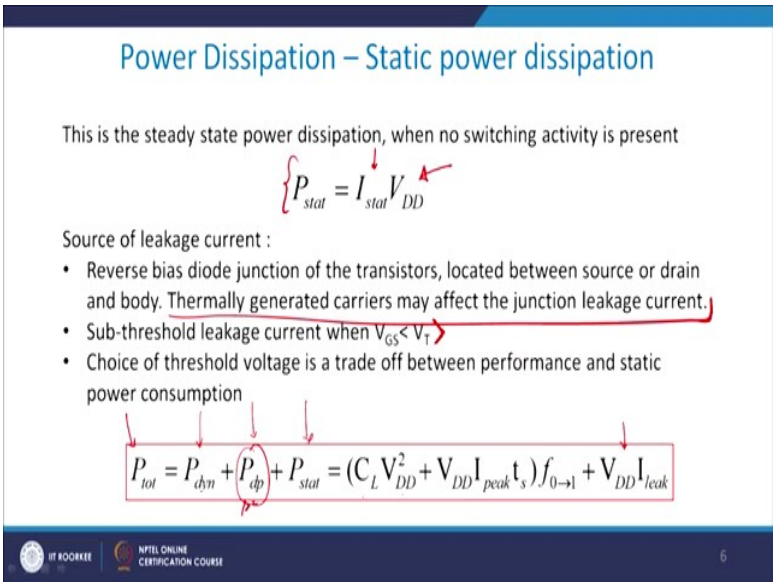
So what is the probability is 3 by 16 here, what is alpha? So if I have now an input cycle which is something like this on the A and let us suppose A and it is something like this on the

A, and B is latch to say 1, then you can find the probability how many transitions are there in the output side from 0 to 1, right.

So if you have 1 here, it means that this is cut off and if you A is giving if A is initially 0 so this is on output will be 0, as it goes high as it goes out this cut offs this switches on, right and since this is already one as this switches on this goes from 0 to 1, so this is 1 so with 1 clock cycle I get one 0 to 1. Similarly, the next clock cycle is get one 0 to 1, so on and so forth so that α multiplied by 3 by 16 will give you the overall probability or the frequency and therefore, that will give you the output characteristics available to you, right, this gives you the frequency of operation of the device.

(Refer Slide Time: 23:15)





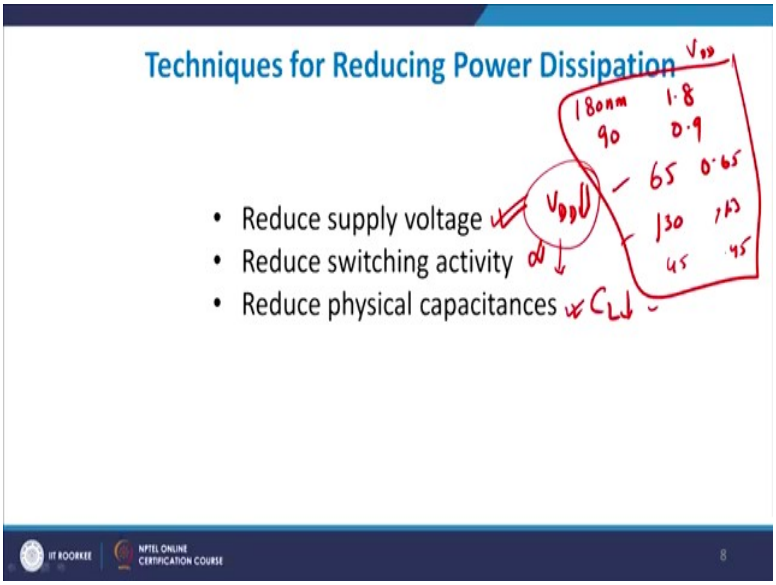
Now, we define a new term which is basically power delay product PDP and it is given by $C_L V_{DD}$ square by 2 at this stage we need not worry about too much about it and we define energy delay product which is PDP multiplied by tp and therefore $C_L V_{DD}$ square by 2 into tp gives you the value of your EDP which is energy delay product,

So we have two types of products which is there with me one is an energy delay product and we have a PDP and we try to optimize the energy sorry we try to optimize the we try to optimize the delay and therefore, if you want to up to reduce the reduce the delay dynamic power dissipation what is what is there in your hand a very good idea to reduce the dynamic power dissipation is to reduce V_{DD}, because if you reduce V_{DD} you have almost a parabolic decrease in the value of your power because it is square you are talking about and that is the reason there is a sudden drop which you will see here in this case.

You can also operate at lower frequency of operation, so in most of the cases when you do not want the power where you do not want a very high frequency of operation you can actually afford to keep your f_{max} low, once you do that your PDP and as well as your total power actually reduces drastically. So these are the few things which you should keep in mind as far as designing is concerned.

Generally PDP is a constant for any system PDP and EDP are a constant, so energy delay product and power delay product are constant which means that if the power dissipation rises if power dissipation becomes high then that can be only accommodated provided your system becomes slow, so tp has to go down, tp going down basically mean that your system is getting faster in this case.

(Refer Slide Time: 25:03)



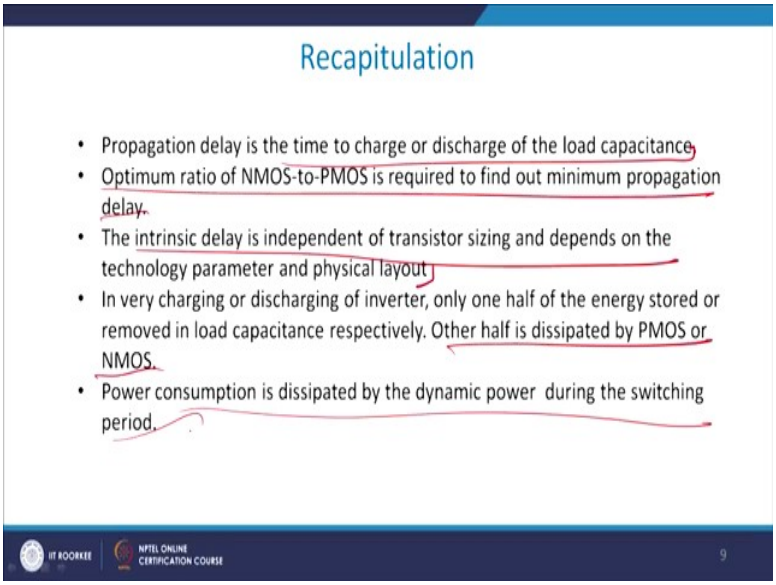
Now this is what you get from the power delay product and energy delay product concept, techniques for reducing power dissipation let me see, we have, we can do it by as I discussed with your reducing power supply, reduce your switching activity α and reduce physical capacitance, so reduce C_L, reduced α, reduce C_L and reduced V_{DD}, once you take care of all these three you will have a reduced power dissipation which is there in this module.

But then the problem is that if you reduce V_{DD} beyond a particular point your current reduces and the time taken to charge or discharge the capacitance becomes large and therefore, the τ becomes a large factor, so you have to optimize the value of V_{DD} to a larger extent.

So the rule of thumb is that typically the rule of thumb is that if you are working with 180 nanometer technology your V_{DD} should be 1.8, you have only 90 nanometer is 0.9, if you want 65, 0.65, if you are working with 130 then 1.3 volts and so on and so forth, if you are working at 45, 0.45 and typically if not in a very layman sort of a rough back-of-the-envelope you do it and you get these are the V_{DD} values.

Best way to do is to reduce V_{DD}, but then if you reduce V_{DD} the price you pay for it is of course that you are ending up having a higher lower current and therefore, a larger τ and therefore the time taken will be also larger in this case.

(Refer Slide Time: 26:30)



So let me therefore recapitulate what we did in this idea, we tried to find out the minimum what is what was the time taken to charge or discharge the load capacitance, we also tried to find out the optimum ratio of NMOS to PMOS for minimum propagation delay.

We saw that the intrinsic delay is independent of the transistor size and depends only on the physical layout and the technology parameter. Now in every charging and discharging as we saw that the inverter actually loses half only half of the energy stored is removed to the load capacitor, other half is dissipated by PMOS and NMOS I think this clear.

So what we do is that for any single cycle take half $C_L V_{DD}$ square is the total energy I take, half I dissipate to the ground and half of it is dissipates through PMOS and NMOS sink. Power consumption is dissipated by dynamic power during switching period. So we have a switching period, how can you reduce power? Three techniques are available reduce your V_{DD}, reduce your α switching activity, or reduce your C_L. Reducing your V_{DD} not a very good idea beyond a particular point because the current will be there for reducing and your delay will be increasing.

So these are the few important takeaways from this from this module, from this lecture of inverter, why was this important? Because now since you know how a inverter, you can optimize its power and delay, you can now therefore do small modules of NAND gate, NOR gate, XOR gate using CMOS technology, right, maybe we can do it in the next time and show it to you how it works out in those domains.

And then optimize using sizing you optimize to get the minimum delay and the highest speed or we do an optimizing of a PDP and EDP of the delay, fine? With this, let me thank you for your patience hearing, okay.