University of Southern California

Viterbi School of Engineering

# EE577A

# VLSI System Design

## Interconnect Modeling

References: syllabus textbook, Professor Massoud Pedram's slides, online resources

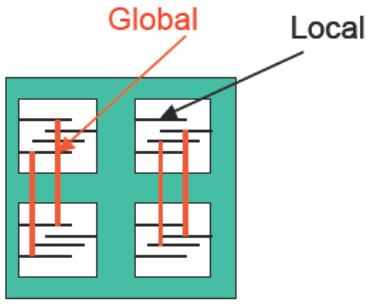
Shahin Nazarian

Spring 2013

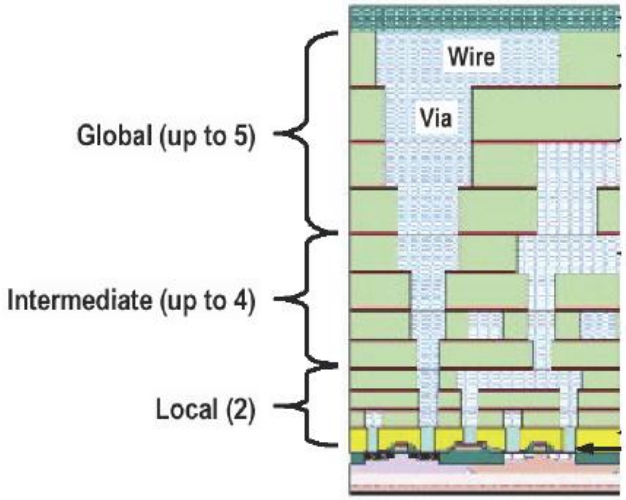
# Background

# Types of Interconnect and Performance Metrics

- Dimension based classification
  - Local
  - Intermediate/semi-global
  - Global
- Function based classification
  - Signaling
  - Clocking
  - Power/ground distribution
- Signaling
  - Propagation delay
  - Power dissipation
  - Data reliability (Noise)
  - Area
- Power Lines
  - Supply reliability



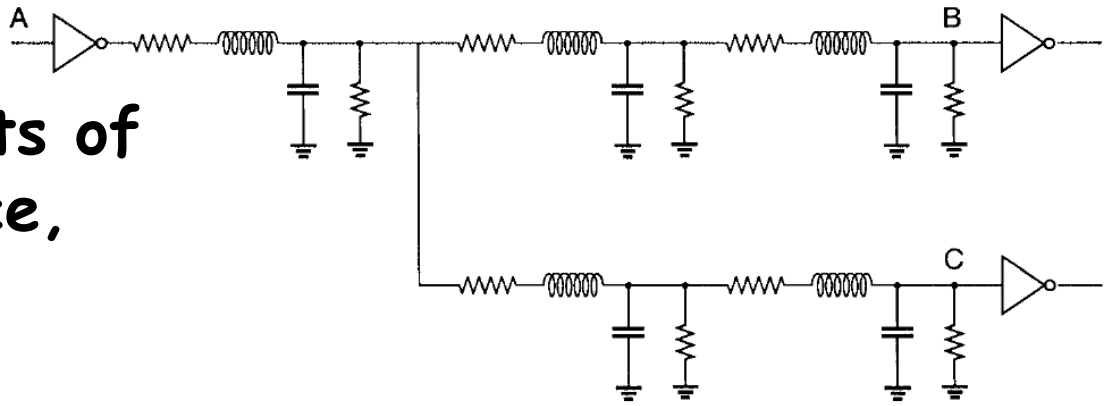
Top View: For Length



Cross Sectional View:
For Height, Width and Spacing

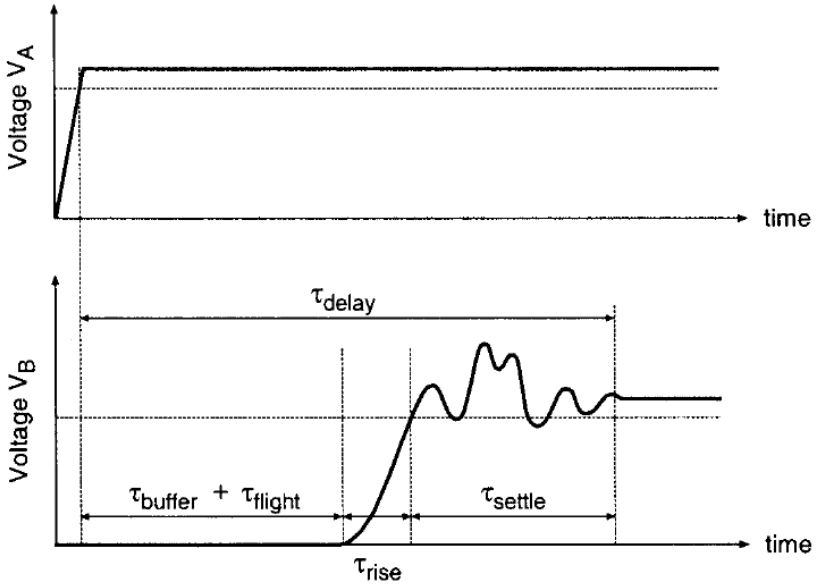
- Reliability
  - Electromigration
- Clocking
  - Timing uncertainty (skew and jitter)
  - Power dissipation
  - Slew rate

# RLCG Model of Interconnection




- RLCG parasitic consists of (resistance, inductance, capacitance and conductance)

- An RLGC interconnection tree with typical signal waveforms at nodes A and B, showing signal delay and various delay components



# When to Use Transmission Line Equations

- If time of flight [determined by light speed] across interconnection line is much shorter than rise/fall time, then wire can be modeled as a capacitive load, or as lumped or as distributed RC network; otherwise it must be modeled as *transmission lines* (with inductance)
  - Simple rule of thumb:

$$\tau_{rise}(\tau_{fall}) < 2.5 \cdot (\frac{l}{\upsilon}) \quad \Rightarrow \{\text{transmission-line modeling}\}$$

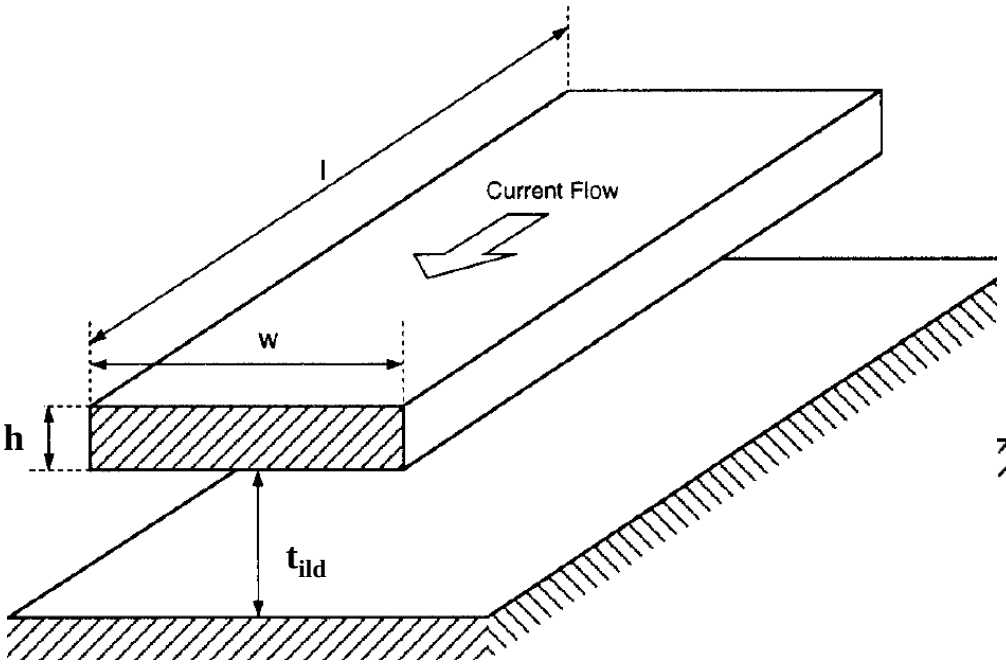
$$2.5 \cdot (\frac{l}{\upsilon}) < \tau_{rise}(\tau_{fall}) < 5 \cdot (\frac{l}{\upsilon}) \quad \Rightarrow \left\{\begin{array}{l}\text{either transmission-line} \\ \text{or lumped modeling}\end{array}\right\}$$

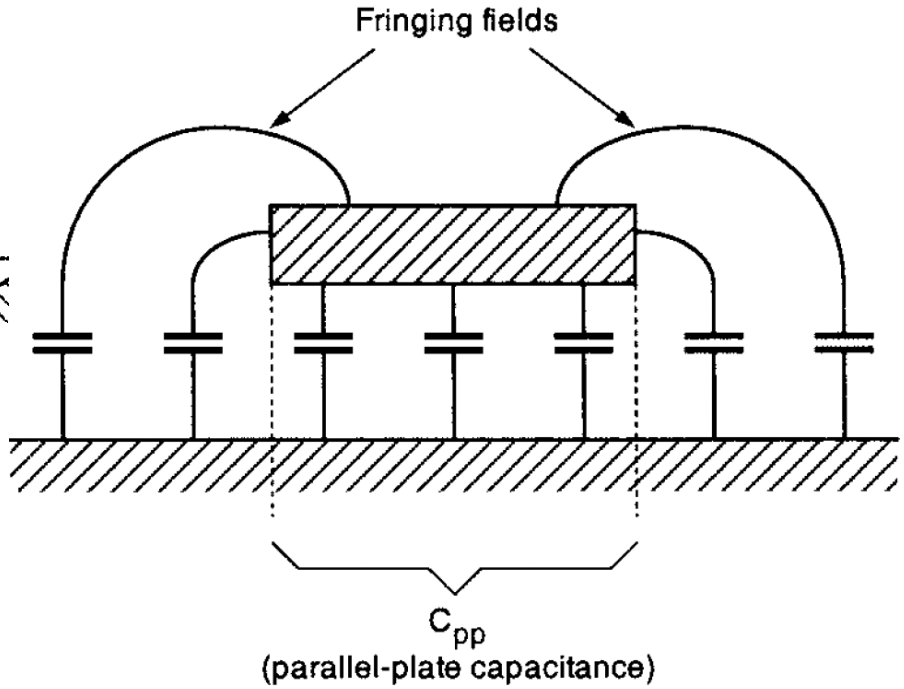
$$\tau_{rise}(\tau_{fall}) > 5 \cdot (\frac{l}{\upsilon}) \quad \Rightarrow \{\text{lumped modeling}\}$$

$l$ is the interconnect line length and $\upsilon$ is the propagtion speed.

Optional: http://en.wikipedia.org/wiki/Velocity_of_propagation

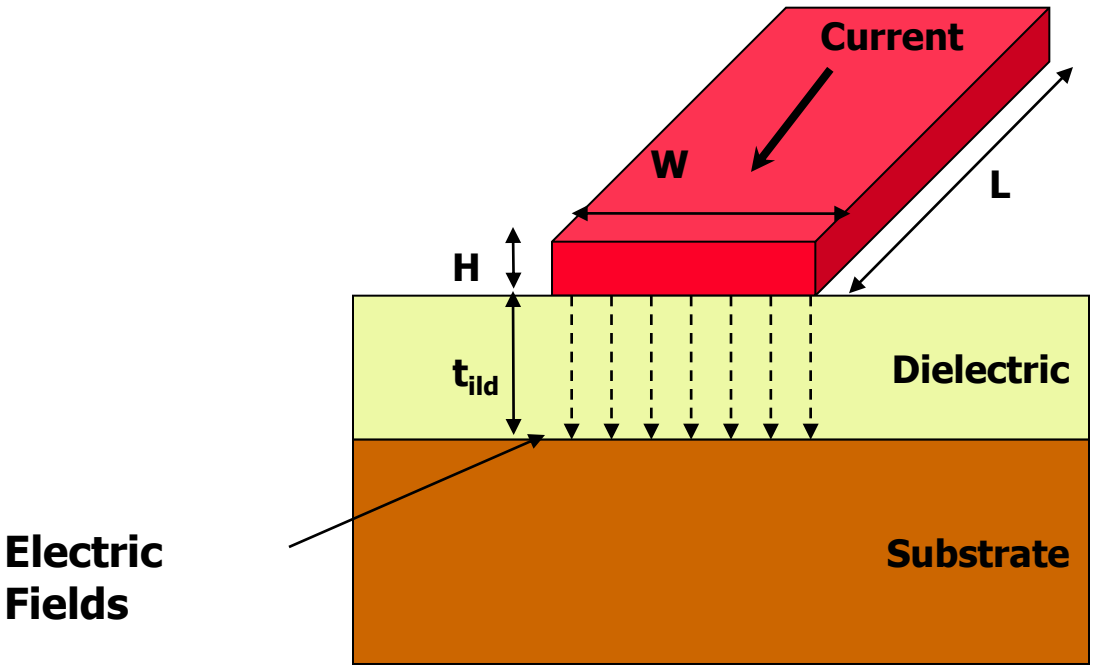
# Interconnect Capacitances



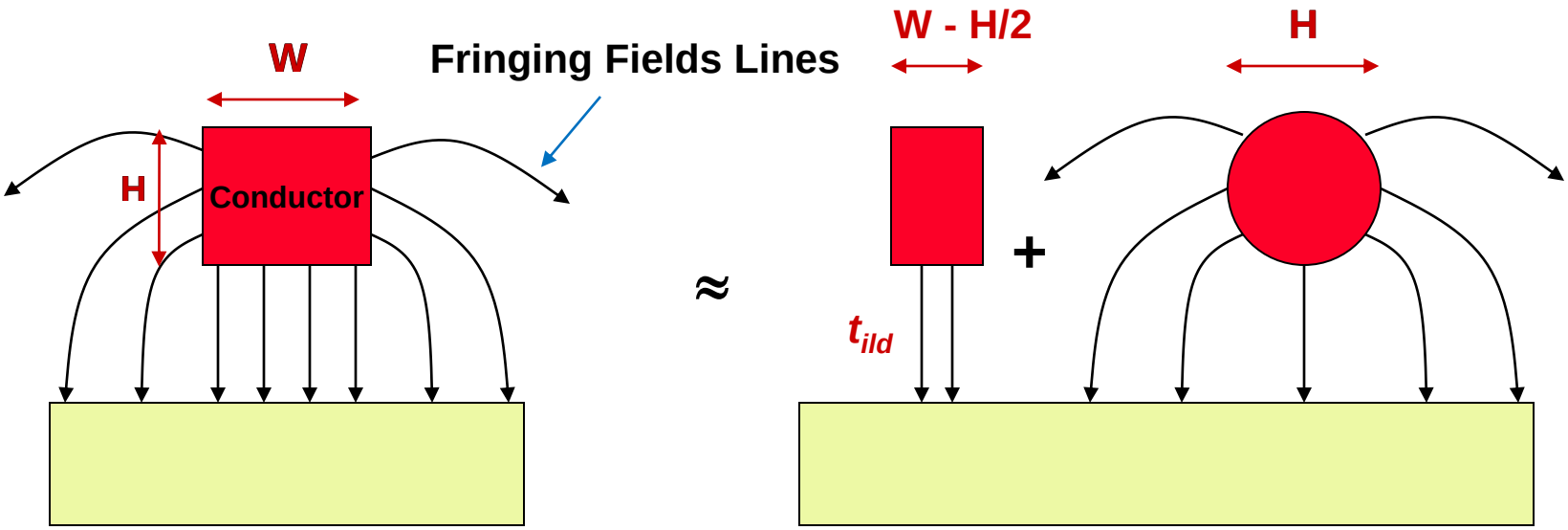

- Interconnect running above substrate
  - *Inter-layer dielectric* (ILD) material is known
  - Influence of fringing electric fields upon the parasitic wire capacitance

# Parallel Plate Capacitance Equation




$$C_{pp} = \frac{\varepsilon_{ild}}{t_{ild}} WL$$

# Wire Capacitance

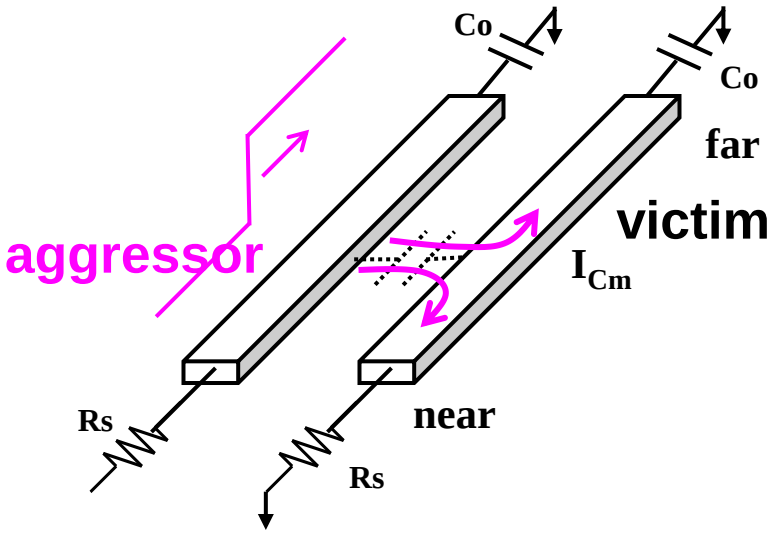


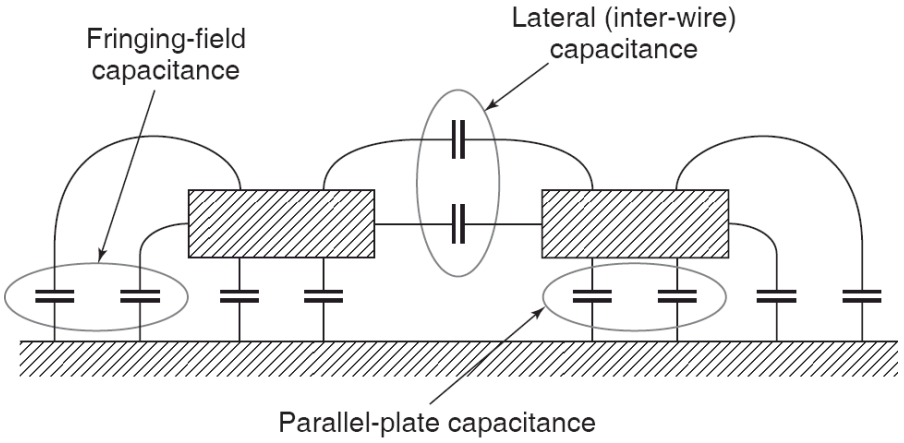

$$C_{wire} = C_{pp} + C_{fringe} = \left( \frac{\left( W - \frac{H}{2} \right) \varepsilon_{ild}}{t_{ild}} + \frac{2\pi\varepsilon_{ild}}{\log\left( \frac{t_{ild}}{H} \right)} \right) L$$

- L denotes the interconnect length

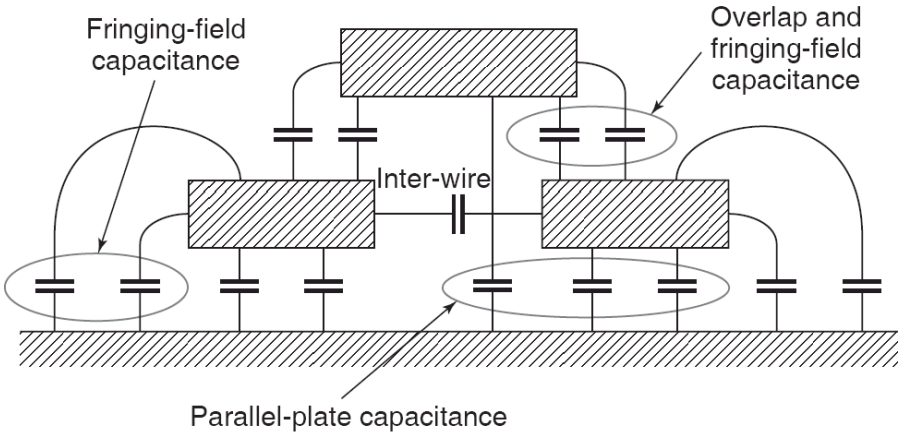
# Capacitive Coupling Components

- Transition in one line can cause noise in another line
  - Signal crosstalk



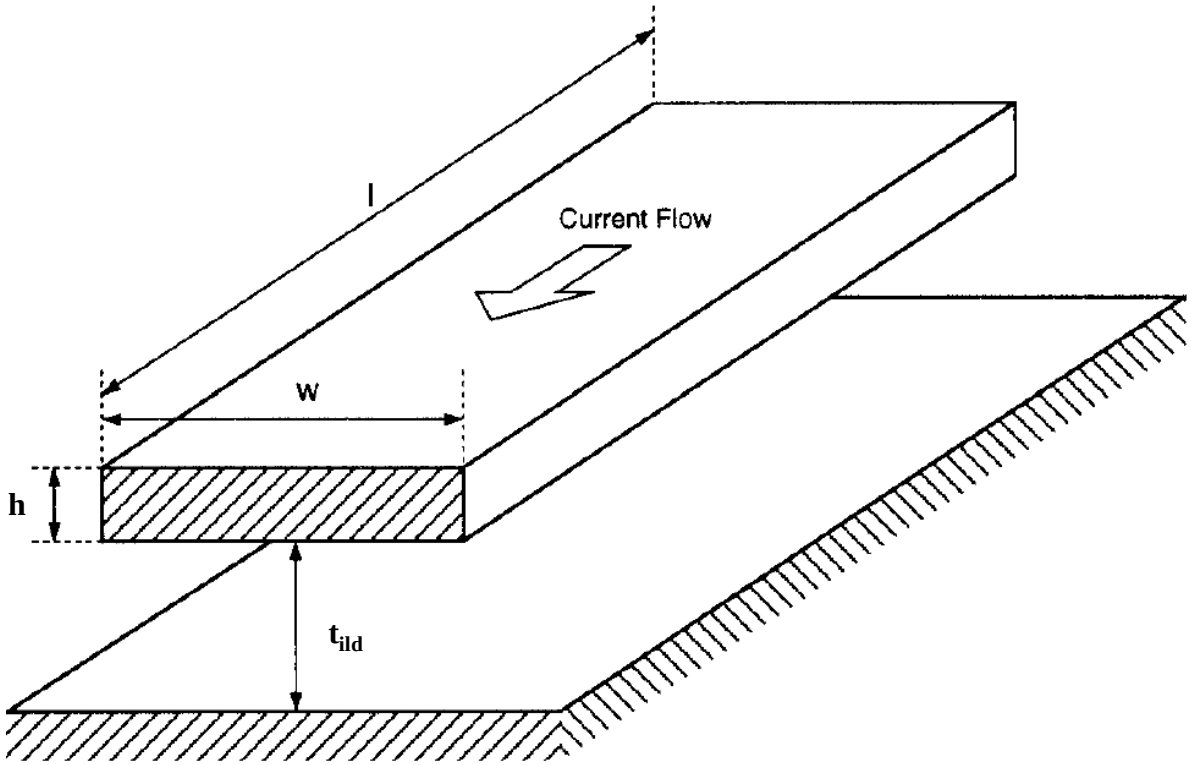




(a)




(b)

# Interconnect Resistance Estimation

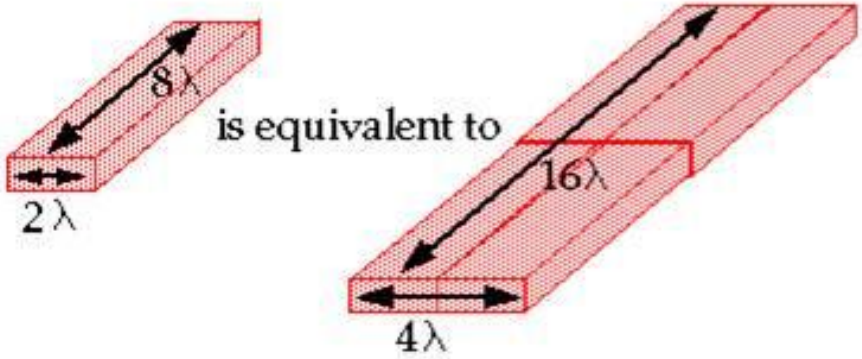



$$R_{wire} = \rho \frac{l}{wh} = R_{sheet} \frac{l}{w} \quad \text{where} \quad R_{sheet} = \frac{\rho}{h}$$

$\rho$ : characteristic resistivity of the interconnect material

$R_{sheet}$ : sheet resistivity of the line ($\Omega$/square)

# Resistivity and Sheet Resistivity (Resistance)




Irregular shapes require more elaborate calculation

Typical sheet resistances of 0.5 $\mu$ to 1.0 $\mu$ processes

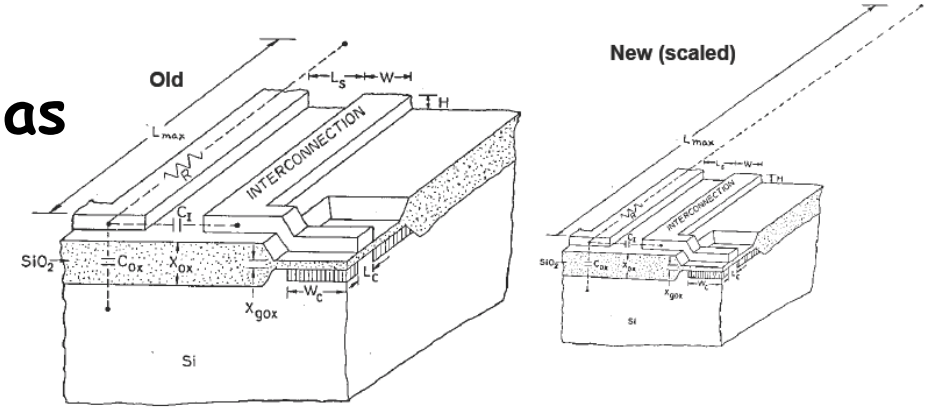
| material | Ω /sq |
|---|---|
| Metal1/Metal2 | 0.07 |
| Metal 3 | 0.04 |
| Poly | 20 |
| Diffusion | 25 |
| n-well | 2K |

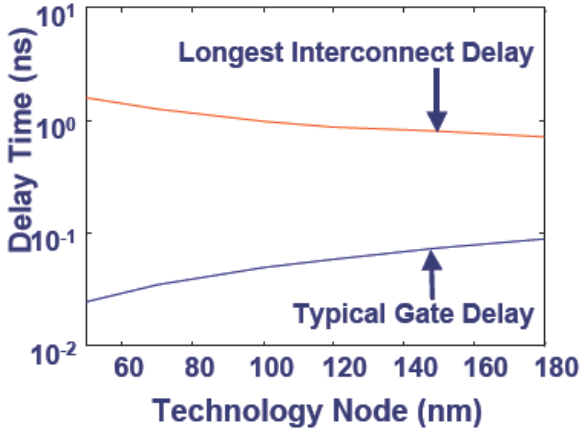
contacts => 0.25 to 20 ohms.

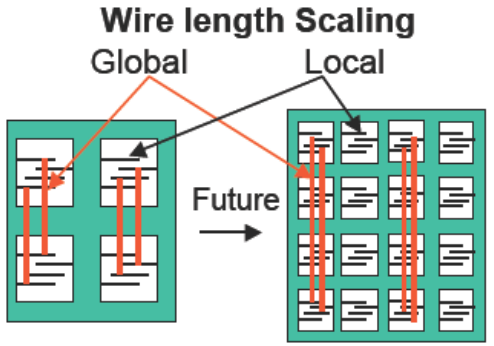
| Material | $\rho$ (Ω-m) |
|---|---|
| Silver (Ag) | $1.6 \times 10^{-8}$ |
| Copper (Cu) | $1.7 \times 10^{-8}$ |
| Gold (Au) | $2.2 \times 10^{-8}$ |
| Aluminum (Al) | $2.7 \times 10^{-8}$ |
| Tungsten (W) | $5.5 \times 10^{-8}$ |

# Interconnect Scaling Scenario

- Chip area increases with each node
- Device dimensions are scaled as discussed before
- Typical scaled wires are:
  - Longer (chip area scaling)
  - Narrower and closer to one another (minimum dimension scaling)
  - Fatter [taller] or look taller (to reduce sheet resistivity)

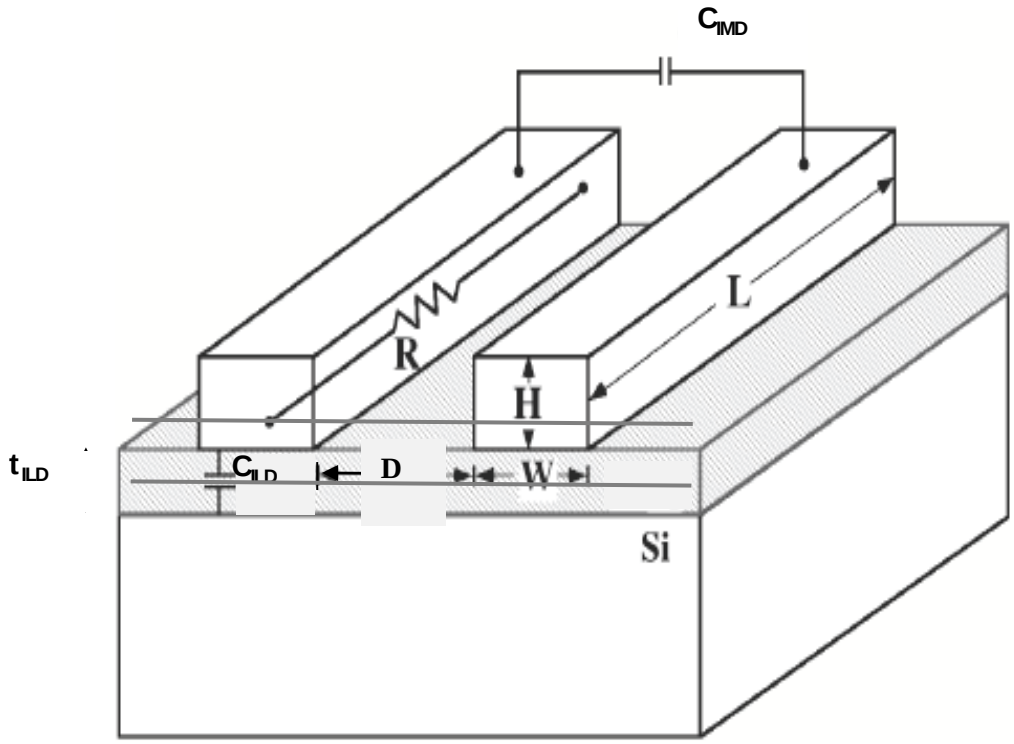


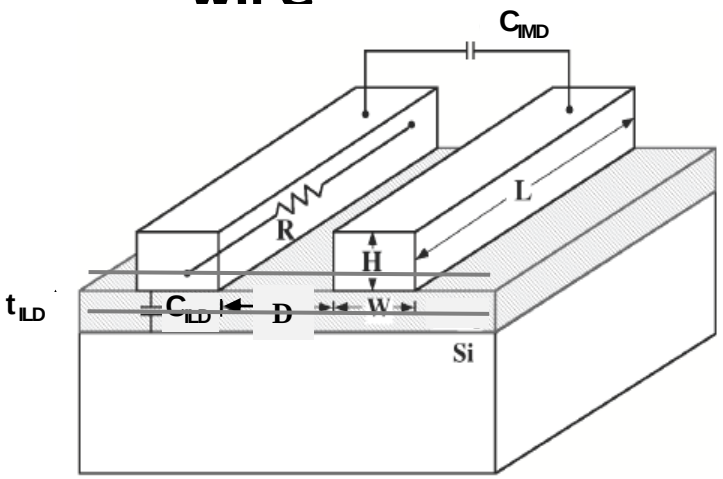
# Scaling for Capacitances and Resistance of a Global Line

- Assume that with technology scaling, $L$ of a <u>global</u> VLSI interconnect line increases, its $H$ remains the same, while its $W$, $t_{ILD}$ and $D$ decrease (see drawing)



# Scaling for Capacitances and Resistance of a Global Line (Cont.)

- Both the *inter-layer* capacitance ($C_{ILD}$) and *inter-metal* capacitance ($C_{IMD}$) increase, although $C_{IMD}$ increases at a faster rate. As a result, $C_{wire}$ increases
- The wire resistance $R_{wire}$ increases even more rapidly
- Notice that the line aspect ratio (defined as H/W) increases. Also note the one-sided fringing field calculation due to the presence of a parallel adjacent wire




$$C_{ILD} = \frac{\left(W - \frac{H}{2}\right) L \varepsilon_{ILD}}{t_{ILD}} + \frac{\pi \varepsilon_{ILD} L}{\log\left(\frac{t_{ILD}}{H}\right)}, \quad C_{IMD} = \frac{HL\varepsilon_{IMD}}{D}$$

$$C_{wire} = C_{ILD} + C_{IMD}, \quad R_{wire} = \rho \frac{L}{WH}, \quad \text{Assume: } \varepsilon_{IMD} = \varepsilon_{ILD}$$

$$\tau_{wire} = 0.69 \rho \varepsilon_{ILD} \left( \frac{\left(W - \frac{H}{2}\right)}{t_{ILD}} + \frac{\pi}{\log\left(\frac{t_{ILD}}{H}\right)} + \frac{H}{D} \right) \frac{L^2}{WH}$$

# Typical VLSI Interconnect Dimensions

## 0.18um technology node

| | width (um) | space (um) | thickness (um) | $t_{ILD}$ (um) | $k_{ILD}$ |
|---|---|---|---|---|---|
| Local | 0.28 | 0.28 | 0.45 | 0.65 | 3.5 |
| Intermediate | 0.35 | 0.35 | 0.65 | 0.65 | 3.5 |
| Global | 0.80 | 0.80 | 1.25 | 0.65 | 3.5 |

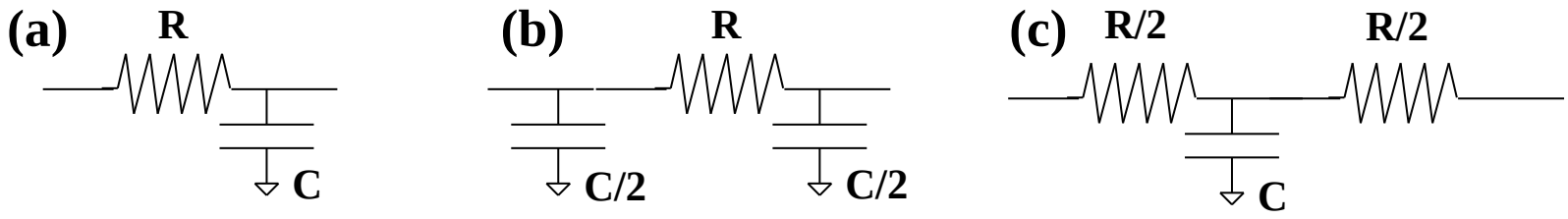
## 90nm technology node

| | width (um) | space (um) | thickness (um) | $t_{ILD}$ (um) | $k_{ILD}$ |
|---|---|---|---|---|---|
| Local | 0.15 | 0.15 | 0.30 | 0.30 | 2.8 |
| Intermediate | 0.20 | 0.20 | 0.45 | 0.30 | 2.8 |
| Global | 0.50 | 0.50 | 1.20 | 0.30 | 2.8 |

## 65nm technology node

| | width (um) | space (um) | thickness (um) | $t_{ILD}$ (um) | $k_{ILD}$ |
|---|---|---|---|---|---|
| Local | 0.10 | 0.10 | 0.20 | 0.20 | 2.2 |
| Intermediate | 0.14 | 0.14 | 0.35 | 0.20 | 2.2 |
| Global | 0.45 | 0.45 | 1.20 | 0.20 | 2.2 |

# Calculation of Interconnect Delay: Lumped RC Delay Model



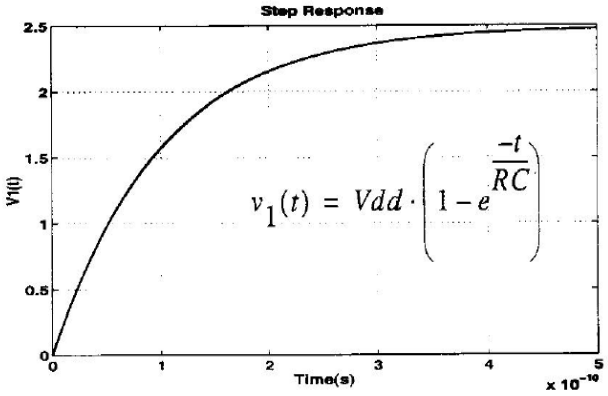

Simple lumped model | π-section model | T-section model

## (a) Simple lumped RC model of an interconnect line

- Assume that the capacitance is initially discharged
- Input is a rising step input pulse at time t=0

$$V_{out}(t) = V_{DD}\left(1 - e^{-\frac{t}{RC}}\right) \Rightarrow \boxed{\tau_{63.2\%V_{DD}} = RC}$$

$$V_{50\%} = V_{DD}\left(1 - e^{-\frac{\tau_{pLH}}{RC}}\right) \Rightarrow \boxed{\tau_{pLH} = 0.69RC}$$




## (b,c) The p- and T-model of the same line, which improve accuracy

# The Elmore Delay (Cont.)

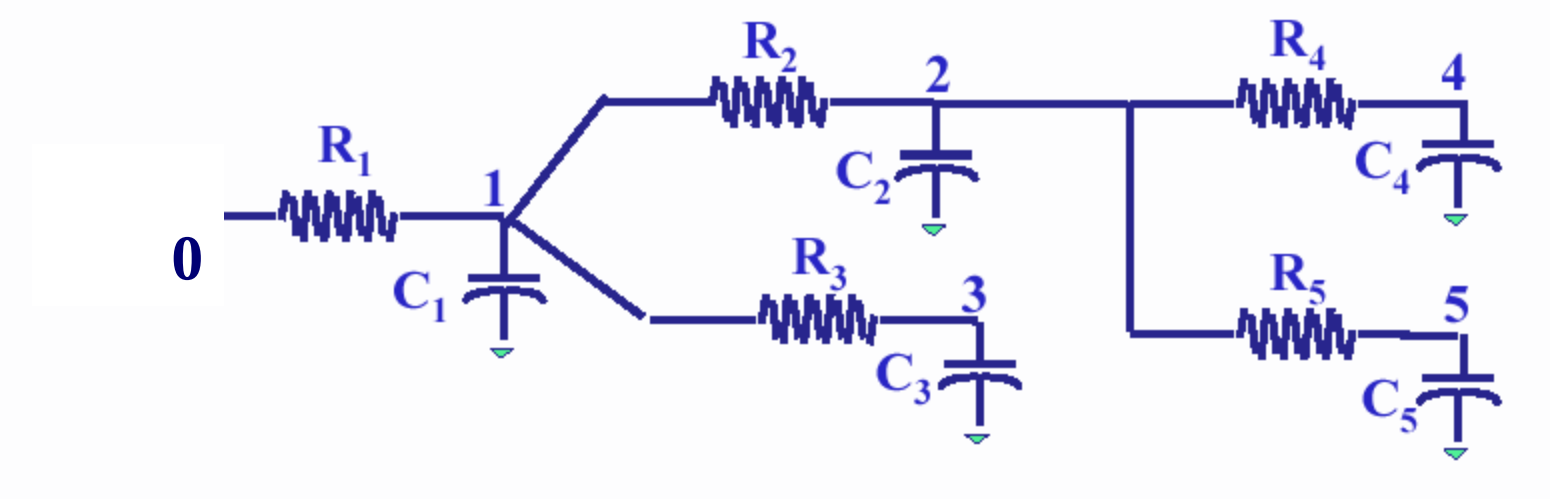
- The Elmore delay is an approximation of actual delay
  - Let $T_i$ denote the subtree rooted at node *i* for *i=1,2,...,N*
  - Let $P_i$ denote the unique path from input node to node i
  - Let $P_{ij} = P_i \cap P_j$ denote the portion of the path which is common between paths $P_i$ and $P_j$
- Assuming input is a step pulse at t=0, the Elmore delay at node *i* is calculated as follows
  - Resistance-oriented Formula:

$$\tau_{E,i} = \sum_{\substack{for\ all \\ j \in P_i}} R_j \sum_{\substack{for\ all \\ k \in T_i}} C_k$$

  - Capacitance-oriented Formula:

$$\tau_{Di} = \sum_{j=1}^{N} C_j \sum_{\substack{for\ all \\ k \in P_{ij}}} R_k$$

# Elmore Delay Calculation Example




- **Resistance-oriented Formula:**

$$\tau_{D4} = R_1(C_1 + C_2 + C_3 + C_4 + C_5) + R_2(C_2 + C_4 + C_5) + R_4(C_4)$$

- **Capacitance-oriented Formula:**

$$\tau_{D4} = C_1(R_1) + C_2(R_1 + R_2) + C_3(R_1) + C_4(R_1 + R_2 + R_4) + C_5(R_1 + R_2)$$

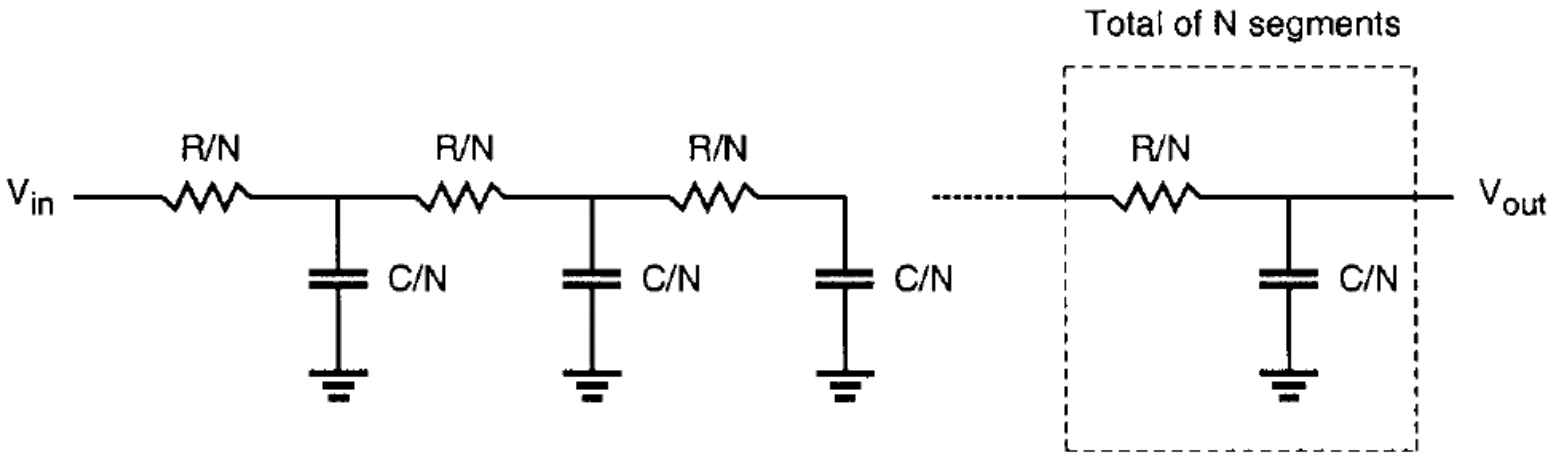
# Elmore Delay Calculation Example (Cont.)

# Distributed RC Ladder Network

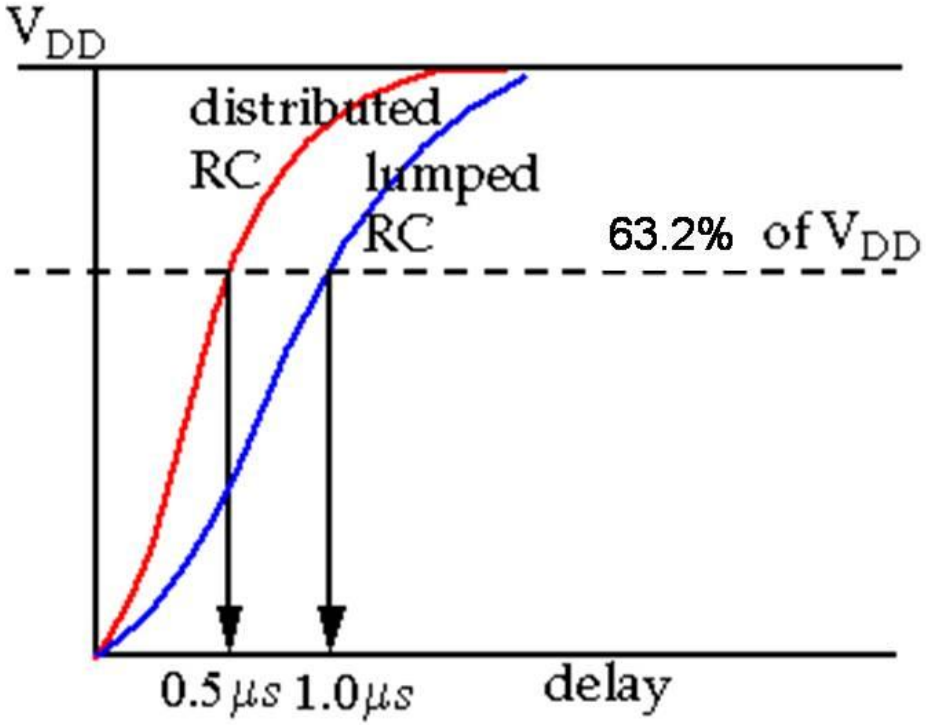
(C) The transient response of an interconnect line can be more accurately represented using an RC ladder network

Let $l$ denote the total length of interconnect, in the limit of N going to infinity, $r$ and $c$ denote resistance and capacitance per unit length of interconnect

$$\tau_{out,63.2\%V_{DD}} = \frac{RC}{2} = \frac{r\,c\,l^2}{2} \quad \text{for } N \to \infty$$



# Comparison Between Lumped and Distributed RC Delay




The lumped version is conservative by a factor of 2.

Usually, conservative models are preferred, particularly in cases which are difficult to approximate accurately otherwise.

Of course, since the distributed model is simple and is more accurate in this case, it is preferred.

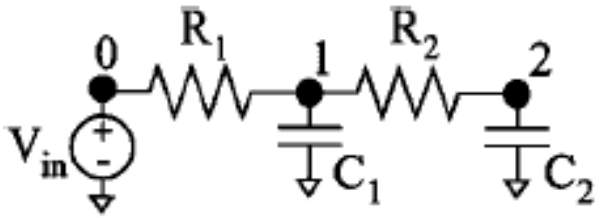
**We estimate propagation delay using RC time constants assuming that the time taken for a signal to reach 63.2% of its final value approximates the switching point of an inverter**
**- Use 0.69RC to calculate 50% prop. delays**

# Elmore Delay as a Bound for RC Trees

- The Elmore delay measure is an *upper bound* on the actual 50% delay of an RC tree response
  - Elmore delay is the true delay corresponding to an infinitely slow ramp
  - This bound approaches the 50% delay point at nodes further downstream from the source in an RC tree. Thus, as one moves away from the source, the Elmore delay becomes a better approximation of the net delay
  - Elmore delay can be inaccurate at the near-end of an RC tree since it ignores the *resistive shielding* effect

$$\tau_{E,1} = R_1(C_1 + C_2)$$

$$\tau_{actual,1} = \begin{cases} R_1(C_1 + C_2) & \text{if} \quad R_2 = 0 \\ R_1 C_1 & \text{if} \quad R_2 = \infty \end{cases}$$



# Elmore Delay as a Bound for RC Trees (Cont.)

- In this course, we will use the following equation for calculating the delay of RC interconnect (lumped RC, RC chain, etc.) which is more a more accurate estimate of the RC delay under step input:

$$\tau_{RC,i} = 0.69 \times \tau_{E,i}$$

# Logic Effective Resistance Estimation

- **Effective resistance of an NMOS transistor in**
  - **Linear region:**

$$R_{ds,n}(ON) = \frac{1}{\frac{\partial I_{ds}(linear)}{\partial V_{ds}}} = \frac{1}{k_n(V_{gs} - V_{t,n} - V_{ds})}$$

As $V_{ds} \rightarrow 0$, $R_{ds,n}(ON) = \frac{1}{k_n(V_{gs} - V_{t,n})}$

  - **Saturation region:**

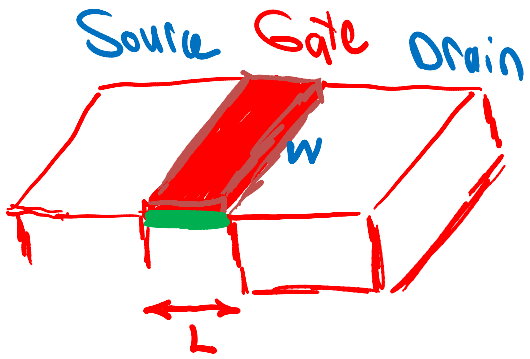
$$R_s(\text{NMOS driver}) = \frac{V_{DD}}{2I_{ds,n}(sat)} = \frac{V_{DD}}{k_n(V_{gs} - V_{t,n})^2}$$

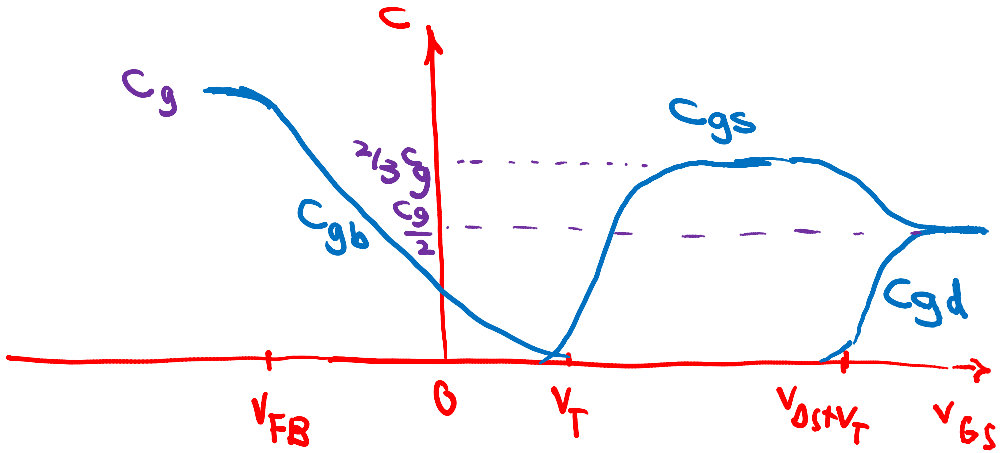
# Logic Effective Resistance (Cont.)

- Verify the saturation formulae

- Exercise: Check the formula on Weste's book, page 154

# Logic Effective Cap Estimation

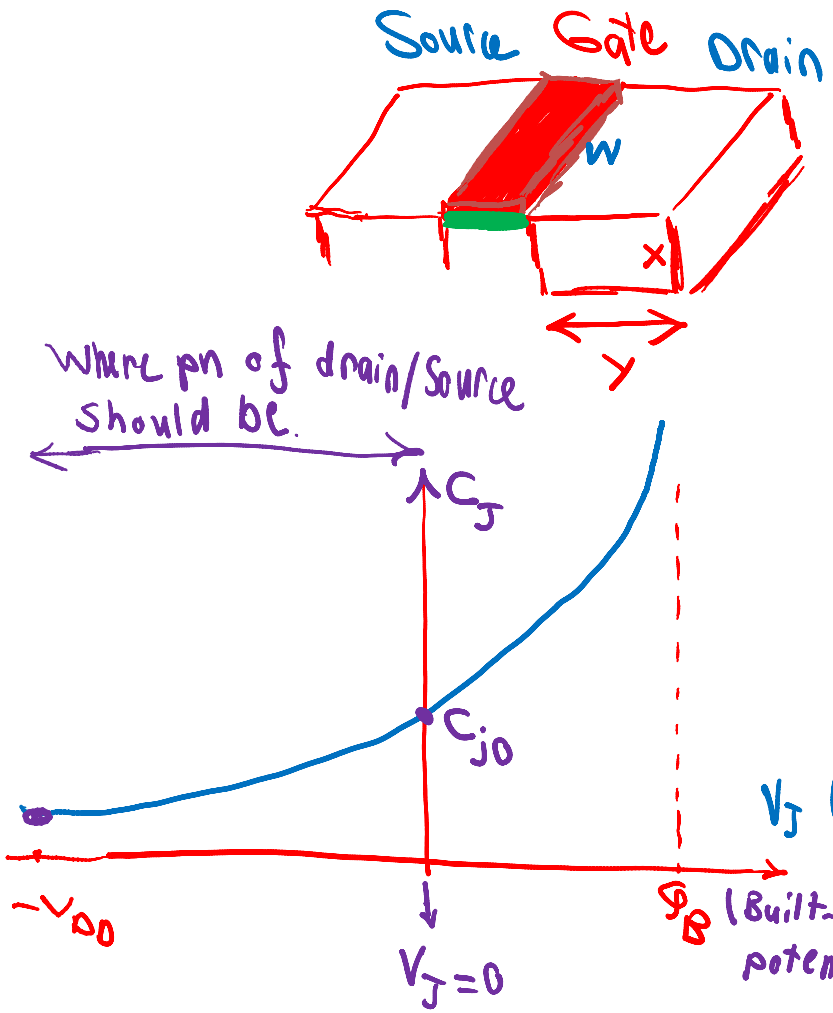



$$C_{gate} = WL\frac{\epsilon_{ox}}{t_{ox}} = WLC_{ox} = WC_g \qquad C_g \triangleq \frac{L\epsilon_{ox}}{t_{ox}}$$

Assuming L and $t_{ox}$ scale at the same rate $\Rightarrow C_g$ remains constant about $1.6\,fF/\mu m$

# Logic Effective Cap Estimation (Cont.)




Approximation $C_{db} = \frac{Q_j(V_2) - Q_j(V_1)}{V_2 - V_1} = \frac{\Delta Q}{\Delta V}$

$C_{db} = W(X+Y)C_{jb} + (2W+Y) \cdot C_{jsw}$ (1)

Simpler approximation:

$C_{db} \propto W \Rightarrow C_{db} = kW$

Example: 130nm process

$C_J(V_J=0) = 0.56\,W$ fF

$C_J(V_J=-1.2) = 0.4W$

equivalent using above approximation $\Rightarrow$ $0.45W$

# Driving an RC Line

- Source resistance of an FET driving a distributed RC line low
  - $R_s$ denotes the equivalent resistance of the driver
  - Typically, it is calculated as:

$$R_s(\text{NMOS driver}) = \frac{V_{DD}}{2I_{ds,n}(sat)} = \frac{V_{DD}}{k_n(V_{gs} - V_{t,n})^2}$$

- Delay from source driver to load is (wire as a single π section):

$$\tau_{out,63.2\%V_{DD}} = R_s(C_{diff} + C_{t,w} + C_L) + R_{t,w}\left(\frac{C_{t,w}}{2} + C_L\right)$$

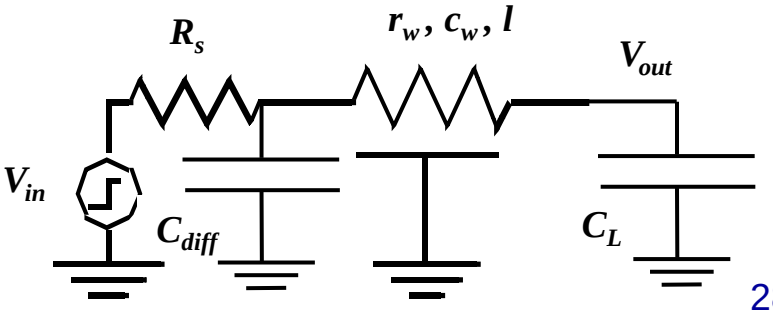
$$= R_s(C_{diff} + c_w l + C_L) + (r_w l)\left(\frac{c_w l}{2} + C_L\right)$$

$$\tau_{out} = R_s(C_{diff} + c_w l + C_L) + 0.69(r_w l)C_L + 0.38 r_w c_w l^2$$

  - Note that 0.38 is an empirical coefficient; it is not equal to 0.69/2 !

$$C_{t,w} = c_w l$$

$$R_{t,w} = r_w l$$



# Calculation of Output Delay of a CMOS Inverter Driving FO3 Load

- **Using a π-section model of each wire segment, we can write:**

implies the delay of driver + wire 2, i.e., $delay_{0\_to\_2}$

$$C_g \equiv C_{g,1} = C_{g,2} = C_{g,3} = (W_n + W_p) L C_{ox}$$

$$C_{wire} \equiv C_{wire,1} = C_{wire,2} = C_{wire,3} = W_{wire} L_{wire} C_{fox}$$

$$C_{diff} = C_{db,n} + C_{db,p}; \quad C_{tot} = C_{diff} + 3C_{wire} + 3C_g$$

$$C_{db,n} = W_n \left(Y_{d,n} + x_j\right) C_{j0,n} K_{eq,n} + (W_n + 2Y_{d,n}) C_{jsw,n} K_{eq,n}(sw)$$

$$C_{db,p} = W_p \left(Y_{d,p} + x_j\right) C_{j0,p} K_{eq,p} + (W_p + 2Y_{d,p}) C_{jsw,p} K_{eq,p}(sw)$$

Method 1: $Delay_{HL,2} = t_{pHL} + \tau_{RC,2} = \dfrac{C_{tot} V_{DD}}{I_{out}(sat) + I_{out}(V_{out} = \frac{V_{DD}}{2})}$

$$+ 0.69 \times R_{wire} C_g + 0.38 \times R_{wire} C_{wire}$$

Alternatively,

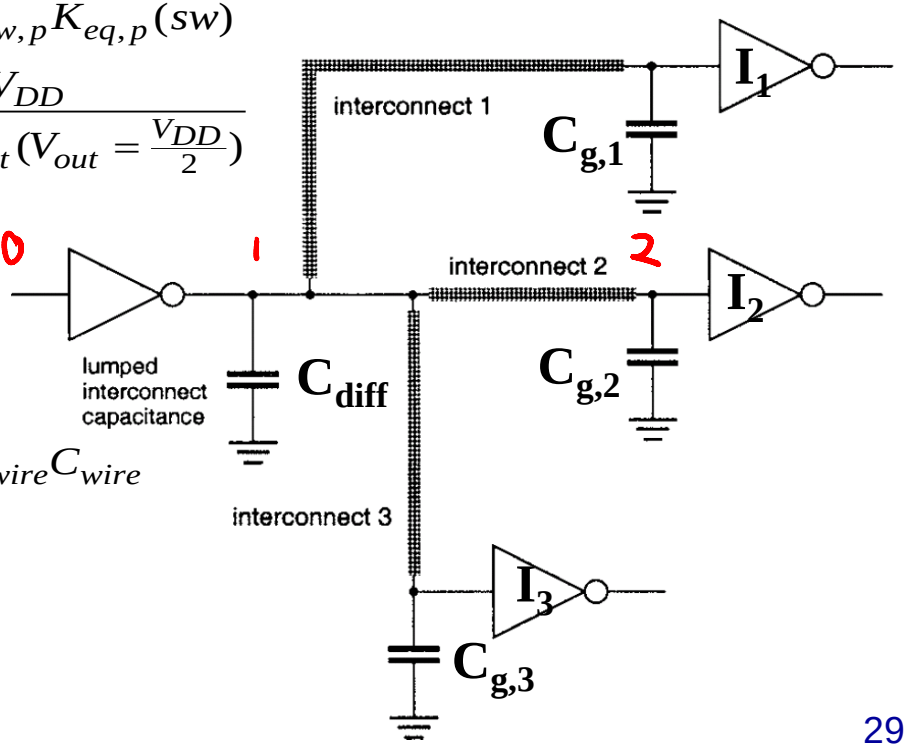
Method 2: $R_{out,n} = \dfrac{V_{DD}}{2 I_{out}(sat)} = \dfrac{V_{DD}}{k_n (V_{DD} - V_{t,n})^2}$

$$Delay'_{HL,2} = R_{out,n}\left(C_{tot} + R_{wire} C_g\right) + 0.38 \times R_{wire} C_{wire}$$

0.69

Clearly, $Delay'_{HL,2} < Delay_{HL,2}$.

Typically, method 1 yields more accurate results.

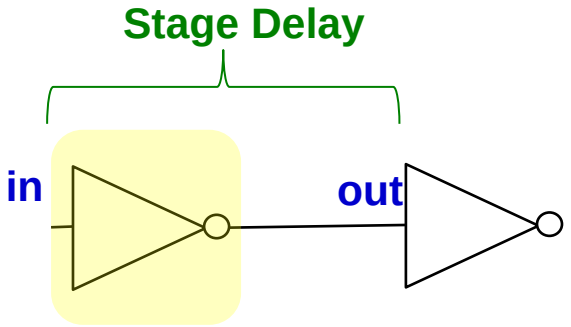


# Example: Transistor Sizing to Minimize Stage Delay

- **Consider a CMOS inverter driving a FO1 load; Assume that wire widths and lengths cannot be optimized (so wire load is a fixed value) . Size the circuit to minimize the stage delay. Model the interconnect as a simple lumped RC**

$$C_{tot} = \alpha_0 + \alpha_n W_n + \alpha_p W_p \ , \quad C_g = (W_n + W_p) L C_{ox}$$

$$V_t \equiv V_{t,n} = \left|V_{t,p}\right|, \quad k_n = \mu_n C_{ox} \frac{W_n}{L}, \quad k_p = \mu_p C_{ox} \frac{W_p}{L}$$




$$Delay_{HL,2} = t_{pHL} + \tau_{RC,2} = \frac{\alpha_0 + \alpha_n W_n + \alpha_p W_p}{\mathbf{G}\, \mu_n W_n} + \gamma_0 + \gamma_1 \left(W_n + W_p\right)$$

$$Delay_{LH,2} = t_{pLH} + \tau_{RC,2} = \frac{\alpha_0 + \alpha_n W_n + \alpha_p W_p}{\mathbf{G}\, \mu_p W_p} + \gamma_0 + \gamma_1 \left(W_n + W_p\right)$$

witl $\mathbf{G} = \frac{C_{ox}\left(V_{DD} - V_{tn}\right)^2}{L V_{DD}}$. Setting $\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p}$ to ensure equal $t_{pHL}$ and $t_{pLH}$ delays, we obtain:

$$Delay_{HL,2} = \delta_0 + \frac{\delta_1}{W_n} + \delta_2 W_n \text{ where } \delta_0 = \frac{\alpha_n + \alpha_p \frac{\mu_n}{\mu_p}}{\mathbf{G}\, \mu_n} + \gamma_0, \delta_1 = \frac{\alpha_0}{\mathbf{G}\, \mu_n}, \delta_2 = \gamma_1 \left(1 + \frac{\mu_n}{\mu_p}\right)$$

- **The minimum delay is achieved exactly when:**

$$\frac{\partial Delay_{HL,2}}{\partial W_n} = 0 \Rightarrow -\frac{\delta_1}{W_n^2} + \delta_2 = 0 \Rightarrow W_n = \sqrt{\frac{\delta_1}{\delta_2}}$$

# Elmore Delay and Transition Time Calculation in RC Trees Under Ramp Input

- Ignoring the resistive shielding effect, in an RC tree with source *s* and destination *i*, assume an Elmore delay of $\tau_{E,i}$ for a step input applied to s. The RC **propagation delay** under a ramp input ($\tau_{RC\text{-}ramp,i}$) with 10-90% transition time, $T_{in,s}$, may be calculated as (Mita et al):

$$\tau_{RC\text{-}ramp,i} = \left(1 - \frac{0.31}{1 + 0.13 \times \left(\frac{T_{in,s}}{\tau_{E,i}}\right)^2}\right) \times \tau_{E,i}$$

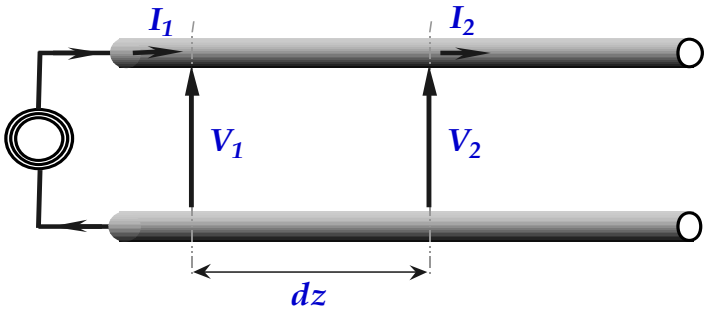
  - Notice that when $T_{in,s}=0$, $\tau_{RC\text{-}ramp,i} = \tau_{RC,i} = 0.69 \times \tau_{E,i}$ and when $T_{in,s} \rightarrow \infty$, $\tau_{RC\text{-}ramp,i} = \tau_{E,i}$
- The 10-90% output transition time $T_{out,i}$ under a ramp input is calculated as:

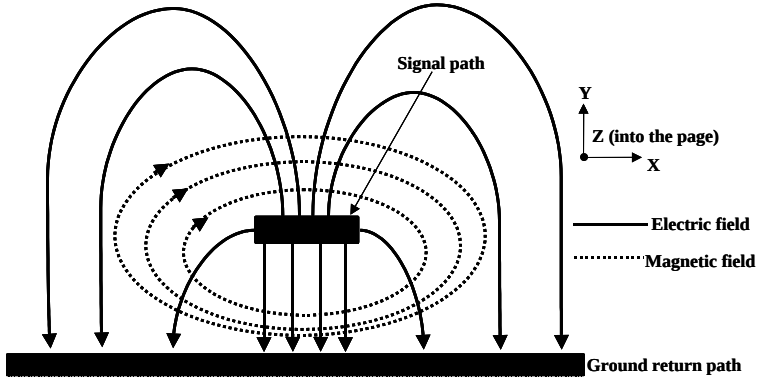
$$T_{out,i} = \sqrt{\left(2.2 \times \tau_{E,i}\right)^2 + {T_{in,s}}^2}$$

  - Notice that when $T_{in,s}=0$, $T_{out,i} = 2.2 \times \tau_{E,i}$ and when $T_{in,s} \rightarrow \infty$, $T_{out,i} \rightarrow T_{in,s}$

# [Optional] Transmission Line Model of VLSI Interconnections

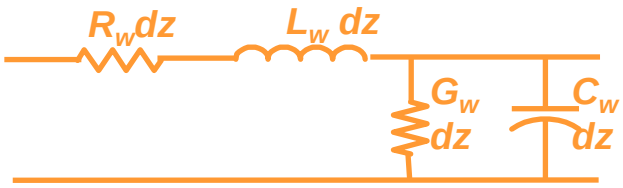
- **We must use *Telegrapher's equations* to derive the waveform propagation in a T-line**




**The signal is really the wave propagating between the conductors.**

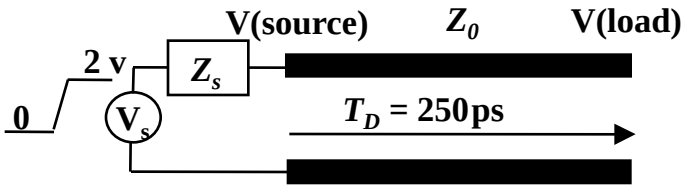
$$\left.\begin{aligned}\frac{\partial}{\partial z}V(z,t) &= -L\frac{\partial}{\partial t}I(z,t) - RI(z,t)\\ \frac{\partial}{\partial z}I(z,t) &= -C\frac{\partial}{\partial t}V(z,t) - GV(z,t)\end{aligned}\right\} \Rightarrow \left\{\begin{aligned}\frac{\partial^2}{\partial z^2}V &= L_wC_w\frac{\partial^2}{\partial t^2}V + (R_wC_w + G_wL_w)\frac{\partial}{\partial t}V + G_wR_wV\\ \frac{\partial^2}{\partial z^2}I &= L_wC_w\frac{\partial^2}{\partial t^2}I + (R_wC_w + G_wL_w)\frac{\partial}{\partial t}I + G_wR_wI\end{aligned}\right.$$




**RLGC model of a section of the T-Line**

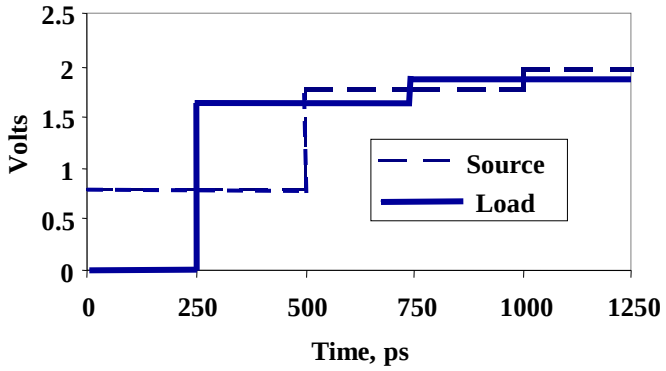
# [Optional] Launching a Waveform in a T-Line

- An RLGC interconnection tree with typical signal waveforms at nodes A and B, showing signal delay and various delay components
  - When a voltage waveform is launched into a transmission line, we may have two cases: under- or over-damped depending on source and characteristic impedances
  - There are multiple reflections from the end of the line before the final waveform is constructed or settles to its steady-state form

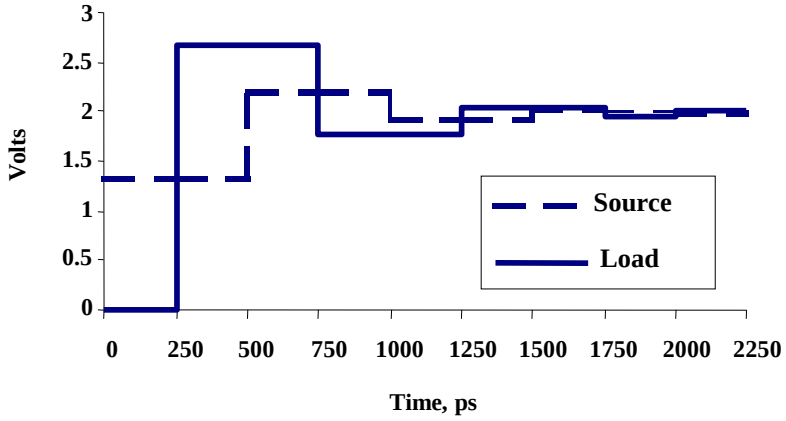



$$Z_0 = \sqrt{\frac{L_w}{C_w}} = 50\ \Omega$$

$$\frac{1}{T_d} \equiv \omega_n = \frac{1}{l\sqrt{L_w C_w}} = 4\times10^9\ rad/\sec$$



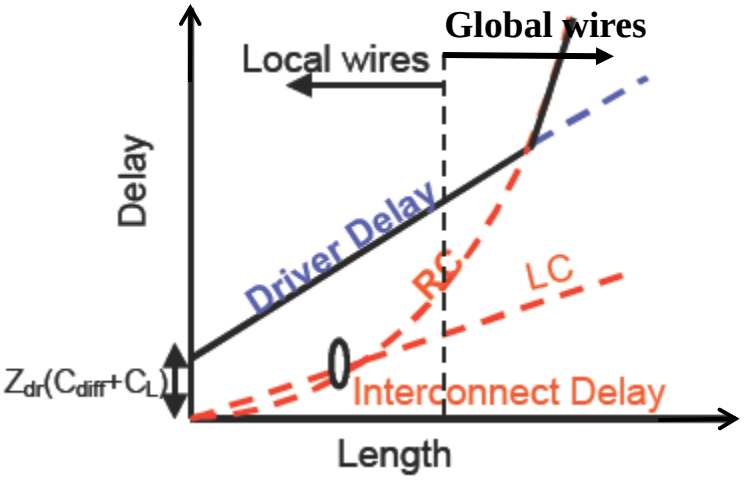

Over-damped case : $Z_s$=75 Ω



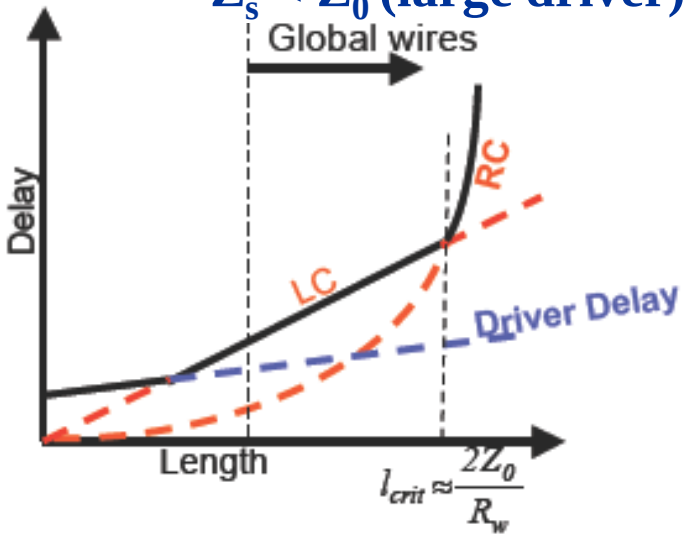

Under-damped case: $Z_s$=25 Ω
The only case that inductance matters!

# [Optional] Delay Calculation for Local and Global Interconnections

**$Z_s > Z_0$ (small driver)**




**$Z_s < Z_0$ (large driver)**




$$Delay_{local} = 0.69 Z_s (C_{diff} + C_w l + C_L)$$

*if* $Z_s > Z_0$ *(weak driver), then*

$$Delay_{global} = 0.69 Z_s (C_{diff} + C_w l + C_L) + 0.69 (R_w l) C_L + 0.38 R_w C_w l^2$$

*else*

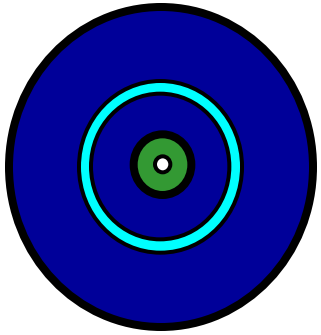
*if* $l < l_{crit}$ *(i.e.,* $\frac{R_w l}{2Z_0} < 1$*), then*

$$Delay_{global} = \sqrt{L_w C_w \left(1 + \frac{C_L}{C_w l}\right)}\, l$$

*else*

$$Delay_{global} = 0.69 (R_w l) C_L + 0.38 R_w C_w l^2$$

# [Optional] Skin Effect



- At DC, if each ring has the same current, which one has more inductance?
- At high frequency, which ring has more impedance?
- If you were an electron, where would you want to be?

- The center path has the highest inductance since it has all the field lines of the outer ring plus the field lines that are between it and the outer ring. In general, the self inductance of the current path will decrease as you approach the outer edge

# [Optional] Skin Effect (Cont.)

- The tendency of alternating current to flow near the surface of a conductor, thereby restricting the current to a small part of the total cross-sectional area and increasing the resistance to the flow of current
- High frequency current flows primarily on the surface of a conductor with current density falling off exponentially with depth into the conductor:

$$J = \exp\left(-\frac{d}{\delta}\right)$$

- The skin depth, d, the depth where the current has fallen off to exp(-1) of its normal value is given by:

$$\delta = \left(\pi f \mu \sigma\right)^{-1/2}$$

- $\sigma = 1/\rho$ is the conductivity of the material, and $f$ is the frequency of the signal