Lecture 2

DSM Interconnect

R. Saleh
Dept. of ECE
University of British Columbia
res@ece.ubc.ca

Overview of Lecture

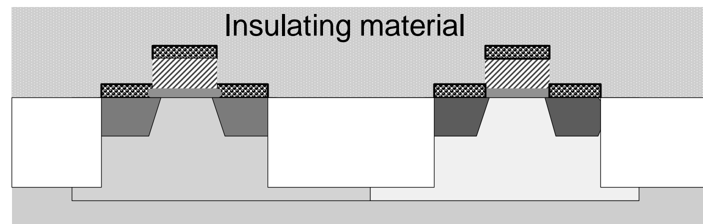
- Deep submicron interconnect issues have been plaguing the integrated circuit designer for about 15 years
- The problems still persist, but there are some well-known solutions to the most critical issues
- This lecture will cover fabrication and reliability issues, followed by the effect of technology scaling on wires: resistance, coupling capacitance and inductance
- We will also discuss solutions to these problems
- References:

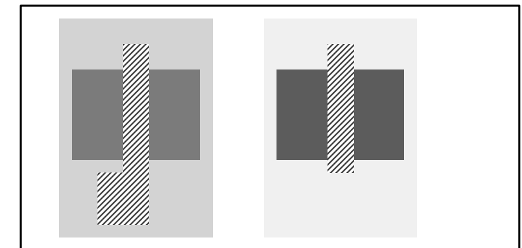
1) "Inteconnect Design", HJS Textbook, Chapter 10

2) "The Future of Wires", Ho et al., Proc. of IEEE, April 2001

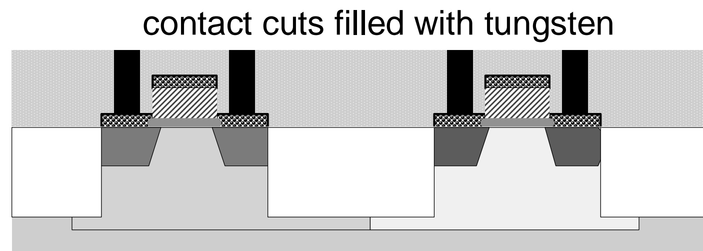
Making Wires

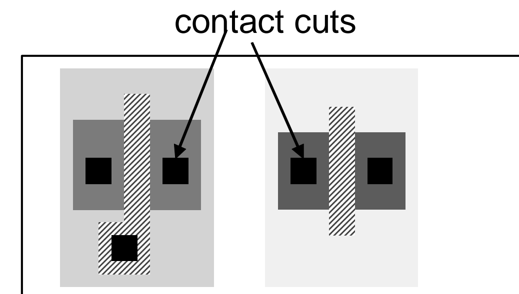
1. Deposit insulator
   may be polished
   to make it flat





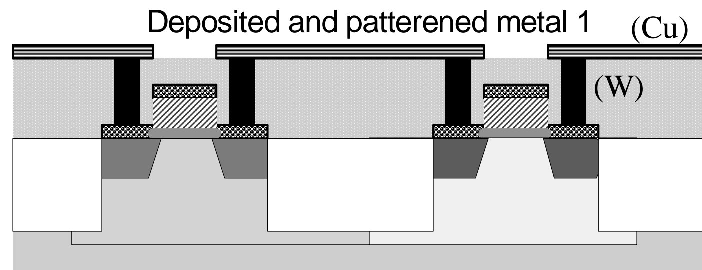
2. Etch contacts to Si
   fill with conductor

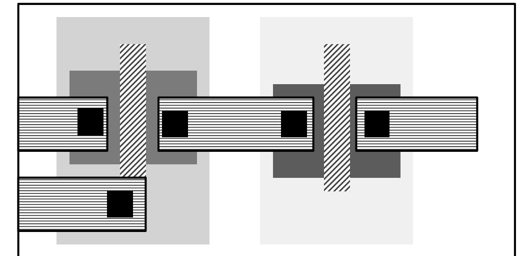




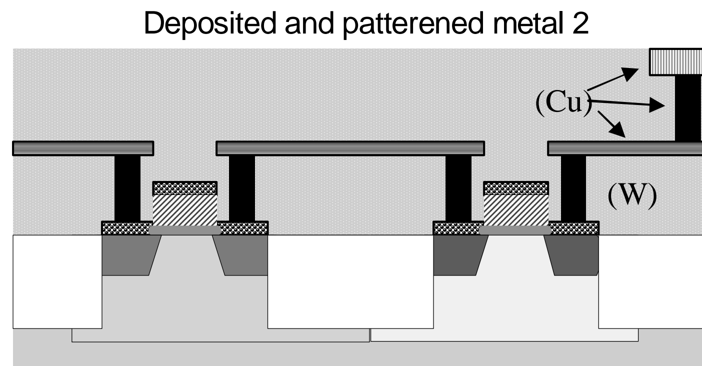
Making Wires (cont'd)

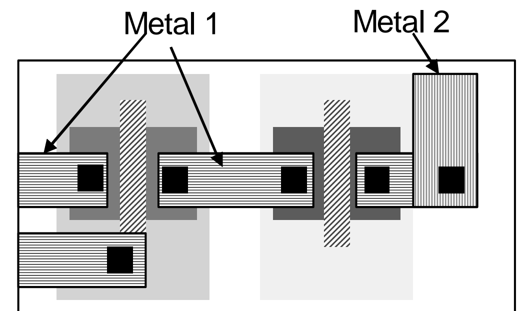
3. Deposit first metal layer and then pattern to provide desired connections



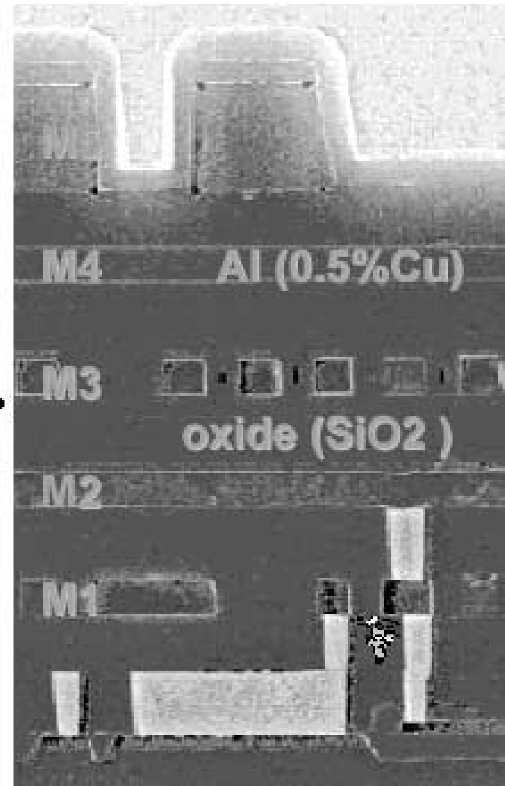


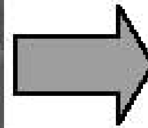
4. Repeat same steps for all subsequent layers

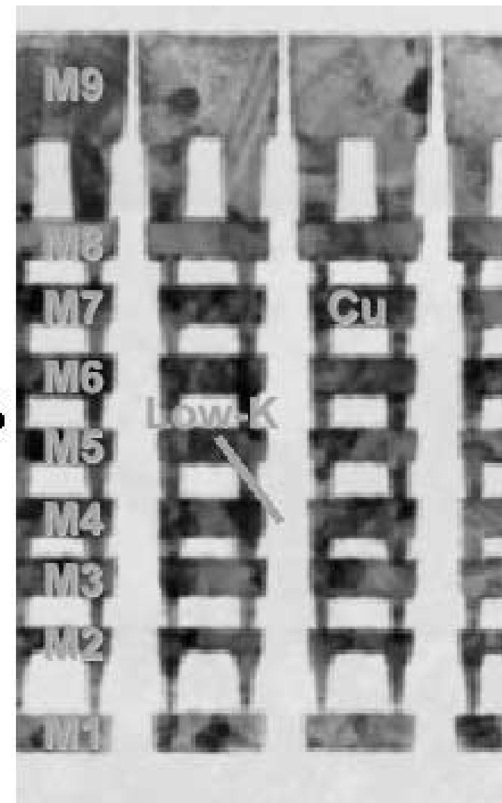




Final Profile in 0.25um vs. 90nm





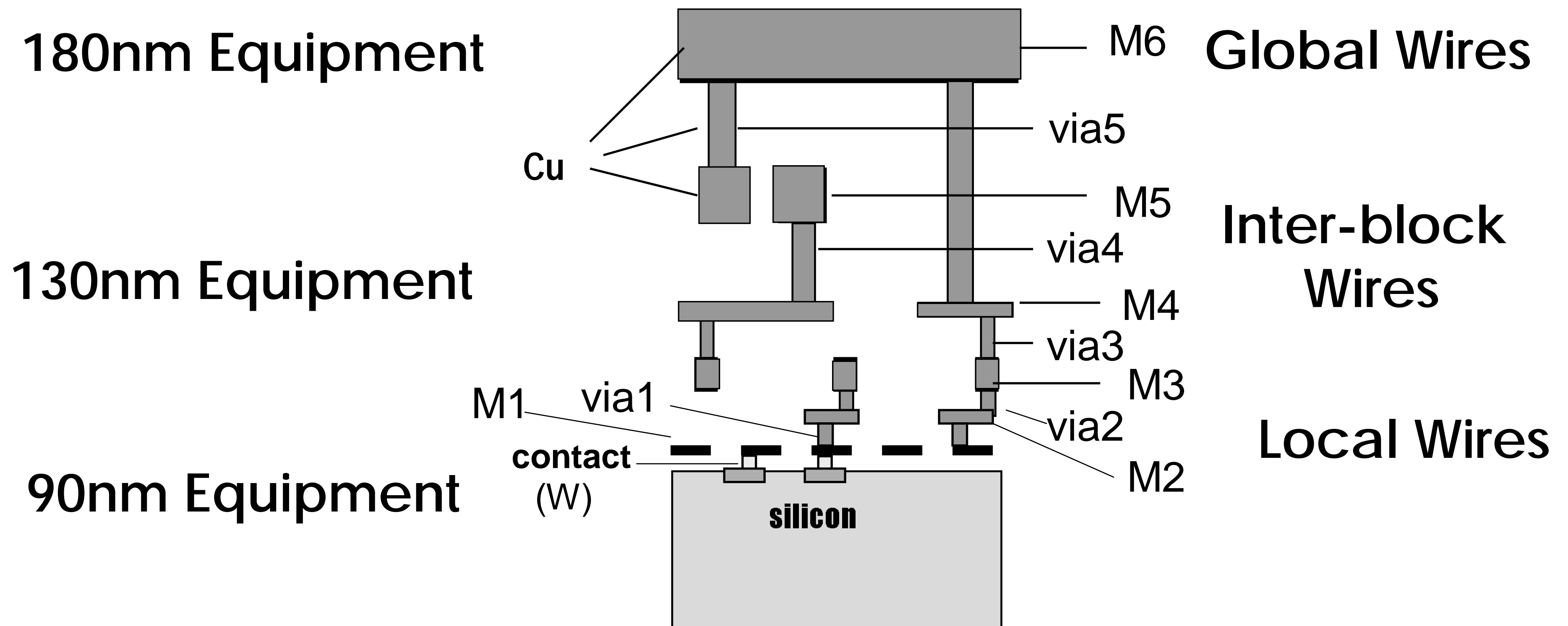


0.25µm CMOS
(Motorola, 1996)

90nm CMOS
(TSMC, 2002)

Purpose of Multi-level Wires

Today, wires are made of copper (both metal and via)

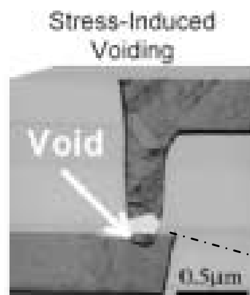


Reliability and Fabrication Issues

Via Reliability

Via Failures at Upper Levels

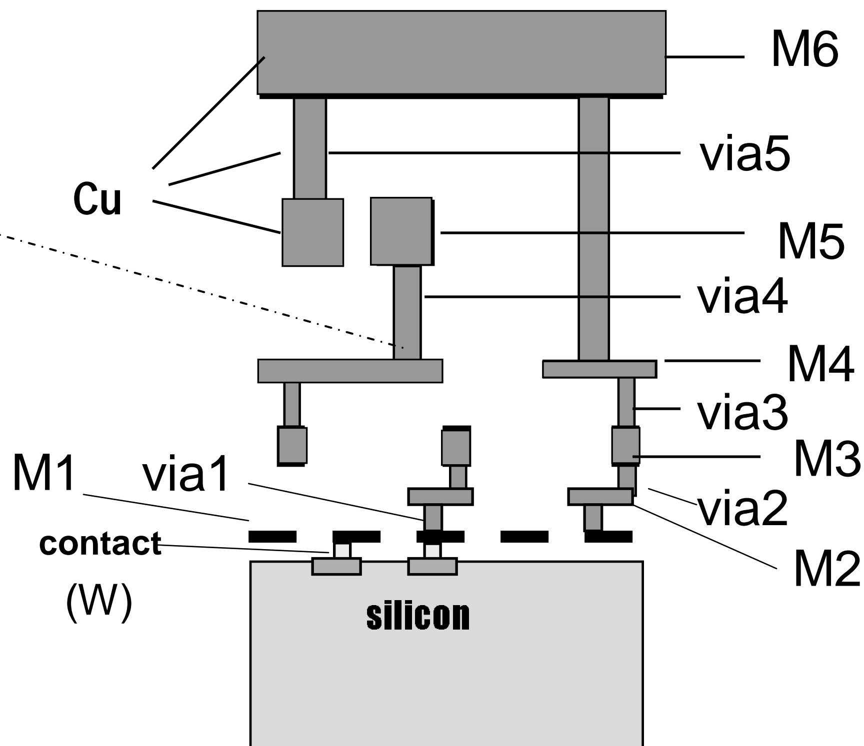
Need to Use Redundant Vias



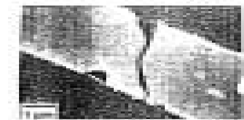
Photolithography

Photolitho-graphic problems at lower levels

Need OPC/PSM



Electromigration

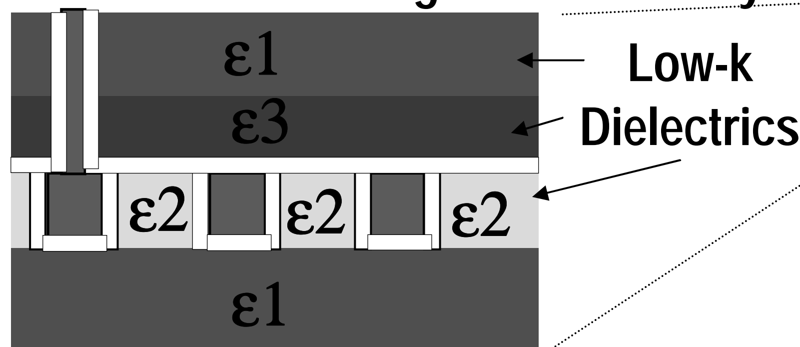


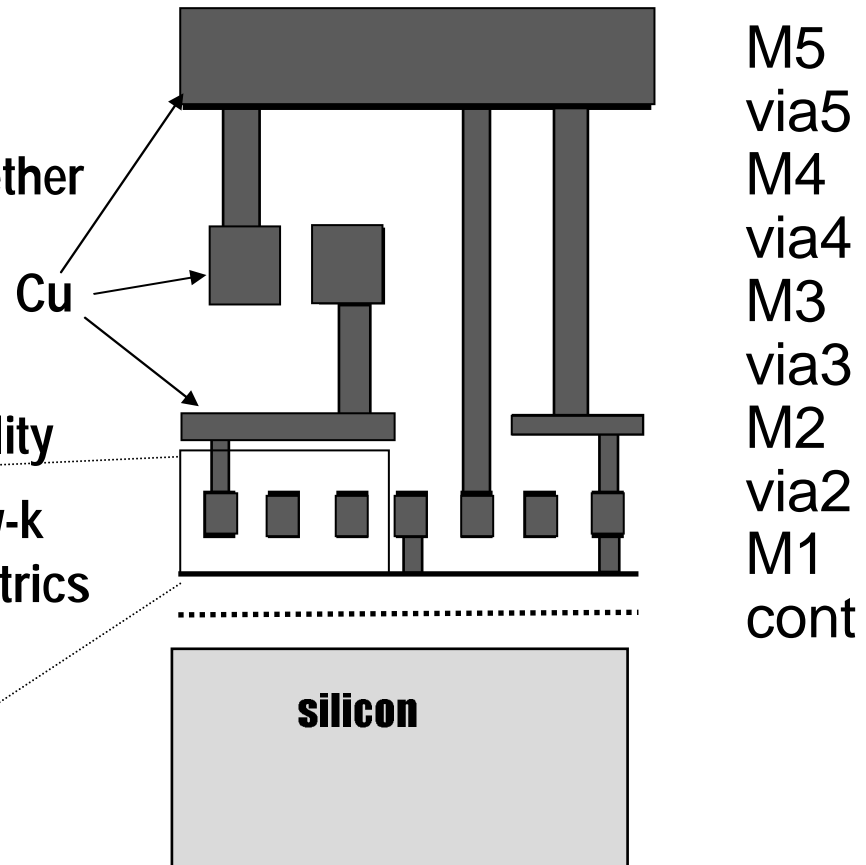
Long-term reliability problems of electromigration

Need to widen wires or reduce current density in wire

Reducing Coupling Capacitance

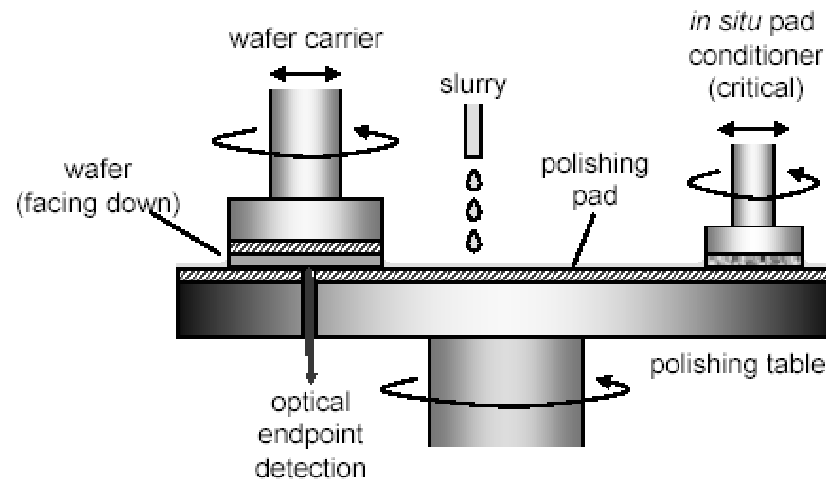
- **Copper and low-k dielectrics**
- **Dual Damascene process**
  - metal and vias fabricated together
  - try to reduce k from 3.9 to 2
    - **lower coupling caps**
  - better electromigration reliability





Multiple Levels of Metal

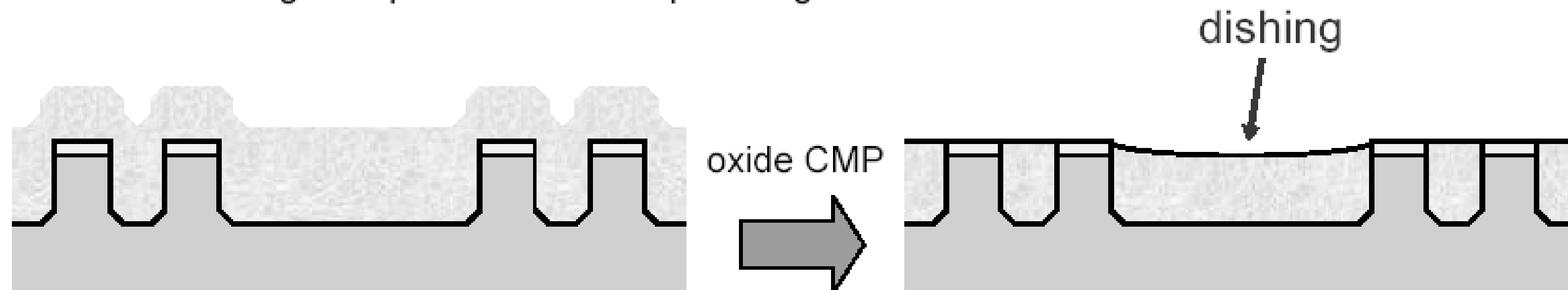
Chemical Mechanical Polishing



- Ideal world for CMP: want *perfect* periodicity of patterns throughout wafer
- Need to throw in dummy features to minimize pattern density variations → optimize planarity
- Polishing pad will flex

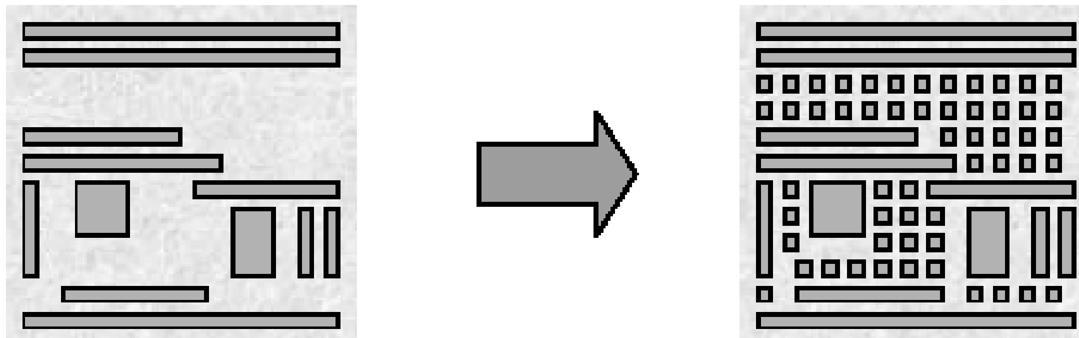
CMP technology pioneered by IBM

- Leveraged expertise from lens polishing

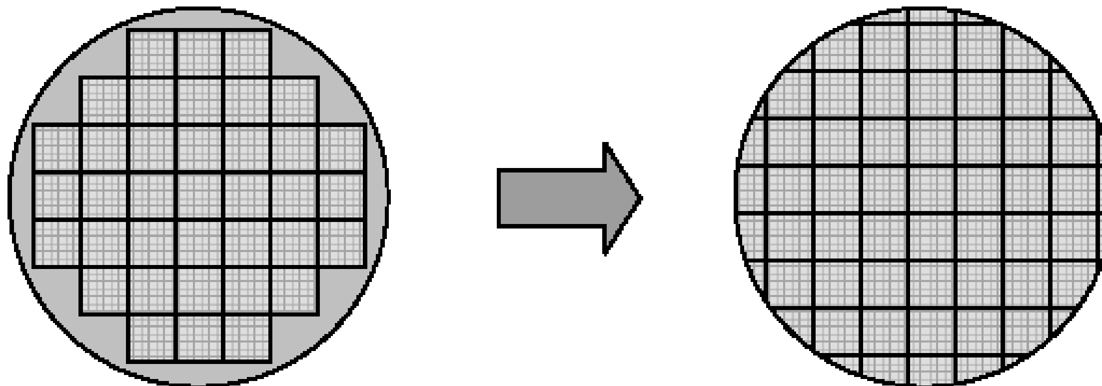


Metal Fill Patterns

- Use dummy metal fill to reduce uneven topography on a layer

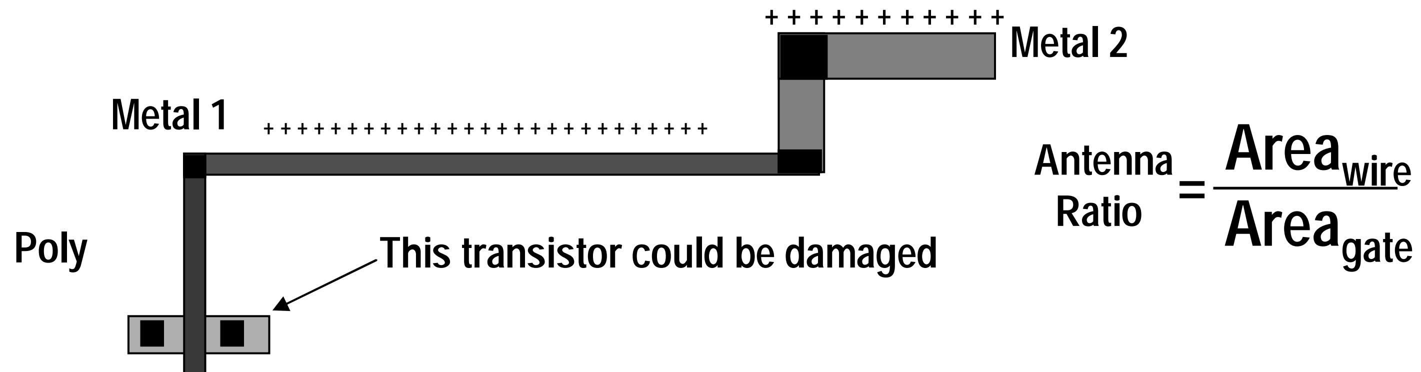


- Also critical to step dummy dies around wafer periphery



Antenna Effects

- As each metal layer is placed on the chip during fabrication, charge builds up on the metal layers due to CMP[1], etc.
- If too much charge accumulates on gate of MOS transistor, it could damage the oxide and short the gate to the bulk terminal

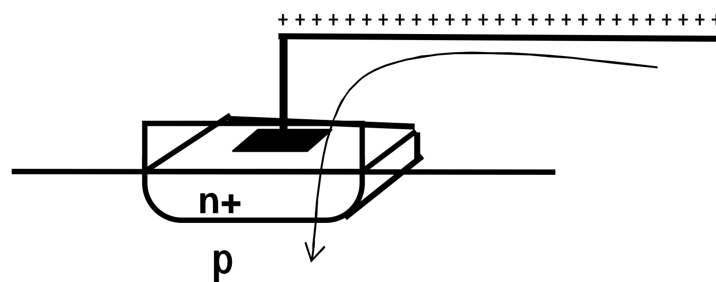


- Higher levels of metal accumulate more charge so they are more troublesome (i.e., metal 5 is worse than metal 1)
- Need to discharge metal lines during processing sequence to avoid transistor damage (becomes a design/layout issue)

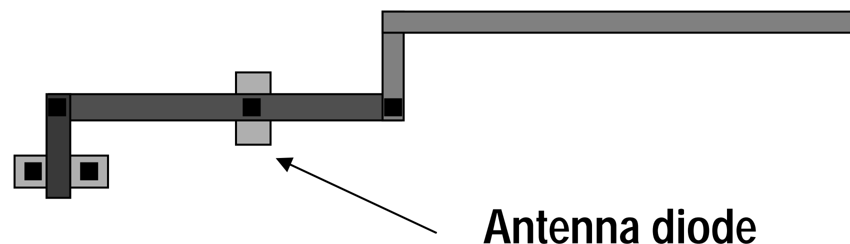
1. CMP is chemical mechanical polishing which is used to planarize each layer before the next layer is placed on the wafer.

Preventing Antenna Effects

- A number of different approaches for antenna repairs:
- Diode Insertion - Make sure all metal lines are connected to diffusion somewhere to discharge the metal lines during fabrication



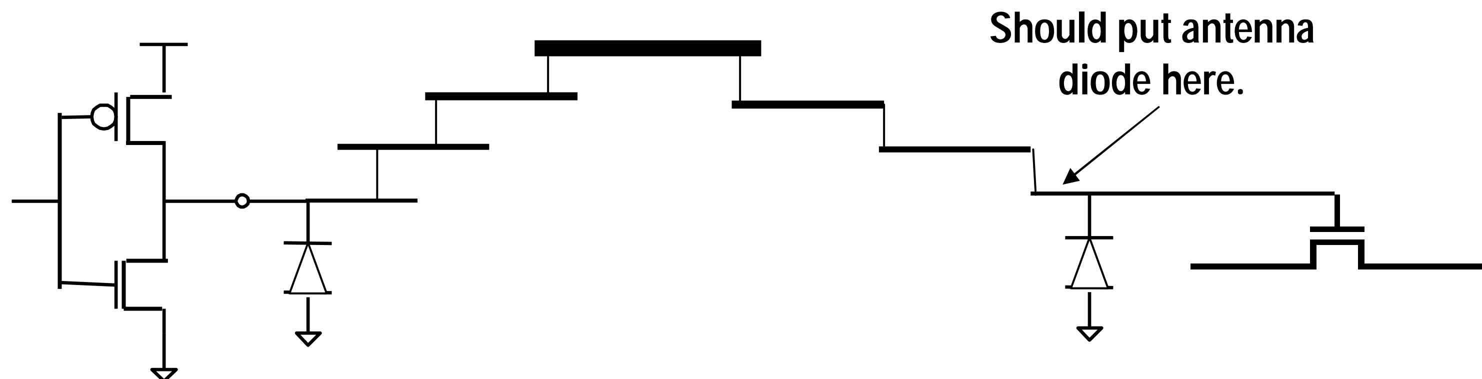
-diodes costs area
- need to optimize number and location
- causes problems for design verification tool



Antenna diode

Preventing Antenna Effects

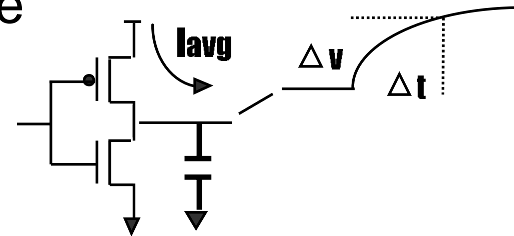
- Note that there are always diodes connecting to source/drain regions of all transistors and charge on each layer is drained before next layer is added…so why are we worried?



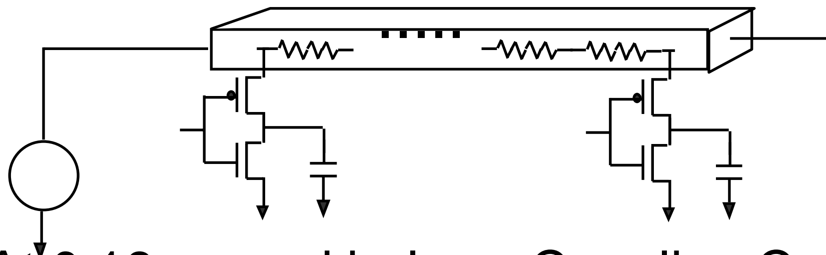
- Gate input of next device may not be connected to a diode until it's too late…charge accumulation on metal exceeds threshold

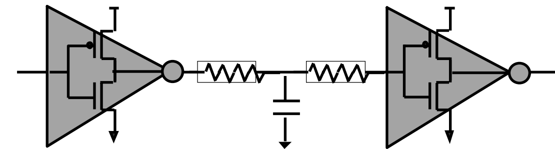
Technology Scaling Effects

- At 0.5um and above: Simple capacitance

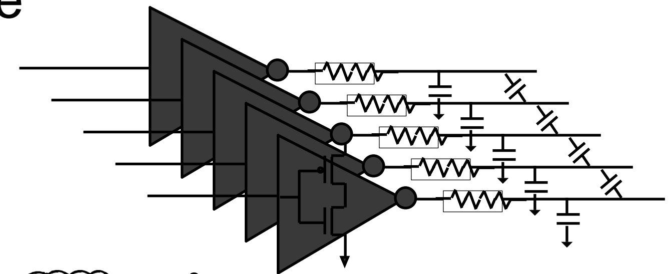


- At 0.35um and below: Resistance

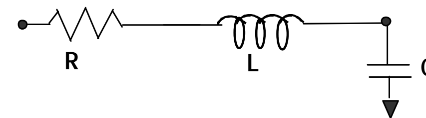




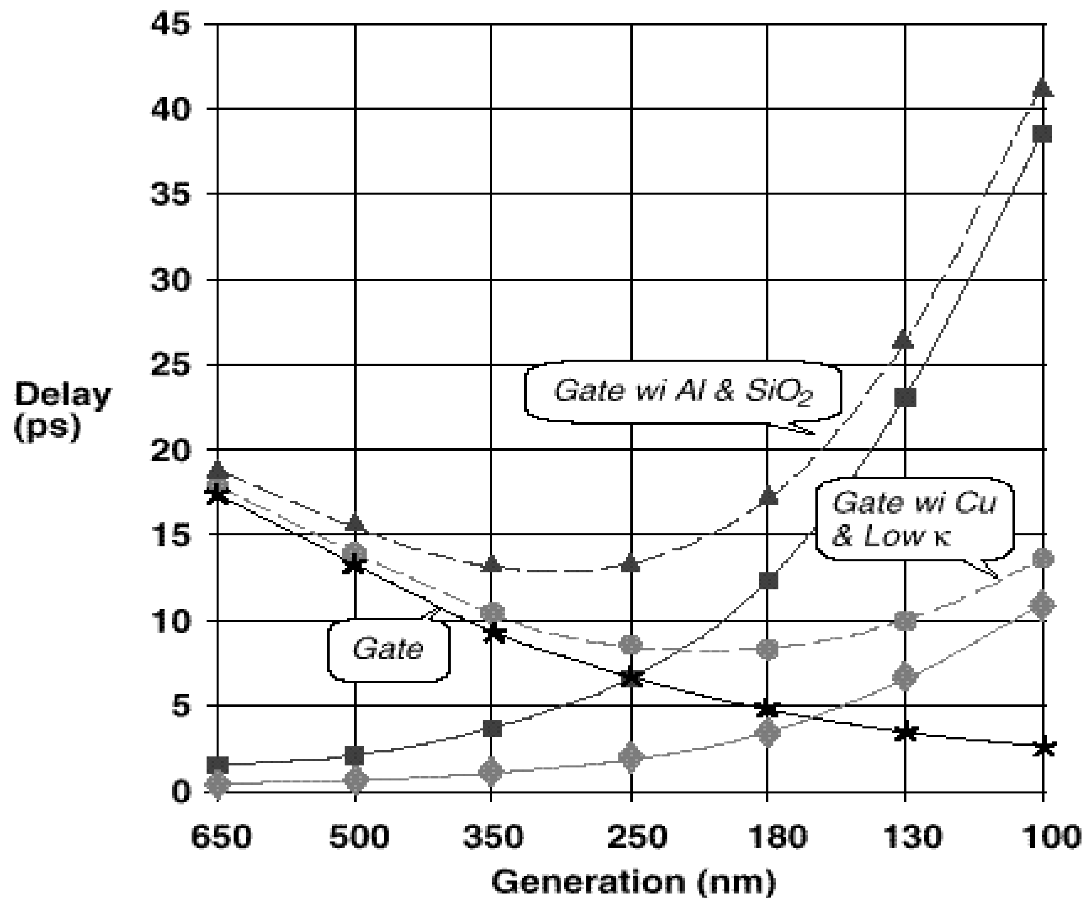
- At 0.18um and below : Coupling Capacitance

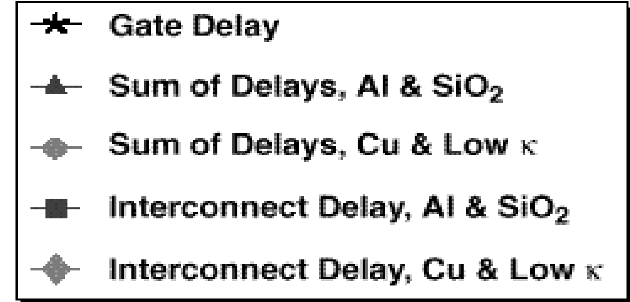


- At 0.10um and below: Inductance



Effect of DSM Interconnect (Circa 1993)

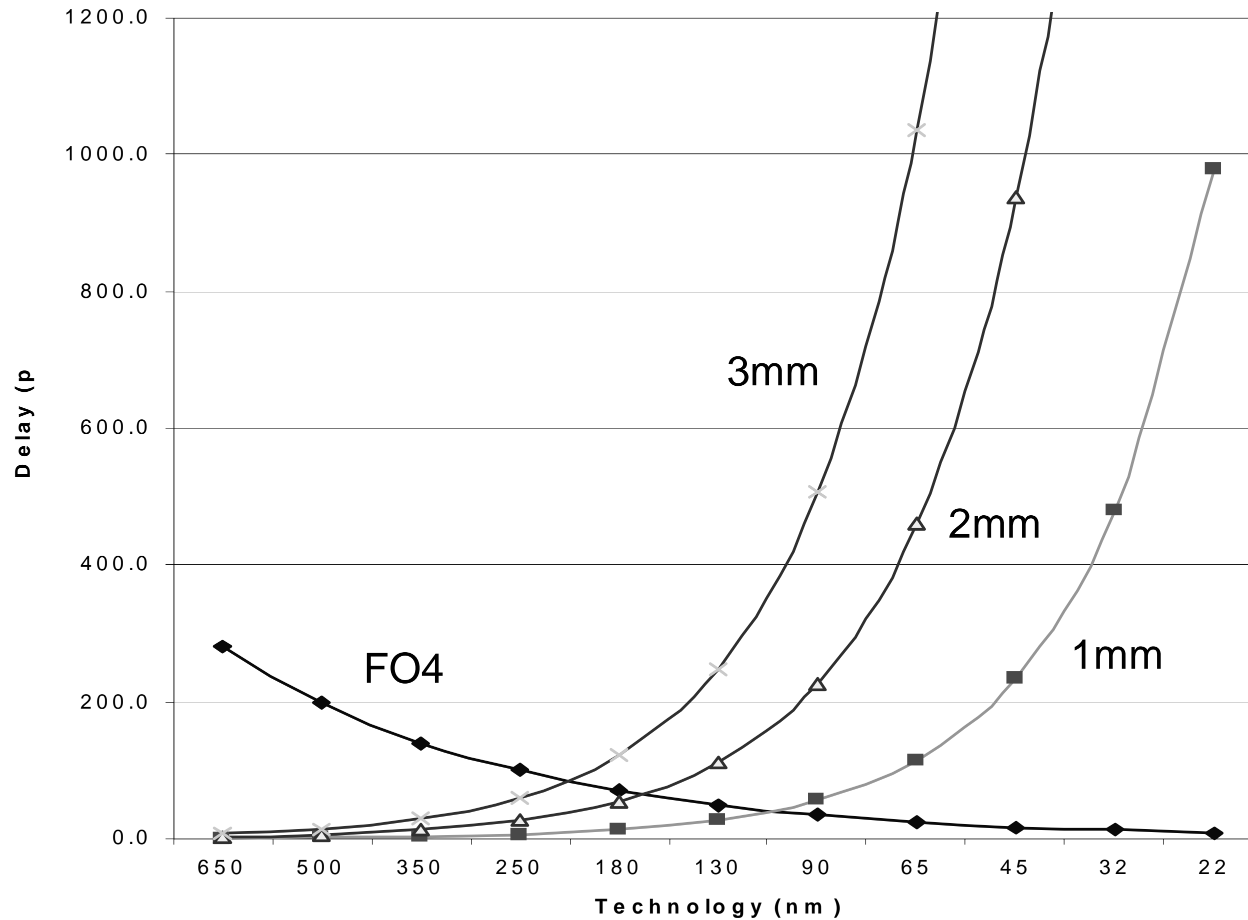




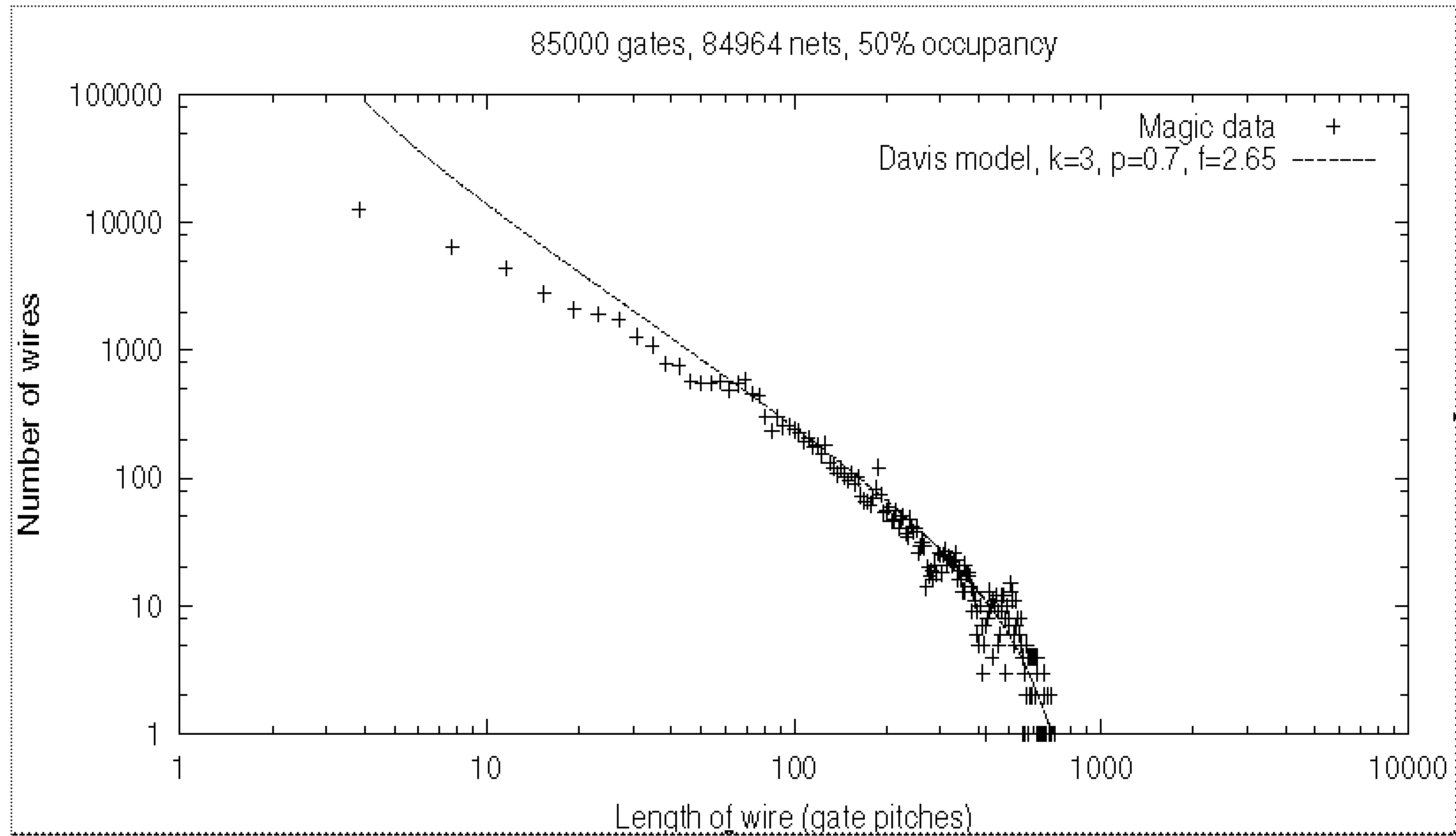
Al	3.0 μΩ −cm
Cu	1.7 μΩ −cm
SiO_2	κ = 4.0
Low κ	κ = 2.0
Al & Cu	.8 μ Thick
Al & Cu Line	43 μ Long

- Cu has much lower resistance and better electromigration characteristics
- Low-k is needed to reduce coupling capacitance

More Realistic Delay Trends

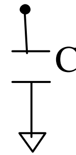


Example of Wire Length Distribution

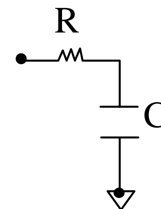


Let’s Look at Wire Models

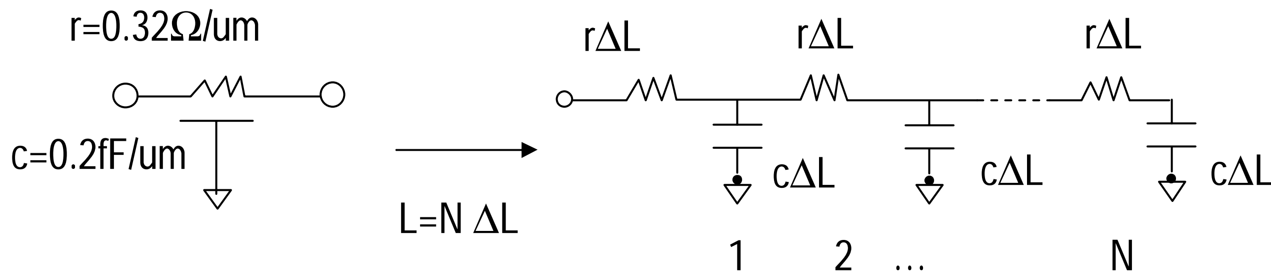
- Very short wires (intra-cell and inter-cell) simply require a grounded capacitance model



- Longer wires require a simple RC models. These are wires that are connections within a block



- Interblock wires are the global wires that have to modeled as a distributed RC model which we convert to a lumped RC ladder



Delay for Long Lines

- What is the delay along the distributed line as a function on length L?

Use Elmore delay $= (r\,\Delta L)(c\,\Delta L) + 2(r\,\Delta L)(c\,\Delta L) + \ldots + N(r\,\Delta L)(c\,\Delta L)$

$= (\Delta L)^2 rc(1 + 2 + \ldots + N)$

$= (\Delta L)^2 rc(N)(N+1)/2 \approx (\Delta L)^2 rcN^2/2$

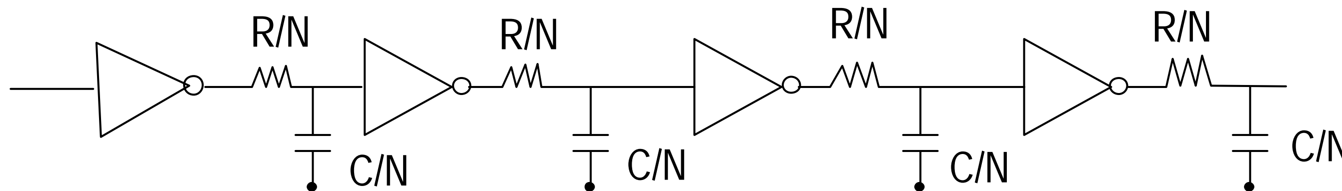
$= L^2\,rc/2$ (measured value is $\approx 0.4rcL^2$)

Table of Delays for different Wire Lengths

Length	Length		Delay (s)
20 um	or	0.02 mm	13 *fs*
200 um	or	.2 mm	1.3 *ps*
1,000 um	or	1 mm	32 *ps*
2,000 um	or	2 mm	128 *ps*
5,000 um	or	5 mm	800 *ps*

Buffer Insertion

- Make long wires into short wires by inserting buffers periodically. Divide interconnect into N sections as follows:



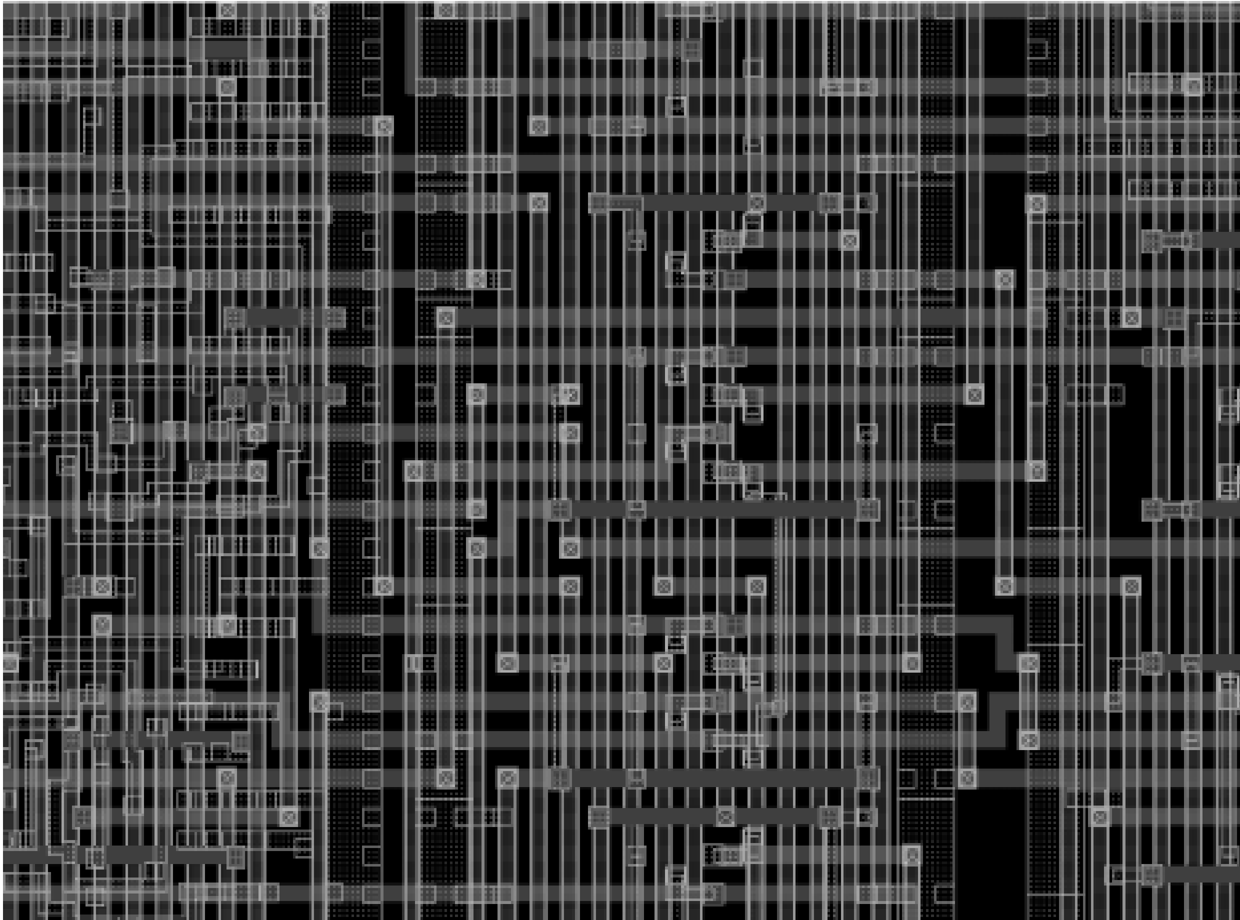
- Delay per stage: $t_{segment}$=Elmore delay of a stage with π model
- Total delay : $t_{total} = N \times t_{segment}$
- What is the optimal number of buffers?

Find N such that $\partial t_{total} / \partial N = 0 \Rightarrow N \approx \mathrm{sqrt}(0.5rcL^2/t_{pbuf})$

For L=5mm case, $N \approx 6$

Therefore, 6 inverters should be inserted (Note that the buffer delay is actually a function of the interconnect length - ignored)

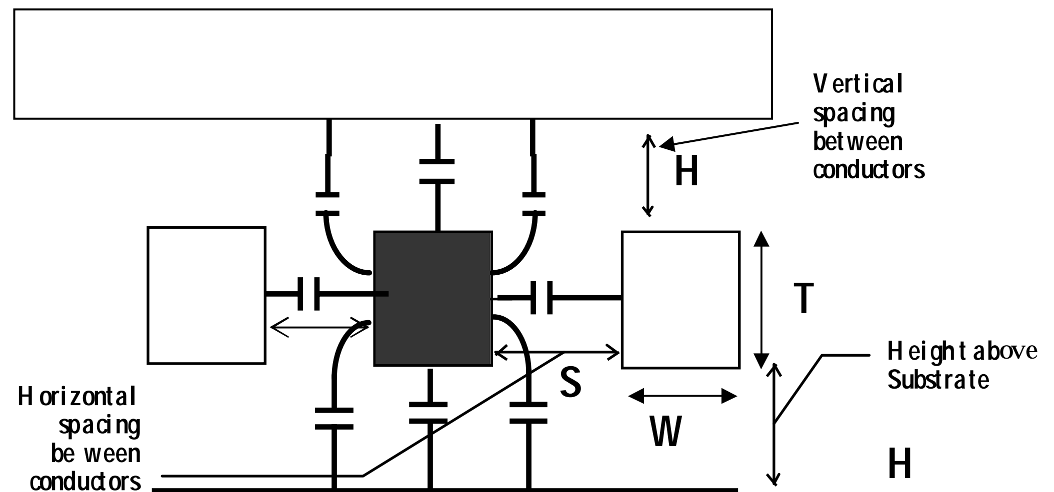
Coupling Between Lines in DSM Layout



- **Over 80% of interconnections in a UDSM chip are parallel crossing lines with 3D effects**

Interconnect Capacitance Profiles

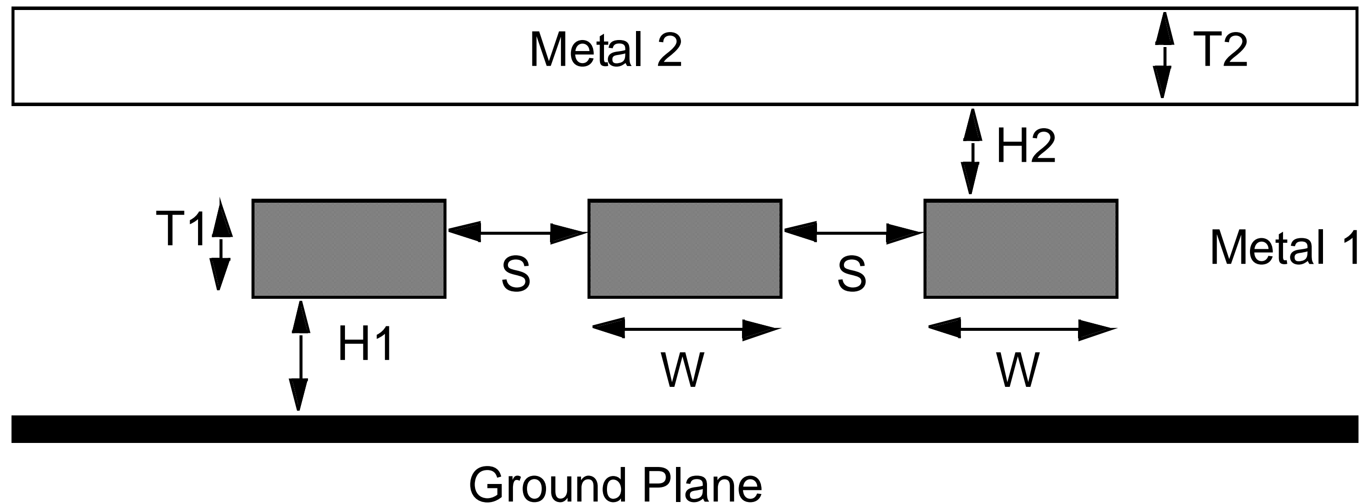
- We decompose the total capacitance into three components:
  - Area capacitance
  - Lateral capacitance
  - Fringe capacitance



$$C_{total} = 2C_{area} + 2C_{lateral} + 2C_{fringe}$$

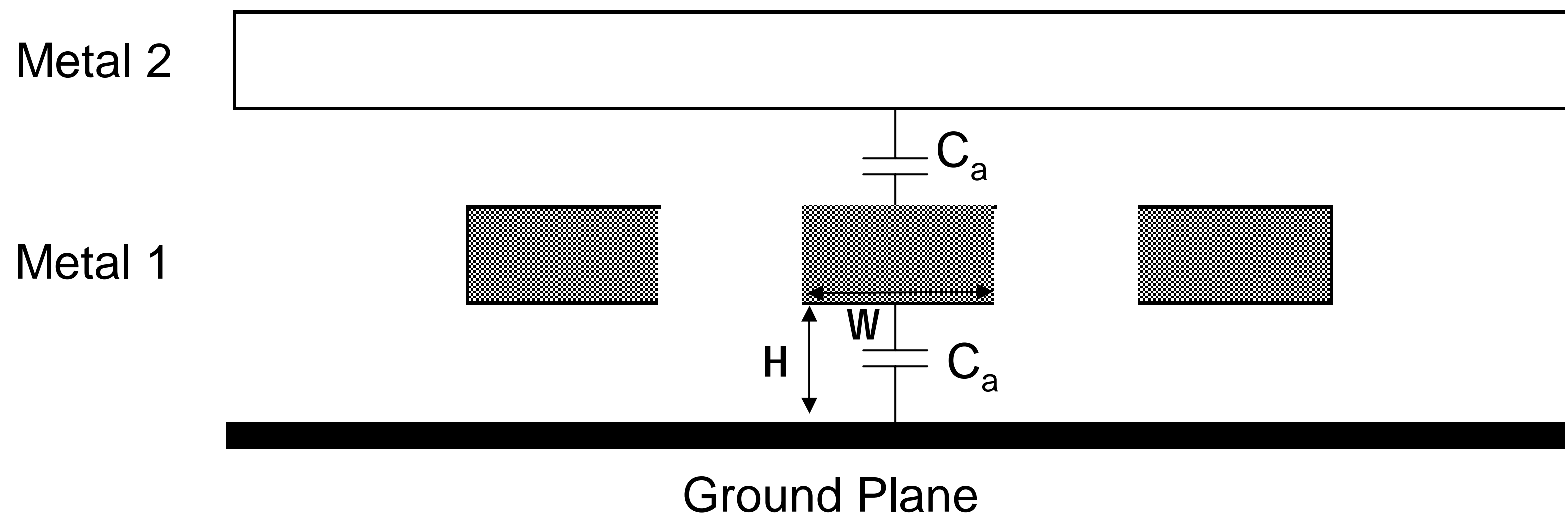
Metal Dimensions

- T=wire thickness, H=vertical wire separation, S=horizontal wire separation, W=wire width, L=wire length



- T and H are fixed parameters based on the fabrication process
- W, S and L are under the designer's control

Area Capacitances

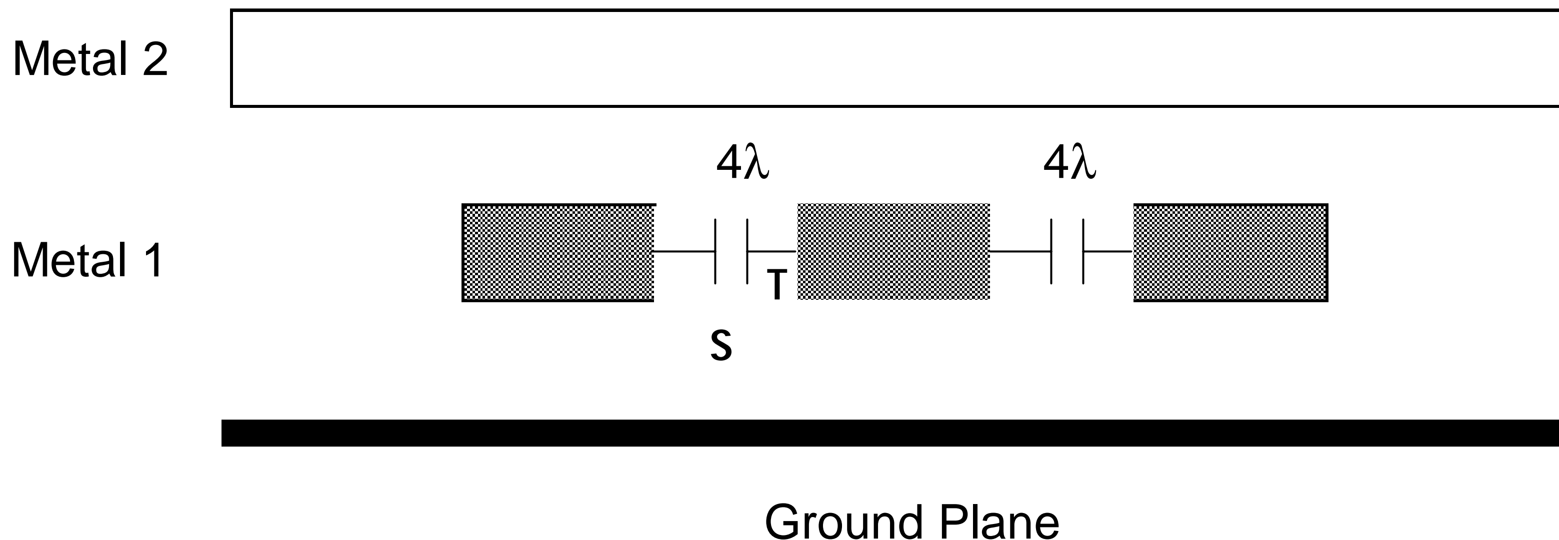


Ground Plane

- Area capacitance per unit length can be simply calculated using:

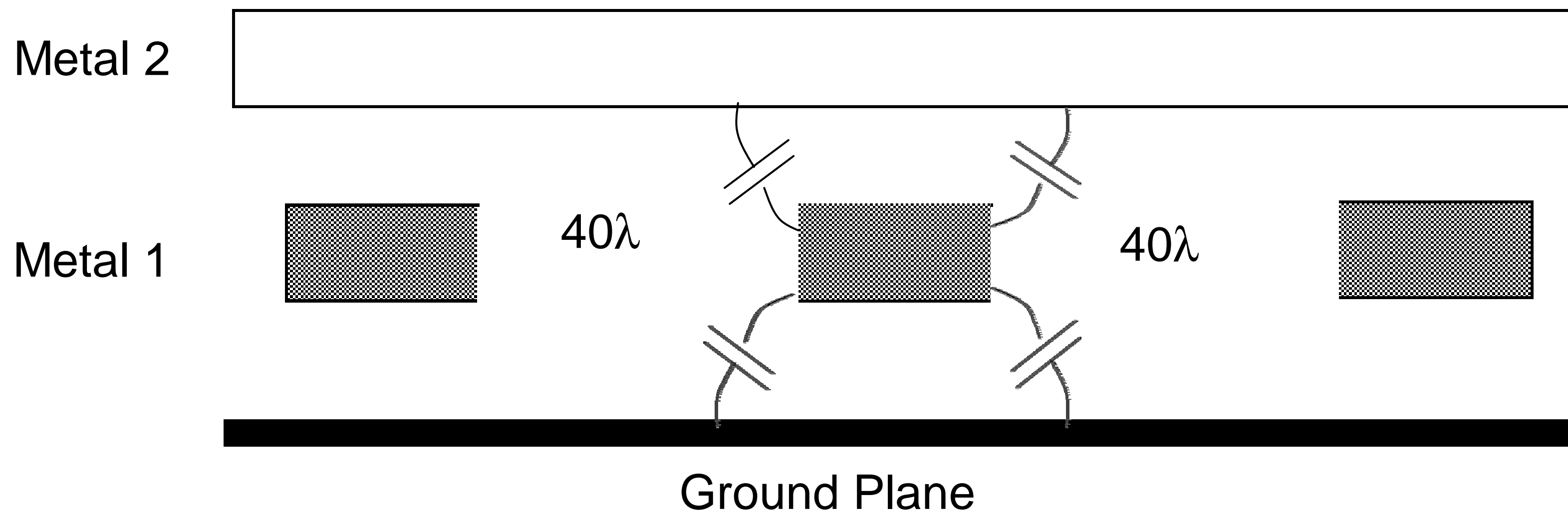
$$C_a = \frac{\varepsilon_{ox}}{t_{ox}} W = 0.035\text{fF/um}\ (W/H)$$

Lateral Capacitances



- Lateral capacitance per unit length for closely spaced wires can be calculated using: $C_L = \frac{\varepsilon_{ox}}{t_{ox}} S = 0.035 fF/um\ (T/S)$
- For widely spaced wires, C_L drops off as 1/S

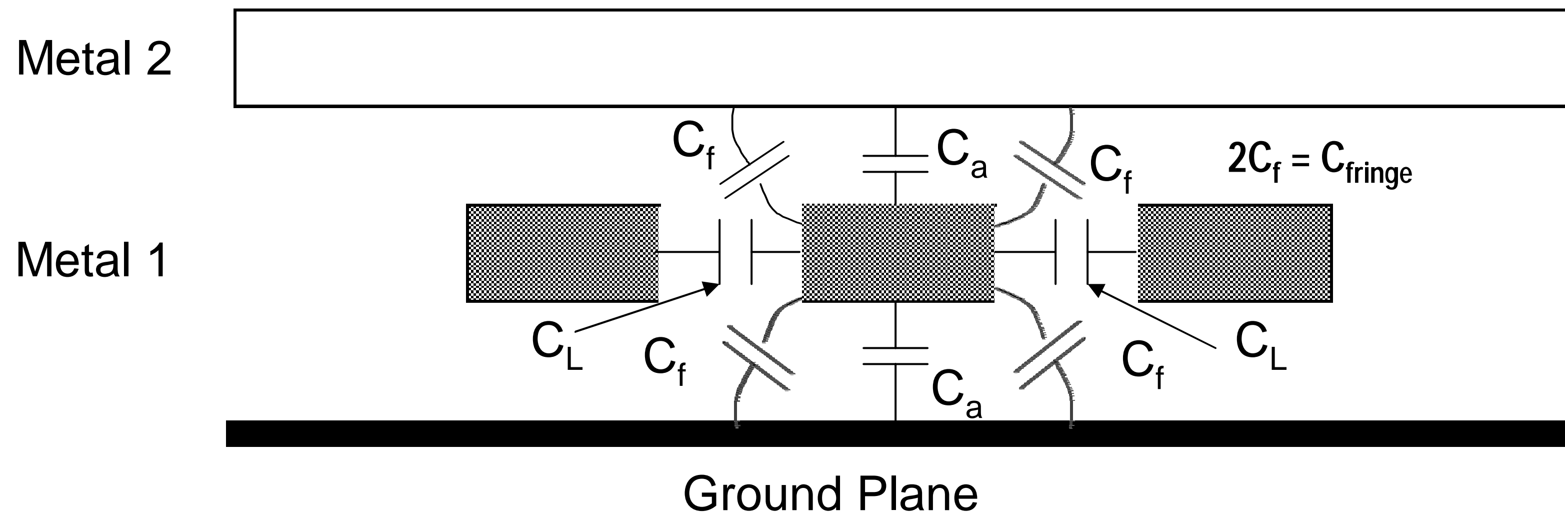
Fringing Capacitances



- Fringing capacitance per unit length for widely spaced wires can be approximated to be (actually depends on H and T which are fixed):

$$C_{fringe} = 0.05 fF/um$$

Total Capacitance



- For closely spaced wires, assume fringe is small

$$C_{total} = 2C_a + 2C_L = 0.2\text{fF/um}$$

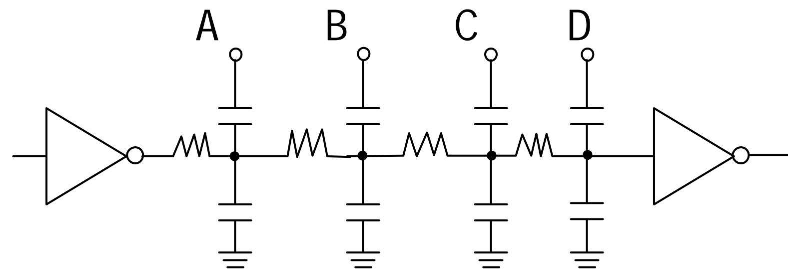
- For widely spaced wires, assume lateral is small

$$C_{total} = 2C_a + 2C_{fringe} = 0.2\text{fF/um}$$

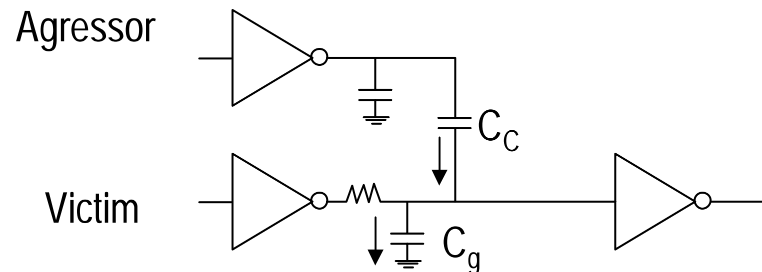
- For medium spaced wires, C_f and C_L will both exist and vary with S

Coupling Effects

- New model of interconnect



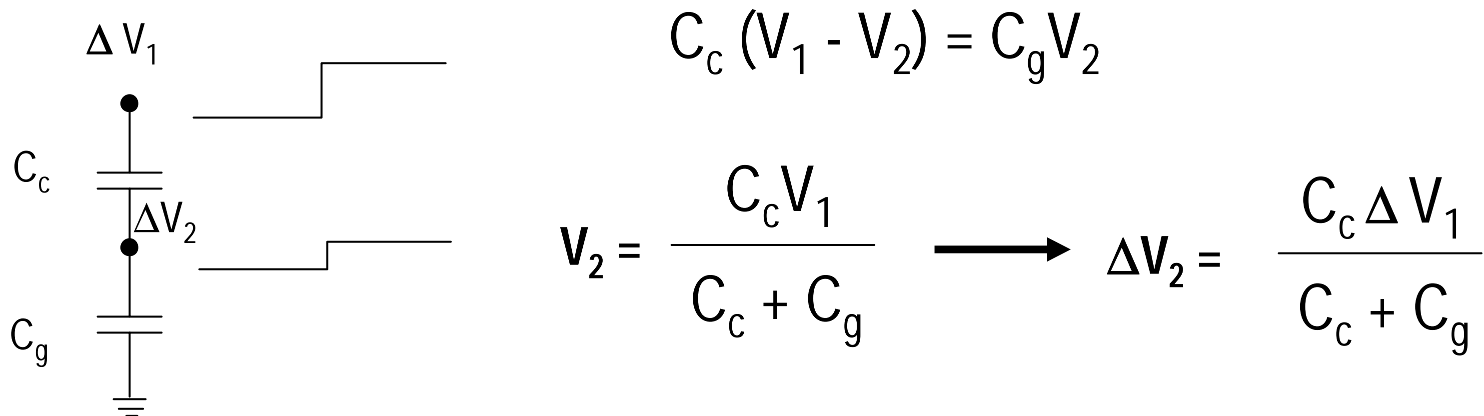
- Each driver connected to A,B,C or D can act as aggressor



- Coupling capacitance could inject noise or affect delay

First-order Noise Analysis

- Assume that aggressor and driver resistances are negligible



$$C_c(V_1 - V_2) = C_g V_2$$

$$V_2 = \frac{C_c V_1}{C_c + C_g} \longrightarrow \Delta V_2 = \frac{C_c \Delta V_1}{C_c + C_g}$$

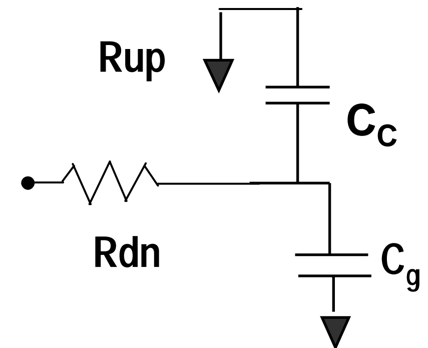
- If V_1 changes by V_{DD}, what change ΔV do we expect to see at the internal node in the worst case?

$$\Delta V_2 = \frac{C_c V_{dd}}{C_c + C_g}$$

First-Order Delay Analysis

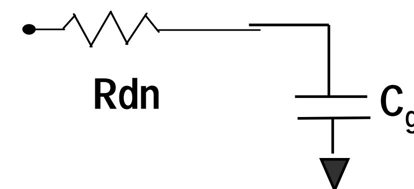
- **If aggressor is not switching**

$$C_{load} = C_C + C_g \qquad \Delta Q = (C_C + C_g)V_S$$



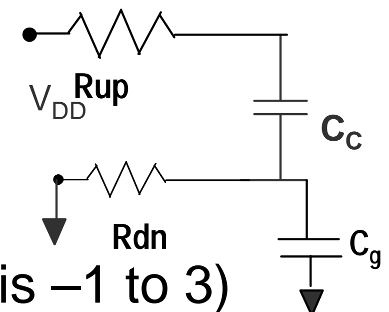
- **If aggressor switches in same direction.**

$$C_{load} = C_g \qquad \Delta Q = C_g \ V_S$$



- **If aggressor switches in opposite direction: "Miller" factor**

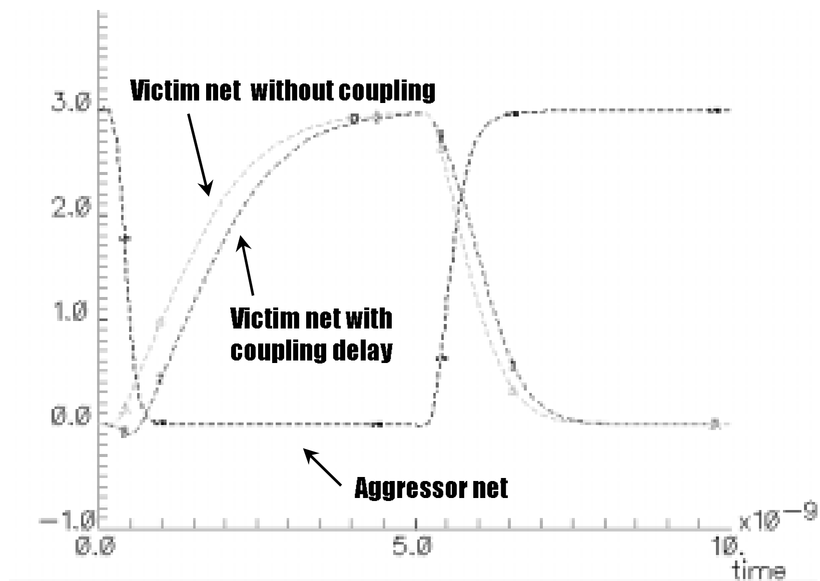
$$C_{load} = 2\ C_C + C_g \qquad \Delta Q = (2\ C_C + C_g)V_S$$



- **Multiplying factor ranges from 0 to 2 (**Actual range is –1 to 3)

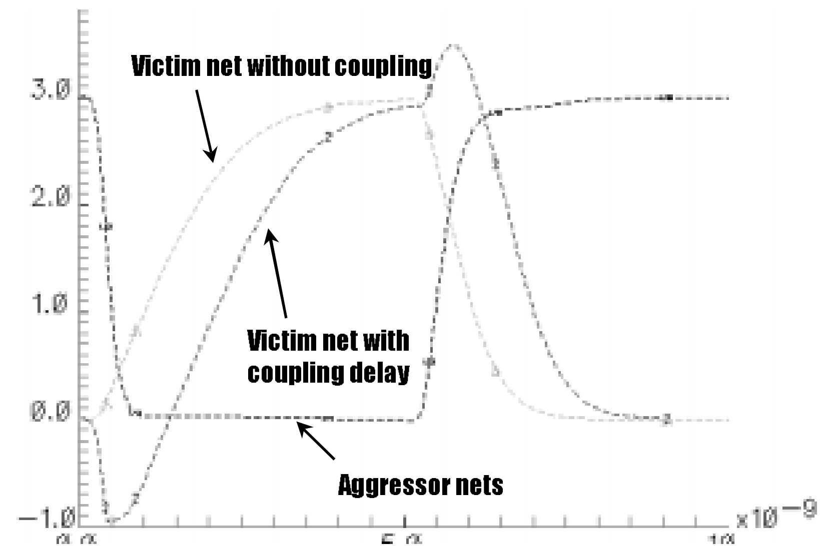
Signal Integrity Effect on Timing

Net delay due to a <u>single</u> coupled aggressor net



Performance impact: 300 picosecond delay (3% of a clock cycle)

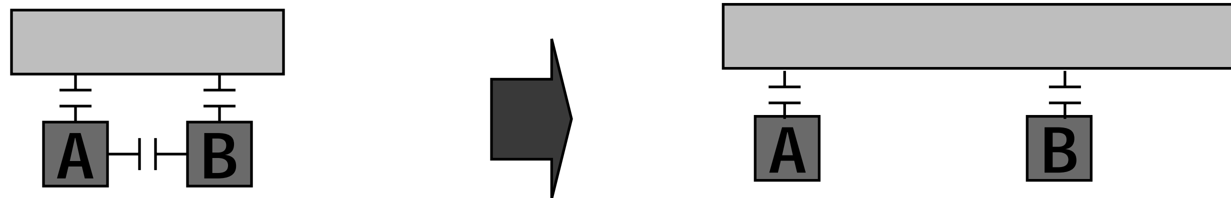
Net delay due to <u>multiple</u> coupled aggressor nets



Performance impact: over 2 nanosecond delay (20+% of a clock cycle)

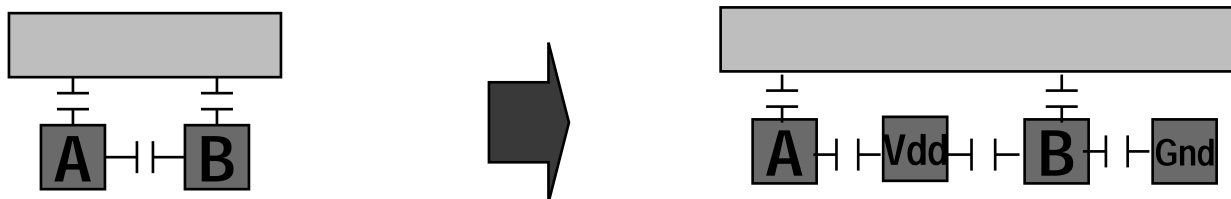
Reducing Coupling Capacitance

- Space out the signals as much as possible, but it cost area.



(a) higher coupling cap./less area (b) lower coup. cap./ more area

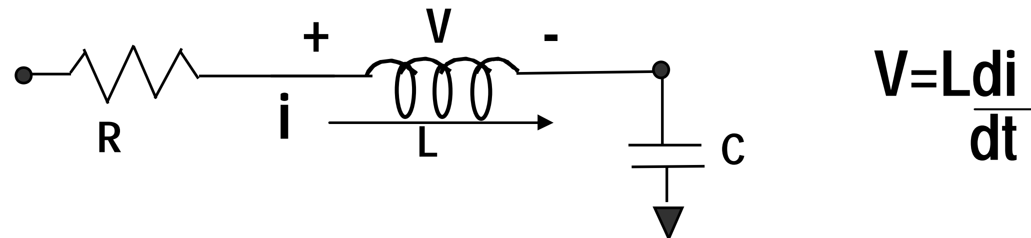
- Use Vdd and Gnd to shield wires wherever required



(a) higher coupling cap./less area (b) higher tot. cap./ more area

Inductance

- Complete interconnect model should include inductance



$$V = L\frac{di}{dt}$$

- With increasing frequency and a decrease in resistance due to wide wires and the use of copper, inductance will begin to influence clocks/busses:

$$Z = R\downarrow + j\omega L\uparrow$$

- Inductance, by definition, is for a loop not a wire
  - **inductance of a wire in an IC requires knowledge of return path(s)**
  - **inductance extraction for a whole chip is virtually impossible...**

Impact of on-chip self-inductance

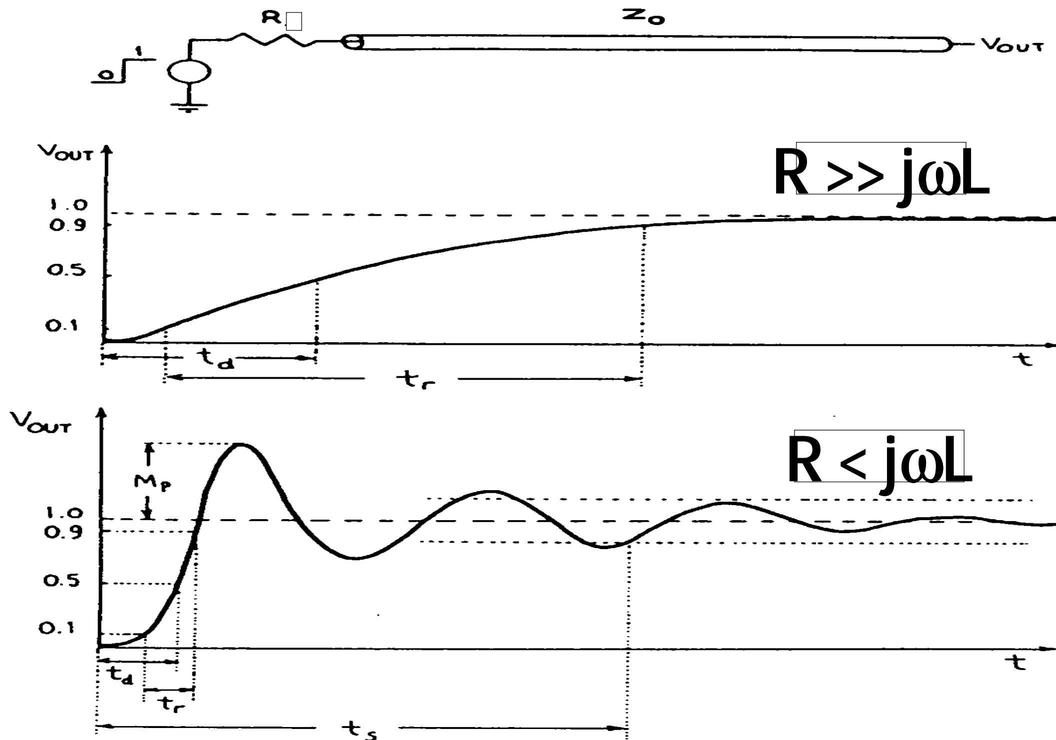


FIGURE 6.9 Waveforms for various R_S and Z_0 combinations with a finite input rise time. The top plot shows the case when $R_S > Z_0$; the output is an exponential waveform. In the bottom plot $R_S < Z_0$; the waveform rings. The delay time t_d, rise time t_r, and settling time t_s are also shown.

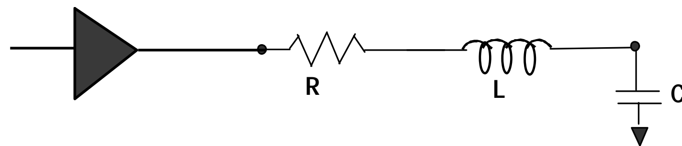
Driver Model + wire

- Most gates behave this way (RC)
- Clocks behave this way (RLC):
  - overshoot/ringing
  - sharp edges
  - reflections

Other Inductance Effects

- For most gates R_{eff} is in the order of kΩ so typically $R >> j\omega L$
  - response is dominant by RC delay for most signals
- Only the large drivers have a small enough R_{on} to allow the inductance to control the dynamic response
  - clocks
  - busses
- For clocks, self-inductance term can dominate the response (especially if shielding is used)
- For busses, mutual inductance term dominates and creates noise events that could cause malfunction
- For power supplies, inductance can also be a problem due to the Ldi/dt drop (in addition to the IR drop) as supplies scale down

Capacitive and Inductive Noise



For most wires, jωL < (Rwire+Rdrive) for the frequency and R of interest. So, for delay, L is not a big issue currently.

But ωL can be ≈ 20 - 30% of R so noise may be seen on adjacent line (mutual coupling)

Dangerous scenario is a combination of localized capacitive coupling noise and long range mutual inductive coupling noise