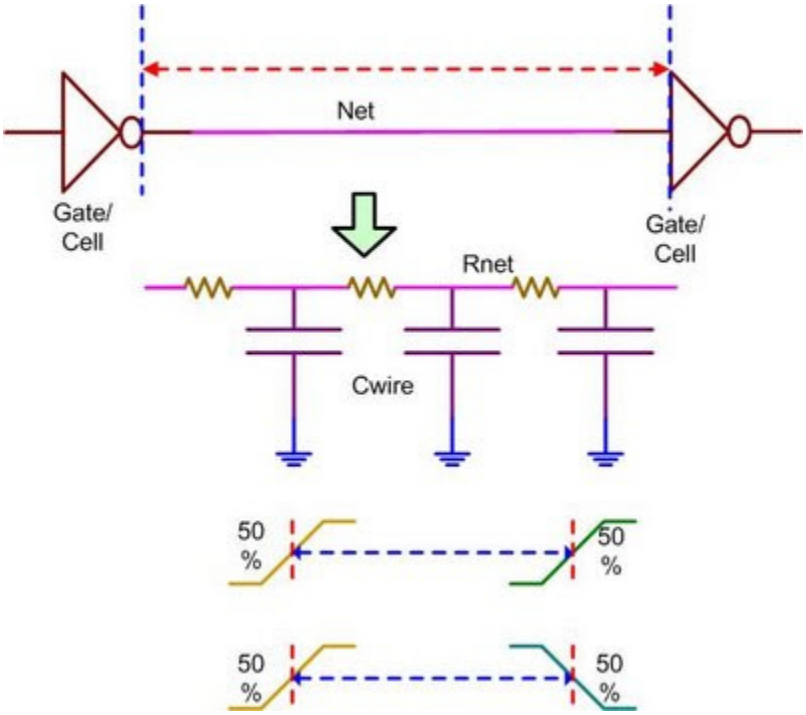
NET DELAY

Net delay is the difference between the time a signal is first applied to the net and the time it reaches other devices connected to that net.

It is due to the finite resistance and capacitance of the net. It is also known as **wire delay.**

Wire delay = function of (Rnet, Cnet+Cpin)

This is output pin of the cell to the input pin of the next cell.



Net delay is calculated using Rs and Cs.

There are several factors which affect net parasitic:

- **Net Length**
- **Net cross-sectional area**
- **Resistively of material used for metal layers (Aluminum vs. copper)**
- **Number of vias traversed by the net**
- **Proximity to other nets (crosstalk)**

Post-layout design is annotated with RCs extracted from layout for better accuracy. Annotated RCs override information from WLM.

Interconnect introduces capacitive, resistive and inductive parasites. All three have multiple effects on the circuit behavior.

1. Interconnect parasites cause an **increase in propagation delay** (i.e. it slows down working speed)
2. Interconnect parasites **increase energy dissipation** and affect the power distribution.
3. Interconnect parasites **introduce extra noise sources**, which affect reliability of the circuit. (Signal Integrity effects)

Dominant parameters determine the circuit behavior at a given circuit node. Non-dominant parameters can be neglected for interconnect analysis.

- Inductive effect can be ignored if the resistance of the wire is substantial enough-this is 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 capacitive only model can be used.
- When the separation between neighboring wires is large or when the wires only run together for short distance, inter-wire capacitance can be ignored, and all the parasitic capacitance can be modeled as capacitance to ground.

~~~~~~

Capacitance

Capacitance can be modeled by the parallel plate capacitor model.

C = (ε / t).WL

Where

ε --> permittivity of dielectric material (SiO2)

t --> thickness of dielectric material (SiO2)

W --> width of wire

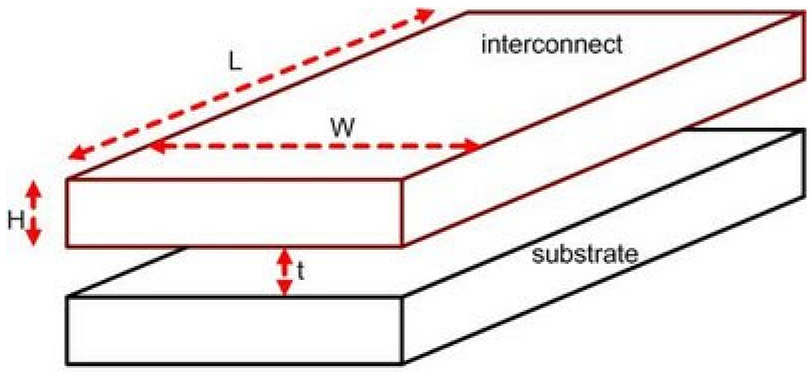
L --> length of wire

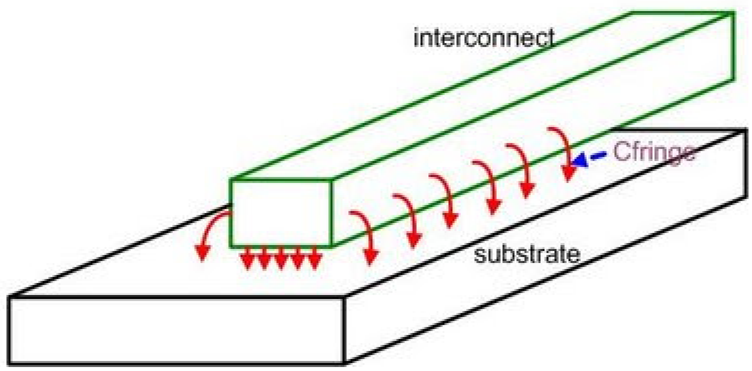
ε --> $\varepsilon_r \varepsilon_o$ where ε_r --> relative permittivity of SiO2

ε_o --> 8.854 x 10-12 F/m; permittivity of free space

As technology node shrinks (scaling), to minimize resistance of the wires, it is desirable to keep the cross section of the wire (WxH) as large as possible. But this increases area.
~~~~~~

Small values of W lead to denser wiring and less area overhead. In advanced process W/H ratio has reduced below unity. Under such circumstances parallel plate capacitance model becomes inaccurate. The capacitance between the sidewall of the wires and substrate called fringing capacitance can no longer be ignored and contributes to the overall capacitance.





Inter-wire capacitance become dominant factor in multilayer interconnect structures. These floating capacitors (not connected to substrate or ground) form a source of noise (cross talk). This effect is more pronounced for wires in the higher interconnect layer, as these are farther away from the substrate.

Generally higher metal layers (i.e. interconnects) have higher thickness (i.e. height) and higher dielectric layers have higher permittivity. Hence these wires display the highest inter-wire capacitance. Hence use it for global signals that are not sensitive to interference. (eg. Supply rails). Or it is advisable to separate wires by an amount that is larger than minimum spacing.

~~~~~~

Resistance

Resistance R= (ρ.L)/ (H.W) = (ρ. L)/ Area
~~~~~~

L --> length

W --> width

ρ --> resistivity (ohm-m)

Since H (height, thickness) is constant for a given technology we can write: R = Rs.(L/W) where Rs=ρ/H ohm/sqare is called "**sheet resistance**".

At very high frequencies "**skin effect**" comes into play such that the resistance becomes frequency dependent. High frequency currents tend to flow primarily on the surface of a conductor, with the current density falling off exponentially with depth into the conductor.

Skin effect is only an issue for wider wires. Since clocks tends to carry the highest frequency signals on a chip and also fairly wide to limit resistance, the skin effect likely to have its first impact on these lines.

~~~~~~

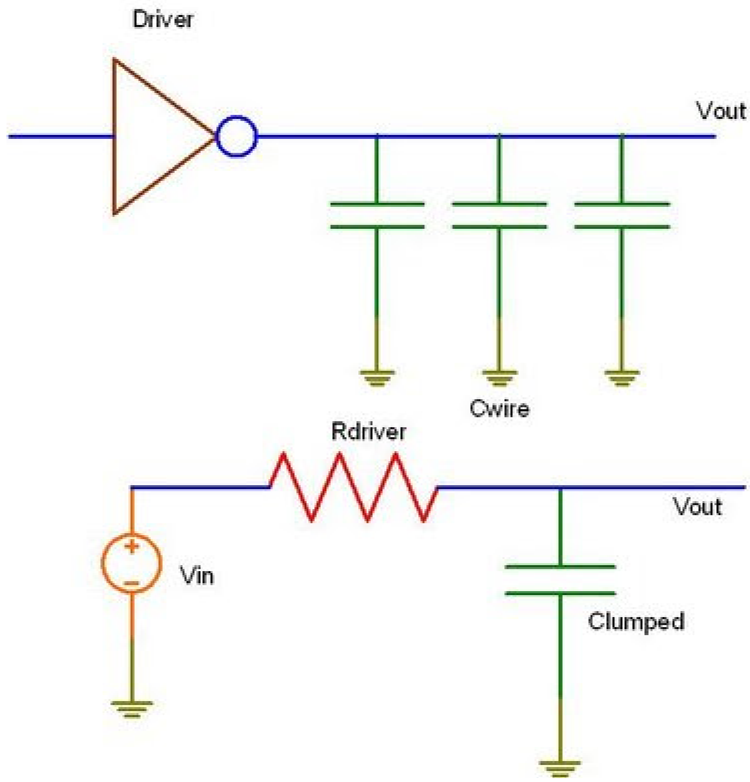
Inductance

With the adoption of low resistance interconnect materials and the increase of switching frequencies to GHz range, inductance starts to an important role. Consequences of on chip inductance include **ringing and overshoot effect, reflection of signals due to impedance mismatch, inductive coupling between lines, and switching noise due to (Ldi/dt) voltage drops.**

~~~~~~

Lumped Capacitor Model

As long as the **resistive component of the wire is small, and 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 capacitance.



The only impact on performance is introduced by the loading effect of the capacitor on the driving gate.

~~~~~~

Lumped RC Model

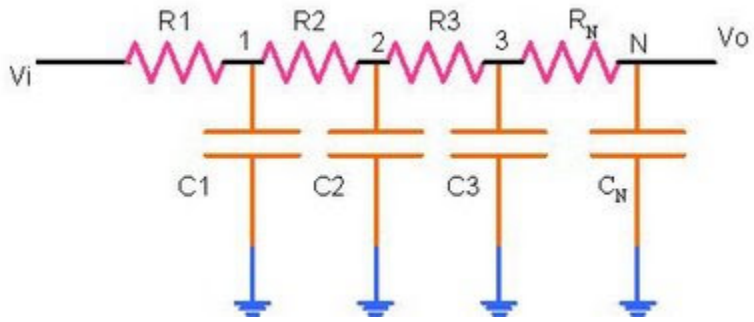
If wire length is more than a few millimeters, the lumped capacitance model is inadequate and a resistive capacitive model has to be adopted.

In lumped RC model the total resistance of each wire segment is lumped into one single R, combines the global capacitive into single capacitor C.

Analysis of network with larger number of R and C becomes complex as network contains many time constants (zeroes and poles). **Elmore delay model overcome such problem**.

~~~~~~

Elmore Delay Model



Properties of the network:

- Has single input node
- All the capacitors are between a node and ground.
- Network does not contain any resistive loops.

"**Path resistance**" is the resistance from source node to any other node.

"**Shared path resistance**" is the resistance shared among the paths from the source node to any other two nodes.

Hence,

Delay at node 1: Tow d1 = R1C1

Delay at node 2: Tow d2= (R1+R2)C2

Delay at node 3: Tow d3 = (R1+R2+R3)C3

In general:

T_{di}=R1C1+(R1+R2)C2+……..+(R1+R2+R3+…..+Ri)Ci

If

R1=R2=R3=….=R

C1=C2=C3=…..C then

T_{di}=RC+2RC+……..+nRC

Thus Elmore delay is equivalent to the first order time constant of the network.

Assuming an interconnect wire of length L is partitioned into N identical segments. Each segment has length L/N.

Then,

T_d=L/N.R.L/N.C+ 2 (L/n.r+L/N.C)+……

=(L/N)2(RC+2RC+…….+NRC)

=(L/N)2. N(N+1)

or $T_d=RC.L^2/2$

=> The delay of a wire is a quadratic function of its length

=> **doubling the length of the wire quadruples its delay**

Advantages

- It is simple
- It is always situated between minimum and maximum bounds

Disadvantages

- It is **pessimistic** and inaccurate for long interconnect wires.

~~~~~~

Distributed RC model

Lumped RC model is always pessimistic and distributed RC model provides better accuracy over lumped RC model.

But distributed RC model is complex and no closed form solution exists. Hence distributed RC line model is not suitable for Computer Aided Design Tools.

The behavior of the distributed RC line can be approximated by a lumped RC ladder network such as Elmore Delay model hence these are extensively used in EDA tools.
~~~~~~

Transmission Line Model

When frequency of operation increases to a larger extent, rise (or fall) time of the signal becomes comparable to time of flight of the net, then inductive effects starts dominating over RC values.
This inductive effect is modeled by Transmission Line models. The model assumes that the

signal is a "wave" and it propagates over the medium "net".

There are two types of transmission models:

Lossless transmission line model: This is good for Printed Circuit Board level design.

Lossy transmission line model: This model is used for IC interconnect model.

Transmission line effects should be considered when the rise or fall time of the input signal is smaller than the time of flight of the transmission line or resistance of the wire is less than characteristics impedance.
~~~~~~

Wire Load Models

Extraction data from already routed designs are used to build a lookup table known as the wire load model (WLM). WLM is based on the statistical estimates of R and C based on "Net Fan-out".

For fanouts greater than those specified in a wire load table, a "slope factor" is specified for linear extrapolation.

wire_load ("5KGATES") {

resistance : 0.000271 **-------------> R per unit length**

capacitance : 0.00017 **-------------> C per unit length**

slope : 29.4005 **---------------------> Used for linear extrapolation**

fanout_length (1, 18.38) **----------> (fanout = 1, length = 18.38)**

fanout_length (2, 47.78)

fanout_length (3, 77.18)

fanout_length (4, 106.58)

fanout_length (5, 135.98)

}

Eg:

Fanout = 7
~~~~~~

Net length = 135.98 + 2 x 29.4005 (slope) = 194.78 ----------> length of net with fanout of 7
Resistance = 194.78 x 0.000271 = 0.05279 units
Capacitance = 194.78 x 0.00017 = 0.03311 units

~~~~~~

Wire load models for synthesis

Wire load modeling allows us to estimate the effect of wire length and fanout on the resistance, capacitance, and area of nets. Synthesizer uses these physical values to calculate wire delays and circuit speeds. Semiconductor vendors develop wire load models, based on statistical information specific to the vendors' process. The models include coefficients for area, capacitance, and resistance per unit length, and a fanout-to-length table for estimating net lengths (the number of fanouts determines a nominal length).

Selection of wire load models in the initial stage (before physical design) depends on the fallowing factors:

1. User specification

2. Automatic selection based on design area

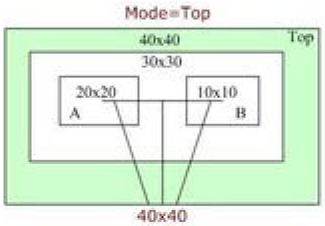
3. Default specification in the technology library

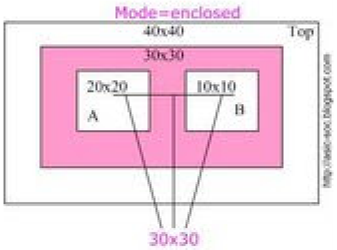
Once the final routing step is over in the physical design stage, wire load models are generated based on the actual routing in the design and synthesis is redone using those wire load models.

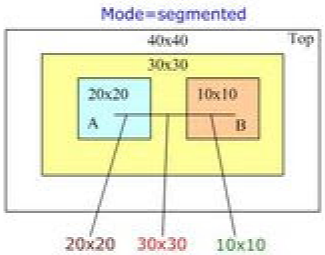
In hierarchical designs, we have to determine which wire load model to use for nets that cross hierarchical boundaries. There are three modes for determining which wire load model to use for nets that cross hierarchical boundaries:

Top:

Applying same wire load models to all nets as if the design has no hierarchy and uses the wire load model specified for the top level of the design hierarchy for all nets in a design and its sub designs.
~~~~~~







Enclosed:

The wire load model of the smallest design that fully encloses the net is applied. If the design enclosing the net has no wire load model, then traverses the design hierarchy upward until we finds a wire load model. Enclosed mode is more accurate than top mode when cells in the same design are placed in a contiguous region during layout.

Use enclosed mode if the design has similar logical and physical hierarchies.

Segmented:

Wire load model for each segment of a net is determined by the design encompassing the segment. Nets crossing hierarchical boundaries are divided into segments. For each net segment, the wire load model of the design containing the segment is used. If the design contains a segment that has no wire load model, then traverse the design hierarchy upward until it finds a wire load model.

~~~~~~

Interconnect Delay vs. Deep Sub Micron Issues

Performances of deep sub micron ICs are limited by increasing interconnect loading affect. Long global clock networks account for the larger part of the power consumption in chips. Traditional CAD design methodologies are largely affected by the interconnect scaling.
~~~~~~

Capacitance and resistance of interconnects have increased due to the smaller wire cross sections, smaller wire pitch and longer length. This has resulted in increased RC delay. As technology is advancing scaling of interconnect is also increasing. In such scenario increased RC delay is becoming major bottleneck in improving performance of advanced ICs.

Here the gate delay and the interconnect delay are shown as functions of various technology nodes ranging from 180nm to 60nm. The interconnect delays shown assumes a line where repeaters are connected optimally and includes the delay due to the repeaters. From the graph it can be observed that with the shrinking of technology gate delay reduces but interconnect delay increases.

~~~~~~

Limits of Cu/low-k interconnects

At submicron level of 250 nm copper with low-k dielectric was introduced to decrease affects of increasing interconnect delay. But below 130 nm technology node interconnect delays are increasing further despite of introducing low-k dielectric. As the scaling increases new physical and technological effects like ***resistivity*** and ***barrier thickness*** start dominating and interconnect delay increases. Introduction of repeaters to shorten the interconnect length increases total area. The vias connecting repeaters to global layers can cause blockage in lower metal layers. Thus as the technology improves material limitations will dominate factor in the interconnect delay. Increasing metal layer width will cause increase in metallization layer. This can't be a solution for the problem as it increases complexity, reliability and cost.

Cu low-k dielectric films are deposited by a special process known as ***Damascene process***. Adhesion property of Cu with dielectric materials is very poor. Under electric bias they easily drift and cause short between metal layers. To avoid this problem a barrier layer is deposited between dielectric and Cu trench. Even though it decreases effective cross section of interconnects compared to drawn dimensions, it improves reliability. The barrier thickness becomes significant in deep submicron level and effective resistance of the interconnect rises further. In addition to this increasing electron scattering and self heating caused by the electron flow in interconnects due to comparable increase in internal chip temperature also contribute to increase interconnect resistance.

Source : http://asic-soc.blogspot.in/2008/10/net-delay.html
~~~~~~