

VLSI Physical Design
Prof. Indranil Sengupta
Department of computer Science and Engineering
Indian Institute Technology, Kharagpur

Lecture – 32
Time Closure (Part 1)

So, in the last few lectures we have been talking about the clocking and the various timing issues that may arise in a design, and what are; some of the constraints like hold time, skew, setup time etcetera they have to be incorporated to have a correct operational system. Today in this week actually we shall be starting our discussion on something called timing closer.

Timing closer means you see we may be having some requirements, which may come from the specifications, that what we really want what kind of timing behaviour is desired what is required from the circuit that we are designing, but it is possible that when we have design the net list, when we are placed routed the blocks the interconnection the nets. Then it may so happen that some of those timing constraints might get violated. Timing closer broadly says that how to evaluate while in a fast way whether any of the constraints are getting violated, and if we see any violation what are the measures we can possibly take or adopt so as those violations can be avoided right. So, this is topic of our discussion next week - timing closer.

(Refer Slide Time: 01:55)



Introduction

- The layout of a chip must satisfy:
 - Geometric constraints (e.g. non-overlapping cells and routability)
 - Timing constraints of the design (e.g. setup and hold constraints)
- The optimization process that meets the above requirements and constraints is often called **timing closure**.
- Integrates placement and routing solutions with specialized methods to improve circuit performance.

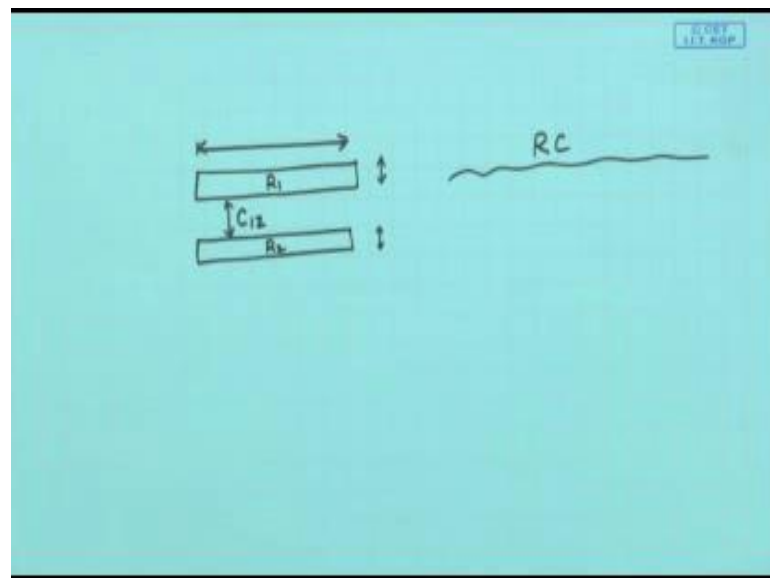
 IIT KHARAGPUR  NPTEL ONLINE CERTIFICATION COURSES



So, see, when the layout of a chip is created; so when I say layout of the chip is created it is through the entire back end design process. So, from the netlist that you get from the front end design, we can do some circuit partitioning, then we can do possibly floor planning and placement routing and then the final layout right, this is what the final layout of the chip will look like and the layout must satisfy 2 constraints. See the first is called geometric constraints, these are something which we shall be discussing in some detail later; geometric constraints says that this cells obviously have to be non overlapping, and there has to be sufficient space in between for routability; like when we say geometric constraints it is not just the cells or the blocks, it is means even it can go down to the lowest level when we consider the final layout of chip.

Now in the final layout of the chip we have the different layers as I had mentioned if you recall, so there the basic features will look like to rectangular regions, and this regions may exist in the poly silicon layer, diffusion layer, metal various layers of metal and so on. So, finally, we have some rectangular regions. So, some of the constraint that may arise at this layer may be like.

(Refer Slide Time: 03:35)



Let us say I have 2 lines may be this may represent interconnections running parallel to each other. There are 2 things what is the width of this lines? Because width of this line will determine what will be the resistance of this lines R 1 and R 2 right. Secondly, the separation between these 2 lines, the separation of these 2 lines will determine the

capacitance between these 2 running wires. So, basically as I mentioned earlier that when we have a long transmission link, the RC delay comes into the picture and there are some approximate ways to model this delays like the lmo delay model.

So, similarly here also when some wires are laid out, there can be some RC delays and this delays will be dependent up on the width of the wires of course, the lengths. So, what will be length of the wires and this separation from some other parallel wire that will determine the inter wire capacitance right fine. So, this will be the first kind of constraint the geometric constraint, and the second kind of constraint we have already talked about earlier that when you have this storage cell in the system, the flip flops or the latches. So, these storage elements will have some inherent setup and hold requirements. So, you recall what was the setup time? So whenever clock edge comes, how much before that I must make my input stable that is the setup time. And the hold times says after the clock edge, how much more time I should keep my data stable otherwise the flip flop may behave incorrectly right.

So, this setup and hold times these are characteristics of the storage elements and these are something which are not under the designers control. Designer can control a clock skew by deferent layout of the clock tree for example, right. So, the other timings you can control, but this setup hold times are something that are characteristic of this storage devices that we use in a design. So, the overall optimisation process that target to meet these requirements is broadly referred to as timing closer.

Now here we shall also see later that timing closer does not only look at a netlist and lets you whether the geometric and timing constraints have been met; well you can also have timing aware placement and routing solutions. So that you create a placement you create a routing such that your final timing constraints or targets are achieved, that is the modern you can say modern the design flow for back end design of a high performance chip or circuit.

(Refer Slide Time: 07:05)

Components of Timing Closure

1. **Timing-driven placement**
 - Minimizes signal delays when assigning locations to circuit elements.
2. **Timing-driven routing**
 - Minimizes signal delays when selecting routing topologies and specific routes.
3. **Physical synthesis**
 - Sizing transistors or gates to decrease the delay or increase the drive strength of a gate.
 - Inserting buffers into nets to decrease propagation delays.
 - Restructuring the circuit along its critical paths.

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, talking about timing closer as I said placement and routing forms the important paths, now broadly there are 3 things here first is of course, the placement; so how to place the blocks, such that the signal delays will meet the timing constraints as per our requirements. Now you see once the block placement is finalised then and then only we can make some realistic and means accurate estimate of the you can say the delay estimate like the interconnection delays, the gate delays or the cell delays and so on.

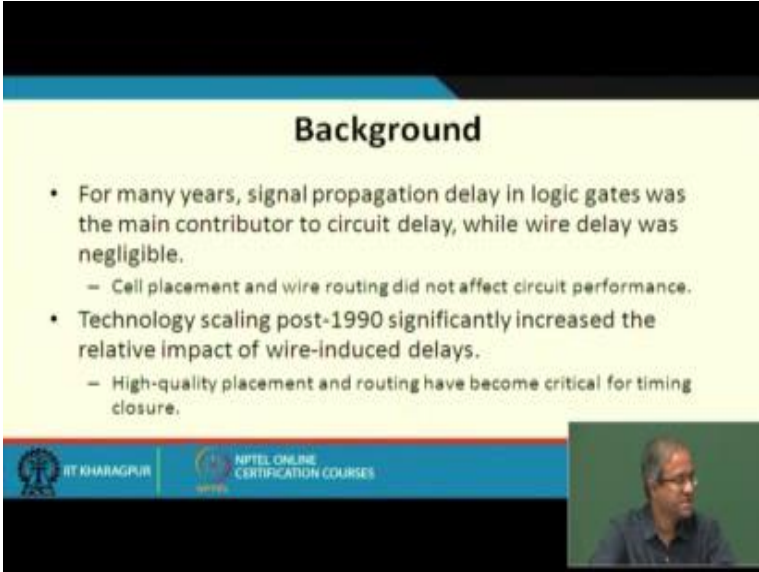
But once the positions are not fixed this delays can also vary; and just now as I had said that the delay is just not a function of how long the wire is, but also on the neighbourhood, what are the other wires which are running in parallel may be on the same layer or on different layers, there will be capacity (Refer Time: 08:13), there will be something like cross talk which also we shall be discussing later. So, all this things also have to be taken care of right. So, this is placement then we have timing driven or timing aware routing. So again how to layout the different nets using metal connections typically, so that the signal delays conform to our constraints or requirements right. What might happen is that after this placement and routing? We may find that some of the delays are still not meeting our requirements, so then we may have to revert to something called physical synthesis.

So, here what we do we apply variety of techniques; so this we shall be discussing some of this later in more detail. These techniques are used to reduce the delay of a particular

net. So, what are the broad approaches, there are several others also the main approaches have mentioned here. So, you can make a transistor or a gate bigger or smaller this is called sizing. So, if you make it bigger, the delay will decrease if you make it smaller the delay will increase.

So, depending on whether you want to decrease or increase the delay of a path, you may properly size a gate or a transistor this is the first technique and the second one we talked about when we discussed at the clock net design that whenever we have a long wire, the propagation delay may be higher. So, it is always good to insert buffers in between so as to decrease the overall propagation delay. And thirdly we may have to restructure a circuit along the critical path, this also we shall be discussing later how we can do it, but broadly these all are techniques using which we can reduce the delay of a circuit.

(Refer Slide Time: 10:26)



Background

- For many years, signal propagation delay in logic gates was the main contributor to circuit delay, while wire delay was negligible.
 - Cell placement and wire routing did not affect circuit performance.
- Technology scaling post-1990 significantly increased the relative impact of wire-induced delays.
 - High-quality placement and routing have become critical for timing closure.

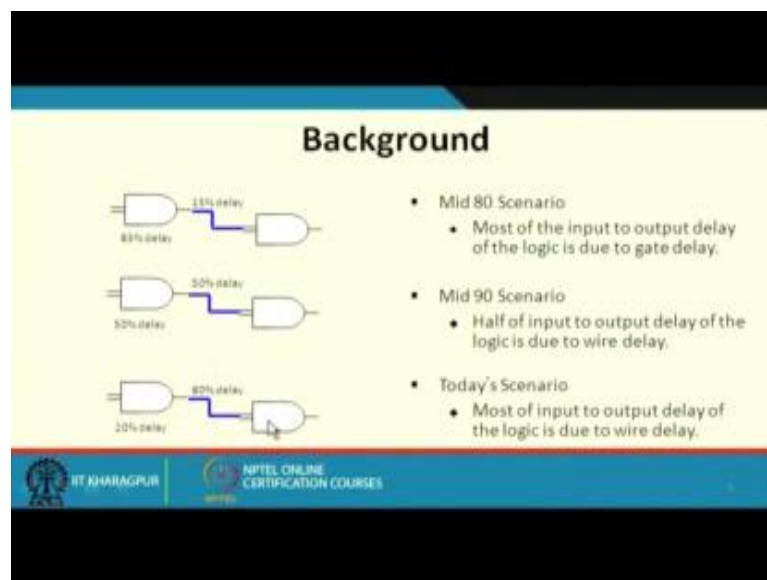
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, what is the main motivation behind this step of timing closer? Because you see in the earlier days of design we did not worry about timing that much, we were more concerned about the functional correctness of the circuit, whether the circuit is working correctly or not and mainly you can say component where the most significant amount of delays took place.

During that period we often ignored the interconnection delay; so what we are saying is that, signal propagation delay was the main contributor to the overall circuit delay; and during those period we had ignored the delay of the interconnection wires, because of

this cell placement and routing did not affect circuit performance because when we evaluated circuit performance, we only talked about the delay of the gates we did not talk about the delay of the interconnection. So, it is immaterial whether the wires are short or wires are longer, we assumed that interconnection delays are negligible as compare to the gate delay. But subsequent to 1990s when the devices started to scale down the MOS transistors, this is called technology scaling. So, here the situations started to change. So, the relative impact of the interconnection delays started to increase, and today we have a situation where the interconnection delay is becoming pre dominant as compare to the gate or cell delays. So, the next slide will give an idea.

(Refer Slide Time: 12:19)

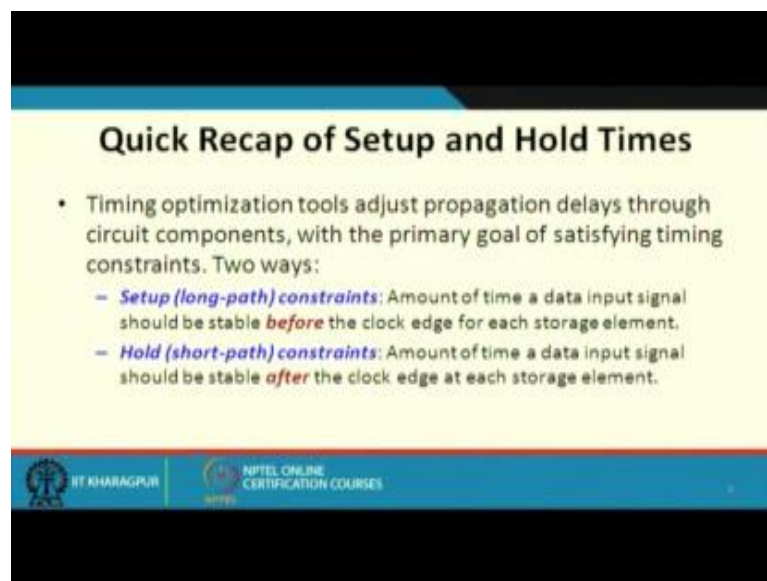


Look at this first diagram just 2 gates the output of 1 gate is driving another gate; so what we are evaluating? We are trying to evaluate that we are applying some input to the first gate, after some time after how much time the corresponding values are coming to the input of the second gate this total delay right. Now in the 80s the scenario was that most of the inputs to output delay the total delay the contributor was the logic gate. Roughly this 85 percent of delay was contributed here, and the rest 15 percent delay was contributed in the interconnections.

But in the mid 90s the interconnection delays well they started to go up because the wires become thinner and also the gates become smaller they become faster. So, now, the contribution becomes roughly 50 percent here and 50 percent there. So, half of the input

to output delay of the logic will be due to the interconnections, and half due to the gate delay. But you see today when you are in to the deep submicron technology domain; here the situation has reversed, now in the interconnections we are expecting about 80 percent delay or even more and while the gate delay is only 20 percent. So, as you can see today in the high performance circuit design or the physical design modelling of interconnection and property routing the interconnections this has become most important, this is of paramount important right. So, this is what the scenario is today.

(Refer Slide Time: 14:20)



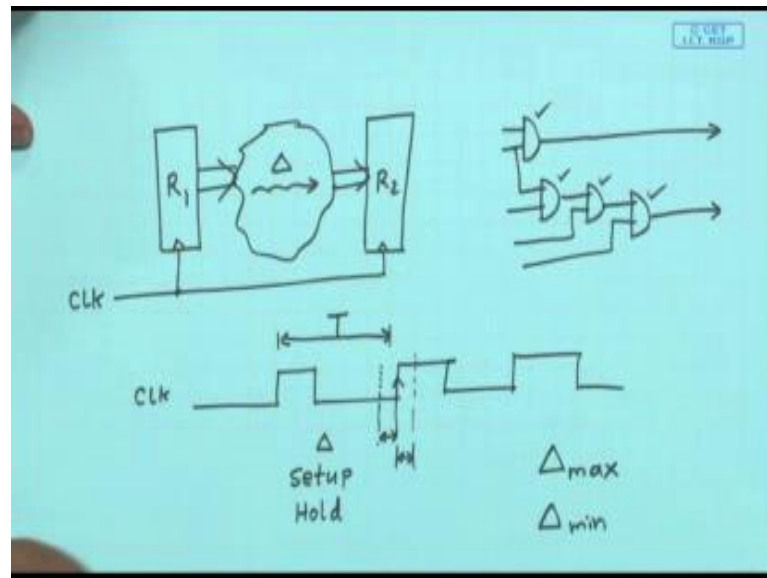
Quick Recap of Setup and Hold Times

- Timing optimization tools adjust propagation delays through circuit components, with the primary goal of satisfying timing constraints. Two ways:
 - **Setup (long-path) constraints:** Amount of time a data input signal should be stable *before* the clock edge for each storage element.
 - **Hold (short-path) constraints:** Amount of time a data input signal should be stable *after* the clock edge at each storage element.

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, let have a quick recap of the setup and hold time, because we shall be using some of the delay related assumptions in our subsequent discussions. So, what we discussed when we discussed the clocking. There we said that when we consider propagation delays through circuit components like when we say propagation delay through circuit components, we are saying that there are some storage elements.

(Refer Slide Time: 14:51)



Lets I have a setup flip flop register R 1. So, I have another setup flip flop register R 2, and there is a combination of circuit in between. So, there is a clock which is feeding both this registers. So, what we said is that the clock will be having certain time period let us say like this, the clock will come this is the clock period, let say capital T. Now what I am saying is that on which means on factor will this time key depend, well of course, it will depend on the delay of this combinational block let us call it delta. So, it will depend on delta, this will also depend on setup time, this will also depend on hold time, because see this combinational circuit will be taking sometime to compute and after that this data will be available at the input of R 2 so that I can apply the clock. But the setup constraint says that before the next clock edge comes, so I must be giving some minimum time before which the input must be stable.

So, not only this delta, I must add this setup time to this total capital T and this hold requirement says not only that even after the clock edge comes, I must maintain the input stable for some more time after that. So, my total clock period should include not only delta, but also the setup and hold times, in this way my total clock period will go up right this was the requirement. So, the timing optimization tools that we use they try to adjust the propagation delays to satisfy these 2 constraints.

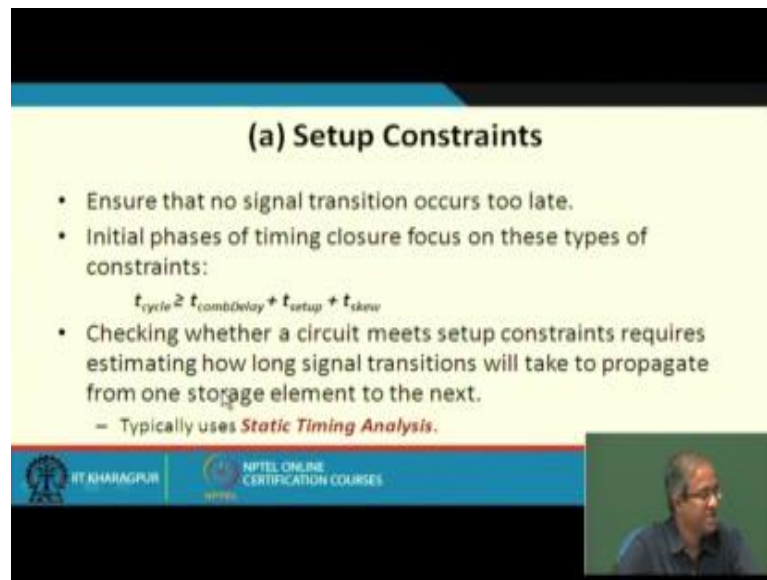
Now you see setup constraint is sometimes called long path constraint. What it says just to recall again, this is the time the data input must be stable before the clock edge, with

respect to every storage element or register. So, why do we call it a long path constraint? Because you see so here we are saying in there with respect diagram that there must be a minimum time here, before which the data must be stable. So, there must be a maximum worst case propagation delay of the combinational circuit that is why I am calling it the longest path. Because there are combinational circuit some path may be short like for example, I may be having a circuit like this, let say let say I have circuit like this there are 2 outputs, the first output is generated only after a delay of a single gate, but the second output is generated after a delay corresponding to 3 gates. So, here when I say the longest path it will be the delay of the second output equivalent 3 gates delays.

So, I must take that in to account, so what is the maximum value of delta? So, what is the value of delta after which the outputs will definitely be stable, because this will become stable much earlier just after a 1 gate delay; so delta must include the longest path right? So, whenever we are considering the setup time that is it. And hold time constraint, why it comes in? So, hold time is the time we must keep the data stable after the clock edge, this tells you that that even after the clock edge I must keep the data stable for some time why? Because if the input data changes before that, maybe the second register will be expecting the previous this need to change data in that same clock cycle instead of going into next clock cycle, because registers will take some time to respond to the clock edge; so whenever a clock edge comes, register does not respond instantaneously, it will take some time.

So, it is said that there must be a minimum timeout here this time; this is the hold time you must keep the data stable. So, that this register can faithfully record and store the input data. If the input starts changing in the mean time, then there can be wrong data getting in. This is the short path constraint is called because the short paths can create a problem here, so you will have to consider 2 cases delta max and delta min. So, when you talk about the setup time we need to consider delta max. So, what is the max maximum time that may be required for the output to become stable, and after that you have to give a time of P setup; and t hold says that not only that the minimum time of this delta, because within that same clock cycle this register 2 times therefore, I have to have this kind of a tolerance.

(Refer Slide Time: 20:52)



(a) Setup Constraints

- Ensure that no signal transition occurs too late.
- Initial phases of timing closure focus on these types of constraints:
$$t_{\text{cycle}} \geq t_{\text{combDelay}} + t_{\text{setup}} + t_{\text{skew}}$$
- Checking whether a circuit meets setup constraints requires estimating how long signal transitions will take to propagate from one storage element to the next.
 - Typically uses *Static Timing Analysis*.

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, that is why this is called long path and short path constraints; the setup and hold constraints. So, just to refresh whatever I have said, in the setup constraint we ensured that no signal transitions occurs too late and the constraints are the total cycle time or the time period must be greater than equal to the combinational circuit delay, the setup time and of course the clock skew if any. So, the initial phases of timing closure normally focuses on the setup constraints. So, what it does is that it checks whether a circuit meets the setup constraints.

So, what do we need to do this? We need to estimate the combinational delay $t_{\text{comb delay}}$ right, because setup and skew are already known I am assuming. So, this is a process we need to carry out. So, how long signal transitions will take to propagate across the combinational circuit, means from 1 storage element to the next. So, the registered are placed in stages and there are combinational circuits in between we need to have a good estimate has to what is the estimate means a delay of the combinational circuit. So, from the output of a storage element how much time you can take for a signal transition to reach the input of the next storage element, and this process is carried out through technique which is called static timing analysis. So, we shall be discussing this in more detail later.

(Refer Slide Time: 22:36)

• What is **Static Timing Analysis**?

- Propagates **actual arrival times** (AAT) and **required arrival times** (RAT) to the terminals of every gate or cell.
- Can quickly identify timing violations, and diagnose them by tracing out critical paths in the circuit that are responsible for these timing failures.
- Models propagation of signal transitions with the worst possible delay.
- Typically excludes **false paths** from the analysis.

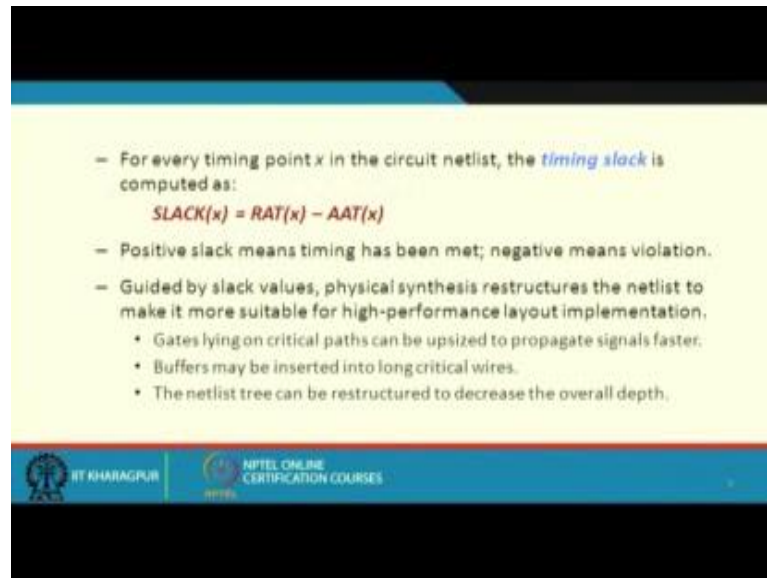
So, static timing analysis essentially it does 2 things, it starts with a given netlist of a gates or transistors. So, a given circuit netlist and it tries to calculate the actual arrival times of the signals and the required arrival times; we shall be going in to detail of this later. Required arrival times may be provided by the user with respect to the output, so we have to calculate them for every terminal or every gate. Now once you do this, so you can identify something call slack which is the difference between RAT and AAT and that can identify the timing violations. For some cell you may find that the actual arrival time is becoming greater than the required arrival time; so there is a timing violation there. So, you may have to adjust the timing in some in some way, so that those timing violations are addressed right.

So, this will see and first static timing analysis normally we model the signal transitions the delays, as the worst case values; and we shall also see there is something called false paths, which are also considered. False paths are like this, you see I have a combinational circuit. So, I am saying that for estimating the minimum clock period required or the maximum clock frequency, I will have to estimate the worst case delay, this I normally do using static timing analysis.

Now what I am saying is that the longest path in the combinational circuit need not necessarily mean the longest delay, because there are may be path, but the circuit may be such that no signal will follow through that path, we shall again see some examples later.

These paths are called false paths, there are path which geometrically you can see there is a path, but with respect to the signal value the gates will work in such a way that signal will never propagate along those long paths. So, if you can identify the false paths a priori, and remove them from a consideration, then your analysis can become much faster this is one important point.

(Refer Slide Time: 25:15)



– For every timing point x in the circuit netlist, the *timing slack* is computed as:

$$SLACK(x) = RAT(x) - AAT(x)$$

– Positive slack means timing has been met; negative means violation.

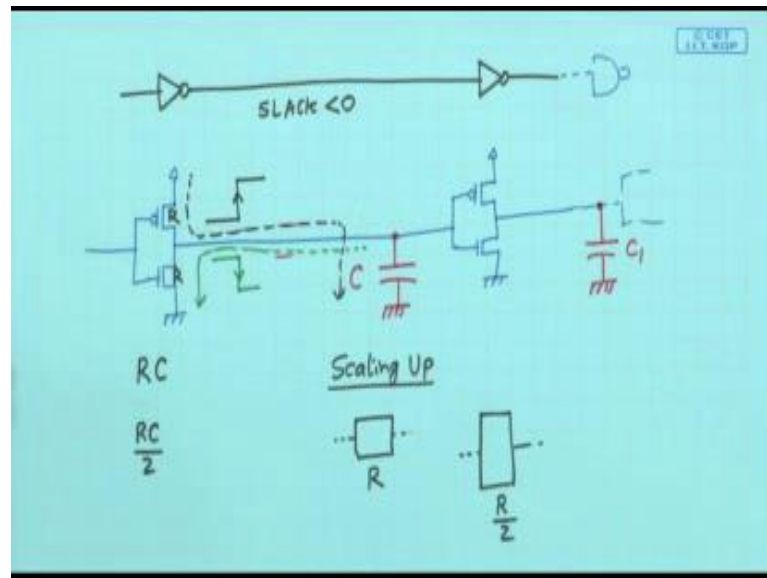
– Guided by slack values, physical synthesis restructures the netlist to make it more suitable for high-performance layout implementation.

- Gates lying on critical paths can be upsized to propagate signals faster.
- Buffers may be inserted into long critical wires.
- The netlist tree can be restructured to decrease the overall depth.

BIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, as I had said once we have calculated the required arrival time and the actual arrival times, for every point x in the circuit this may denote the input a gate, the output of a gate, some pins you can calculate this slack. So, a positive slack means your required arrival time is greater than the actual time actual arrival, which means the timing violation not there. But if rat is smaller; that means this is negative, which means there is the timing violation. Now if there is a timing violation then we have to go for physical synthesis restructuring, where we modify the netlist so that the actual arrival time can be reduced so that this slack value can be again made 0 or positive. Now there are a few ways in which you can do that I just explain a few of them, the first technique what is maintain here is that gates lying on the critical path can be upsized to propagate signals faster. So, what I mean is something like this.

(Refer Slide Time: 26:34)



Let say there is gate here say an inverter, there is another gate here. Suppose this is the path which is creating the problem, for which the slack value is negative right. So, we are trying to reduce the delay of this path. So, how we can do this let us see, this inverter if you look at the CMOS level realization of an inverter, just it is a p trip PMOS transistor this is an NMOS. So, the input is coming here, the output is going from here similarly you have another such. So, your circuit is like this, may be this is driving some other gate right some other gate it is driving, so it would be driving something else.

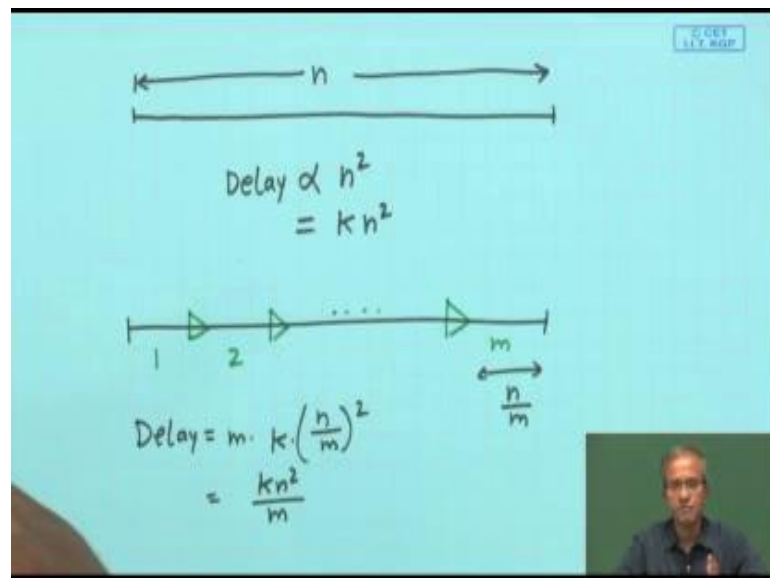
So, the way we model this is that you see these are interconnection line, there will be some parallel lines here, there will be capacitances there will also be gate capacitance because it is ultimately driving the gate. So, if you thing about the gate MOS transistor there is a that looks like a parallel plate capacitor, gate on top and the substrate at bottom substrate is normally grounded. So, essentially what you can model is that as if this is driving a capacitor a capacitive loads, similarly when I say this will is driving a some other gate, so here also there will be some capacity blow let us call it C_1 right. So, when I say that the signal at the output of a gate is going from 0 to 1, let say what does this mean? This means the capacitor is charging from 0 to high voltage via this path.

So, this p type transistor is coming in to the picture, and similarly when I say that the output of this gate is going from 1 to 0, which means the capacitor was already in the charge state is 1 high, it is discharging through this path. So, now, this discharging path

will be like this. So, either the PMOS transistor or the NMOS transistor will be conducting, so if I just assume for the sake of argument, the resistance of these 2 transistors when in the conducting state they are equal, so the charge discharging delay will be RC product of that. Now in this method what we say that we are scaling up a transistor, scaling up means let say earlier my channel or the gate which was above the channel was like this, this is 1 terminal this is 1 terminal. Now what I make I make it wider say by 2 times. So if I make it wider what will happen? If the resistance of this was R this resistance will become R by 2.

So, if I scale up the transistor make them bigger let us say R by 2, R by 2 then my delay will become R C by 2. Similarly if I want to increase the delay I can make it narrower I can make it 2 R. So, just by scaling the transistor we can have a very nice way of controlling the delays of the different segments of the wires right; this is exactly what we mean by the first point. Gates lying on the critical paths can be upsized to propagate signal faster, the second point I talked about while discussing clocks let me again look at it once more, buffers may be inserted into long critical wires.

(Refer Slide Time: 30:44)

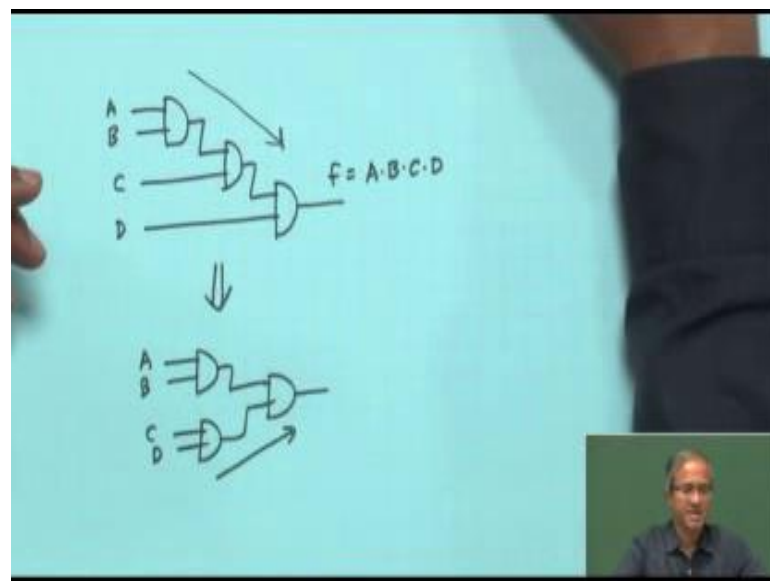


Let say there is a long wire of length n ; with respect to the lmo delay model if we consider distributed resistances and capacitances, the delay is propositional to the square of the length. So, let say it is equal to some constant let say K into n square. So, what we are saying is that to speed up the circuit what the proposal is that this was the hold line,

now you insert some buffers in between. So, break it up into some segments, let us say there are m segments 1 2 up to m . So now, delay of every segment will be how much? Total length was n , so width of each segment will be n divide by m because there are m such pieces I have made. So, delay of each segment will be K into n by m whole square and there are m such segments so multiplied by m , this will be the delay here.

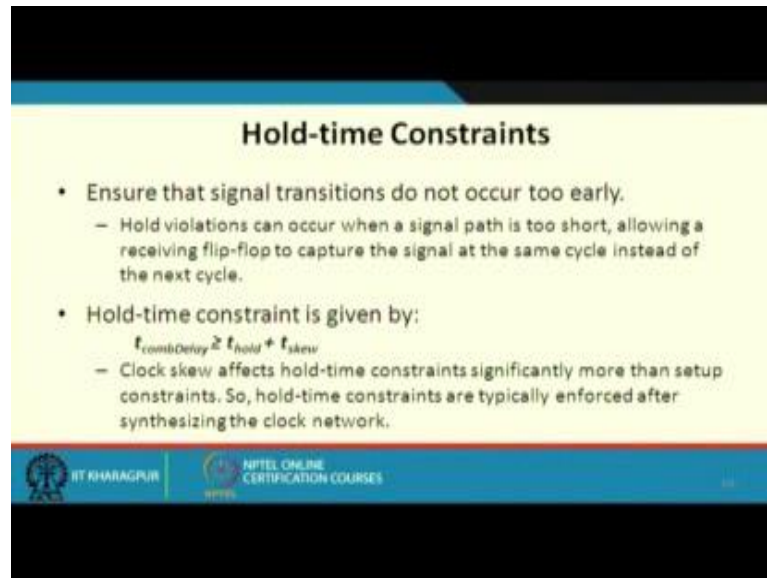
So, if you make a short calculation this becomes $K n^2$ divide by m . So, roughly the delay decreases m time's right. So, you see if you insert buffers in between these a nice way to reduce the delay of course, the price we have paying is some additional area for the buffers right this is of course, trade our design trade of your making, you are investing a little bit of hardware to make this circuit faster and thirdly of course, the netlist tree you can restructure to decrease the overall depth, this is very simple I am giving a very simple example.

(Refer Slide Time: 32:49)



Suppose you have a netlist like this, just a 4 input nand function say if the inputs are A B C and D, the output produces is A B C D. So, you can basically restructure it like this these 2 are functionally equivalent. So, here the delay was 3, now the delay was 2; 2 levels of gates. So, this kind of restructuring of netlist is also a very simple way to reduce the delay of a circuit; now we shall see other method also later, so how we can enhance the delay of the circuit that we shall be discuss in later.

(Refer Slide Time: 33:50)



Hold-time Constraints

- Ensure that signal transitions do not occur too early.
 - Hold violations can occur when a signal path is too short, allowing a receiving flip-flop to capture the signal at the same cycle instead of the next cycle.
- Hold-time constraint is given by:
 - $$t_{\text{combDelay}} \geq t_{\text{hold}} + t_{\text{skew}}$$
 - Clock skew affects hold-time constraints significantly more than setup constraints. So, hold-time constraints are typically enforced after synthesizing the clock network.

BIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And with respect to the hold time constraints, this I mention that these are the early signal transitions that create a problem. So, for hold time constraints this also we discussed earlier, that to address the hold time constraint that means, your within the permissible limit you must ensure that your combination circuit delay must always be greater than equal to hold plus skew time, and this combination delay now is the minimum possibly not the maximum possibly this is greater than equal to. So, the minimum possible delay must be greater than equal to time this must be ensured.

So, with this we come to the end of this lecture, we just continue with our discussion in the next lecture.

Thank you.