INTERCONNECT:

Interconnect in an integrated circuit are physical connections between two transistors and/ or the external surroundings.

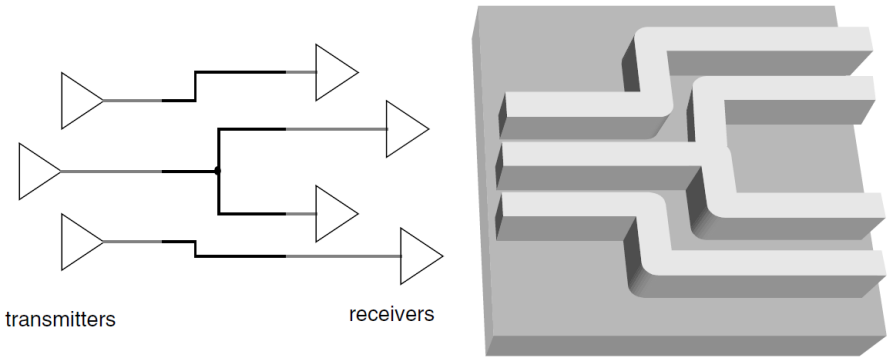
- An electronic circuit designer has multiple choices in realizing the interconnections between the various devices that make up the circuit.
- Here the start of the art processes offers multiple layers of aluminium or copper, and at least one layer of polysilicon. Even the heavily doped n^+ and p^+ diffusion layers are typically used for the realization of source and drain regions can be employed for wiring purposes. These wires appear in the schematic diagrams of electronic circuit as simple lines with no apparent impact on the circuit performance.

These wiring of integrated circuits forms a complex geometry that introduces the following parasitics:

1. Capacitive Parasitics
2. Resistive Parasitics and
3. Inductive Parasitics

The capacitive, resistive and inductive parasitics have multiple effects of the circuit's behaviour i.e.

- They all cause an increase in propagation delay, or equivalent, a drop in performance.
- They all have an impact on the energy dissipation and the power distribution.
- They all cause the introduction of extra noise sources, which affect the reliability of the circuit.

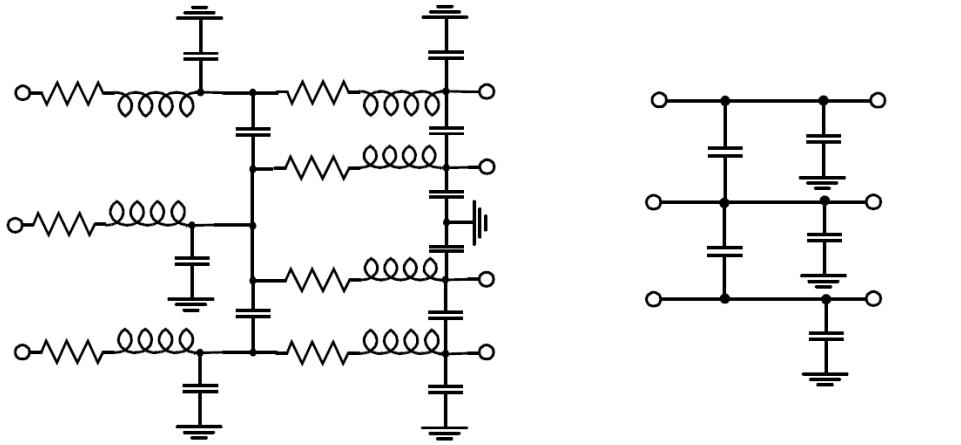


SCHEMATIC AND PHYSICAL VIEWS OF WIRING OF BUS-NETWORK

It is important that the designer has a clear insight in the parasitic wiring effects, their relative importance, and their models. This is best illustrated with the simple example as shown above. Each wire in a bus network connects a transmitter (or transmitters) to a set of receivers and is implemented as a link of wire segments of various lengths and geometries. Assume that all segments are implemented on a single interconnect layer, isolated from the silicon substrate and from each other by a layer of dielectric material. Be aware that the reality may be far more complex.

Analyzing the behavior of this schematic, which only models a small part of the circuit, is slow and cumbersome. Fortunately, substantial simplifications can often be made, some of which are enumerated below,

- Inductive effects can be ignored if the resistance of the wire is substantial- this is for instance the case for long Aluminum wires with a small cross-section- or if the rise and fall times of the applied signals are slow.
- When the wires are short, the cross-section of the wire is large, or the interconnect material used has a low resistivity, a capacitance- only model can be used (figure FOR WIRE PARASITICS WITH CAPACITANCE ONLY shown below)
- The separation between neighbouring wires is large, or when the wires only run together for a short distance, inter-wire capacitance can be ignored, and all the parasitic capacitance can be modeled as capacitance to ground. Obviously, the latter problems are the easiest to model, analyze, and optimize.



WIRE PARASITICS (WITH THE EXCEPTION OF INTER-WIRE RESISTANCE AND MUTUAL INDUCTANCE) *WIRE PARASITICS WITH CAPACITANCE ONLY*

WIRE MODELS FOR PARASITICS

The various interconnect parameters whose values can be estimated, simple models to evaluate their impact, and a set of rules- of- thumb to decide i.e. when and where a particular model or effect should be considered are:

1. Capacitance Parameter
2. Resistance Parameter
3. Inductance Parameter

CAPACITANCE INTERCONNECT PARAMETER:

The capacitance of such a wire is a function of its shape, its environment, its distance to the substrate, and the distance to surrounding wires. An accurate modeling of the wire capacitance(s) in a state-of-the-art integrated circuit is a non-trivial task and is even today the subject of advanced research.
In capacitance parameter there are two types of capacitance occurring i.e.

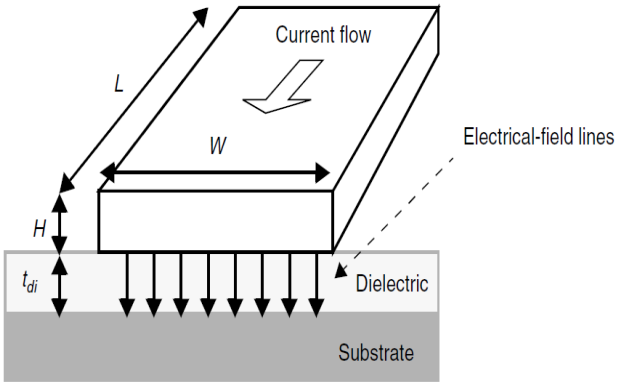
1. Parallel plate Capacitance and
2. Fringe Capacitance

Parallel plate Capacitance:

Consider first a simple rectangular wire placed above the semiconductor substrate, as shown in figure below. If the width of the wire is substantially larger than the thickness of the insulating material, it may be assumed that the electrical-field lines are orthogonal to the capacitor plates, and that its capacitance can be modeled by the *parallel-plate capacitance model.* Under those circumstances, the total capacitance of the wire can be approximated as,

$$c_{int} = \frac{\varepsilon_{di}}{t_{di}} WL$$

Where W and L are respectively the width and length of the wire, and t_{di} and ε_{di} represent the thickness of the dielectric layer and its permittivity. SiO2 is the dielectric material of choice in integrated circuits, although some materials with lower permittivity, and hence lower capacitance, are coming in use.

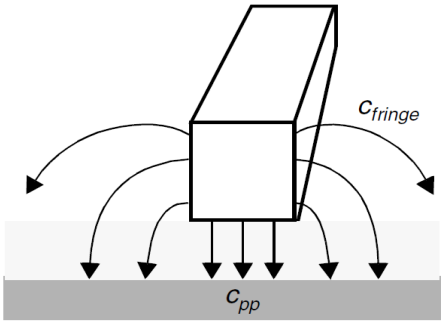


PARALLEL-PLATE CAPACITANCE MODEL OF INTERCONNECT WIRE

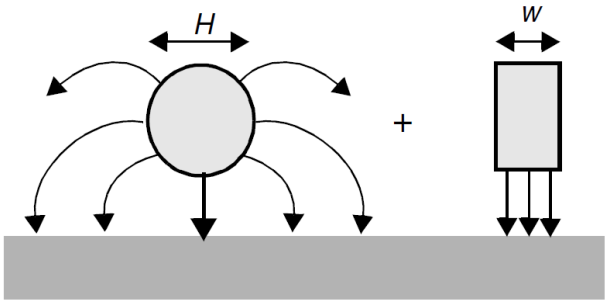
In actuality, this model is too simplistic. To minimize the resistance of the wires while scaling technology, it is desirable to keep the cross-section of the wire ($W \times H$) as large as possible. On the other hand, small values of W lead to denser wiring and less area overhead. As a result, we have over the years witnessed a steady reduction in the W/H- ratio, such that it has even dropped below unity in advanced processes.

Fringe/ Fringing Capacitance:

The capacitance between the side-walls of the wires and the substrate, called the *fringing capacitance*, as shown below.



FRINGING FIELDS/ THE FRINGING-FIELD CAPACITANCE



MODEL OF FRINGING-FIELD CAPACITANCE- DECOMPOSES THE CAPACITANCE INTO TWO CONTRIBUTIONS: A PARALLEL-PLATE CAPACITANCE, AND A FRINGING CAPACITANCE, MODELED BY A CYLINDRICAL WIRE WITH A DIAMETER EQUAL TO THE THICKNESS OF THE WIRE

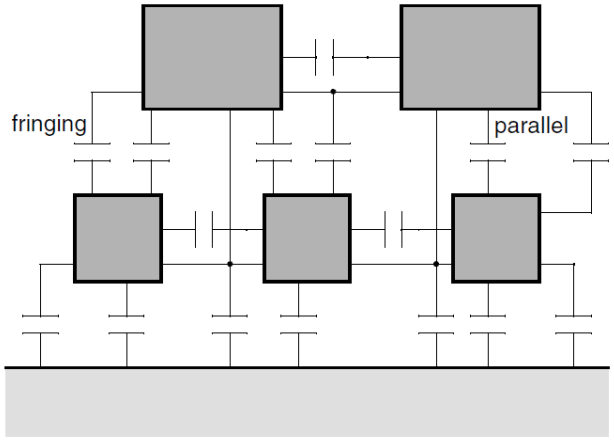
Therefore, the parallel plate capacitance and fringing capacitance constitutes the overall capacitance. Which is given as,

$$c_{wire} = c_{pp} + c_{fringe} = \frac{w\varepsilon_{di}}{t_{di}} + \frac{2\pi\varepsilon_{di}}{\log(t_{di}/H)}$$

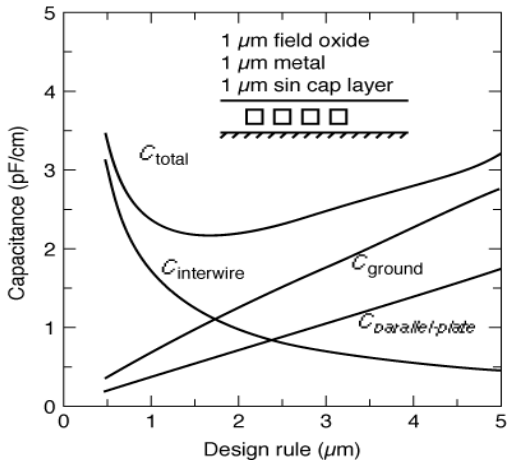
With $w = W - H/2$ a good approximation for the width of the parallel-plate capacitor.

CAPACITANCE COUPLING/ CAPACITANCE COUPLING EFFECT:

Assuming that a wire is completely isolated from its surrounding structures and is only capacitively coupled to ground, becomes untenable. This is illustrated in figure, where the capacitance components of a wire embedded in an interconnect hierarchy are identified. Each wire is not only coupled to the grounded substrate, but also to the neighbouring wires on the same layer and on adjacent layers. The main difference is that not all its capacitive components do terminate at the grounded substrate, but that a large number of them connect to other wires, which have dynamically varying voltage levels, these floating capacitors causes crosstalk and a negative effect to the circuit also.



CAPACITIVE COUPLING BETWEEN WIRES IN INTERCONNECT HIERARCHY



INTERCONNECT CAPACITANCE AS A FUNCTION OF DESIGN RULES

Inter- wire capacitances become a dominant factor in multi- layer interconnect structures. This effect is more important for wires in the higher interconnect layers, as these wires are farther away from the substrate. The increasing contribution of the inter- wire capacitance to the total capacitance with decreasing feature sizes is illustrated by graphical figure as shown, which plots the capacitive components of a set of parallel wires routed above a ground plane, it is assumed that dielectric and wire thickness are held constant while scaling all other dimensions. When W becomes smaller than 1.75 H, the inter-wire capacitance starts to dominate.

Wiring Capacitances for 0.25 µm CMOS Technology:

The table rows represent the top plate of the capacitor, the columns the bottom plate. The area capacitances are expressed in aF[1]/µm[2], while the fringe capacitances (given in the shaded rows) are in aF/µm.

	Field	Active	Poly	Al1	Al2	Al3	Al4
Poly	88						
	54						
Al1	30	41	57				
	40	47	54				
Al2	13	15	17	36			
	25	27	29	45			
Al3	8.9	9.4	10	15	41		
	18	19	20	27	49		
Al4	6.5	6.8	7	8.9	15	35	
	14	15	15	18	27	45	
Al5	5.2	5.4	5.4	6.6	9.1	14	38
	12	12	12	14	19	27	52

Inter- Wire Capacitance per unit wire length for different interconnect layers of 0.25 µm CMOS Technology Process:

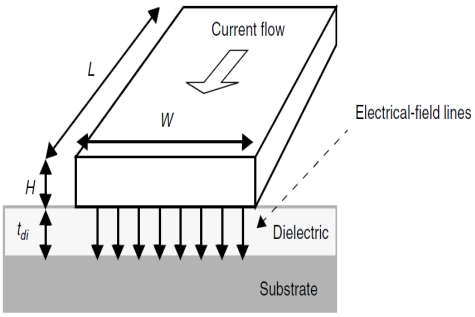
The capacitances are expressed in aF/mm, and are for minimally-spaced wires

Layer	Poly	Al1	Al2	Al3	Al4	Al5
Capacitance	40	95	85	85	85	115

RESISTANCE INTERCONNECT PARAMETER:

The resistance of a wire is proportional to its length L and inversely proportional to its cross- section A. The resistance of a rectangular conductor as shown in figure below can be expressed as,

$$R = \frac{\rho L}{A}\ ohm = \frac{\rho L}{H\,W}\ ohm$$



[1] *Attofarad (aF)- 1 aF is equal to 10^{-18} C/V*

Where:

ρ = resistivity

A = HW = area of cross section of the rectangular wire

If L = W, i.e. square of resistive material, then

$$R = \frac{\rho}{H} = R_s \; in \; {}^{ohm}/_{square} \; \textbf{(Sheet Resistance)}$$

the *sheet resistance* of the material, having units of Ω/ sq. This expresses that the resistance of a square conductor is independent of its absolute size, as is apparent from $R = R_S \frac{L}{W}$. To obtain the resistance of a wire, simply multiply the sheet resistance by its ratio (L/ W).

Resistivity of Commonly used Conductors/ Interconnect Resistance:

Material	ρ (Ω-m)
Silver (Ag)	1.6×10^{-8}
Copper (Cu)	1.7×10^{-8}
Gold (Au)	2.2×10^{-8}
Aluminum (Al)	2.7×10^{-8}
Tungsten (W)	5.5×10^{-8}

Aluminum is the interconnect material most often used in integrated circuits because of its low cost and its compatibility with the standard integrated- circuit fabrication process. Unfortunately, it has a large resistivity compared to materials such as Copper. With ever- increasing performance targets, this is rapidly becoming a liability and top- of- the- line processes are now increasingly using Copper as the conductor of choice.

Typical values of the Sheet Resistance of various Interconnect Materials using 0.25 µm CMOS Technology:

Material	Sheet Resistance (Ω/❑)
n- or p-well diffusion	1000 – 1500
n^+, p^+ diffusion	50 – 150
n^+, p^+ diffusion with silicide	3 – 5
n^+, p^+ polysilicon	150 – 200
n^+, p^+ polysilicon with silicide	4 – 5
Aluminum	0.05 – 0.1

From the table, we conclude that Aluminum is the preferred material for the wiring of long interconnections. Polysilicon should only be used for local interconnect. Although the sheet resistance of the diffusion layer (*n*+, *p*+) is comparable to that of polysilicon, the use of diffusion wires should be avoided due to its large capacitance and the associated *RC* delay.

INDUCTANCE INTERCONNECT PARAMETER:

The inductance of a section of a circuit states that a changing current passing through an inductor generates a voltage drop ΔV.

$$\Delta V = L\frac{\mathrm{d}i}{\mathrm{d}t}$$

On-chip inductance include ringing and overshoot effects, reflections of signals due to impedance mismatch, inductive coupling between lines, and switching noise due to *Ldi*/*dt* voltage drops.

It is possible to compute the inductance a wire directly from its geometry and its environment. A simpler approach relies on the fact that the capacitance *c* and the inductance *l* (per unit length) of a wire are related by the following expression,

$$cl = \varepsilon\mu$$

With ε and μ respectively the permittivity and permeability of the surrounding dielectric.

Other interesting relations, obtained from Maxwell's laws, can be pointed out. The constant product of permeability and permittivity also defines the speed ν at which an electromagnetic wave can propagate through the medium,

$$\nu = \frac{1}{\sqrt{lc}} = \frac{1}{\sqrt{\varepsilon\mu}} = \frac{c_0}{\sqrt{\varepsilon_r\mu_r}}$$

c_0 equals the speed of light (30 cm/ nsec) in a vacuum.

Considering a lumped RLC model we get $Z_{RL} = R + J\omega L, where\ \omega = 2\pi f$.

$$If\ R \gg J\omega L, then\ inductance\ effect\ is\ not\ important\ and\ .$$

$$If\ R \ll J\omega L, then\ inductance\ effect\ will\ be\ observed\ in\ signal\ reduction\ in\ resistance$$

Dielectric constants and wave-propagation speeds for various materials used in electronic circuits; (The relative permeability μ_r of most dielectrics is approximately equal to 1)

Dielectric	ε_r	Propagation speed (cm/nsec)
Vacuum	1	30
SiO_2	3.9	15
PC board (epoxy glass)	5.0	13
Alumina (ceramic package)	9.5	10

INTERCONNECT MODELING:

As we know the parasitic elements have an impact on the electrical behaviour of the circuit and influence its delay, power dissipation, and reliability. To study these effects requires the introduction of electrical models that estimate and approximate the real behaviour of the wire as a function of its parameters. These models vary from very simple to very complex depending upon the effects that are being studied and the required accuracy.

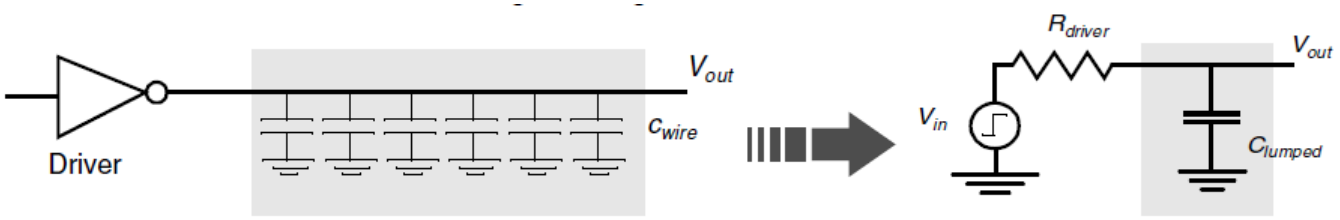
The types of interconnect modelling are:

1. Lumped Model
2. Lumped RC Model- The Elmore Delay
3. Distributed RC line Model/ Distributed *rc* line Model
4. Transmission Line Model

Lumped Model:

The circuit parasitics of a wire are distributed along its length and are not lumped into a single position. Yet, when only a single parasitic component is dominant, when the interaction between the components is small, or when looking at only one aspect of the circuit behaviour, it is often useful to lump the different fractions into a single circuit element. The advantage of this approach is that the effects of the parasitic then can be described by an ordinary differential equation.

As long as the resistive component of the wire is small and the switching frequencies are in the low to medium range, it is meaningful to consider only the capacitive component of the wire, and to lump the distributed capacitance into a single capacitor as shown in figure. It is observed that in this model the wire still represents an equipotential region, and that the wire itself does not introduce any delay. The only impact on performance is introduced by the loading effect of the capacitor on the driving gate. This capacitive lumped model is simple, yet effective, and is the model of choice for the analysis of most interconnect wires in digital integrated circuits.



DISTRIBUTED VERSUS LUMPED CAPACITANCE MODEL OF WIRE. $C_{LUMPED} = L \times C_{WIRE}$, WITH L THE LENGTH OF THE WIRE AND C_{WIRE} THE CAPACITANCE PER UNIT LENGTH. THE DRIVER IS MODELED AS A VOLTAGE SOURCE AND A SOURCE RESISTANCE R_{DRIVER}

The operation of this simple RC network is described by the following ordinary differential equation,

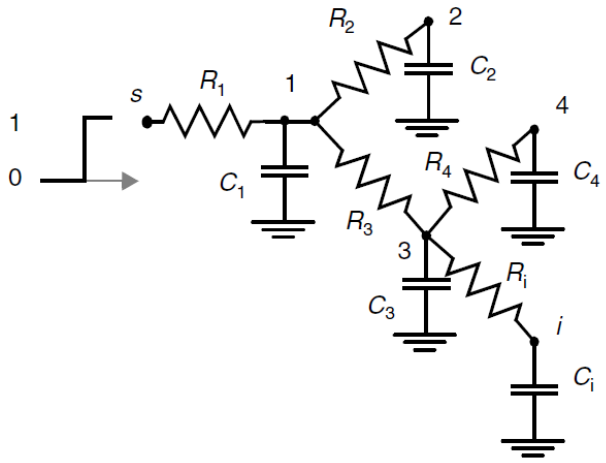
$$C_{lumped}\frac{dV_{out}}{dt} + \frac{V_{out} - V_{in}}{R_{driver}} = 0$$

Lumped RC Model/ The Elmore Delay:

On-chip metal wires of over a few mm length have a significant resistance. The equipotential assumption, presented in the lumped-capacitor model, is no longer adequate, and a resistive-capacitive model has to be adopted.

A first approach lumps the total wire resistance of each wire segment into one single *R* and similarly combines the global capacitance into a single capacitor *C*. This simple model, called the ***lumped RC model***, is pessimistic and inaccurate for long interconnect wires, which are more adequately represented by a *distributed rc-model*. Yet, before analyzing the distributed model, it is worthwhile to spend some time on the analysis and the modeling of lumped *RC* networks for the following reasons:

1. The distributed *rc*-model is complex and no closed form solutions exist. The behaviour of the distributed *rc*- line can be adequately modeled by a simple *RC* network.
2. A common practice in the study of the transient behavior of complex transistor-wire networks is to reduce the circuit to an *RC* network. Having a means to analyze such a network effectively and to predict its first-order response would add a great asset to the designers tool box.



TREE- STRUCTURED RC NETWORK

An interesting result of this particular circuit topology is that there exists a unique resistive path between the source node *s* and any node *i* of the network. The total resistance along this path is called the *path resistance* R_{ii}. For example, the path resistance between the source node *s* and node 4 equals,

$$R_{44} = R_1 + R_3 + R_4$$

The definition of the path resistance can be extended to address the *shared path resistance* R_{ik}, which represents the resistance shared among the paths from the root node *s* to nodes *k* and *i*:

$$R_{ik} = \sum R_j \Rightarrow (R_j \in [path(s \rightarrow i) \cap path(s \rightarrow k)])$$

Here,

$$\boldsymbol{R_{i4} = R_1 + R_3 \text{ while } R_{i2} = R_1}$$

Assume now that each of the *N* nodes of the network is initially discharged to GND, and that a step input is applied at node *s* at time $t = 0$. The Elmore delay at node *i* is then given by the following expression:

$$\tau_{Di} = \sum_{k=1}^{N} C_k R_{ik}$$

Therefore, the Elmore delay is equivalent to the first-order time constant of the network (or the first moment of the impulse response). The designer should be aware that this time- constant represents a simple approximation of the actual delay between source node and node *i*. Yet in most cases this approximation has proven to be quite reasonable and acceptable. It offers the designer a powerful mechanism for providing a quick estimate of the delay of a complex network.

The RC delay of a tree structured network is given as,

$$\tau_{Di} = R_1C_1 + R_1C_2 + (R_1 + R_3)C_3 + (R_1 + R_3)C_4 + (R_1 + R_3 + R_i)C_i$$

i.e. using

$$\tau_{Di} = \sum_{k=1}^{N} C_k R_{ik}$$

We can compute the Elmore delay for node *i*

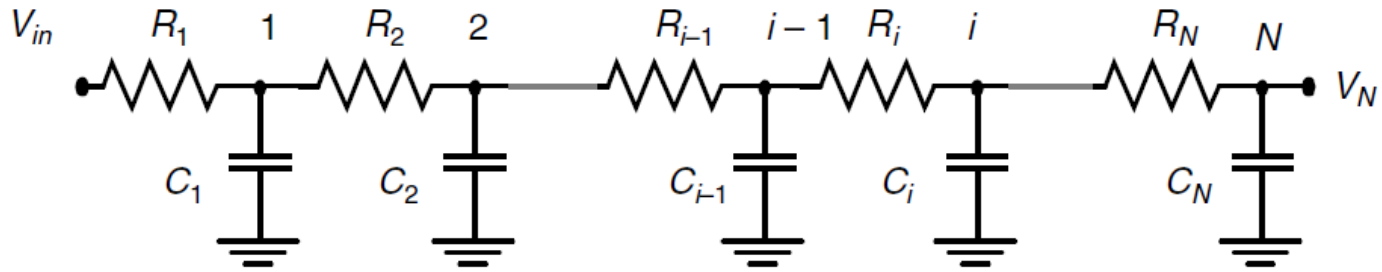
RC Chain/ The Elmore RC Chain Delay:

As a special case of the *RC* tree network, let us consider the simple, non- branched *RC* chain (or ladder) shown in figure. This network is worth analyzing because it is a structure that is often encountered in digital circuits, and also because it represents an approximative model of a resistive-capacitive wire. The Elmore delay of this chain network can be derived with the aid of

$$\tau_{Di} = \sum_{k=1}^{N} C_k R_{ik}$$

As

$$\tau_{DN} = \sum_{i=1}^{N} C_i \sum_{j=1}^{i} R_j = \sum_{i=1}^{N} C_i R_{ii}$$



RC CHAIN MODEL

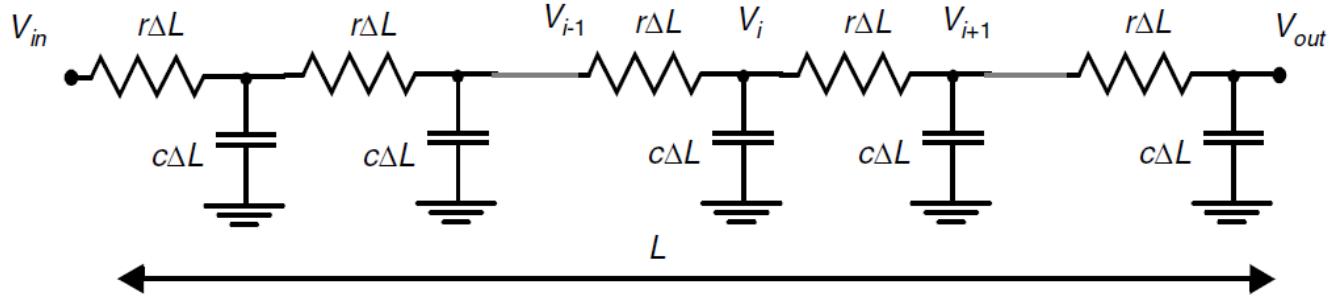
The component of node 1 consists of C_1R_1 with R_1 the total resistance between the node and the source, while the contribution of node 2 equals $C_2(R_1 + R_2)$. The equivalent time constant at node *2* equals $C_1R_1 + C_2(R_1 + R_2)$. τ_i of node *i* can be derived in a similar way.

$$\tau_{Di} = C_1R_1 + C_2(R_1 + R_2) + \ldots + C_i(R_1 + R_2 + \ldots + R_i)$$

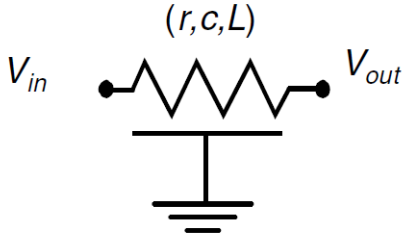
Thus, the Elmore delay formula has proven to be extremely useful. Besides making it possible to analyze wires, the formula can also be used to approximate the propagation delay of complex transistor networks. The evaluation of the propagation delay is then reduced to the analysis of the resulting *RC* network. More precise minimum and maximum bounds on the voltage waveforms in an *RC* tree have further been established.

Distributed RC line Model/ Distributed *rc* line Model:

A distributed *rc* line model is a more appropriate model as shown below which has, *r* and *c* stand for the resistance and capacitance per unit length.



DISTRIBUTED RC LINE MODEL



SCHEMATIC SYMBOL FOR DISTRIBUTED RC LINE

The voltage at node *i* of this network can be determined by solving the following set of partial differential equations:

$$c\Delta L\frac{\partial V_i}{\partial t} = \frac{(V_{i+1} - V_i) + (V_{i-1} - V_i)}{r\Delta L}$$

The correct behavior of the distributed *rc* line is then obtained by reducing ΔL asymptotically to 0. For $\Delta L \rightarrow 0$, the above equation becomes the well-known *diffusion equation*:

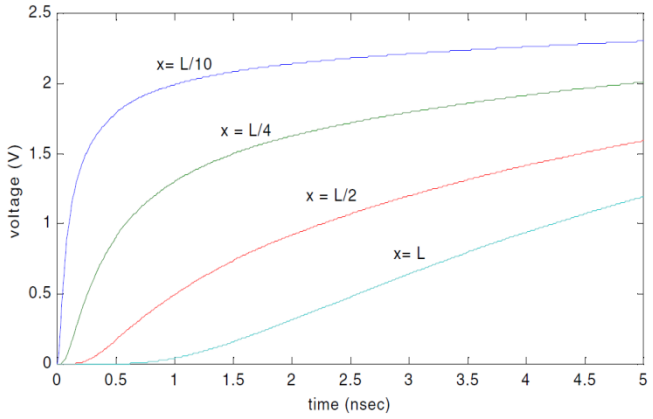
$$rc\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2}$$

Where *V* is the voltage at a particular point in the wire, and *x* is the distance between this point and the signal source. No closed form solution exists for this equation, but approximative expressions such as the formula written below can be derived:

$$V_{out}(t) = 2erfc(\sqrt{\frac{RC}{4t}}) \qquad t \ll RC$$

$$= 1.0 - 1.366e^{-2.5359\frac{t}{RC}} + 0.366e^{-9.4641\frac{t}{RC}} \qquad t \gg RC$$

The graph below shows the response of a wire to a step input, plotting the waveforms at different points in the wire as a function of time. It is observable how the step waveform "diffuses" from the start to the end of the wire, and the waveform rapidly degrades, resulting in a considerable delay for long wires. Driving these *rc* lines and minimizing the delay and signal degradation is one of the trickiest problems in modern digital integrated circuit design.

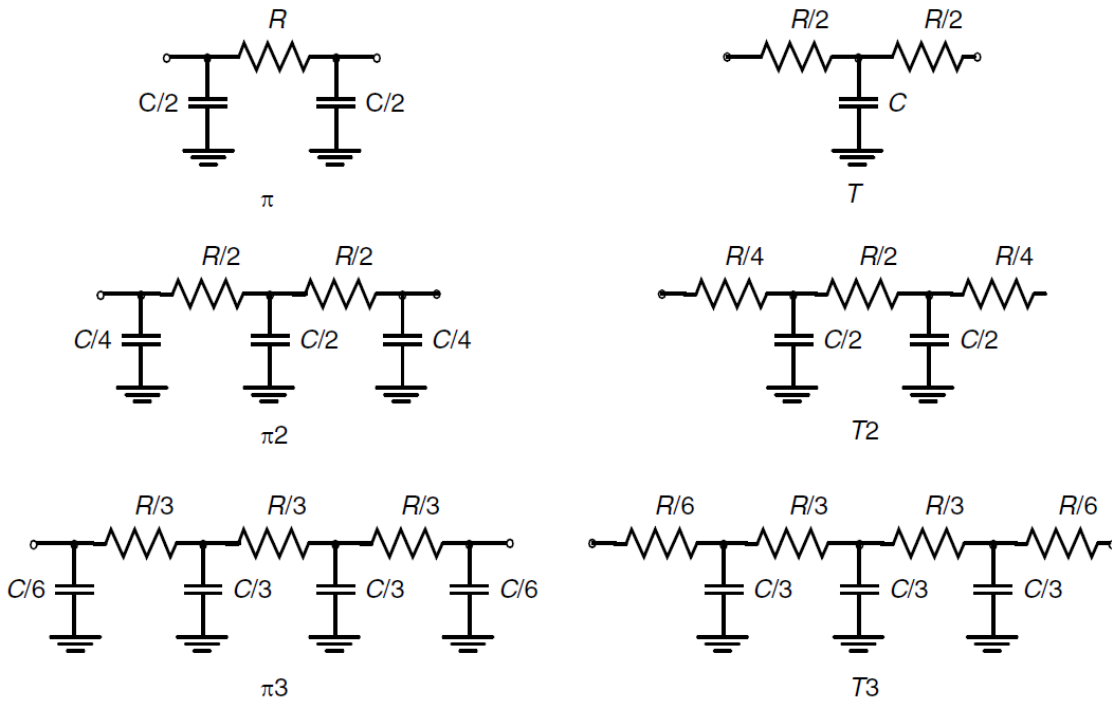


SIMULATED STEP RESPONSE OF RESISTIVE-CAPACITIVE WIRE AS A FUNCTION OF TIME AND PLACE

***rc* delays should only be considered when the rise (fall) time at the line input is smaller than *RC*, the rise (fall) time of the line.**

$$t_{rise} < RC$$

With R and C the total resistance and capacitance of the wire. When this condition is not met, the change in signal is slower than the propagation delay of the wire, and a lumped capacitive model suffices.



SIMULATION π AND T MODELS FOR DISTRIBUTED RC LINE

Step response of lumped and distributed *RC* networks- Points of Interest:

Voltage range	Lumped *RC* network	Distributed *RC* network
$0 \rightarrow 50\%$ (t_p)	0.69 *RC*	0.38 *RC*
$0 \rightarrow 63\%$ (τ)	*RC*	0.5 *RC*
$10\% \rightarrow 90\%$ (t_r)	2.2 *RC*	0.9 *RC*
$0\% \rightarrow 90\%$	2.3 *RC*	1.0 *RC*

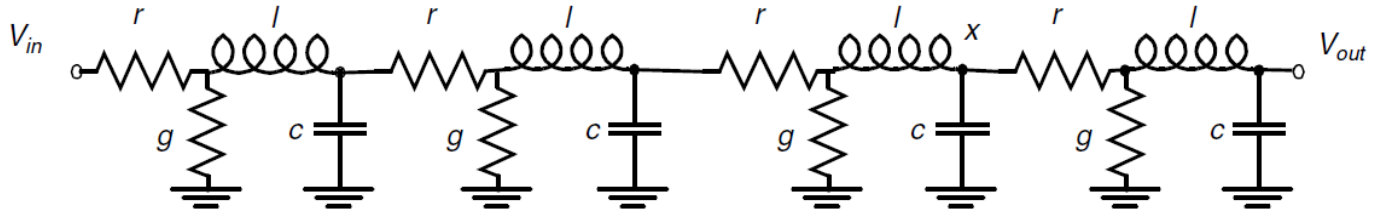
The Transmission Line Model:

Similar to the resistance and capacitance of an interconnect line, the inductance is distributed over the wire. A distributed *rlc* model of a wire, known as the transmission line model, becomes the most accurate approximation of the actual behaviour.

The transmission line has the prime property that a signal propagates over the interconnection medium as a *wave*. This is in contrast to the distributed *rc* model, where the signal *diffuses* from the source to the destination governed by the diffusion equation i.e.

$$rc\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2}$$

In the wave mode, a signal propagates by alternatively transferring energy from the electric to the magnetic fields, or equivalently from the capacitive to the inductive modes.



A LOSSY TRANSMISSION LINE

Consider the point *x* along the transmission line of figure as shown above at time *t*. The following set of equations holds:

$$\frac{\partial v}{\partial x} = -ri - l\frac{\partial i}{\partial t}$$

$$\frac{\partial i}{\partial x} = -gv - c\frac{\partial v}{\partial t}$$

Assuming that the leakage conductance *g* equals 0, which is true for most insulating materials, and eliminating the current *i* yields the *wave propagation equation*,

$$\frac{\partial^2 v}{\partial x^2} = rc\frac{\partial v}{\partial t} + lc\frac{\partial^2 v}{\partial t^2}$$

where *r*, *c*, and *l* are the resistance, capacitance, and inductance per unit length respectively.

COPING WITH INTERCONNECT:

As till now we have concentrated on the growing impact of interconnect parasitics on all design metrics of digital integrated circuits. As mentioned, interconnect introduces three types of parasitic effects i.e. capacitive, resistive, and inductive- all of which influence the signal integrity and degrade the performance of the circuit. While so far we have concentrated on the modeling aspects of the wire, we now analyze how interconnect affects the circuit operation, and we present a collection of design techniques to cope with these effects with considering each parasitics- this is referred to as *coping with interconnect*.

CAPACITIVE PARASITICS:

- **Capacitance reliability and Cross talk:**

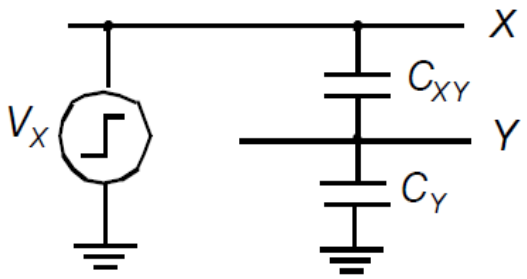
An unwanted coupling from a neighbouring signal wire to a network node introduces an interference that is generally called *cross talk*. The resulting disturbance acts as a noise source and can lead to hard-to-trace intermittent errors, since the injected noise depends upon the transient value of the other signals routed in the neighbourhood. In integrated circuits, this inter signal coupling can be both capacitive and inductive.

Capacitive cross talk is the dominant effect at current switching speeds, although inductive coupling forms a major concern in the design of the input-output circuitry of mixed-signal circuits. The potential impact of capacitive crosstalk is influenced by the impedance of the line under examination. If the line is floating, the

disturbance caused by the coupling persists and may be worsened by subsequent switching on adjacent wires. If the wire is driven, on the other hand, the signal returns to its original level.

- **Floating Lines:**

Approaching capacitive parasitic with respect to capacitive coupling to a floating line.



CAPACITIVE COUPLING TO A FLOATING LINE

Considering the circuit shown as above, where line X is coupled to wire Y by a parasitic capacitance C_{XY}. Line Y sees a total capacitance to ground equal to C_Y. Assuming that the voltage at node X experiences a step change equal to ΔV_X. This step appears on node Y attenuated by the capacitive voltage divider.

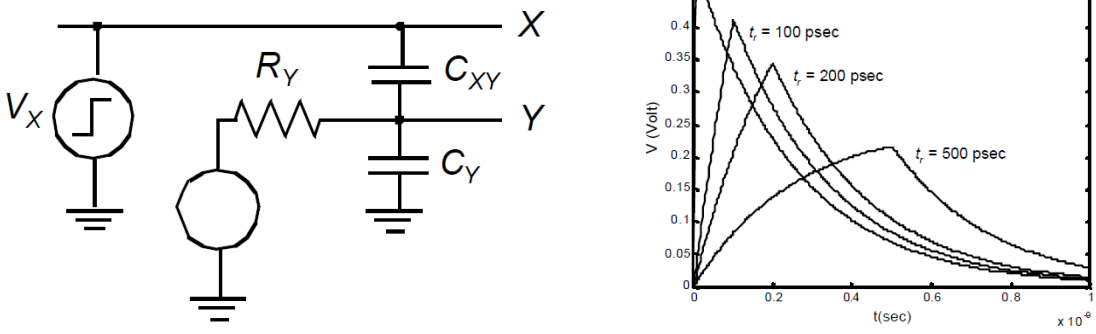
$$\Delta V_Y = \frac{C_{XY}}{C_Y + C_{XY}} \Delta V_X$$

Circuits that are particularly susceptive to capacitive cross talk are networks with low- swing pre charged nodes, located in adjacent to full- swing wires (with $\Delta V_X = V_{DD}$).

Examples of these are dynamic memories, low swing on chip busses and some dynamic families.

To address the cross talk issue, level- restoring device or keepers are a must in dynamic logic.

- **Driven Lines:**



CAPACITIVE COUPLING TO A DRIVEN LINE AND ITS VOLTAGE RESPONSE

As seen from the figure, if the line Y is driven with a resistance R_Y, a step on line X results in a transient on line Y. The transient decays with a time constant $\tau_{XY} = R_Y(C_{XY} + C_Y)$. The actual impact on the victim line is a strong function of the rise- fall time of the interfering signal.

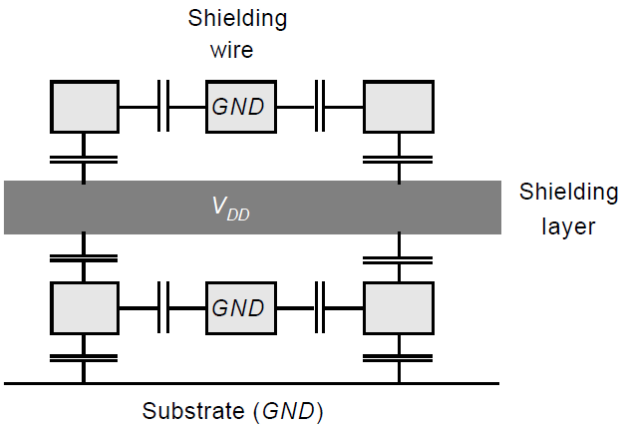
If the rise time is comparable or larger than the time constant, the peak value of disturbance is diminished. This can be observed in the response figure.

Obvious, keeping the driving impedance of a wire and hence τ_{XY} low goes a long way towards reducing the impact of capacitive cross talk. The keeper transistor added to a dynamic gate or pre charged wire is an excellent example of how impedance reduction helps to control noise.

Therefore, the impact of cross talk on the signal integrity of driven nodes is rather limited. The resulting glitches may cause malfunctioning of connecting sequential elements, and should therefore be carefully monitored. The most important effect is an increase in delay.

Design Techniques to Deal with Capacitive Cross talk:

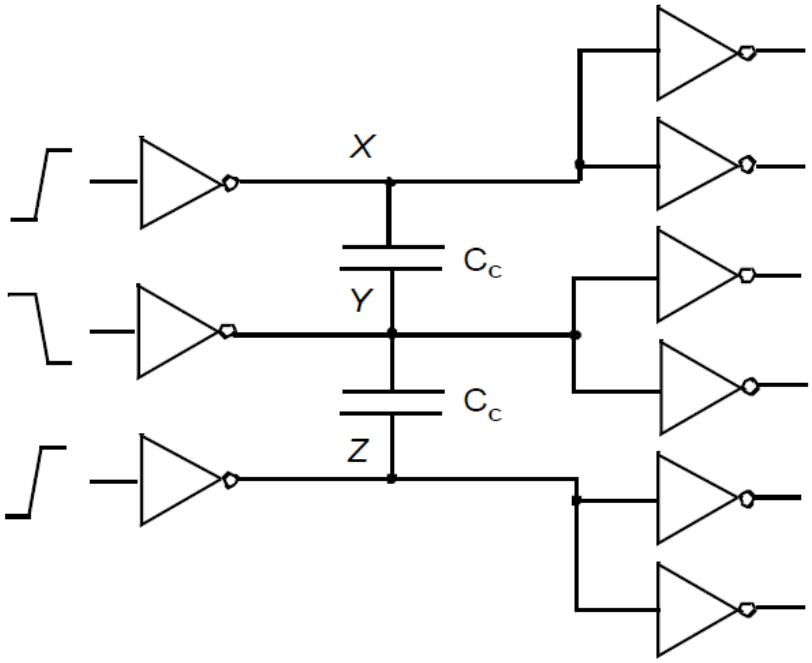
1. If possible avoid floating nodes, nodes sensitive to cross talk problems such as pre charged busses, should be equipped with keeper devices to reduce the impedance.
2. Sensitive nodes should be well separated from full swing signals.
3. Making the rise- fall time as large as possible subjection to timing constraints.
4. Use differential signalling in sensitive low swing wiring networks. This turns the cross talk signal into a common mode noise source that does not impact the operation of the circuit.
5. To keep the cross talk minimum, do not allow the capacitance between the two signal wires to grow too large.
6. If necessary provide shielding wire- GND or V_{DD} between the two signals as show below. This effectively turns the interwire capacitance into a capacitance to ground and eliminates interference. An adverse effect of shielding is the increased capacitive load.



CROSS SECTION OF ROUTING LAYERS ILLUSTRATING THE USE OF SHIELDING TO REDUCE CAPACITIVE CROSS TALK

7. The interwire capacitance between signals on different layers can be further reduced by addition of extra routing layers.

- **Impact of Cross talk on Propagation Delay (With respect to CMOS):**



IMPACT OF CROSS TALK ON PROPAGATION DELAY

The circuit schematic illustrates of how capacitive cross talk may result in a data-dependent variation of the propagation delay. Assume that the inputs to the three parallel wires X, Y, and Z experience simultaneous transitions. Wire *Y* (called the victim wire) switches in a direction that is opposite to the transitions of its neighbouring signals *X* and *Z*. The coupling capacitances experience a voltage swing that is double the signal swing, and hence represent an effective capacitive load that is twice as large as C_c- the by now well known *Miller effect*.

Since the coupling capacitance represents a large fraction of the overall capacitance in the deep-submicron dense wire structures, this increase in capacitance is substantial, and has a major impact on the propagation delay of the circuit. Observe that this is a worst-case scenario. If all inputs experience a simultaneous transition in the same direction, the voltage over the coupling capacitances remains constant, resulting in a zero contribution to the effective load capacitance.

The total load capacitance C_L of gate *Y*, hence depends upon the data activities on the neighbouring signals and varies between the following bounds:

$$C_{GND} \leq C_L \leq C_{GND} + 4C_c$$

with C_{GND} the capacitance of node *Y* to ground, including the diffusion and fan out capacitances.

Design Techniques for Circuit Fabrics with Predictable delay:

With cross talk making wire-delay more and more unpredictable, a designer can choose between a number of different methodology options to address the issue, some of which are,

1. Evaluate and improve: After detailed extraction and simulation, the bottlenecks in delay are identified, and the circuit is appropriately modified.
2. Constructive layout generation: Wire routing programs take into account the effects of the adjacent wires, ensuring that the performance requirements are met.
3. Predictable structures: By using predefined, known, or conservative wiring structures, the designer is ensured that the circuit will meet his specifications and that cross talk will not be a show stopper.

- **Capacitive Load (With respect to CMOS):**

The increasing values of the interconnect capacitances, especially those of the global wires, emphasize the need for effective driver circuits that can (dis)charge capacitances with sufficient speed. This need is further highlighted by the fact that in complex designs a single gate often has to drive a large fan-out and hence has a large capacitive load.

Typical examples of large on-chip loads are busses, clock networks, and control wires. The latter include, for instance, reset and set signals. These signals control the operation of a large number of gates, so fan-out is normally high. Other examples of large fan-outs are encountered in memories where a large number of storage cells is connected to a small set of control and data wires.

The capacitance of these nodes is easily in the multi-pico farad range. The worst case occurs when signals go off-chip. In this case, the load consists of the package wiring, the printed circuit board wiring, and the input capacitance of the connected ICs or components.

Typical off-chip loads range from 20 to 50 pF, which is multiple thousand times larger than a standard on-chip load. Driving those nodes with sufficient speed becomes one of the most crucial design problems.

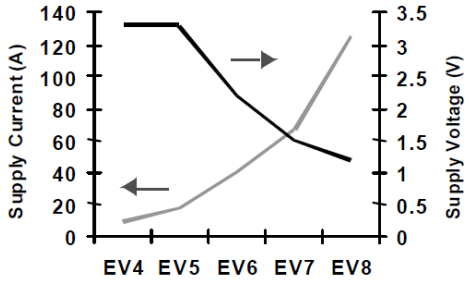
The main secrets to the efficient driving of large capacitive loads are:

1. Adequate transistor sizing is instrumental when dealing with large loads.
2. Partitioning drivers into chains of gradually-increasing buffers helps to deal with large fan out factors.

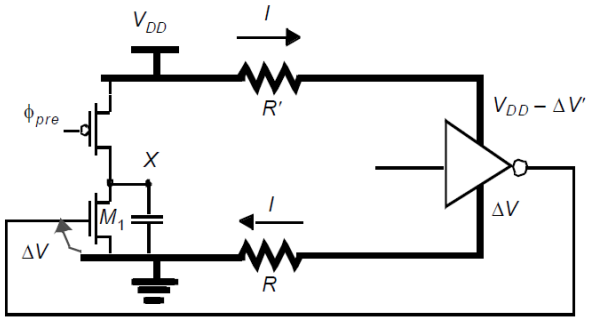
RESISTIVE PARASITICS:

➢ Resistance and Reliability- Ohmic Voltage Drop:

Current flowing through a resistive wire results in an ohmic voltage drop that degrades the signal levels. This is especially important in the power distribution network, where current levels can easily reach amperes as shown below.



EVOLUTION OF POWER SUPPLY CURRENT AND SUPPLY VOLTAGE



OHMIC VOLTAGE DROP ON THE SUPPLY REDUCES NOISE MARGIN

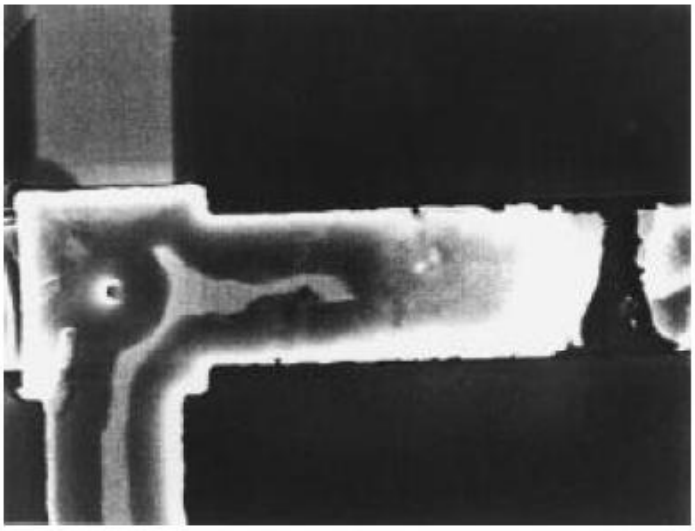
Consider a 2 cm long V_{DD} or *GND* wire with a current of 1mA per μm width. This current is about the maximum that can be sustained by an aluminum wire due to *electromigration* and assuming a sheet resistance of 0.05 Ω/sq, the resistance of this wire (per μm width) equals 1 kΩ. A current of 1 mA/μm would result in a voltage drop of 1 V. The altered value of the voltage supply reduces noise margins and changes the logic levels as a function of the distance from the supply terminals. This is demonstrated by the circuit shown above, where an inverter placed far from the power and ground pins connects to a device closer to the supply.

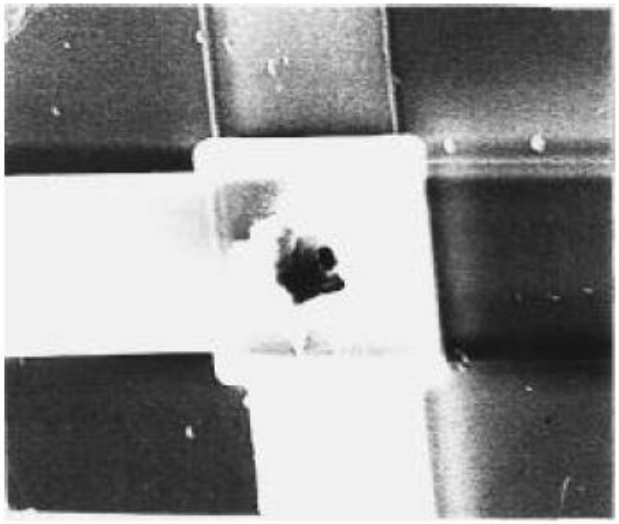
The difference in logic levels caused by the *IR* voltage drop over the supply rails might partially turn on transistor M_1. This can result in an accidental discharging of the pre charged, dynamic node *X*, or cause static power consumption if the connecting gate is static. In short, the current pulses from the on-chip logic, memories and I/O pins cause voltage drops over the power- distribution network and are the major source for on- chip power supply noise. Beyond causing a reliability risk, IR drops on the supply network also impact the performance of the system. A small drop in the supply voltage may cause a significant increase in delay.

The most obvious problem is to reduce the maximum distance between the supply pins and the circuit supply connections which is most easily accomplished through a structured layout of the power distribution network. A number of on- chip power distribution networks with peripheral bonding.

➢ Electromigration:

The current density (current per unit area) in a metal wire is limited due to an effect called *electromigration*. A *direct* current in a metal wire running over a substantial time period, causes a transport of the metal ions. Eventually, this causes the wire to break or to short circuit to another wire. This type of failure will only occur after the device has been in use for some time.





Line Open Failure *Open Failure in Contact Plug*

ELECTROMIGRATION RELATED FAILURE MODES

The rate of the electromigration depends upon the temperature, the crystal structure, and the average current density. The latter is the only factor that can be effectively controlled by the circuit designer. Keeping the current below 0.5 to 1 mA/ μm normally prevents migration. This parameter can be used to determine the minimal wire width of the power and ground network.

Signal wires normally carry an ac- current and are less susceptible to migration. The bidirectional flow of the electrons tends to anneal any damage done to the crystal structure. Most companies impose a number of strict wire-sizing guidelines on their designers, based on measurements and past experience.

Electromigration effects are proportional to the average current flow through the wire, while IR voltage drops are a function of the peak current.

From designing point of view, at the technology level, a number of precautions can be taken to reduce the migration risk i.e.

1. To add alloying elements (such as Cu or Tu) to the aluminum to prevent the movement of the Al ions.
2. To control the granularity of the ions.
3. The introduction of new interconnect materials is a big help as well. For instance, the use of Copper interconnect increases the expected lifetime of a wire with a factor of 100 over Al.

- **Resistance and Performance- RC Delay:**

The delay of a wire grows quadratically with its length. Doubling the length of a wire increases its delay by a factor of four. The signal delay of long wires therefore tends to be dominated by the *RC* effect. This is becoming an ever larger problem in modern technologies, which feature an increasing average length of the global wires, at the same time that the average delay of the individual gates is going down. This leads to the rather bizar situation that it may take multiple clock cycles to get a signal from one side of a chip to its opposite end.

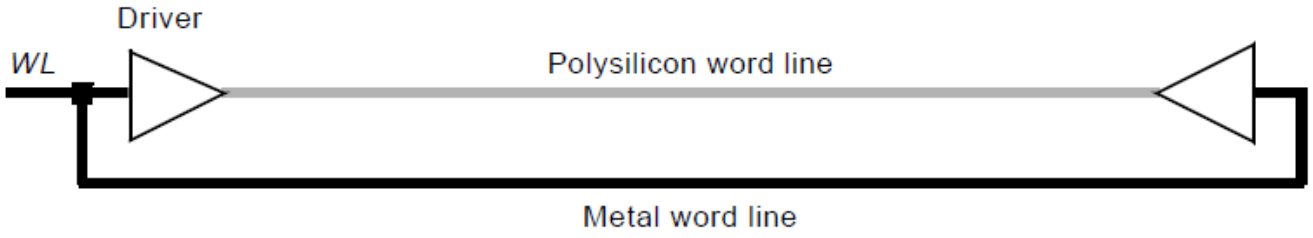
Providing accurate synchronization and correct operation becomes a major challenge under these circumstances. Therefore the different design techniques to cope with the delay imposed by the resistance of the wire are,

- **Better Interconnect Materials:**

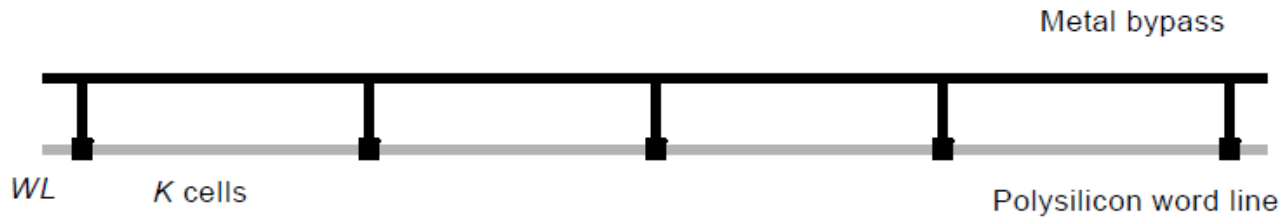
Use better interconnect materials when they are available and appropriate. The introduction of silicides and Copper have helped to reduce the resistance of polysilicon (and diffused) and metal wires, respectively, while the adoption of dielectric materials with a lower permittivity lowers the capacitance. Both Copper and low- permittivity dielectrics have become common in advanced CMOS technologies.

Here the designer should be aware that these new materials only provide a temporary respite of one or two generations, and do not solve the fundamental problem of the delay of long wires. Innovative design techniques are often the only way of coping with the latter.

Sometimes, it is hard to avoid the use of long polysilicon wires. A good example of such circumstance are the address lines in memories, which must connect to a large number of transistor gates. Keeping the wires in polysilicon increases the memory density substantially by avoiding the overhead of the extra metal contacts. The polysilicon- only option unfortunately leads to an excessive propagation delay. One possible solution is to drive the word line from both ends, as shown in Figure. This effectively reduces the worst-case delay by a factor of four. Another option is to provide an extra metal wire, called a *bypass*, which runs parallel to the polysilicon one, and connects to it every *k* cells as shown in figure. The delay is now dominated by the much shorter polysilicon segments between the contacts. Providing contacts only every *k* cells helps to preserve the implementation density.



(a) Driving the word line from both sides



(b) Using a metal bypass

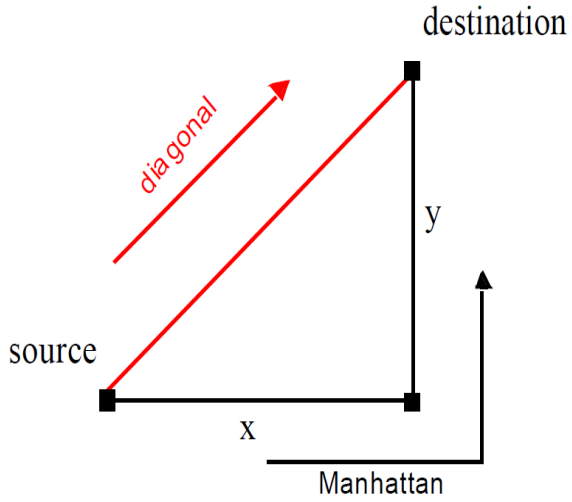
APPROACHES TO REDUCE THE WORD LINE DELAY

- **Better Interconnect Strategies:**

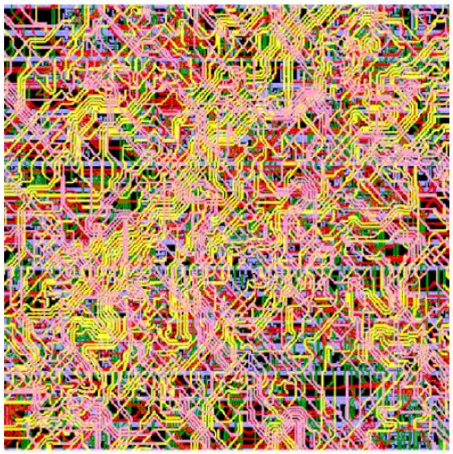
The length of the wire being a prime factor in both the delay and the energy consumption of an interconnect wire, any approach that helps to reduce the wire length is bound to have an essential impact.
There are two wiring strategies i.e. the Manhattan- Style Routing and Diagonal- Style Routing.

- In Manhattan style routing, interconnections are first routed along the one of the preferred directions, followed by a connection in the other direction as shown.

- In Diagonal style routing less size of the wire length is required, on comparison to Manhattan 29% in best case. And the use of 45°lines is ironical in integrated circuits. The main issues of diagonal routing are its complexity, impact on tools and masking concerns.



MANHATTAN VS DIAGONAL ROUTING

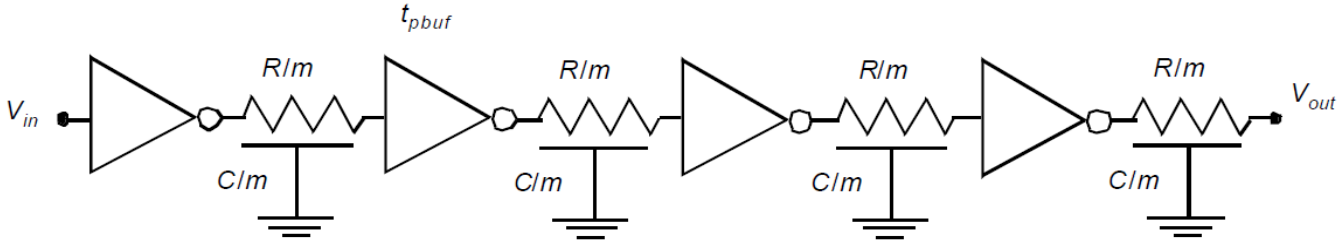


LAYOUT EXAMPLE OF 45° LINES (FOR UNDERSTANDING- NO NEED TO DRAW)

Earlier Manhattan routing was preferred because of the issues of diagonal routing inspite of its features. Now diagonal routing is preferred due to its features i.e. less wire length and 45° lines, its issues of complexity, impact on tools and masking concerns are easily overcomed nowadays by using CAD tools (Computer Aided Design Tools) like Cadence. Therefore the impact on wiring is quite tangible, a reduction of 20% in wire length, resulting in higher performance, lower power dissipation and smaller chip area.

- **Introducing Repeaters/ Buffer Insertion for very long wires:**

The most popular design approach to reducing the propagation delay of long wires is to introduce intermediate buffers, also called *repeaters*, in the interconnect line as shown below.



REDUCING RC INTERCONNECT DELAY BY USING REPEATERS

Making an interconnect line *m* times shorter reduces its propagation delay quadratically, and is sufficient to offset the extra delay of the repeaters when the wire is sufficiently long. Assuming that the repeaters have a fixed delay t_{pbuf}, we can derive the delay of the partitioned wire.

$$t_p = 0.38 rc\left(\frac{L}{m}\right)^2 m + m t_{pbuf}$$

The optimal number of buffers that minimizes the overall delay can be found by setting $\frac{\partial t_p}{\partial m} = 0$,

$$m_{opt} = L\sqrt{\frac{0.38rc}{t_{pbuf}}} = \sqrt{\frac{t_{pwire(unbuffered)}}{t_{pbuf}}}$$

yielding a minimum delay of,

$$t_{p,\,opt} = 2\sqrt{t_{pwire(unbuffered)} t_{pbuf}}$$

and is obtained when the delay of the individual wire segments is made equal to that of a repeater.

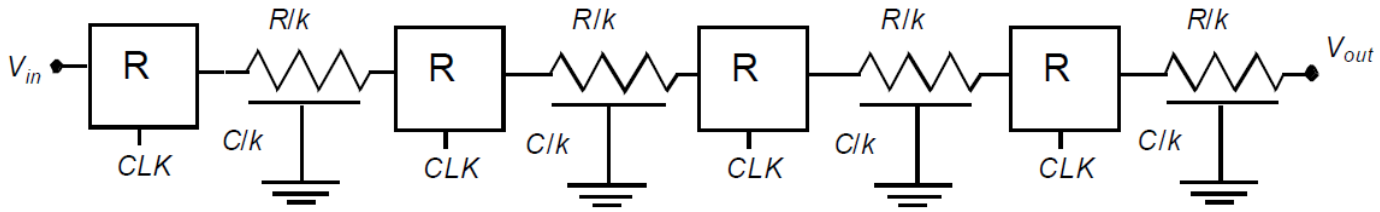
- **Optimizing the Interconnect Architecture:**

Even with buffer insertion, the delay of a resistive wire cannot be reduced below the minimum dictated by Equation,

$$t_{p,\,opt} = 2\sqrt{t_{pwire(unbuffered)} t_{pbuf}}$$

Long wires hence often exhibit a delay that is longer than the clock period of the design. For instance, the 10 cm long Al wire of comes with a minimum delay of 4.7 nsec, even after optimal buffer insertion and sizing, while the 0.25 μm CMOS process featured in this text can sustain clock speeds in excess of 1 GHz (this is, clock periods below 1 nsec). The wire delay all-by-itself hence becomes the limiting factor on the performance achievable by the integrated circuit. The only way to address this bottleneck is to tackle it at the system architecture-level.

Wire pipelining is a popular performance- improvement technique in this category which improves the throughput performance of logic modules with long critical paths. Similar approach can be used to increase the throughput of a wire, as is illustrated in figure below.



WIRE PIPELINING IMPROVES THE THROUGHPUT OF A WIRE

The wire is partitioned in *k* segments by inserting registers or latches. While this does not reduce the delay through the wire segment, it takes *k* clock cycles for a signal to proceed through the wire, it helps to increase its throughput, as the wire is handling *k* signals simultaneously at any point in time. The delay of the individual wire segments can further be optimized by repeater insertion, and should be below a single clock period.

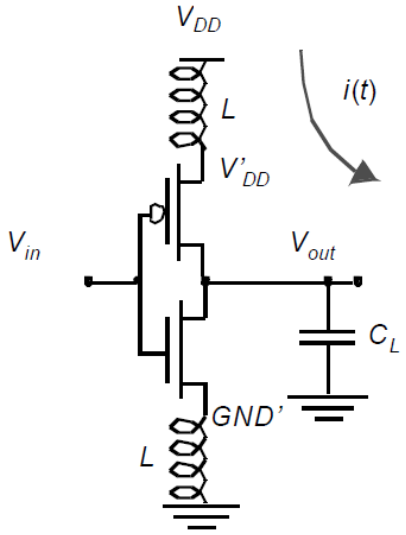
This is only one example of the many techniques that the chip architect has at her disposal to deal with the wire delay problem. The most important concern from this is that the wires have to be considered early on in the design process, and can no longer be treated as an afterthought as was most often the case in the past.

INDUCTIVE PARASITICS:

Interconnect wires also exhibit an inductive parasitic. An important source of parasitic inductance is introduced by the bonding wires and chip packages. Even for intermediate- speed CMOS designs, the current through the input- output connections can experience fast transitions that cause voltage drops as well as ringing and overshooting, phenomena not found in *RC* circuits. At higher switching speeds, wave propagation and transmission line effects can come into the picture.

- **Inductance and Reliability- $L\frac{di}{dt}$ *Voltage Drop*:**

During each switching action, a transient current is sourced from (or sunk into) the supply rails to charge (or discharge) the circuit capacitances as shown. Both V_{DD} and V_{SS} connections are routed to the external supplies through bonding wires and package pins and possess a non ignorable series inductance. Hence, a change in the transient current creates a voltage difference between the external and internal (*V'DD*, GND') supply voltages. This situation is especially severe at the output pads, where the driving of the large external capacitances generates large current surges. The deviations on the internal supply voltages affect the logic levels and result in reduced noise margins.

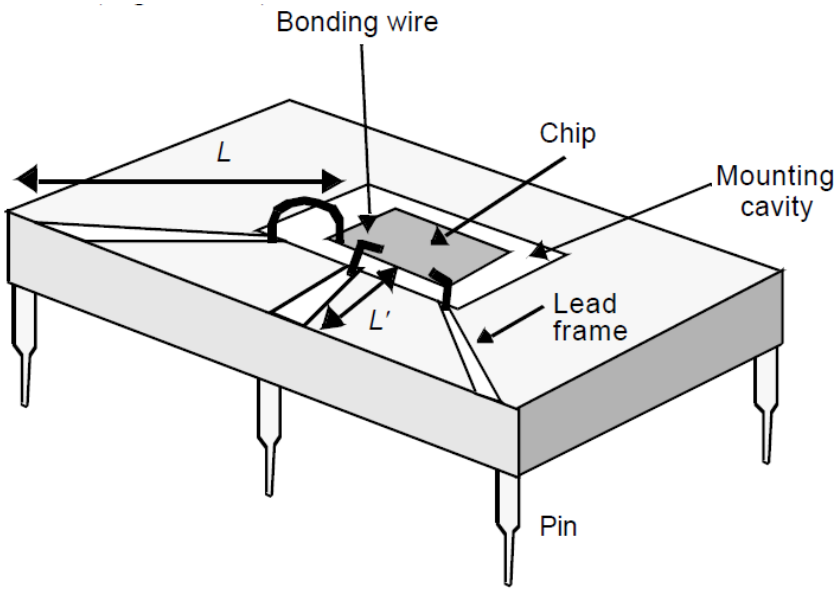


INDUCTIVE COUPLING BETWEEN EXTERNAL AND INTERNAL SUPPLY VOLTAGES

In an actual circuit, a single supply pin serves a large number of gates or output drivers. A simultaneous switching of those drivers causes even worse current transients and voltage drops. As a result, the internal supply voltages deviate in a substantial way from the external ones. For instance, the simultaneous switching of the 16 output drivers of an output bus would cause a voltage drop of at least 1.1 V if the supply connections of the buffers were connected to the same pin on the package. Improvements in packaging technologies are leading to ever- increasing numbers of pins per package. Packages with up to 1000 pins are currently available. Simultaneous switching of a substantial number of those pins results in huge spikes on the supply rails that are bound to disturb the operation of the internal circuits as well as other external components connected to the same supplies.

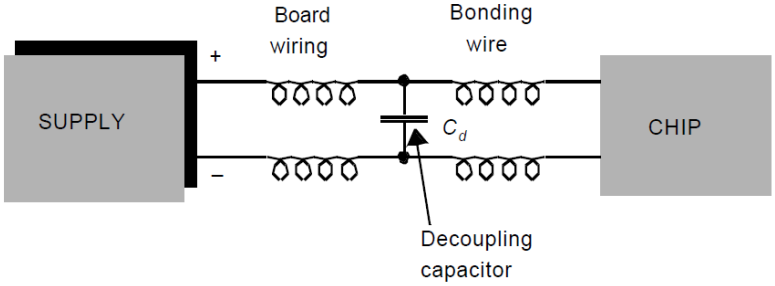
Design techniques to address $L\frac{di}{dt}$ *Voltage Drop* problem:

1. Separate pins for I/O pads and chip core. Since the I/O drivers require the largest switching currents, they also cause the largest current changes. Therefore, it is wise to isolate the core of the chip where most of the logic action occurs, from the drivers by providing different power and ground pins.
2. Multiple power and ground pins in order to reduce the $\frac{di}{dt}$ per supply pin, we can restrict the number of I/O drivers connected to a single supply pin.
3. Careful selection of positions of the power and ground pins on the package. The inductance of pins located at the corners of the package is substantially higher as shown below.



THE INDUCTANCE OF A BONDING WIRE/ PIN COMBINATION DEPENDS UPON THE PIN POSITIONS

4. Schedule current consuming transitions so that they do not occur simultaneously.
5. Increase the rise and fall times of the off-chip signals to the maximum extent allowable, and distributed all over the chip, especially under the data busses.
6. Use advanced packaging technologies such as surface-mount or hybrids that come with a substantially reduced capacitance and inductance per pin.
7. Adding decoupling capacitances on the board. These capacitances, which should be added for every supply pin, act as local supplies and stabilize the supply voltage seen by the chip. They separate the bonding- wire inductance from the inductance of the board interconnect as shown below. The bypass capacitor, combined with the inductance, actually acts as a low- pass network that filters away the high-frequency components of the transient voltage spikes on the supply lines.



DECOUPLING CAPACITORS ISOLATE THE BOARD INDUCTANCE FROM THE BONDING WIRE AND PIN INDUCTANCE

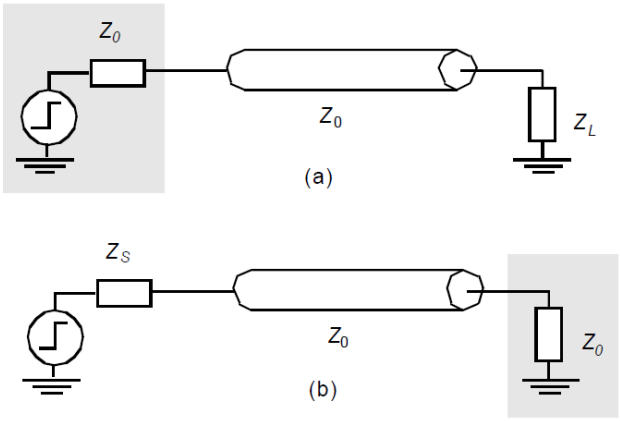
➢ Inductance and Performance- Transmission Line Effects:

When an interconnection wire becomes sufficiently long or when the circuits become sufficiently fast, the inductance of the wire starts to dominate the delay behaviour, and transmission line effects must be considered. This is more precisely the case when the rise and fall times of the signal become comparable to the time of flight of the signal waveform across the line as determined by the speed of light. As advancing technology increases line lenghts and switching speeds, this situation is gradually becoming common in fastest CMOS circuits as well, and transmission- line effects are bound to become a concern of the CMOS designer as well.

Some of the techniques to minimize the impact of the transmission line behaviour are:

o Termination:

To avoid the negative effects of transmission-line behaviour such as ringing or slow propagation delays, the line should be terminated, either at the source (series termination), or at the destination (parallel termination) with a resistance matched to its characteristic impedance Z_0.



MATCHED TERMINATION SCENARIOS FOR WIRES BEHAVING AS TRANSMISSION LINES: (a) SERIES TERMINATION AT THE SOURCE, (b) PARALLEL TERMINATION AT THE DESTINATION

The two scenarios- series and parallel termination as shown are depicted in figure. Series termination requires that the impedance of the signal source is matched to the connecting wire. This approach is appropriate for many CMOS designs, where the destination load is purely capacitive. The impedance of the driver inverter can be matched to the line by careful transistor sizing.

o Shielding

If we want to control the behaviour of a wire behaving as a transmission line, we should carefully plan and manage how the return current flows. A good example of a well-defined transmission line is the coaxial cable, where the signal wire is surrounded by a cylindrical ground plane. To accomplish similar effects on a board or on a chip, designers often surround the signal wire with ground (supply) planes and shielding wires. Being shielding, adding shielding makes the behaviour and the delay of an interconnection a lot more predictable. Yet even with these precautions, powerful extraction and simulation tools will be needed in the future for the high-performance circuit designer.

Assignment:

1. Explain RC versus lumped C.
2. Explain interwire capacitance and cross talk.
3. Illustrate the behavior of various transmission line terminations.