

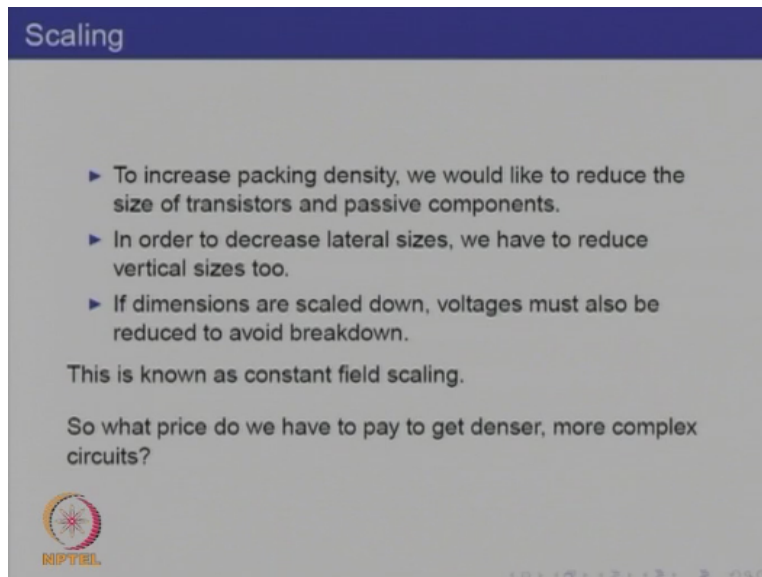
Advanced VLSI Design
Prof. D. K. Sharma
Department of Electrical Engineering
Indian Institute of Technology – Bombay

Lecture - 14

Interconnect Aware Design: Impact of scaling, buffer insertion and inductive peaking

In VLSI design, we normally worry about logic stages. However, as technologies have been scaled down the role of interconnects has become increasingly important. In this module, we shall see what is the impact of interconnects on modern VLSI design and what are the techniques that we use to mitigate the effects of large interconnect delay. To put the whole topic in perspective let us quickly revise the ideas of device scaling.

(Refer Slide Time: 01:00)



The slide is titled "Scaling" and contains the following text:

- ▶ To increase packing density, we would like to reduce the size of transistors and passive components.
- ▶ In order to decrease lateral sizes, we have to reduce vertical sizes too.
- ▶ If dimensions are scaled down, voltages must also be reduced to avoid breakdown.

This is known as constant field scaling.

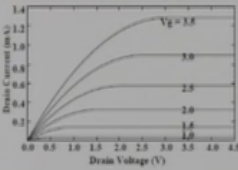
So what price do we have to pay to get denser, more complex circuits?

The slide also features the NPTEL logo in the bottom left corner and navigation icons in the bottom right corner.

To increase packing density, we would like to reduce the size of transistors and passive components. In order to decrease the lateral sizes, we have to reduce the vertical sizes as well. If dimensions are scaled down, voltages must also be reduced to avoid breakdown. This is the usual constant fields scaling. So, what price do we have to pay to get denser, more complex circuits.

(Refer Slide Time: 01:32)

MOS model




- ▶ For $V_{gs} \leq V_T$,
 $I_{ds} = 0$
- ▶ For $V_{gs} > V_T$ and $V_{ds} \leq V_{gs} - V_T$,
 $I_{ds} = K \left[(V_{gs} - V_T)V_{ds} - \frac{1}{2}V_{ds}^2 \right]$
- ▶ For $V_{gs} > V_T$ and $V_{ds} > V_{gs} - V_T$,
 $I_{ds} = \frac{K}{2}(V_{gs} - V_T)^2$

$$K \equiv \mu C_{ox} \frac{W}{L}$$

$$C_{ox} \equiv \frac{\epsilon_{ox}}{t_{ox}}$$

(Gate capacitance C_{ox} is per unit area)



Here is a very simple first order in fact zero order model for an MOS transistor and we know that in this model for gate source voltages less than the threshold voltage the current is 0. As the gate voltage exceeds threshold voltage but before the drain source voltage becomes equal to $V_{gs} - V_T$ or the saturation voltage. The current is given by this equation which is the linear region and it is $V_{gs} - V_T$ into $V_{ds} - \text{half } V_d \text{ square}$.

Once V_{ds} exceeds this saturation voltage then it is given by K by $2 V_{gs} - V_T$ whole square. In digital circuitry the on current of transistor is typically given by this equation. However, whatever the regime please notice that the current is this K factor multiplied by a term which is square in voltages. The details of this may be unimportant because when we scale voltages we scale all voltages.

And essentially what it means is that because this has dimensions of V square and if all voltages are scaled then this term will scale as V square.

(Refer Slide Time: 03:03)

Consequences of Scaling

All dimensions and voltages divided by the factor $S(> 1)$.

Device area	$\propto W \times L : (\downarrow S)(\downarrow S)$	$\downarrow S^2$
C_{ox}	$\epsilon_{ox}/t_{ox} : \text{const}/(\downarrow S)$	$\uparrow S$
C_{total}	$\epsilon A/t : (\downarrow S^2)/(\downarrow S)$	$\downarrow S$
V_{DS}, V_{GS}, V_T	Voltages : $(\downarrow S)$	$\downarrow S$
I_d	$\mu C_{ox}(W/L)(\propto V^2) : (\uparrow S)(\text{const})(\downarrow S^2)$	$\downarrow S$
Slew Rate $\frac{dV}{dt}$	$I/C_{total} : (\downarrow S)/(\downarrow S)$	const.
Delay	$V/\frac{dV}{dt} : (\downarrow S)/(\text{const})$	$\downarrow S$
Static Power	$V \times I : (\downarrow S)(\downarrow S)$	$\downarrow S^2$
dynamic power	$C_{total} V^2 f : (\downarrow S)(\downarrow S^2)(\uparrow S)$	$\downarrow S^2$
Power delay product	delay \times power $(\downarrow S)(\downarrow S^2)$	$\downarrow S^3$
Power density	power/area : $(\downarrow S^2)/(\downarrow S^2)$	const.

Let us now look at the impact of what happens as we scale down all dimensions and voltages we are going to divide all these quantities by some factor S which is the scaling factor and S is supposed to be greater than one. That means things will become smaller it is clear that the device area which is the product of the width and the length both the width and the length will be scaled down by the factor S bringing down the area by S square.

The term C_{ox} which occurs in the transistor equation is actually the capacitance per unit area and it is given by epsilon divided by t_{ox} which is a constant divided by S . So, the overall value of C_{ox} goes up by the factor S . The total capacitance is inclusive of the area and the area goes down by S square as we have seen here. Whereas the thickness of various things thin oxide, thick oxide goes down by factor S .

And as a result the overall value of the total capacitance goes down by the factor S . All voltages will be scaled down as we had said earlier by this factor S . So, let us look at the current the drain current. Drain current is given by this expression which is $\mu C_{ox} W$ by L and a term which is of the order of V square, μ is a material constant. We have already seen that C_{ox} goes up as this.

W by L being a ratio remains constant with scaling and because the voltages are scaled down by S this term will scale down by S square. Therefore overall the current will also go down like this.

So, therefore the voltage scale down by S and so does the current. Let us look at the slew rate, slew rate is the rate of change of voltage at the output and it is given roughly by I divided C . We have seen that I goes down by S and C also goes down by S we have seen it here.

So, overall the slew rate remains constant. The total delay is the total amount of voltage change that we require in going from zero to one or one to zero divided by slew rate. Now, the total voltage is scaled down the slew rate remains the same and therefore the total delay will go down by a factor S . This is in fact good news that means not only do we get smaller transistors we get faster transistors.

The static power which the product of voltage and current goes down by the factor S square and the dynamic power which is C times V square times frequency noticed that because the delay has gone down by S . We will now operate our transistors at a frequency which is higher by S . So, the total capacitance goes by S this we had seen here. V square goes down by S square and the frequency goes up by S because the delays have gone down by S .

So, we are going to operate our transistor at higher frequencies even at this enhanced frequency you can see that the dynamic power also scales down by S square. Therefore total power which is the sum of these two will also scale by S square. A measure of the goodness of a technology is the power delay product and because the delay goes down by S and power goes down by S square it in fact improves by a very considerable factor which is S cube.

We must not get carried away by the improvements in speed and power alone. We need to worry about the power density, power density determines the overall temperature at which this device will operate and fortunately they both reduce by the same area. The power goes down by S square so does the area and as a result the power density remains constant. So, let us summarize what is the effect of all of this when you scale down the technology?

(Refer Slide Time: 08:10)


Impact of scaling

- ▶ Improved packing density: $\uparrow S^2$
- ▶ Improved speed: delay $\downarrow S$
- ▶ Improved power consumption: $\downarrow S^2$

However ...
The above improvements apply to active circuits.

What about passive components?

Also, reduced voltages imply a lower signal to noise ratio.



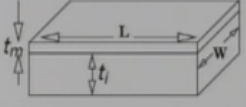
NPTEL

The packing density improves it goes up by a square because the area of each transistor is smaller. The speed improves because the delay goes down by S and power consumption improves because it goes down by S square. All of these are extremely welcome changes and it is indeed the motivation for the continual scaling down of technologies that we have seen over the last so many decades.

However, all these formula were derived from the equations of active circuits coming from transistor characteristics and capacitance. What about passive components? Also we must be aware that as the voltage is reduced the signal to noise ratio will also go down. Because the noise remains the same and the signal is proportional to the voltages that we have.

(Refer Slide Time: 09:11)


Concern: Interconnect Delay



$$R = \rho \frac{L}{W t_m}, \quad C = \epsilon \frac{LW}{t_i}$$

$$\text{Charge Time} \approx RC = \rho \epsilon \frac{L^2}{t_m t_i}$$

- ▶ To first order, delay is independent of W . This is because increasing W reduces resistance but increases capacitance in the same ratio.
- ▶ Unfortunately W is the only parameter that the circuit designer can decide! (L is fixed by the distance between the points to be connected, ρ , ϵ , t_m and t_i are decided by the technology).



Let us do a very simple order of magnitude competition for interconnect. What I show here is a line of length L of width W and with a metal whose thickness is t_m . This metal line is running over an insulator whose thickness is t_i and we assume that there is a ground plane below. The resistance of this line is the resistivity, times the length divided by the area of cross section which is W multiplied by the thickness of the metal.

On the other hand, the capacitance is given by the parallel plate capacitance of this and that is given by the dielectric constant multiplied by the area which is W times L divided by the thickness of the insulator now which is t_i . Therefore the charge time it is not equal but roughly equal of the order of RC times constant. And if we multiply these two things together what we get is $\rho \epsilon L^2$ divided by $t_m t_i$.

The factor W cancels. That means to first order delay is in fact independent of W and that is understandable if you increase W the resistance will go down. Because now you have a wider length but the capacitance will increase in the same ratio because of the same reason that the W is larger. Unfortunately, W is the only parameter that the circuit designer could decide L is fixed by the distance between the points to be connected you cannot change that.

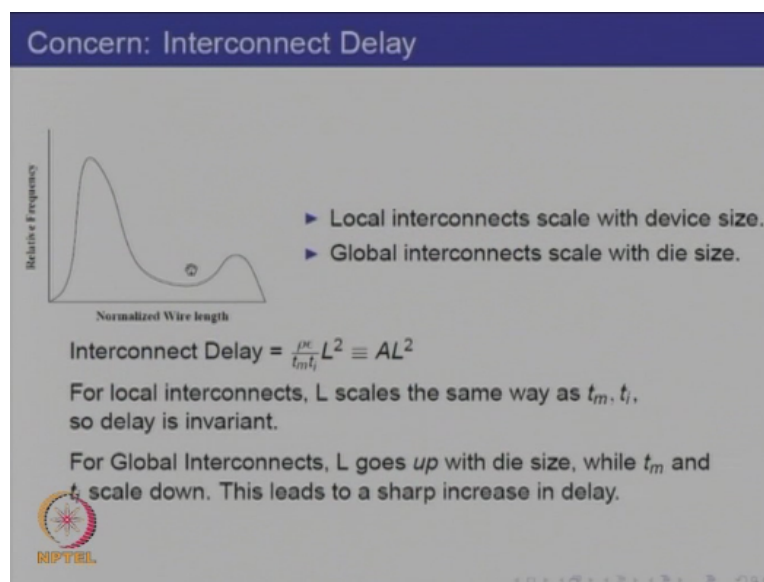
ρ and ϵ are material properties and t_m and t_i are decided by the technology they are not in the hands of the designer. The only factor which could have been controlled by the VLSI

designer is in fact how wide a line you will lay out and as we see that to first order that does not matter. In fact, this is a very simple model in reality if you take this source resistances and various things into account.

The width does matter a little bit so the VLSI designer has to be aware of what widths to use. But by and large it is a low order dependence on W. On the other hand, the dependence on L is square that means if you double the length of a wire the delay will not just double it becomes four times and that is very disappointing. That means the delay of a long wire will in fact increase very fast in it.

Now, let us look at the statistics of the kind of wires that we normally have on an integrated circuit.

(Refer Slide Time: 12:24)



This is just a notional diagram essentially you have a very large number of wire length which are distributed around some peak here. This is a histogram; the x axis shows the normalized wire length. The wire length is normalized to the diagonal of the di sides. Therefore, it will essentially spread from a range which is roughly zero to one. We will hardly have wires which are much longer then the diagonal of the entire di.

The y axis plots the number of wires which are of this length. It turns out that a very large

number of wires are distributed around this somewhat small length and these represent the local interconnects when you make flip flops when you take signals from one stage to the immediate next one. This is roughly the span of lens for which there will be wires. However, in addition to that there are these wires.

These wires are the global interconnects. These are the supply wires the clock distribution, the buses, the global signals, the reset signal, all such signals which has to run over the entire chip and there are a few of these because a bus typically may be very wide. It may have eighth, sixteen, 32 or these day 128 wires and these buses will be of very long normalize length.

Why we scale the technology it turns out that these length will scale down if the devices are smaller the interconnects which are immediate will be smaller on the other hand these global interconnects are scaled not by the transistor size but by the die size. So, the interconnect delay which we had derived earlier as $\rho \epsilon / t_{mi}$ into L^2 . We put this as a constant as AL^2 and its scales as L^2 .

For local interconnects as we change the technology L scales the same way as t_m times t_i . So, the delay is invariant if we scale down L by the factor S and scale down t_m and t_i by the same factor S both the numerator and the denominator will be divided by S^2 . So, the delay is invariant. This is not very good our active circuitry is becoming faster but our interconnects remains the same speed even as we scale the technology.

However, this is still manageable what is alarming is in fact the global interconnect. In case of global interconnects L actually goes up with die size while t_m and t_i scale down as a result even in a modern scale down technology the global interconnect delay will show a very sharp increase and this is indeed alarming. Our electron excess becomes much faster, our power has gone down but at the same time the delay in global interconnects goes up very, very sharply indeed.

And this is one of the major problems of modern VLSI design. To summarize, the global interconnect delay can be the determining factor for the speed of an integrated system.

(Refer Slide Time: 16:21)


Buffer Insertion

Global Interconnect delay can be the determining factor for the speed of an integrated system.

The L^2 dependence of interconnect delay is a source of particular concern.

This problem can be somewhat mitigated by buffer insertion in long wires.

We define some critical wire length and when a wire segment exceeds this length, we insert a buffer.



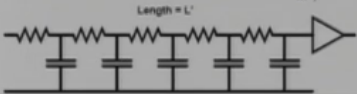
The L^2 dependence of interconnect delay is a source of particular concern. The problem can be somewhat mitigated by buffer insertion in long wires. This is a technique that we will see and in order to do this what we do is that all right if the delay is increasing super linearly that means if the length becomes double the delay becomes four times. Then let us not have long wires.

Let us divide the wire into short segments and let us buffer the signal after each short segment. So, we define some critical wire length and when a wire segment exceeds this length we insert a buffer.

(Refer Slide Time: 17:14)

Repeater Insertion in Voltage Mode

What is the optimum wire length after which we should insert a buffer? (Wire Delay = $\rho c \frac{L^2}{tm} = AL^2$)



Let the wire segment length = L' .
 Segment wire delay = AL'^2 .
 Let buffer delay = τ


For n segments, there will be $n-1$ buffers, and $L = nL'$.

$$\Delta = nAL'^2 + (n-1)\tau = \frac{L}{L'}AL'^2 + (\frac{L}{L'} - 1)\tau = ALL' + (\frac{L}{L'} - 1)\tau$$

Putting the derivative with respect to $L' = 0$ for optimization,

$$AL - \frac{L}{L'^2}\tau = 0, \text{ so } AL'^2 = \tau$$

L' should be so chosen that the wire segment delay = τ .
 Total delay is proportional to n and so, is linear in L .



So here is a model of this notice that this is not multiple segments. This is just one segment but we model it as a distributed RC and this entire thing is the segment length which we denote L' here. The delay of even this multi segment length will be proportional to L'^2 and next we put a buffer here. We assume that the delay of the buffer is τ . So, notice that while we are buffering the signal we are incurring the delay introduced by this buffer.

Let us assume that this segment length of the wire is L' . In that case the segment delay of this wire which constitutes just one segment of the wire long wire is AL'^2 where L' is the length of the segment and the buffer delay is τ . So, therefore the total delay of one stage of this will be $AL'^2 + \tau$ and for n segments they will be $n - 1$ buffers and L will be n times L' .

So, the total delay which I represent with this capital delta that is n times $AL'^2 + n - 1$ into τ . Now, n is nothing but L by L' we have divided the length L into n segments of length L' . Therefore, n can be put down in terms of L and L' also. So, we substitute for n this L by L' and finally we get for delay as this AL times L' , + L divided by $L' - 1$ into τ .

Now, we want to know what is the optimum segment length to minimize this overall delay? After all, if we have very small segments then we will have a large number of them. The total delay introduced by the buffer will then dominate the delay. On the other hand, if we segment which are too long then the L square dependence will make the wire delay too much and therefore there is an optimization an optimal length L' .

To find this optimal length we put the derivative of this total delay term with respect to L' zero for optimization and when you take the derivative of this what you get is $A L'$ from the first term and minus L by L'^2 from this term times of course τ . This is a constant and we will give you zero on derivation. What this means is that AL'^2 should be equal to τ . $A L'^2$ is in fact just the wire delay of the segment.

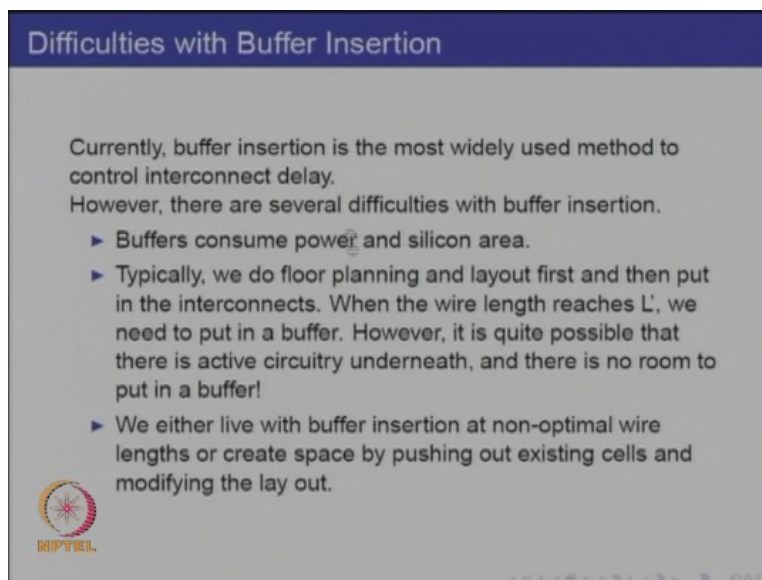
So, therefore we have a very simple formula that L' should be so chosen that the wire

segment delay is equal to the buffer delay. Now the total delay will be proportional to n and therefore is linear in L . This segment delay is not two τ twice the delay suffered by the wire but overall the total delay is linear in L the multiplier this linear proportionality constant has doubled.

But it remains linear and therefore over longer wire lengths we will gain a lot if we use this trick. As a result, buffer insertion has in fact become the dominant technique for laying out long wires in modern technologies. These buffers are typically wide buffers. So the τ is small and then unfortunately what it means is that they consume power and they consume silicon area. However, in current technologies that is the only way to go.

Otherwise global delays will pull down the performance of your system by a very large amount.


(Refer Slide Time: 21:45)



Difficulties with Buffer Insertion

Currently, buffer insertion is the most widely used method to control interconnect delay. However, there are several difficulties with buffer insertion.

- ▶ Buffers consume power and silicon area.
- ▶ Typically, we do floor planning and layout first and then put in the interconnects. When the wire length reaches L' , we need to put in a buffer. However, it is quite possible that there is active circuitry underneath, and there is no room to put in a buffer!
- ▶ We either live with buffer insertion at non-optimal wire lengths or create space by pushing out existing cells and modifying the layout.

 NIPTEL

So, these are the problems with buffer insertions. Let us look at this buffer consume power and silicon area. In a typical design what we do is, floor planing and layout the unit cells of our design first and then put in the interconnects. Now while putting down these interconnects we roll this wire out and when the wire length reaches L' we need to put in a buffer. However, it is quite possible that when we reach the length L' on this wire.

We find that there is active circuitry underneath on silicon and there is no room to put in a buffer.

Now, what do we do we either live with buffer insertion at non optimal wire length that means we use an L prime which is too small or too long at this point. Because we cannot put down a buffer at this length or we insist that our interconnect will be optimal and whatever is obstructing.

Putting a buffer here will be moved out and therefore the existing cells must be relayed out and we must remodify the layout this leads to a cycle we keep on adjusting our floor plan that changes the wire delays. That changes the wire length and finally it becomes very difficult to get closer of our design. So, this is the big problem of modern VLSI design with global interconnects.

Global interconnects often includes data buses and this is yet another problem that means these data buses need to be bidirectional. Now as long as we just had a wire there is no problem the wire can be bidirectional however as soon as we put in a buffer this defines a direction because the buffer has an input and an output. So, for example suppose you have a bus connecting a processor and a CAS or on chip memory.

Now the buffer insertion will fix the direction of data flow. Unfortunately, the data flow is in both directions you need to read from the memory as well as write to it. So, what do you do either you have redundant buses one for reading and one for writing that means your cost both in terms of area and power and complexity goes up by a factor of two. Or you replace the buffer with bidirectional transceiver.

However, this is not such an easy solution the transceiver need a control signal which will determine in which direction it conveys the data and where does this control signal come from these transceivers are all along the wire and therefore this control signal must be run along with bus that means now we must run more wires in the bus. Also this control signal must not be slower than the data.

In fact, it should be faster than the data so that the direction of the transceiver is set before the data arrives. That means this wire must also be buffer. If it is unidirectional there is no problem

but if it is bidirectional we have major headache here because then we need to decide how to set the direction of the buffers which convey the control signal. So, that means bidirectional buffer inserted wires are extremely complicated to design.

They are expensive both in terms of area silicon area certainly. We are have not doubled the wire length by putting transceiver but the silicon area will indeed double and we must put in additional wires which will carry the control signal for turning around the direction of these transceiver. Notice that buses can at times be very wide and that means each buffer of the control signal must drive.

Say 16, 32, 64 or 128 line buffers in parallel and this is indeed a tall order and you pay by having a slower bus and one which consumes a lot of area as well as power. So buffer insertion is unfortunately the only solution available but practically it comes with lot of disadvantages. Modern VLSI design research has therefore for concentrated on how to improve the performance of the interconnect and this constitutes what we call interconnect aware design.

One of the solution is in fact to change our design style and lay down the wires first. So, this is like designing the roads in a city first and then use up the area for active circuitry. The active circuitry area left must be adequate to accommodate exactly the practical silicon circuits that you will put there if it is too large the area between the interconnects then you will waste silicon area. If it is too small you would not be able to fit in the functional circuits.


That you want to put in after you have laid down these wires already. So, even that solution is not optimal is not without its problems. The other solution is to come up with ways of conveying data which is different from what is used right now and we shall now look at some of these techniques. There are other problems as well which are common there is a serious signal integrity problem because of electro static coupling between long wires.

(Refer Slide Time: 27:55)

Concern: Signal Integrity

As interconnect wire separation is reduced ...

- ▶ There is a serious signal integrity problem because of electrostatic coupling between long wires.
- ▶ Inter-signal interference can lead to unpredictable delay variations.
- ▶ Grounded shielding wires must often be inserted to avoid interference.
- ▶ This leads to extra capacitance and CV^2f power loss.




Inter-signal interference capacitive coupling of signals going down one wire. Coupling to a wire which is adjacent to it they can lead to unpredictable delay variations. The delay variations are data dependent and cannot be anticipated at the time of design to avoid this we might use grounded shielding wires between wires but then that leads to extra capacitance and CV square f power losses not to speak of area.

So, there are solutions but these solutions are generally expensive.

(Refer Slide Time: 28:36)

Concern: Timing closure

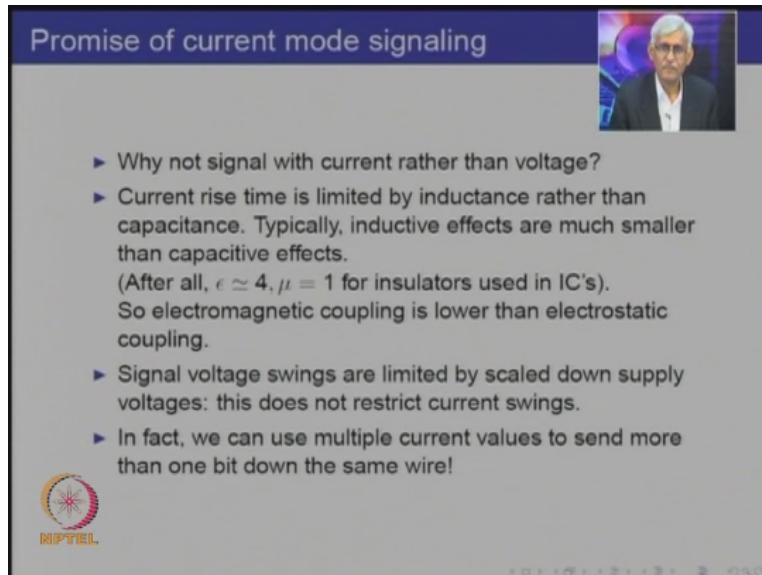
- ▶ Global interconnects are placed *after* active circuit design and layout is complete.
- ▶ One has to anticipate the wire length, and then design the active circuits to meet total delay specifications.
- ▶ If the actual wire length is different from what was anticipated, one has to re-design the active circuits after layout.
- ▶ After a fresh layout, wire lengths and hence, delays are changed.
- ▶ This leads to a design-layout-redesign iteration known as Timing Closure. This iteration becomes longer and longer when total delays are dominated by interconnect delay.



Timing closer we have already discussed that if global interconnects are placed after active circuit design then we lead to a closer problem that means you change the layout that changes

wire lengths, that changes delays and then you must redesign the circuits and so on round and round till you meet your timing specification.

(Refer Slide Time: 29:00)



The slide is titled "Promise of current mode signaling" and features a small portrait of a man in the top right corner. The main content is a list of points:

- ▶ Why not signal with current rather than voltage?
- ▶ Current rise time is limited by inductance rather than capacitance. Typically, inductive effects are much smaller than capacitive effects.
(After all, $\epsilon \simeq 4$, $\mu = 1$ for insulators used in IC's).
So electromagnetic coupling is lower than electrostatic coupling.
- ▶ Signal voltage swings are limited by scaled down supply voltages: this does not restrict current swings.
- ▶ In fact, we can use multiple current values to send more than one bit down the same wire!

The slide also includes the NPTEL logo in the bottom left corner and navigation icons in the bottom right corner.

To get around this problem we come up with a new suggestion. We say why is signals zeros and one by voltage? Why should we have that a low voltage is zero and a high voltage is one. Why cannot we signal a zero or one by the presence or absence of a current. The current rise time is limited by inductance rather than capacitance and typically inductive effects are much smaller than capacitive effects, inductance effects are becoming important now with scaling down.

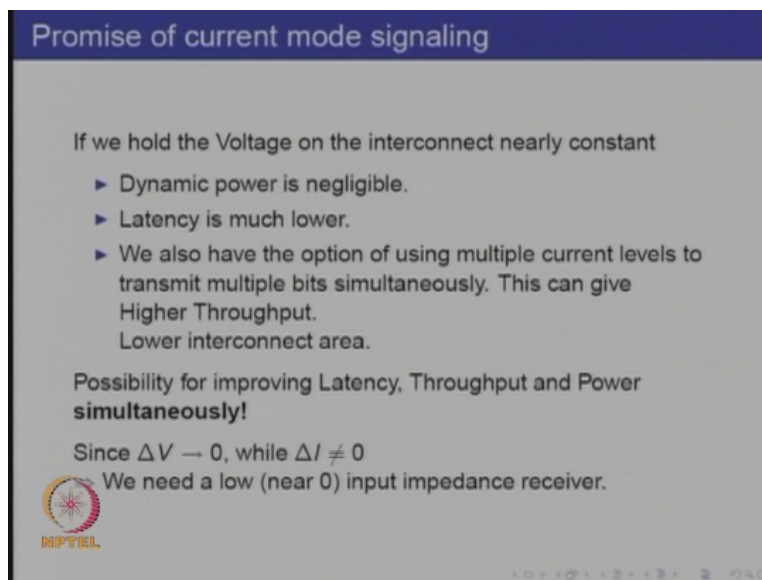
But still the inductive effects are much smaller than the capacitive effects after all the value of epsilon is close to four for most of the material average. This is the silicon dioxide dielectric constant. So, epsilon is of the order of four but practically none of the materials that we use are magnetic and therefore the value of mu is of the order of one. And therefore the inductive effects tend to be somewhat smaller than the capacitive effects.

Signal voltage swings are limited by scale down supply voltages remember we got higher performance by scaling down voltages and therefore the voltage swings are limited. On the other hand, the current swings are not limited we can still drive a transistor a much harder and take much larger current from it. So, therefore we are not limited in the head room for signal to noise ratio if we signal using current rather than voltage.

Indeed, so much so that we can use multiple current values to send more than one bit down the same wire that means the amount of current could be zero unit, one unit, two units or three and that would represent a two-bit information being carried down the same wire. So, therefore the current mode signaling appears very promising indeed and indeed this is some work that many researchers in the world have been doing and this includes our group here at IIT, Bombay.

So, let us see how we use this current mode signaling.

(Refer Slide Time: 31:19)




Promise of current mode signaling

If we hold the Voltage on the interconnect nearly constant

- ▶ Dynamic power is negligible.
- ▶ Latency is much lower.
- ▶ We also have the option of using multiple current levels to transmit multiple bits simultaneously. This can give Higher Throughput. Lower interconnect area.

Possibility for improving Latency, Throughput and Power simultaneously!

Since $\Delta V \rightarrow 0$, while $\Delta I \neq 0$
We need a low (near 0) input impedance receiver.

 NIPTEL

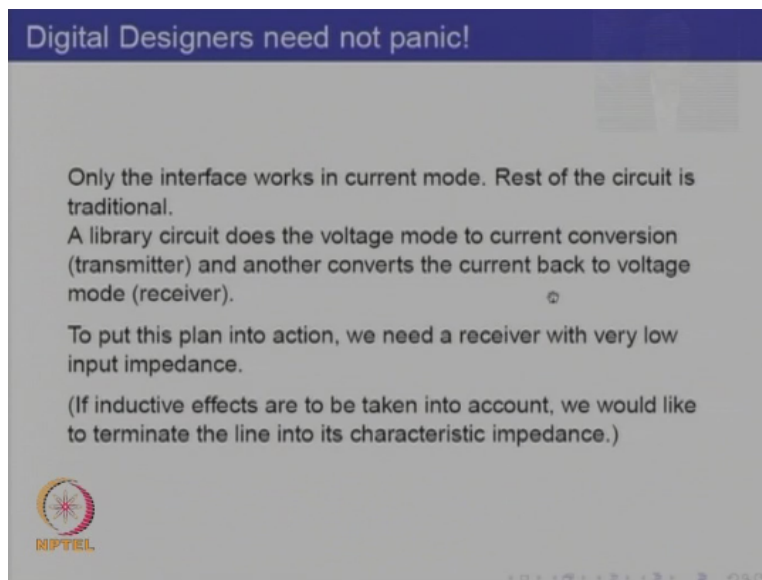
The suggestion is as following. If you hold the voltage on the interconnect nearly constant in fact, we will not be able to do it so. We will later see that we will use low swing signaling. But just to argue the point let us say that the voltage on the interconnect is nearly constant. That means the dynamic power will be negligible. In fact, if you hold it exactly constant then the dynamic power will be zero. Correspondingly, the latency will be much lower.

We also have the option of using multiple current levels to transmit multiple bits simultaneously. This can give Higher Throughput and Lower interconnect area because there are fewer wires for carrying the same amount of information. Thus we have the possibility for improving latency throughput and power simultaneously. Normally, we tend to trade off one against the other.

However, this is easier said than done. How are you going to change the current without changing the voltage because Ohm's law connects the two in fact if we say that ΔV tends to zero while Δi is non zero that means the input resistance at the receiver should approach zero otherwise as you change the current across this input impedance a voltage will develop across this which will change in proportion to this current.

Therefore, in current mode what we do is to terminate the line with a lot resistance rather than using the traditional high input impedance of CMOS design. So, let us see what happens when we use this technique.

(Refer Slide Time: 33:10)



However, the digital designers who are used to doing traditional design might panic. However, what we are proposing is that we will design the interconnect as a library element. It will be designed using more or less analog techniques. However, the digital designers need not panic as far as their voltage levels are concerned these voltage levels will be standard. They will swing from rail to rail and these will be digital signals.

All the overhead of converting this rail to rail signal to a lower voltage and then converting it back to a rail to rail to digital signal the power as well as the delay involved in these operations will in fact be charged to the current mode of transmission signal. And if in spite of this overhead if it comes out better then indeed it is a very good replacement for the traditional buffer inserted

drivers.

(Refer Slide Time: 34:22)

Zero input impedance circuit

Low r_{in} amps are used for photo-detectors. ¹

$$i_1 = g_{mn1}V_1 = g_{mp1}(V - V_2)$$

$$i_2 = g_{mn2}V_1 = -g_{mp2}V_2$$

$$V_2 = -\frac{g_{mn2}}{g_{mp2}}V_1 = -\frac{g_{mn2}}{g_{mp2}}\frac{i_1}{g_{mn1}}$$

$$i_1 = g_{mp1}V + \frac{g_{mp2}/g_{mn1}}{g_{mp2}/g_{mp1}}i_1$$

define $\Gamma \equiv \frac{g_{mn2}/g_{mn1}}{g_{mp2}/g_{mp1}}$ then, $i_1(1 - \Gamma) = g_{mp1}V$

This gives $r_{in} = (1 - \Gamma)/g_{mp1}$

NPTEL C.-K. Kim et al, "High Injection Efficiency Readout Circuit for Low Resistance Infrared Detector", IEE Electronic Letters, 35, 1507, 1999.

So let us look at this particular idea the main problem is how do we get a very low input termination resistance? There are many ideas for these the simplest of these would only actually put a passive resistor as the end of the line. However, there are some smart solutions which have been proposed for this and this is one of them. This is taken from a paper by C K Kim and other offers and it was originally designed not for interconnects.

But for read out circuits for low resistance infrared detectors. This paper originally appeared in 1999 the reference is given here. this is configuration by the way is known as a beta multiplier. This is a slight modification of a beta multiplier. In a beta multiplier these two transistors will have different geometries and the larger of these will have a resistor in this arm. However, in this case we have the match and there is no resistor.

You can think of it as a P type current mirror driving and N-type current mirror below. The circuit above the half looks like a P type current mirror. The circuit below the half looks like an N-type current mirror. So, we connect a P type current mirror to an N-type current mirror and they essentially sustain a value of current within each other. The reference current of this is in fact the current output of P type current mirror.

And this is echoed back by this transistor and becomes the reference current of the other. So is the result if the multiplying factor is truly one then this current will in fact be sustained. We can actually look at the input impedance of this arrangement we propose to connect this point marked V to the interconnect, long interconnect line and therefore the input impedance is the impedance seen here to (∞) (36:47)

To calculate that we compute the ratio of the small signal fluctuation v_1 to the small signal fluctuation i_1 . This can be easily analyzed for example i_1 is $g_{m,n1}$ that is the g_m of this n channel transistors times V_1 where the AC or the fluctuating voltage across this transistor. Notice that this is diode connected and therefore the current is g_m times V_1 where V_1 is the voltage at the drain as well as at the gate.

But you look at the P channel transistor then the same current i_1 which is in series is given by g_m of this transistor, $g_{m,p1}$ times the difference of fluctuation at this point and this point that is the gate-source voltage of this. So, therefore this current i_1 is also $g_{m,p1}$ times $V - V_2$, V is the fluctuation here, V_2 is the fluctuation here and therefore this is given by this. Similarly, these two transistors are in series and therefore convey the same current.

And this current i_2 is given by g_m and two times v_1 . Look at the n channel transistor, the gate fluctuation is v_1 this is connected diode style here therefore the fluctuation here is V_1 and therefore the current through this transistor the fluctuation in current through this transistor is given by i_2 equal to $g_{m,n2}$ times v_1 remember v_1 is the small signal fluctuation in voltage not the DC voltage.

And this is equal to $-g_{m,p2} v_2$ because this is a P channel transistor V_2 is the fluctuation here. The current fluctuation here will have opposite sign $2 V_2$ here, an increasing voltage here. This being a P channel transistor will decrease the current through this. Now because this is a current mirror these two currents are equal and we can solve these equations to get eventually the input impedance of this.

So by eliminating the unknown voltages here which is v_1 here. We finally get i_1 in terms of V

and eliminate V_1 V_2 in terms of V_1 o. So, i_1 is $g_{m1} v +$ this ratio of ratios g_{m2} by g_{m1} divided by g_{m2} divided by g_{m1} times i_1 . If we collect all terms in i_1 then we can get the input impedance and finally the input impedance will come out as $1 - \gamma$ divide by g_{m1} .

The net impact of this is that by bringing γ very close to 1 we can bring the input impedance of the circuit very close to zero. In fact, we can put the input impedance to any desired value by setting the value of γ appropriately. And γ is nothing but g_{m2} divided by g_{m1} the whole thing divided by g_{m2} divided by g_{m1} . In short, it is the ratio of the mirror ratio here.

Mirror ratio in the n channel transistor divided by the mirror ratio of the P channel transistor. Now, it turns out that this is a very robust quantity.

(Refer Slide Time: 40:35)

Robustness of design

In saturation,

$$I_d = \frac{1}{2} \mu C_{ox} \frac{W}{L} (V_g - V_T)^2$$

So, $g_m = \mu C_{ox} \frac{W}{L} (V_g - V_T) = \sqrt{2 \mu C_{ox} \frac{W}{L} I_d}$

$$g_{m2}/g_{m1} = \sqrt{\frac{(W/L)_{n2} I_2}{(W/L)_{n1} I_1}}$$

$$g_{mp2}/g_{mp1} = \sqrt{\frac{(W/L)_{p2} I_2}{(W/L)_{p1} I_1}}$$

Therefore $\Gamma \equiv \frac{g_{m2}/g_{m1}}{g_{mp2}/g_{mp1}} = \sqrt{\frac{(W/L)_{n2}/(W/L)_{n1}}{(W/L)_{p2}/(W/L)_{p1}}}$

NIPTEL

If you derive the value of g_m it is given by $\sqrt{2k}$ times I_d if the transistors are assumed to be saturated and when you substitute for this the value of γ actually turns out to be this number which is dependent only on geometry. Notice that this is independent transistor parameters and therefore it will be independent of process variations. It is independent of K and mobility and therefore it will be independent of temperatures.

Neither k nor V_T nor any other transistor parameter appears in the expression for γ as a


result it is set statically once and for all by the geometries. And therefore it can be defined very, very robustly thus we choose an input impedance that we need, set the appropriate value of gamma and now we can make a receiver which has the desired input impedance which can go all the way down to zero and once we have a low input impedance.

Essentially a large current fluctuation will lead to a very small voltage fluctuation on the line and then we will reach the consequent advantages.

(Refer Slide Time: 41:51)

Receiver Design - Input stage

- ▶ Input resistance controlled by geometry of transistors
- ▶ Interconnect voltage held fixed
- ▶ Input resistance insensitive to process variations

 NIPYEL

This current which then comes to the beta multiplier as we had seen can be further mirrored and then we can detect this current by converting it to voltage or by current comparator.

(Refer Slide Time: 42:03)

Reduced swing signaling

- ▶ In reduced swing voltage mode signaling, the line is not terminated in a low impedance.
- ▶ Current mode signaling terminates the line in a low impedance.

This reduces the time constant, increases bandwidth.
 However, this also leads to static power consumption.

Indeed, low swing signaling can also be used for voltage mode so we should be careful when comparing the advantages of low swing by itself and low swing voltage versus, load string current. What is the difference? Indeed, I can design buffers in which this voltage swing is not rail to rail. These are like analogs circuits with gain of the order of one and now I can have a driver where this swing much smaller than V_{dd} .

In that case I shall still reap the benefits of reducing the dynamic power. But what I sense at this end is then a voltage and this buffer will then need to restore the small signal voltage here to a full rail to rail signal here. In this case of current mode everything is exactly the same I drive it. I get my full swing input and I drive it by a low swing driver. I have the same passive line which is RC

But I now terminated it in a load resistance which is small because of this termination the swing at this particular point will continue to remain low but also it will bring down the RC time constant of this entire line. As a result, the bandwidth of this line will be higher and I will be able to use much higher data rates on this line if I am using current mode. So, it is the combination of this a low terminating resistance and low swing signaling that the current mode of signaling.

Shows extreme promise and indeed some of the best results in terms of power and delay for standard technologies not going to optical and so on are shown by technologies which use this

trick. And this is an upcoming technique which is used widely to signal at low energy cost and at very high speeds.

(Refer Slide Time: 44:34)

The slide is titled "Improving Current Mode Signaling" and features a small portrait of a man in the top right corner. The main content includes a circuit diagram and a list of points.

The circuit diagram, labeled "Low Swing Current Mode Receiver", shows a "Low swing Driver" connected to a "Line" (represented by a resistor and capacitor in series). The "Line" is then connected to a "Receiver" which is terminated with a resistor R_L . The driver and receiver are both connected to a common ground.

Current mode signaling

- ▶ Consumes Static Power
- ▶ Direct Trade-off between speed and static power

Possible Improvements

- ▶ Inductive Peaking
- ▶ Dynamic Over-driving

The NPTEL logo is visible in the bottom left corner of the slide.

Now we have seen the basic premise of current mode signaling. We would now like to optimize it further. The current mode signalling can be modeled as a low swing driver which drives this distributed RC line and which is then terminated in a low resistance amplifier. This receiver or amplifier takes a low swing voltage at the input and converts it to a full rail to rail traditional digital signal.

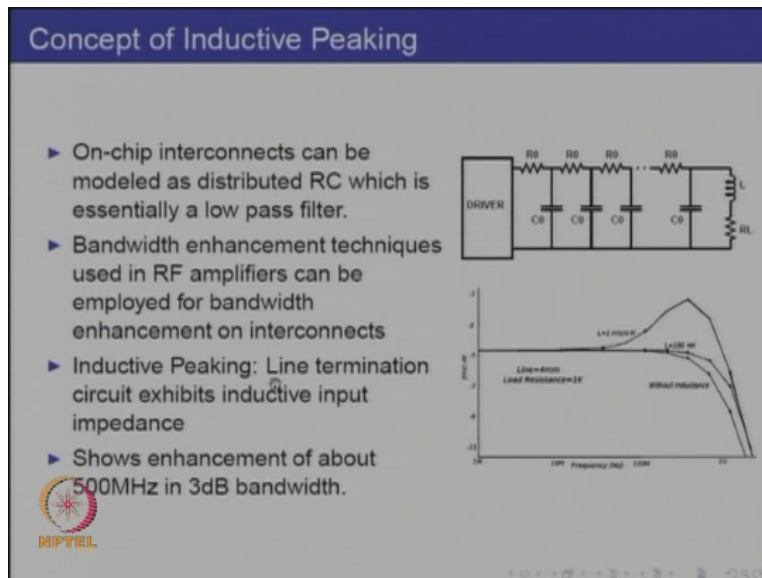
So notice that in this link the input is a traditional rail to rail digital signal and so is the output. Therefore, if we design this link very well once and for all and put it in a library then our overall VLSI design style need not change. On the other hand, if this link is faster and consumes much less power than we have an improved VLSI working with us actually it turns out that interconnects actually consume between half to three quarters of the total power consumed by a modern large VLSI.

So, if we reduce the delay and reduce the power we shall have an overall impact on the performance of the circuit. This does not mean that current mode signaling has no disadvantages at all. We must be aware of these and we must know when current mode signaling can in fact be used and this is the list of things that we have to worry about. Notice that because a current is

made to flow it will consume static power.

And this power will be independent of the data rate that means at low data rates this may not be such a good idea. There is a direct trade-off between speed and static power. If we have to make this link faster than the static current which flows through this must be increased. There are two tricks which are now suggested for improving this kind of circuitry and they include inductive peaking and dynamic overdriving and we shall now look at both of this in turn.

(Refer Slide Time: 47:13)



Let us look at inductive peaking first and let us first understand the motivation of using inductive peaking. As we have seen on chip interconnect can be modeled as distributed RC which is essentially a low pass filter. That means the problem that we have is that the digital signal provided by our driver is passing through a low pass filter. This low pass filter then has this kind of roll off at high frequencies.

If at the end apart from the terminating resistance, I put a small inductor in that case the voltage developed across this will have a high frequency behavior remember the voltage across the inductor the inductive impedance will be ω times L and when I pass a current through this the voltage will be I times ω times L which is linearly proportional to ω . Higher the frequency, higher is this voltage.

So essentially by suggesting that you put an inductive termination you are saying that you will counter the low pass nature of this interconnect with the high pass value of this load. So, this is the basic premise on which we are making this solution. This is not a new idea and this idea has been used in RF amplifiers for a long time. The only program is that putting inductor on chip is not practical.

Inductors are large devices and they consume a lot of area and therefore a traditional inductor may not be possible to understand the problem what we do is first find out what is the amount of inductance that we may put in a practical circuit to expect improvements in device speeds. And then find out if this much of inductance can be generated electronically that means as an active inductor rather than putting a traditional spiral inductor which in fact has a low Q.

When we do static analysis with different values of L, not caring how we are going to get this value of L. We noticed that an enhancement of about 500 megahertz in 3 dB bandwidth is possible. I would warn you that this x axis is on an algorithmic scale and it turns out so these three curves are with zero in between and large inductance put here and therefore the cut off frequency is pushed out.

And while this looks very marginal on the scale that is because this is a long scale and indeed the 3dB bandwidth is increased by about 500 megahertz where the actual limit is of the order of a couple of megahertz. So, this improvement is substantial indeed in short what this passive analysis tells us is that if we can produce of the order of tens to hundreds of Nanohenry of inductance.

Somehow in that case the kind of improvement in data speed that we can expect neutralizing the low pass behavior of this inductor can be substantially indeed. Our next endeavor then will be how to generate an artificial electronically generated inductance which is of this value and we shall look at it in the next lecture. So just to summarize we have looked at the scaling behavior of interconnects.

We notice with considerable alarm that the delay of the interconnects goes up as L square and as

a result the overall delay will be dominated in fact by the delay of the long interconnect, short interconnects the delay remains more or less constant and while in the long term it may be of concern. In short term perhaps we can live with it. But long wires like clock distribution, like buses very long wire which carry high speed signals are a great concern.

The traditional solution for that problem has been buffer insertion. We have looked at the optimization problem of that buffer insertion and we came to the conclusion that it is worthwhile to divide a long wire into short wire segments and the optimum segment length is such that these segment wire delay is equal to the buffer delay. However, this solution had lots of drawbacks and we have been looking at low swing current mode signaling as a replacement.

For that buffer insertion method and we have looked at one possible way in which we can counter the low pass behavior of this long interconnect wire by terminating it into a high pass termination which requires inductors of the order of tens to hundreds of nanohenries. We shall see in the next lecture how we can generate these tens to hundreds of nanohenries and what are the other ways of countering the low pass behavior of this wire model.

This is where we conclude this particular lecture.