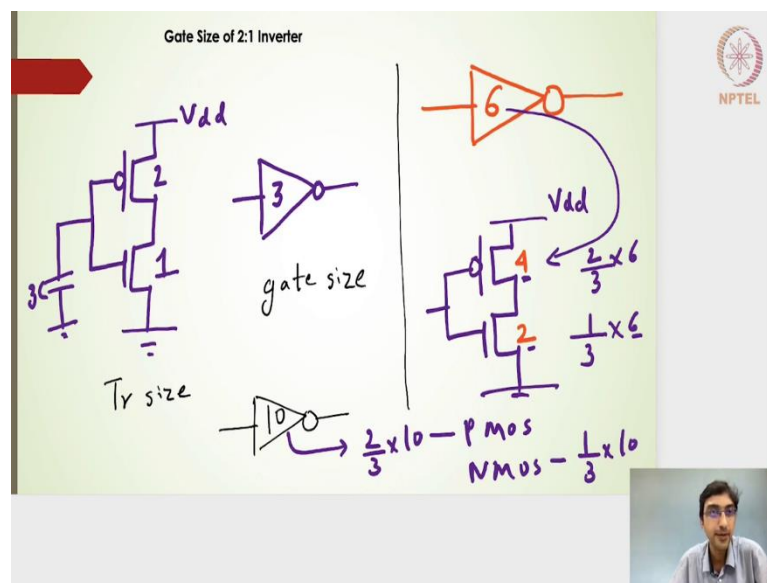


Design and Analysis of VLSI Subsystems
Dr. Madhav Rao
Department of Electronics and Communication Engineering
Indian Institute of Information Technology, Bangalore

Lecture - 34
Optimizing Gate Size

Hello students, welcome to this lecture on optimizing the sizes for a single path digital circuit. In a digital circuit we might have multiple paths, that is something we will analyze those circuits and then components of the size of those circuits later on. But to begin with, I think we will do a single path digital circuit and then try to optimize the sizes of those gates as well as the transistors to obtain a minimum delay. This lecture will talk about, how do we go about optimizing the sizes. Let us proceed or begin further.

(Refer Slide Time: 00:57)



I need to draw a gate size of 2:1 inverter. Let us say that, I have an inverter. I have a PMOS and then an NMOS circuit. I am going to draw those, V_{dd} and then the ground rail and then I have a bubbled gate which represents the PMOS circuit. Suppose this is the size of a 2:1 inverter, which is generally our benchmark inverter, what it means is at the input side, I will have a capacitance of $3C$. This particular inverter, which is of 2:1 is actually represented in the form of the size of 3. I am going to write it as a gate size, a representation says that it is an inverter, and then the size is 3 and then this is the transistor level sizing which is 2:1.

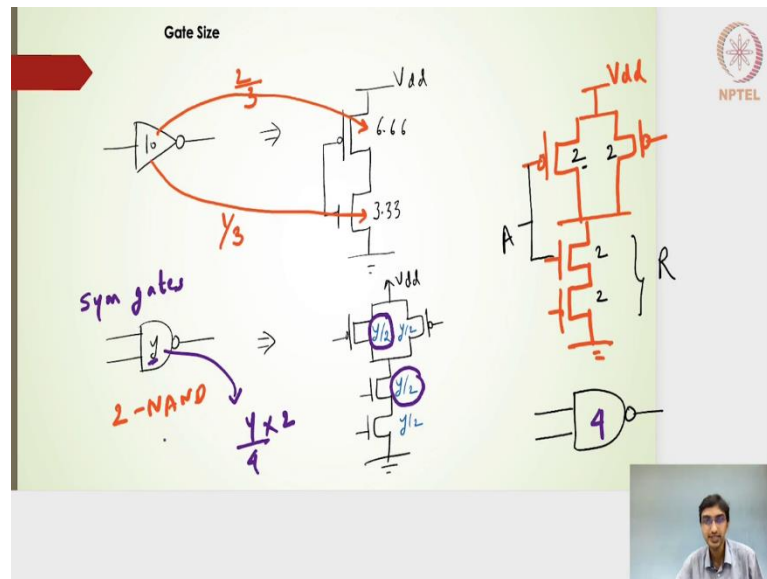
This is how we will represent a gate size from the transistor. Let me say that I have an inverter here and its gate size is 6. What it represents is, on the transistor level. Let me draw the transistor level PMOS and NMOS, that is the V_{dd} and then the ground rail and then this is the PMOS transistor and then the NMOS transistor, it sees a capacitance of $6C$ at the input side.

We know that in a long channel current model, we need twice the size of the PMOS transistor as that of the NMOS, so that to leverage the beta values to be same. That I will get the same current, rising current and the falling current which will give me the falling resistance and the rising resistance to be same. In that sense, what we want is from this 6 size or whatever the $6C$ capacitance, we need to accommodate $\frac{2}{3} \times 6 = 4$. That will be the size of 4 which will go to the PMOS side and then the size of 2 will go to the NMOS side.

What it means is, I will do is $\frac{1}{3} \times 6$ will go towards the NMOS side and then $\frac{2}{3} \times 6$ value will go towards the PMOS side and that is how I represent the size of 4:2. In this case, I have a different size, let me take up an inverter let us say it has a size of 10, what it means is, I will need $\frac{2}{3} \times 10$ on the PMOS side.

This will go to the PMOS side, the size the PMOS transistor width will be $\frac{2}{3} \times 10$ and NMOS will have $\frac{1}{3} \times 10$. Which will be nothing but 6.66 and 3.33 respectively for PMOS and NMOS. I hope this is clear. This particular notation of the gate size and then from the gate size allocating the necessary or the required widths to the PMOS and NMOS transistor is what we are going to follow from here on.

(Refer Slide Time: 04:44)



Moving further. Let us say that if I have a size of 10 here, that is what I had put it previously, size of 10 here. $\frac{2}{3}$ parts will go to the PMOS side and then the $\frac{1}{3}$ parts will go to the NMOS side. Because I want 2:1 ratio, so that in a long channel model, I will get the same current the rising current as well as the falling current.

That is the reason why we want the 2 parts, the double the parts, to be going towards the PMOS side and then single parts to be going towards the NMOS side. If I draw the 2 input NAND gate, this is the series on the pull down side and then I will have it 2 transistors which are in parallel on the PMOS side.

This will be my V_{dd} and then the other rail will be ground. PMOS transistor I will just represent it in a bubbled transistor form, so as to put the width. What we have seen is, we have used 2:2 here, 2:2 here.

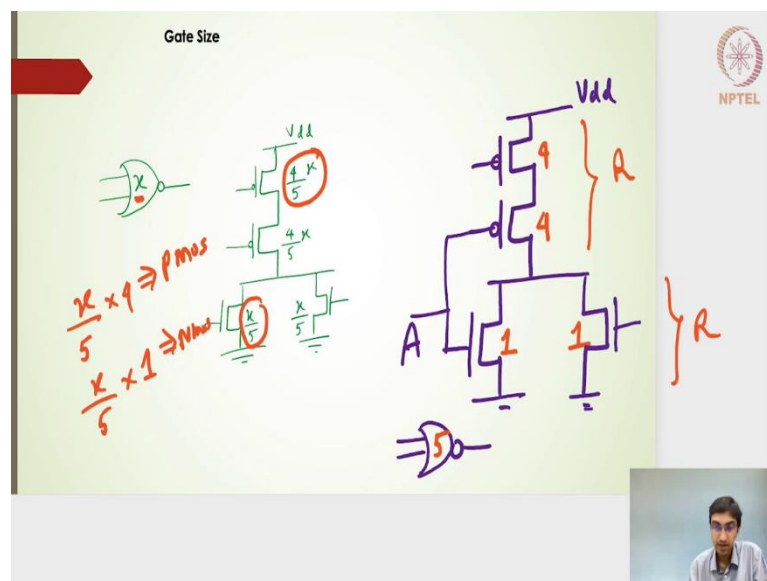
This 2 will give me the falling resistance of R, matches with that of the 2:1 inverters, the falling resistance. Under worst case condition either of these transistors will be on and thereby will give me the rising resistance of R, which matches with that of the 2:1 inverters rising resistance.

Here the input side, if I pick up this being my transistor A or the input A which goes to this particular transistor and then this particular transistor. If I see that, the size of the input size of the 2 input NAND gate is nothing but 4. Where 2 parts are going to the PMOS and then 2 parts are allocated to the NMOS, equal parts.

If I have a size of y here, total should be y here, that means that out of 4 parts, 2 parts of the y will go to the PMOS 1 and then out of the 4 parts, 2 parts will go to the NMOS 1. That is the reason why we have done the 4 parts and then 2 parts will be going towards the NMOS and then 2 parts will be going towards the PMOS. If I create 4 parts, $\frac{y}{4} \times 2$, turns out to be $\frac{y}{2}$ and that is what we have it here, $\frac{y}{2}$ here and then the $\frac{y}{2}$ here.

It is basically creating 4 parts and then 2 parts allocated to PMOS and NMOS respectively and again, we are at this point of time we are assuming symmetric gates. Let me write it down symmetric gates what it means is, both the inputs sees the same gate size. Both the inputs sees the same capacitance, input capacitance. Later on in the future slides, we will see the asymmetric gates, but at this point of time, we are assuming that the gates are all symmetric. All the input sees a same capacitance, hope this is clear, for a 2 input NAND gate.

(Refer Slide Time: 07:58)



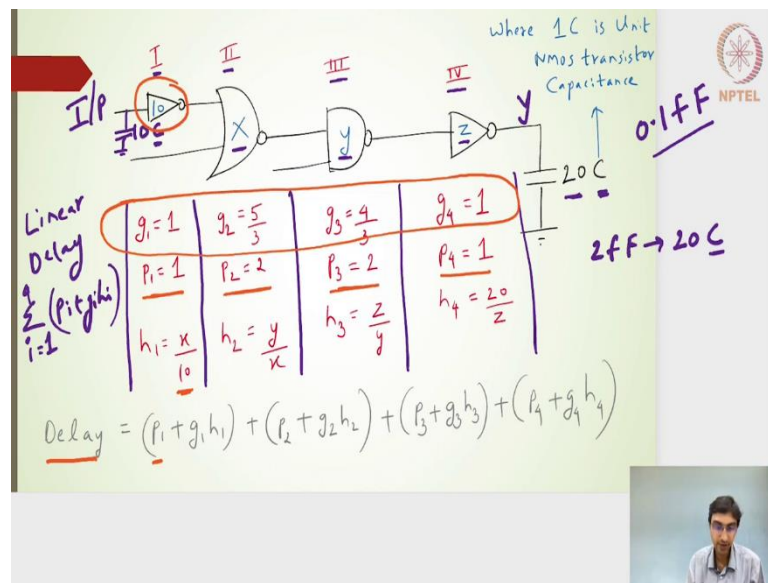
Moving further, 2 input NOR gate 2 input NOR gate again, 2 of the transistors on the pull up side to be in series and then 2 of them will be in parallel on the pull down side. This is my V_{dd} and this is my ground and this is my PMOS, this is my PMOS.

Let me put up the size here, I need a size of 4 here, on the 2 input NOR gate side and a size of 1 here. So, that the equivalent rising resistance is R matches with that of the 2:1 inverter and then on the worst case condition, 2 of the parallel NMOS transistors will give

me a falling resistance of R. What we will do is, for each of this inputs, if I connect these 2 inputs together, this will be my the first input A and the other one will be the B. So, the gate size put together it will be 5. This will be represented as a NOR gate of size 5, where 4 parts are going to the PMOS and then 1 part is allocated to the NMOS.

Given an x size here, we will create 5 parts and then 4 parts will be going to the PMOS side and the remaining 1 part, $\frac{x}{5} \times 1$ will be going towards the NMOS side. That is what we have $\frac{x}{5}$ here and then $\frac{4x}{5}$ and again it is a symmetrical gate. Both the inputs should see the same capacitance of x. So, $\frac{4x}{5}$ will be given to the other input PMOS transistor and then $\frac{x}{5}$ will be given to the other inputs NMOS transistor, put together the overall capacitance seen at the input side will be x, hope this is clear.

(Refer Slide Time: 10:07)



Moving forward, let us take this particular single path example. I have now 4 gates, this is my input this is my output. I am going to represent it as y and then I have 4 gates, the 1st one is the inverter, the 2nd one is a 2 input NOR gate, 3rd one is the 2 input NAND gate and then the fourth one is the inverter again, which is connected to a load capacitance of $20C$.

Where C represents $1C$ unit NMOS capacitance. If I choose a unit NMOS transistor, the width will be $100nm$. Whatever its parasitic capacitance, that will be the $1C$. If we know the unit transistor its capacitance should be nothing but close to $1.1fF$.

We have converted this load capacitance into 20 multiplied by the unit NMOS transistor capacitance. Whatever is the capacitance, if it is actually 2fF we are going to represent it as 20C, where C represents the unit NMOS transistors, capacitance which is nothing but 0.1fF. Hope this is clear. Then we have allocated the size of x, y and z in this particular gates.

At the input side the inverter has a overall gate size as 10, that means in this particular input, it sees a capacitance of 10C, where C represents nothing but the same unit NMOS transistor capacitance. Now, that is why we have written down everything in the form of a unit NMOS transistor, because when we find out what is x, y and z we can easily relate it to what should be the transistors capacitance.

What we are going to do now is, we are going to write an expression for the delay and what we ultimately want to do is find out this x value, find out this y value and find out this z value, my overall delay for this particular circuit will be minimum. The intention of this particular task is to find an optimum value of x, y and z such that, my overall delay will be very very less should be minimum.

For this particular path from input to output, we are going to use a linear delay model. I need to know individual stages, whereas the individual stages I have written it as 1 stage, 2nd stage, 3rd stage and 4th stage, individual stages the logical effort which is nothing but g and then the parasitic normalized parasitic p and then the h value which is electrical effort or the fanout h.

Then what we can do is, we can do a summation of all the 4 stages, $\sum_{i=1}^4 p_i + g_i h_i$ which will give me the overall delay. That is why, I have written this g value, individual stages g, p and h values, individual stages g, p and h values, individual stages g, p and h values. Now let us try to understand what is this g value and p value and h value, is it correct?

For an inverter the logical effort is nothing but 1. I have written it as 1, for a 2 input NOR gate the logical effort is 5/3, for a 2 input NAND gate, the logical effort is 4/3 and for an inverter again it is nothing but 1. This particular portion is, that is something we have taken it or we had seen that earlier. This is done the normalized parasitic for the NOR and the NAND gates is equal to the number of inputs.

If it is 2 input NOR gate, we will have the normalized parasitic as 2, if we have a 2 input NAND gate, the normalized parasitic is 2, for inverter the normalized parasitic value is nothing but 1. Finally, the electrical effort for the individual stages, the electrical effort as per the definition it is nothing but the input capacitance divided by whatever it sees the input capacitance here, on this particular first stage if I consider, it is nothing but the input capacitance which is loaded the input capacitance of the second stage which is loaded into the first stage. That will be nothing but $\frac{xC}{10C}$, it will be $\frac{x}{10}$.

Similarly, the h value for the 2nd stage will be nothing but the input capacitance of the 3rd stage which is loaded into the 2nd stage, it will be $\frac{yC}{xC}$. So, $\frac{y}{x}$ and for the 3rd stage it will be $\frac{zC}{yC}$. It will be $\frac{z}{y}$ and then the last stage the 4th stage it will be $\frac{20}{z}$.

Hope you know the g, p and h of the 4 stages are individually validated and verified. What should be the overall delay for this 4 stages, it will be nothing but the individual stages $p + gh$ and then the summation of all the 4 stages will give me the overall delay. That is what I have written here,

$$\text{Delay} = (p_1 + g_1h_1) + (p_2 + g_2h_2) + (p_3 + g_3h_3) + (p_4 + g_4h_4)$$

If I know the value of x, y and z I should be able to find out the overall delay for this particular 4 stages, although it will be a normalized delay, but that is perfectly fine. Because, once I get the normalized delay, multiplied by $3RC$ will give me the absolute delay of any critical path. Hope this is clear.

(Refer Slide Time: 16:14)

Delay = $\left(1 + \frac{x}{10}\right) + \left(2 + \frac{5y}{3x}\right) + \left(2 + \frac{4z}{3y}\right) + \left(1 + \frac{20}{z}\right)$

Delay = $6 + \frac{x}{10} + \frac{5y}{3x} + \frac{4z}{3y} + \frac{20}{z}$ → Sum

Moving forward, I have this particular critical path circuit and then the delay of this is written here.

$$\text{Delay} = \left(1 + \frac{x}{10}\right) + \left(2 + \frac{5y}{3x}\right) + \left(2 + \frac{4z}{3y}\right) + \left(1 + \frac{20}{z}\right)$$

Overall delay is nothing but,

$$\text{Delay} = 6 + \frac{x}{10} + \frac{5y}{3x} + \frac{4z}{3y} + \frac{20}{z}$$

(Refer Slide Time: 17:45)

Minimum Delay

I II III IV

Delay = 6 + \sum g_i h_i

Delay = 6 + \frac{x}{10} + \frac{5y}{3x} + \frac{4z}{3y} + \frac{20}{z}

g_i h_i = a \times b \times c \times d = a

I am going to concentrate on the two aspect. Delay what we had got was

$$\text{Delay} = 6 + \sum_i^4 g_i h_i$$

$$\text{Delay} = 6 + \frac{x}{10} + \frac{5y}{3x} + \frac{4z}{3y} + \frac{20}{z}$$

If I concentrate on this particular portion, what it see this 6 is a constant, we cannot vary this because 6 is already a value that has been given. If I somehow vary this x value, let us say if I increase this x value. I will have an increased value here, but I will have a overall $\frac{y}{x}$ value is going to decrease. Overall $\frac{5y}{3x}$ is going to decrease. We do not know what values of x we will get an optimized summation results, if I increase an x or should I decrease an x. If I decrease the x, this particular quantity will decrease, but this quantity will increase. Similarly, if I increase the y here, this particular component will decrease, however this particular $\frac{5y}{3x}$ is going to increase.

Similarly, if I decrease or increase the z value, this particular component will also be changing. There lies an optimum value of x, y and z such that we should be able to get a minimum value of this particular summation.

(Refer Slide Time: 20:14)

The slide contains the following handwritten content:

- Equation: $F = \prod_i g_i h_i = g_1 h_1 \cdot g_2 h_2 \cdot g_3 h_3 \cdot g_4 h_4$
- Equation: $= 1 \cdot \frac{x}{10} \cdot \frac{5y}{3x} \cdot \frac{4z}{3y} \cdot \frac{1 \cdot 20}{z}$
- Equation: $F = \frac{400}{90} = \frac{40}{9}$ (labeled as Product)
- Text: "Path Effort" with an arrow pointing to the F symbol.
- Text: "Sum is unknown", "Product is known"
- Text: "To minimize the Sum of n^{th} term, where Product is known, then the individual term should be $(1/n)^{\text{th}}$ root of the product."
- Equation: $a \cdot b \cdot c \cdot d = \frac{40}{9}$, $a = \left(\frac{40}{9}\right)^{\frac{1}{4}} = b = c = d$
- Equation: $a + b + c + d = ?$
- Text: $a \cdot b \cdot c \cdot d = 16$, $a = 16^{\frac{1}{4}} = 2$, $a + b + c + d = 8$

That is what we need to find out. Let us say, I have this particular value as a 1 into $\frac{x}{10}$ as a, $\frac{5y}{3x}$ as b, this is my b value and $\frac{4z}{3y}$ as c and $\frac{20}{z}$ as d value. Now, if I look closely into this and if I do a product of a, b, c, d that means I am going to do a product. Product is nothing but given by this particular sign.

The product of all the inputs, I am going to write it as product of $g_i h_i$ which is nothing but a x b x c x d. If I do that, I am going to have a product which is free of the variables x, y, z. And the reason is very simple, If I look into each of these variables and if I multiply this, it will get cancelled.

The x variable will get cancelled with that of the b variable. If I multiply ab, I will get the x variable will be cancelled out. If I multiply ab you know bc, I know that the y variable will get cancelled. If I multiply c and d I know that the z variable will get cancelled. I will get a definite product value is what I will get.

That is what I am going to show you in the next slide. What I have done is, I have taken the product of this a, b, c, d, I have also written it in the form of $g_i h_i$, remember that $g_i h_i$ is the individual effort delays. Let me write it down here, individual effort delay and the product of the individual stage effort delay will be nothing but $g_i h_i$ multiplied for all the 4 stages.

$$F = \prod_i^4 g_i h_i = g_1 h_1 \cdot g_2 h_2 \cdot g_3 h_3 \cdot g_4 h_4$$

$$= 1 \frac{x}{10} \frac{5}{3} \frac{y}{x} \frac{4}{3} \frac{z}{y} \frac{20}{z}$$

$$F = \frac{400}{90} = \frac{40}{9}$$

There is a new term, there is a new definition for this called the path effort. If I do a product of all the individual stage effort delays, What is an effort delay? It is nothing but $g_i h_i$, that is an individual stage effort delays. If I do a $g_i h_i$ and then multiply with all the stages, I will get an definite value and that we will call it as an F value which is nothing but the path effort.

For that particular path, from the input to output how many number of stages is there, , that many number of stages if I pick the electrical the effort delays. If I pick the effort delays and then multiply those effort delays, I will get the path effort. The path effort is nothing but the product of all stages effort delay and will give me a definite value.

It will give me a value of $\frac{40}{9}$. What we have got is, we do not know the value of x, y, z but the product of all those terms we got it as $\frac{40}{9}$. If I have $a+b+c+d$, this particular value I do not know, but if the product of a, b, c, d if I can find out, if it is $\frac{40}{9}$ in this particular case.

What should be that minimum a value? What should be that minimum b value? What should be that minimum c value? and what should be that minimum d value? Is the question, if I have the minimum value of a, b, c, d, I will get the summation of that to be a minimum value. In fact, I should rephrase this.

If I want to know, what should be the minimum of the sum of a, b, c, d given the product of a, b, c, d, that is what we need to calculate. What should be that minimum sum of this a, b, c, d given the product of a, b, c, d as $\frac{40}{9}$. It turns out that, if all a, b and c and d all have the same values, $a = b = c = d$, then I will get the summation of this a, b, c and d to be the minimum value.

In that case if $a = b = c = d = \frac{40}{9}$, then if all of them has to be equal, then what I will do is a is equal to nothing but $\frac{40}{9}$ and then to the $\left(\frac{40}{9}\right)^{1/4}$. Then, I can easily say that $a + b + c + d$ is minimum and you can take any kind of an example here. Let us say that, the product of a, b, c, d is actually 16.

In that case $a b c d = 16$. In that case, my $a = 16^{1/4}$ turns out to be nothing but the value of 2. The summation if I use $a = b = c = d = 2$, then only I will get the $a + b + c + d = 8$ which is the minimum one. I can always choose a value of a or b or c or d as nothing but different value than 2.

Let us say that if I choose a value of 4 for a, instead of a 2 value if I choose a 4 and then if I choose a b value of 2 and if I choose a c value of 2 and then d value as 1, my product will be 16, that that would not change. But if I do a summation of that $4 + 2 + 2 + 1 = 9$, which is larger than that value of 8.

what we are saying is, if the product value is known and if I can actually take out, the root of that, whatever you know, if there are 4 variables unknowns, then it will be having 1 by 4th root will be equal to that of the individual variables. The summation of those, will be the minimum. In this particular case, $\frac{40}{9}$ if I do a 1 by 4th root, I will be able to find out the individual $g_1 h_1 \cdot g_2 h_2 \cdot g_3 h_3 \cdot g_4 h_4$ allocated to $\frac{40}{9}$ and then the 1 by 4th root of that.

To generalize this, what I have stated here is to minimize the sum of nth terms. In this case I have 4 terms, because there are 4 stages, but if there are n stages and if I want to minimize the sum of those n stages, where the product is known then the individual term should be 1 by nth root of the product. This is kind of a very general statement applicable in the linear delay model for the n stages.

Tomorrow, I have a 6 number of stages, then I am going to find out the path effort which is nothing but the multiplication of all the individual stages effort delay and then do 1 by 6th root for that particular f value and then find out the individual effort delays values. Hope this is clear.

(Refer Slide Time: 28:30)

Handwritten notes on a green background:

- Left side: $f_1 = f_1 = g_1 h_1$, $f_2 = g_2 h_2$, $f_3 = g_3 h_3$, $f_4 = g_4 h_4$
- Right side: $f_i = g_i h_i = \left(\frac{40}{9}\right)^{1/4} = 1.4519$
- Center: $\text{delay}_{\min} = \sum p_i + N \frac{F^N}{N}$
- Bottom left: $g_4 h_4 = 1.4519$
- Bottom left calculations: $1.20 = 1.4519 \Rightarrow z = 13.77$, $\frac{4}{3} z = 1.4519 \Rightarrow y = 12.65$, $\frac{5}{3} y = 1.4519 \Rightarrow x = 14.52$
- Right side: $F = \prod_{i=1}^4 g_i h_i$
- Bottom right: $\text{delay} = \sum p_i + N \frac{F^N}{N}$

My individual f_i value f_1, f_2 and f_3 I know that it will be nothing but $\left(\frac{40}{9}\right)^{1/4}$ which will be nothing but 1.4519. If I actually calculate, 1.4519 to individual $g_4 h_4$ and then $g_3 h_3, g_2 h_2$ and $g_1 h_1$, I should be able to find out the individual variables of x, y and z . It turns out to be nothing but 14.52, 12.65, 13.77 which will give us the summation of the $g_i h_i + 6$ value which is coming from the normalized parasitic delay to be minimum. What we have done is, we have taken the F value which is nothing but the product of the individual effort delays for the 4 stages.

The overall minimum delay is nothing but,

$$\text{delay} = \sum p_i + NF^{1/4}$$

What we have done earlier was, F_i if I make all the stages see the same effort delay, then we know that the summation of that will be minimum. Now if all the effort delays are same then 4 multiplied by the effort delays will give me the overall value plus the parasitic of 6 which we have calculated earlier should give me the overall delay, that is what I have written here. The overall delay minimum delay in this particular case for this particular single stage circuit is,

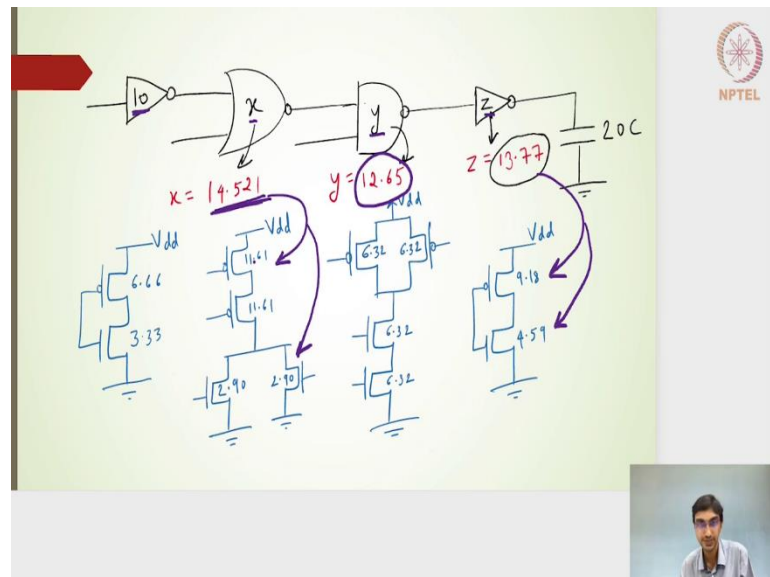
$$\text{delay}_{\min} = \sum p_i + NF^{1/N}$$

where n is the number of stages and $F^{1/N}$ if I have n such stages.

$F^{1/N}$ will give me the individual stage effort delay and then multiplied by n will give me the overall summation of all the n stages effort delay. Alright, what it also says is, without even finding out the gate sizes, without even identifying what is x , what is y , what is z that is the gate sizes we can actually estimate the minimum delay based on the number of stages, which is a very crucial component we are not identifying what is x , y and z .

But before that, if individual stage c is the same effort delay then I can easily find out what is the minimum delay, no need to calculate the x , y and z values to find out what is the minimum delay of that particular path, hope you have realized that, moving ahead.

(Refer Slide Time: 31:54)



If I have this x value calculated, if I have this y value calculated, if I have this z value calculated, for an inverter we know that 2:1 inverter I need to make this 3 parts of this where the 2 parts will go to the PMOS side and then 1 part will go to the NMOS side, that is what I have done 9.18 will be the size of the PMOS and then 4.59 will be the size of the NMOS.

Similarly, for a 2 input NAND gate we will make 4 parts, 2 parts will go to the PMOS and then 2 parts will go to the NMOS and for the 2 input NOR gate we will make 5 parts out of that 4 parts will go to the PMOS and then 1 part will go to the NMOS. Lastly, this size of 10 is nothing but 2 parts will go to the PMOS and then 1 part will go to the NMOS. We

will make 3 parts and 2 and 1 parts of it which will be allocated to PMOS and NMOS respectively.

Hope this sizing is clear now. The sizing of x, y and z which is the gate sizing and how it has been distributed appropriately to its respective NMOS and PMOS transistors, hope this is clear.

(Refer Slide Time: 33:10)

The image shows a handwritten derivation on a green background. At the top right is the NPTEL logo. The equations are as follows:

$$\text{delay}_{\min} = \sum p_i + N F^{1/N}$$

$$= 6 + 4 F^{1/4}$$

$$= 6 + 4(1.4519)$$

$$\text{delay}_{\min} = 11.8016$$

Below this, it says: "If $R C = 1 \text{ ps}$ ", then:

$$\text{Delay}_{\min} = 11.8016 \times 3RC$$

$$\text{Delay} = 35.4 \text{ ps}$$

A small video inset of a person is visible in the bottom right corner of the slide.

The last one what should be the minimum delay here? It is nothing but, I am going to call this as a minimum delay and then underneath it I am going to write it as minimum which will be nothing but,

$$\begin{aligned} \text{delay}_{\min} &= \sum p_i + N F^{1/N} \\ &= 6 + 4 F^{1/4} \\ &= 6 + 4(1.4519) \\ \text{delay} &= 11.8016 \end{aligned}$$

11.806 is the minimum delay for that particular stages, with the load of 20C and then with the input inverter gate size as 10. The overall delay, the absolute value of the delay the minimum delay of course, the absolute value will be,

$$\text{delay} = 11.8016 \times 3RC$$

If RC is nothing but 1ps, I will get an overall delay of 35.4ps, hope this is clear to everyone. In this particular lecture what we had seen is we started with understanding the gate sizes and how it is kind of distributed to the individual transistors sizing, the width of the transistors, especially on the PMOS side on the NMOS sides for different gates and then using that particular notation we applied the linear delay model.

Then try for a one particular single path, whatever number of the stages is there, we were able to estimate the minimum x value minimum no, the optimum x value, the optimum y value and optimum z value. That the overall delay turns out to be minimum, remember that what we have identified the optimum sizes of the gates. Once we identify the optimum sizes of the gates, we were able to distribute it and allocate the optimum width for the individual transistors.