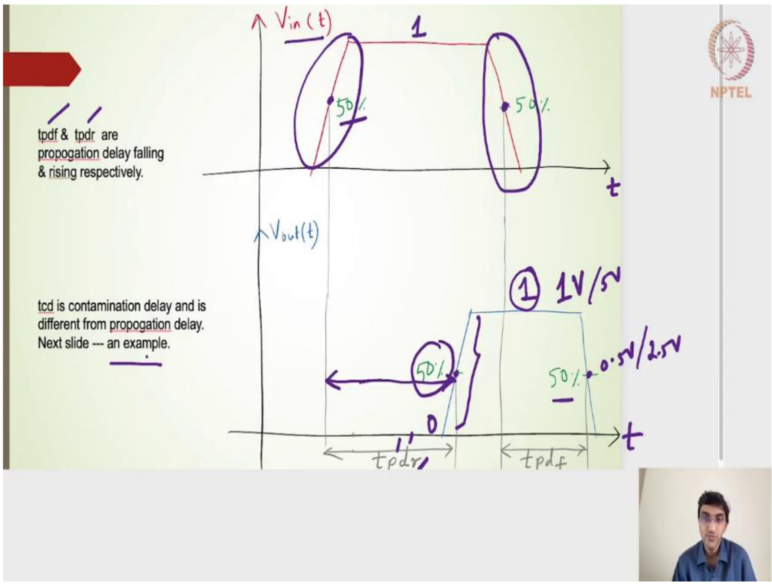
Design and Analysis of VLSI Subsystems
Dr. Madhav Rao
Department of Electronics and Communication Engineering
International Institute of Information Technology, Bangalore

Lecture – 20
Introduction to Delay in CMOS

Hello students, welcome to this the next module of this particular course. This talks about the delay more significantly here. This particular module will talk about the delay which means the performance of the circuit also. I have written it as a CMOS Delay which it means that it is the CMOS related circuit, and it is delay analysis for the digital circuits which we have designed using the CMOS circuit families.

(Refer Slide Time: 00:47)



tpdf & tpdr are propogation delay falling & rising respectively.

tcd is contamination delay and is different from propogation delay. Next slide --- an example.

Moving ahead what we will do is, we will quickly look at some of the definitions and then go back to our primitive digital circuits which is the inverter circuit. This is a definition of the propagation delay here what I have drawn. Let me pick up my pointer, this is the propagation delay falling and then propagation delay rising.

If I have an input signal here, this is the input signal, there is a little bit of a ramp signal and then a steady state signal of logic 1 and then it falls back. This is the input signal with respect to the time domain for a given circuit. Let us we do not know what circuit it is, but if you apply this particular input to that particular circuit.

The output of the circuit behaves like, let us take an example the output of the circuit behaves like this or gives this kind of a response. Then you can see that the output is kind of much much it arrives after some time with respect to the input signal. The output rises initially, and then it reaches to a steady state logic 1 and then it falls back.

In that case, if I want to define what is the propagation delay rising and propagation delay falling. Whenever the output starts rising, this particular the rising the output is actually rising from 0 to the logic 1 here. We consider that as the rising definition. Then the propagation delay definition turns out that whenever the duration to which the input whenever the input is at 50 percent and the duration till the output is 50 percent, we say that this is the propagation delay, and that is the propagation delay the p and d stands for that.

Then the rising one is with respect to the rising output. The propagation delay rising the term $\mathrm{t_{pdr}}$ which is nothing but the propagation delay rising defines the point where the input is 50 percent and then the output rising is 50 percent. Similarly, the propagation delay falling will be when the output is actually falling and it has reached the 50 percent of its falling output.

What the really the 50 percent means? Whenever it has gone to a 1, that digital logic 1 which means it is 1 volts or any other 5 volts, and the 50 percent represents either 0.5 volts and if it is 5 volts on this side. If it is 5 volts, then it should be 2.5 volts, so the 50 percent of the maximum voltage. Whenever it starts falling, whenever it has reached to the 50 percent of the maximum voltage, the duration from the input 50 percent to that of the output 50 percent will be my propagation delay falling. Hope you know this is clear.

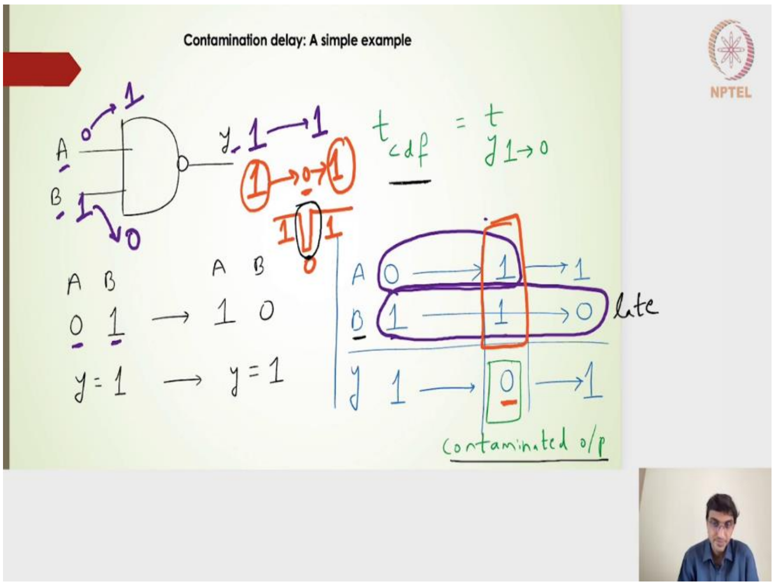
The input here I have taken a kind of a ramp signal, some kind of a rising signal here and then some kind of a falling signal here. But let us say that if it is a step input, it is a completely a step input then in that case the output voltage here if it turns out to be very very same as what I have drawn here.

Then for a step input, it will be nothing but wherever the step input has begun, and then the output whenever it has reached 50 percent of the rising 1. We will take that as the propagation delay rising and then we have another definition called as the contamination delay falling, and then which we will look into the next slide.

But most often we will come across the delay or the performance, it is basically the propagation delay falling and or the propagation delay rising, and also the average of that for a circuit. In a circuit if it has multiple paths, then we take the critical path.

The critical path being the path which has more number of gates we take that as a critical path and then determine its propagation delay characteristics. It could be the both the propagation delay rising and then the propagation delay falling for that particular critical path for the given circuit.

(Refer Slide Time: 05:27)



Contamination delay: A simple example

Let us proceed further and then see the contamination delay example. Just to see the example it is very similar to what we had done for the propagation delay falling and rising. It is again nothing but that 50 percent of the input 50 percent of the output and then the duration of that will give me the contamination delay. But the scenario is slightly different here.

Let me take an example here of a two input NAND gate. My inputs are A and B here, and output is y. Let us take a very simple example of 0, 1, if I provide a 0 and 1 onto the NAND gate, if it is 0 and 1, my output will be nothing but for NAND gate it will be nothing but 1. Now, tomorrow if I change from 0 to 1 here, and then from B 1 to 0 here, my output should stay to 1, it should not change.

But what happens is at the input side if A changes from 0 to 1 quickly, but B does not change quickly as the A has changed. B takes some amount of time to change from 1 to 0. In that case, this particular middle portion let me pick up a different color. This particular middle portion where A and B, when the inputs of a two input NAND gate is 1 and 1, my output is actually 0, it actually goes from 1 to 0 to 1.

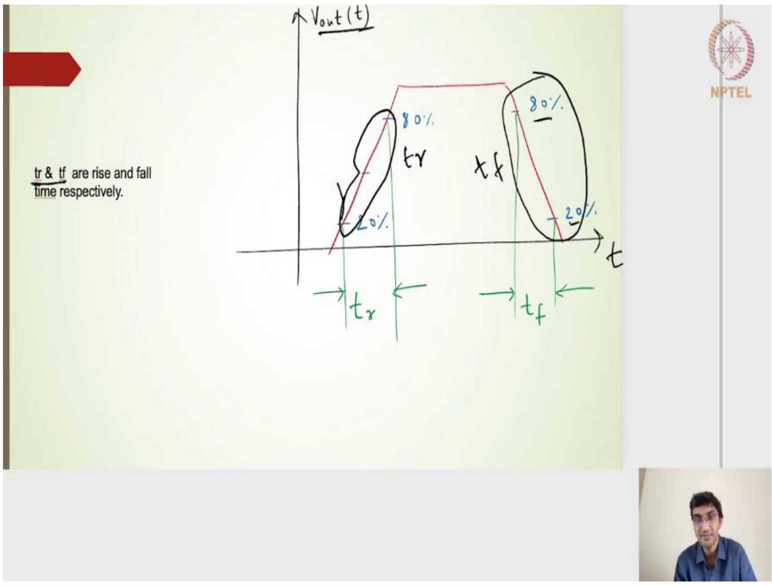
This is a contaminated output, this is not really the output we are looking for. The output is actually the steady state output we are looking for is 1 and 1, it stays to be 1, but there is a contaminated output. There is a 1 here and then there is a contaminated output for a short duration and then goes back to 1. If I draw the timing diagram, this is what we will get.

This particular contaminated output as the name says it is a contamination, and we really do not want this, but it is there in the system because of whatever the design we have done, and really because the B is reaching 0 late. This is actually reaching 0 very very late, and that is why we are getting this contaminated output.

Now, with respect to the input and then on the output, if I draw that 50 percent the input rising and then 50 percent of the output rising with respect to this contaminated output, we should then say that that is a contamination delay rising. Similarly for the contamination delay falling, it will be nothing but with respect to the rising input and then the or rather the output it is falling. The 50 percent of that will be my contamination delay falling output.

Remember that for both propagation delay falling or contamination delay falling, it is always the output which is falling. Then similarly contamination delay rising and propagation delay rising, it is the output it is defined with respect to the output which is a rising. Hope, it is clear.

(Refer Slide Time: 08:49)

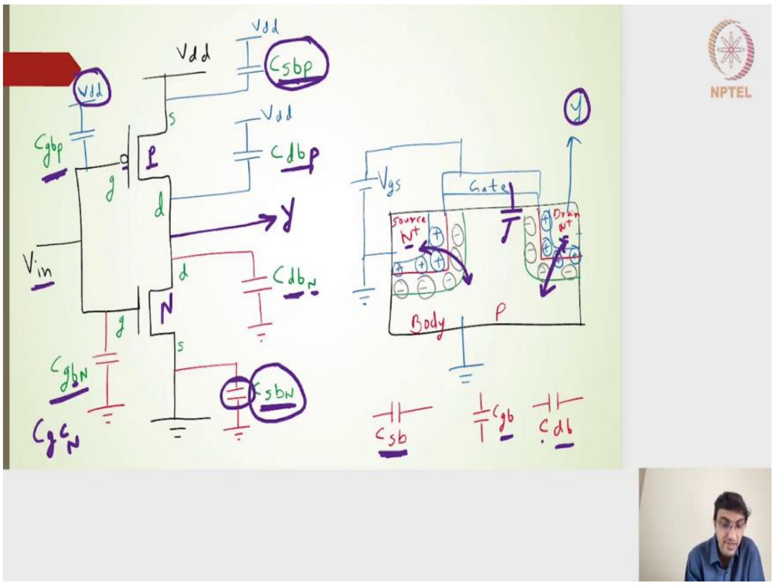


tr & tf are rise and fall time respectively.

Moving ahead there is one more definition called as the rise time and then the fall time which I have specified here t_r and t_f which is nothing but the rise time and fall time and with respect to the output signal, again this is the time axis and then the y-axis is the output axis. The rise time is nothing but when the output signal is rising you and then the duration it takes to this particular portion from 20 to 30, 20 to 80 percent is my rise time. Then from the output when it is falling from 80 to 20 that will be my fall time.

Do not get confused with the rise time and the fall time with that of and then mix with that of the propagation delay falling and rising. The propagation delay of falling and rising is with respect to the 50 percent of the output voltage level, whereas the rise time and fall time is that is with respect to the 20 to 80 percent. Do remember that in when you are calculating or estimating the rise time and the fall time for different circuits.

(Refer Slide Time: 09:56)



Coming back, we have done this definition. What is the use of this particular definition? Of course, it helps us in characterizing the delay of a circuit, especially the propagation delay falling and rising which will help us to characterize the delay of the circuits. The contamination delay falling and rising is kind of very important.

I you remember the contaminated output, it should not get captured into the next set of circuits especially when we are doing the sequential circuits, we have to characterize the setup time and then the hold time appropriately such that this contaminated output does not get carried forward, that is one thing.

The rise time and the fall time it is generally defined with respect to the input characteristics, to the input signal characteristics. Suppose I have an input of 20, the rise time so and so, how should the output behave. For a step input, I think the rise time will be 0.

But generally, what happens is whenever we are giving a signal or whenever we are getting the signal from the clocks or whenever we are generating a new signal, generally, we will have some kind of a ramp signal. Then similarly if I want to bring the output to some other level, we will have an input which is generated which will have some non-zero fall time, that is the use of the all these particular parameters.

But now how do we fit in or how do we characterize an inverter circuit with all these parameters? I have to go back to my inverter circuit, understand the parasitic capacitance elements and then see whether I can calculate or whether I can characterize some of this delay parameters, that is why we have come back to our the standard primitive inverter circuit.

Although, it looks little bit overwhelming, let me try to evaluate this, what I have drawn. I have an inverter here, it is nothing but the PMOS circuit, this is PMOS and this is NMOS. PMOS looking at this particular bubble signal we will know that this is a PMOS circuit and then the NMOS is the gates are tied together and then given to this as an input alright.

Let us say that for an NMOS, I have drawn an NMOS here N^+ and N^+ and let us say that this is my output signal. I have labeled it as y, this is the y here in the cross section and there is a gate and then there is of course the oxide and then of course, there is a body. Body is grounded, the source is also grounded.

Now, how many capacitances do we have? One capacitance I can easily see the depletion capacitances coming from drain to body. This particular capacitance I am talking about drain to body in the NMOS side. I have written drain to body capacitance on the NMOS side and I also have a capacitance on this particular side which is source to body on the NMOS side.

C_{sbn}, n represents the NMOS, s and b represents source and body. Similarly, here the capacitance C_{dbn} represents the drain to body capacitance and then the NMOS here. On this particular side, this particular capacitance if I look at it very very closely, I know that the source is anyways grounded, and then the body is also grounded. On both the terminals of this particular capacitance, it is grounded alright.

Similarly, if I look into the gate side, the gate side I will have a capacitances here, the gate to body capacitance or gate to the channel capacitances which we had seen in the previous module that as the $C_{sbn_{gs}}$ changes or as the V_{ds} changes, we can also appropriate a portion that into C_{gs} and C_{gd}.

I am going to take it as a particular value and then put it as a C_{gb} value C_{gc} or C_{gb}, where g b represents the gate body, we can also call it as C_{gc} channel with respect to the NMOS, but here I have taken it as C_g gate to body. For an NMOS transistor, we have three capacitances which we have denoted it.

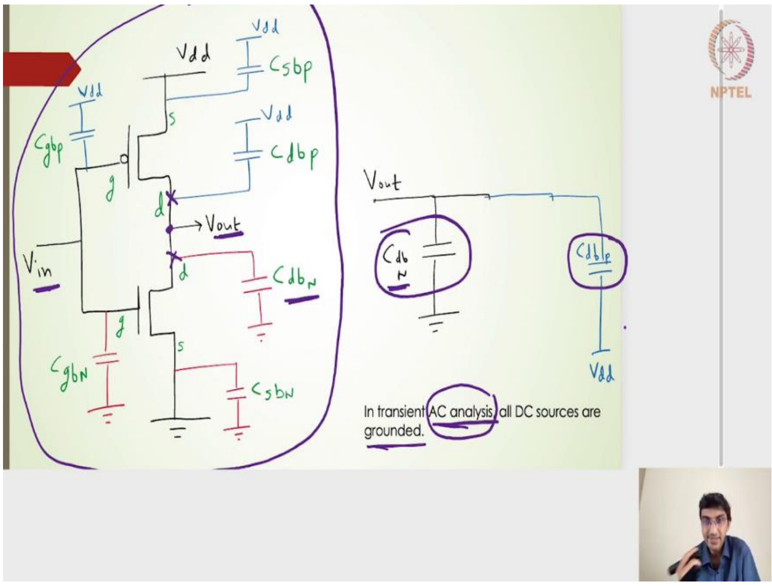
One is the input capacitance or the gate capacitances. The other two are the parasitic capacitances or nothing but the diffusion capacitance is also called as the parasitic capacitances, and depletion capacitances alright. Similarly, we will get for the PMOS, we will get three capacitances; one on the gate side, and then one on the drain side, and then one on the source side.

The body for the PMOS is now connected to the V_{dd}. Remember that the body of the PMOS is connected to the V_{dd}, source is connected to the V_{dd}. My source to body capacitances on both the terminals I will see the V_{dd} connected. On the drain side, it is connected to the body I mean rather the capacitances of the drain to body capacitances for the PMOS is denoted here.

Whereas on the input side, we will have the gate to body capacitances of the PMOS. The body is connected to the V_{dd} and that is why the V_{dd} terminal is here, and the other side is connected to the gate. Now I have the capacitances, looking into the capacitances, if I have any kind of a capacitances in the circuit, we know that there will be charging and discharging effect.

The charging and discharging profile of 50 should give me the propagation delay of falling and propagation delay in rising values, that is why the capacitance is realizing the capacitances in a digital circuit is very very crucial. This is the some of the definitions which we have anyway seen C_{sb} represents source to body capacitance, gate to body capacitances, and then drain to body capacitances. Hope this is clear.

(Refer Slide Time: 16:20)



In transient AC analysis all DC sources are grounded.

Moving ahead, I have drawn here this particular portion of the circuit is taken from the previous slide. On the right-hand side, I have drawn two capacitances, let us have a closer look at it the C_{dn} here at the output side.

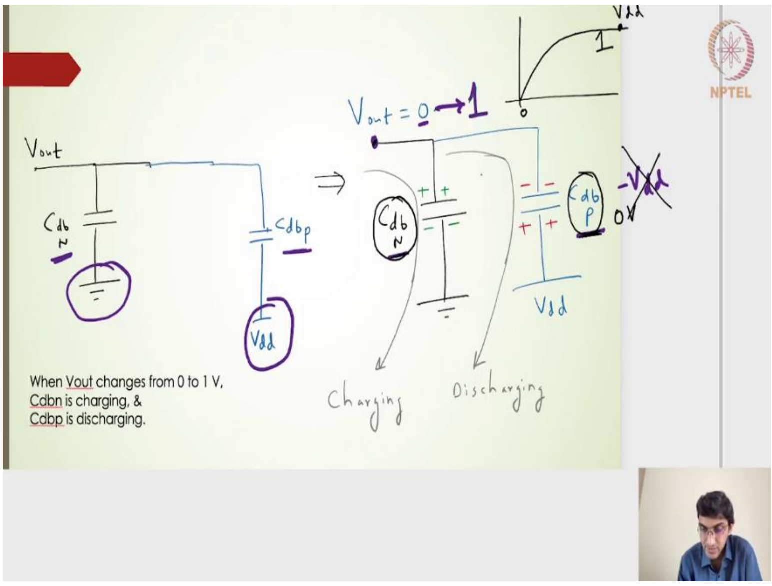
This is my output for an inverter, we know that the drains of the PMOS and the drains of the NMOS are tied together, that is the output and then one capacitance is C_{db} drain to body capacitances of NMOS and then drain to body capacitance of the PMOS. This is the drain to body capacitance of the PMOS, and then this is the drain to body capacitance of the NMOS. On one side, it is connected to the ground, the other side it is connected to the Vdd.

I have also written that in the transient AC analysis, all the DC sources are you know if I want to do a transient analysis especially if I am looking at the capacitances charging and discharging profile or rather the output whether it is rising or decreasing. Then estimating the propagation delay falling or rising looking at that particular 50 percent point in the output voltage profile, then actually we are doing a transient analysis.

This is basically the transition from 0 to 1 volts transition from any change in the input voltage should give me a change in the output voltage to reach to a steady state value. Before reaching the steady state value, we are doing that kind of a transient analysis.

For any kind of an easy analysis or a transient analysis what we say is the DC sources should be considered as a ground for any kind of an AC or a transient analysis that is one way of looking at it, and then trying to estimate the propagation delay parameters.

(Refer Slide Time: 18:19)



When Vout changes from 0 to 1 V,
Cdbn is charging, &
Cdbp is discharging.

But, what really happens? We will see. When the output voltage is actually 0 for this particular circuit of the drain to body capacitance for the PMOS and NMOS, for the NMOS it is connected to the ground here, for the PMOS it is actually connected to the V_{dd} alright.

When it is actually 0, this particular capacitance C_{db} drain to body capacitance of the NMOS is on both the terminals it is 0, it is not at all charged. The charge for that particular capacitance is 0, the voltage is actually 0. Whereas, the C_{db} here for the PMOS, the output voltage is 0, and on the other terminal it is Vdd, it is kind of completely charged to minus V_{dd} value.

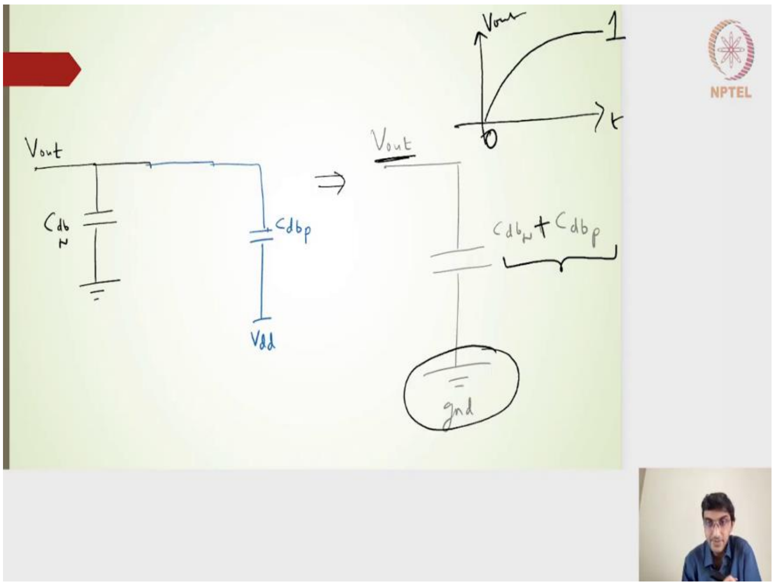
If I take a KVL, this particular V_{out} value turns out to be 0 when V_{out} was 0. Now, when we switch to 1 here, when the output voltage switches to 1 for an inverter when the output voltage switches to 1, that means that the input side there is a change from 1 to 0 at the input side, and that is why we get an output voltage change of 0 to 1.

When the output voltage changes from 0 to 1 here, this particular capacitance the C_{db} capacitances, the C_{db} capacitances this particular capacitance starts charging. This particular output voltage of course when it changes from 0 to 1 slowly it charges. My output voltage profile is going to have some kind of capacitor charging profile from 0 to 1 for this particular capacitance.

This particular capacitance when one says once the output voltage reaches to 1, that means, the V_{dd} value here. What it really happens is the capacitance is here, whatever the charge it has accumulated, it should actually go away. What it means is the voltage across the C_{db} of PMOS the drain to body PMOS should be the 0 volts.

What it really means is this particular capacitance this is actually discharging and this particular capacitance the NMOS drain to body capacitance is charging, the PMOS drain to capacitance is actually discharging, that is what is happening.

(Refer Slide Time: 20:46)



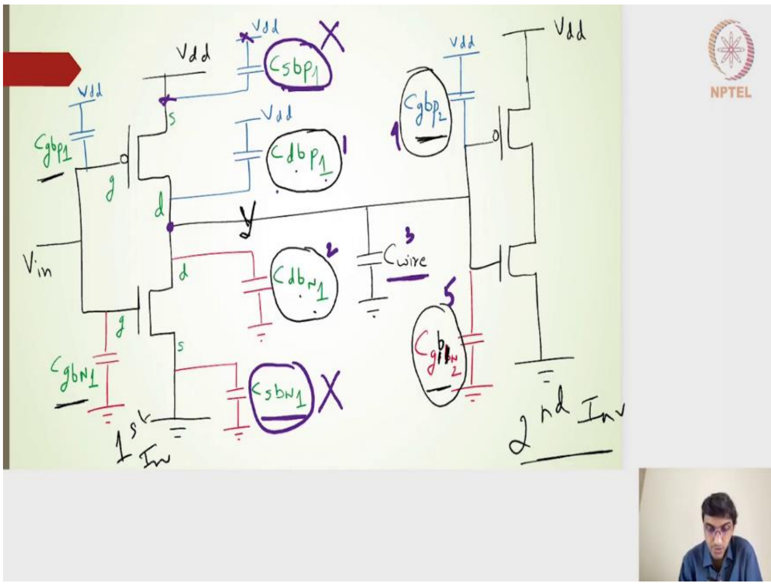
In fact, what we can say is effectively we can consider both the capacitances because both the capacitances are in parallel. If I actually take it with respect to the ground, now both the capacitances will be parallel because effectively we are actually supplying the same kind of a charges on the both the sides.

If I go back to the slide number 7, you will see that here the moment it goes higher the output voltage goes higher. It is providing the positive charges. Here also it is providing the positive charges, so that on both the sides of this capacitance it will neutralize, and then it is kind of considered that to be completely discharged.

Here also it is supplying the positive ions, and thereby we will get this capacitance to be charged. It is actually in a way both the capacitances, we can consider that it is actually charging effectively it is having the same process, one is discharging, another one is charging, but effectively the process is the same.

Output voltage is actually increasing from 0 to 1 whatever the profile may be an exponential profile or whatever the linear profile, it starts from with respect to time. It starts from 0 and then eventually reaches to a value of one volts. But effectively it is coming from both the capacitances, that is why we have drawn in equivalent capacitances of the two parallel capacitances. The equivalent of the two parallel capacitance is nothing but the summation of those two capacitances.

(Refer Slide Time: 22:28)



We come back to this particular circuit of an inverter where we notice that I can actually take capacitance of drain to body of PMOS, and drain to body capacitance of the NMOS, I can actually have an equivalent capacitances and tie it together as $C_{dbp} + C_{dbn.}$ Here one more addition is done. The output actually here, the output of this particular capacitance is connected to another inverter.

It is actually connected to another inverter where the another inverters input capacitances C_{gs} of course, this b is not there oh sorry, this s is not that this b is there C_{gbn2}, this is the another inverter, this is the second inverter, and my this is the first inverter.

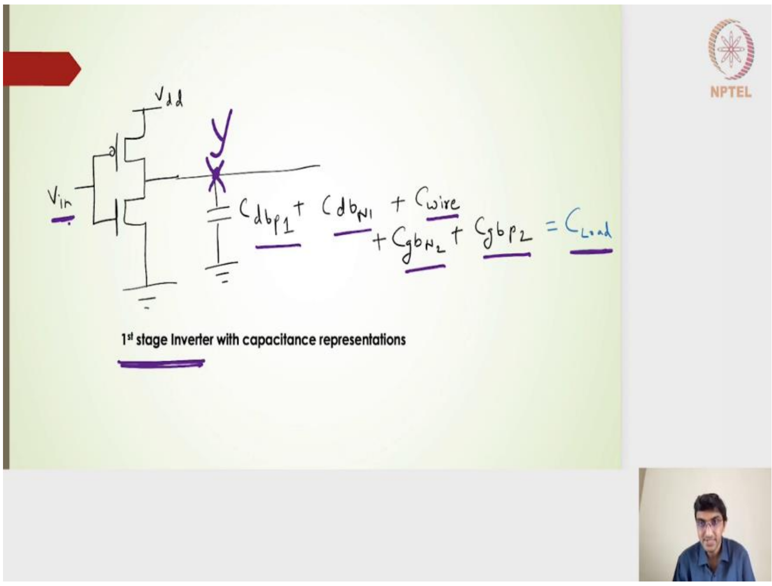
Here also we had the input capacitance of gate to body of the PMOS and NMOS, and I have written a subscript of 1 representing this is the first inverter. The second inverter, we will have the same notation C_{gbp2}, and C_{gbn2} for the second inverter.

Now, at the output y node, I now have four capacitances $C_{gdbp1,}C_{dbn1,}C_{gbp2}$ and C_{gbn2} and I have also included the wire capacitances here. A wire can always be modeled as a capacitance, assuming that there is a copper wire going along. Then it is completely the region is completely encapsulated or ensulated with the silicon dioxide.

On the bottom side, there is one more wire that is going around. We will have this as the wire is kind of modeled as a capacitances. I have also considered or accommodated in this particular C_{wire} model capacitance. Now, I have at the output node, I have this 5 capacitances. 1 capacitance, 2 capacitance, 3 capacitance, 4 capacitances, and then the 5 fifth capacitances, all put together I can actually have C_{load} capacitance at the output node.

Remember that, another information the source to body of NMOS, source to body of the PMOS are connected to the same potential. Even if the output changes here from 0 to 1 or 1 to 0, this will not have any effect. This particular capacitance, we can actually ignore this particular capacitance on both the terminals it is ground and if the output changes there would not be any effect, this also we can ignore. The 5 capacitances, I can consider at the output node.

(Refer Slide Time: 25:20)


Vdd
y
Vin
$C_{dbp1} + C_{dbN1} + C_{wire} + C_{gbN2} + C_{gbp2} = C_{Load}$
NPTEL

1st stage Inverter with capacitance representations

This is what we have. Finally, we have the first stage inverter. And if I really want to find out how much time it has taken at the output node, if given the change in the input the output changes. What is the propagation did, they falling and or the propagation delay rising, it is actually dependent on all this five capacitances 1, 2, 3, 4 and 5 which we which we put it as AC load capacitance.

This is about the first stage of course there will be a second stage also because we have drawn that. If I actually take the second stage output, if I actually probe the second stage output, I should be able to characterize the propagation delay falling and rising for the second stage as well.

But, we are at this particular point of time, we are interested in the first stage output how it is changing with respect to the change in the input voltage. Hope you know up till this point of time it is clear. What we had seen was completely a different module on the delay, we understood the definition of the propagation delay and the contamination delay falling and then rising.

Then we started looking at the inverter circuit, understood all the capacitances the input capacitances and the parasitic or the depletion capacitances and especially how the depletion capacitances for the output node can be made as an equivalent capacitances because all these capacitances are in parallel. For the transient or an AC analysis even if it is connected to the $\mathrm{V_{dd}}$, we consider that to be I mean all the DC sources are considered or represented to be the ground potential.

Finally, we have the first stage inverter even if it is even if it is cascaded to this next stage inverter, the first stage inverter output sees five capacitances altogether including the wire capacitance. What we have done is we have determined an equivalent capacitance called the $\mathrm{C_{load}}$ capacitance.