Digital IC Design

Lec 3: MOS/Wire RC for Transient Time

黃柏蒼 Po-Tsang (Bug) Huang

bughuang@nycu.edu.tw

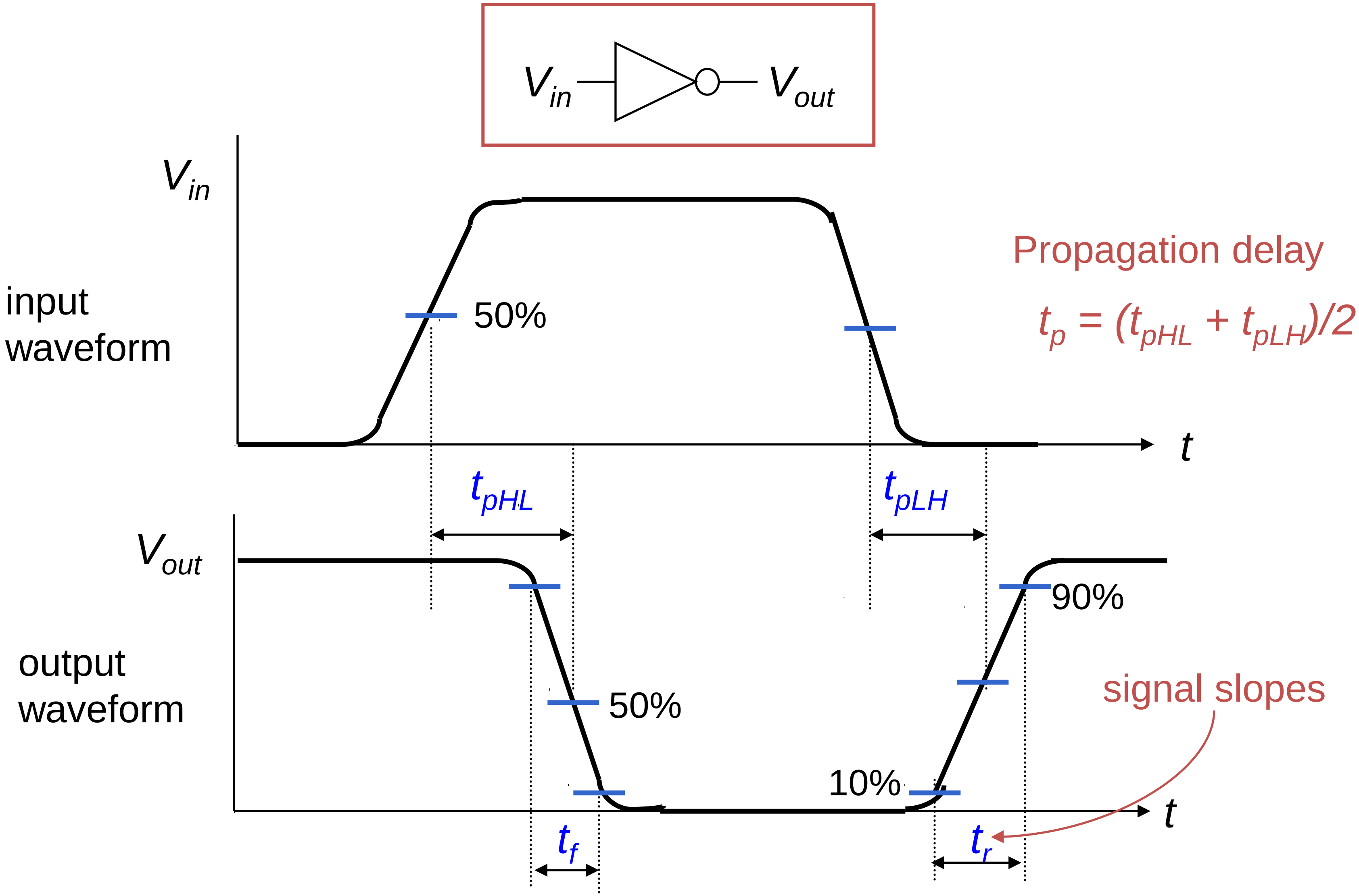
International College of Semiconductor Technology

National Chiao Tung Yang Ming University

國立陽明交通大學

NATIONAL YANG MING CHIAO TUNG UNIVERSITY

Delay Definitions



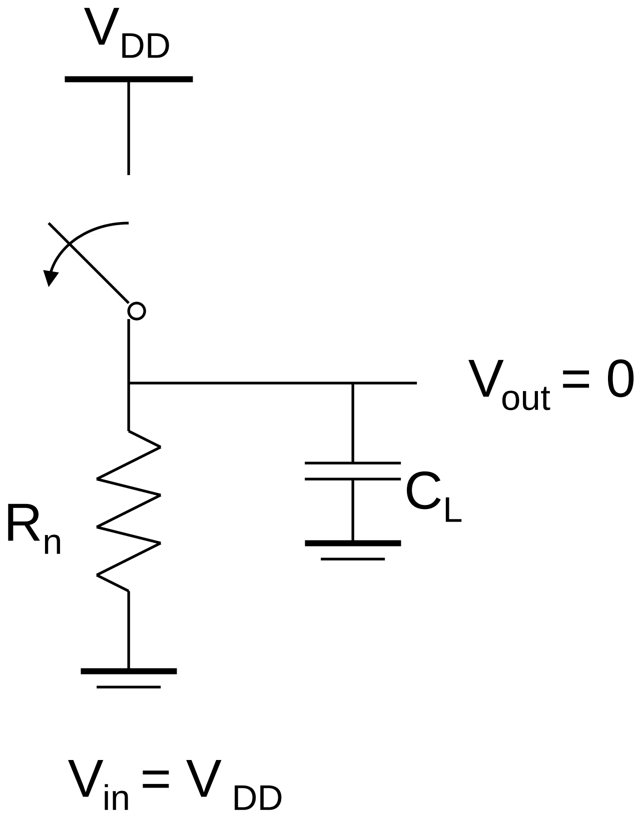
Propagation delay

$t_p = (t_{pHL} + t_{pLH})/2$

signal slopes

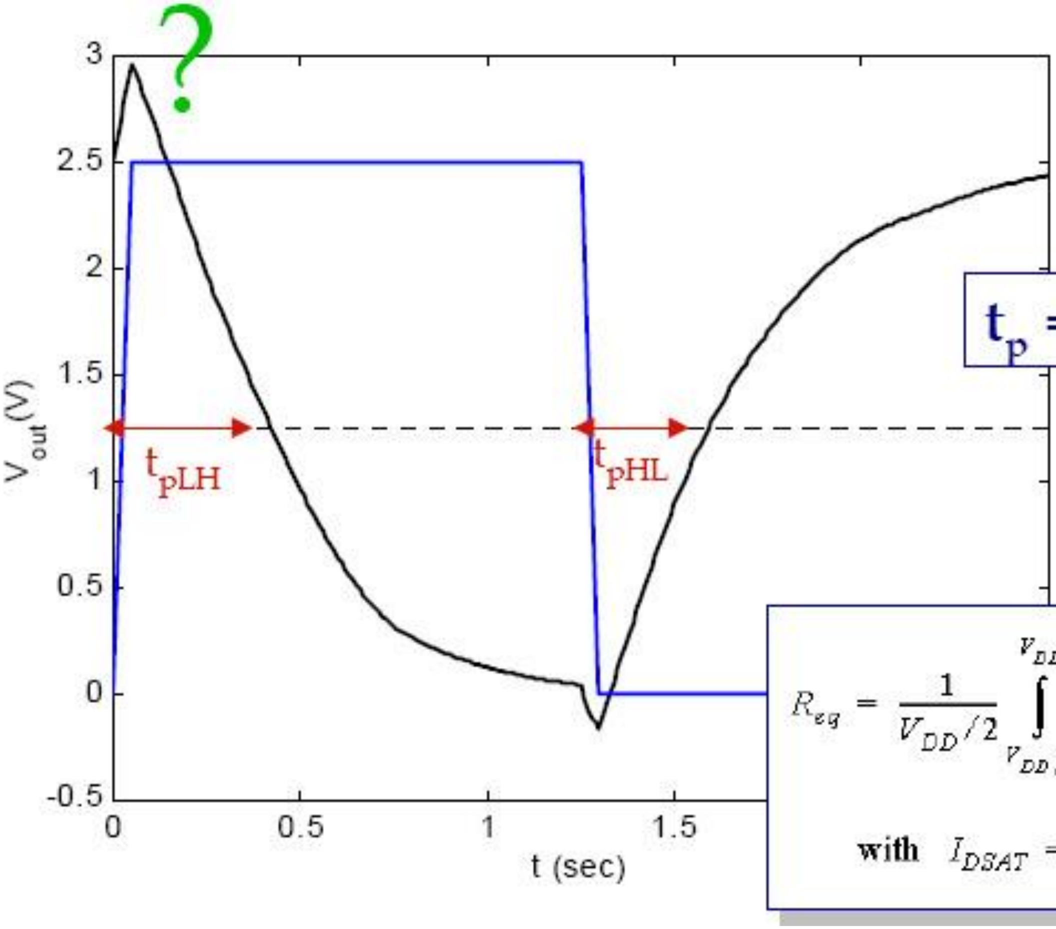
CMOS Inverter : Dynamic

- Transient, or dynamic, response determines the maximum speed at which a device can be operated.



$$\boxed{t_{pHL} = f(R_n, C_L)}$$

Transient Response



$$t_p = 0.69\ C_L\ (R_{eqn}+R_{eqp})/2$$

$$R_{eq} = \frac{1}{V_{DD}/2}\int_{V_{DD}/2}^{V_{DD}} \frac{V}{I_{DSAT}(1+\lambda V)}dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right)$$

$$\textbf{with}\quad I_{DSAT} = k'\frac{W}{L}\left((V_{DD}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}\right)$$

Inverter Propagation Delay (Designer)

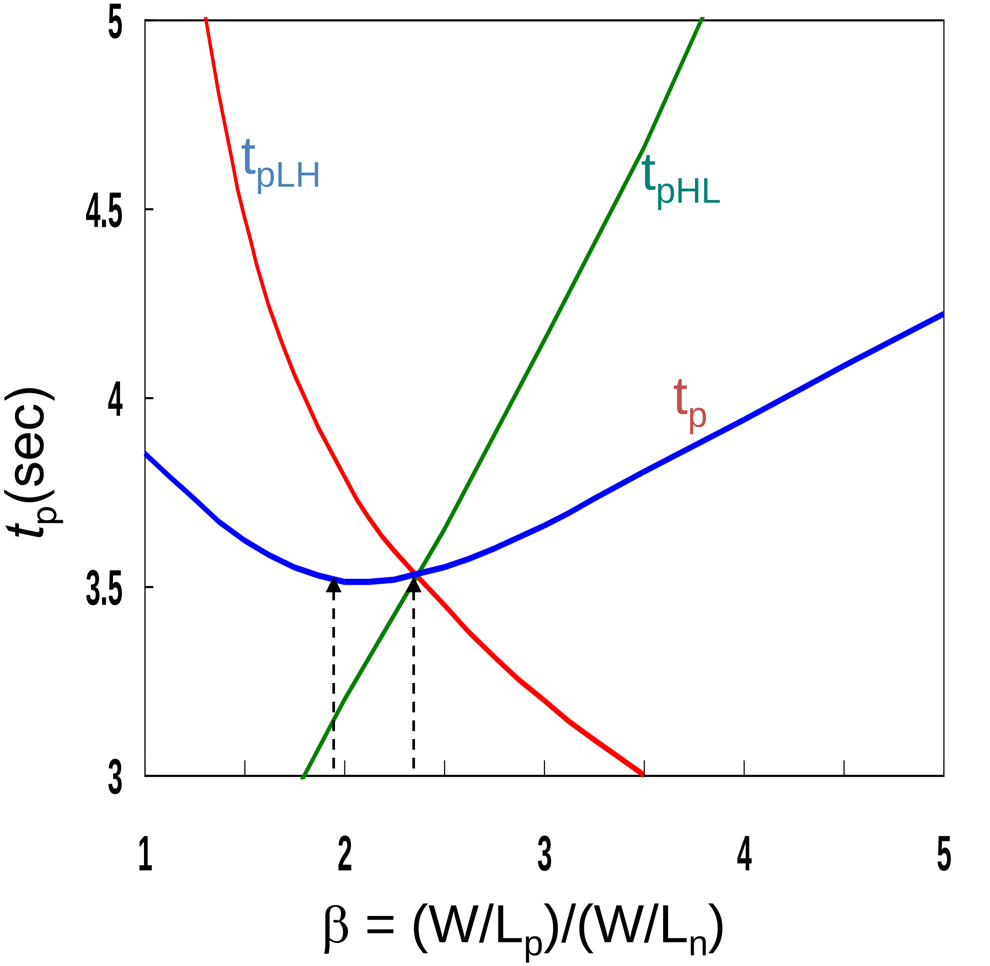
- To see how a designer can optimize the delay of a gate have to expand the R_{eq} in the delay equation

$$t_{pHL} = 0.69\ R_{eqn}\ C_L$$

$$= 0.69\ (3/4\ (C_L\ V_{DD})/I_{DSATn}\ )$$

$$\approx 0.52\ C_L\ /\ (W/L_n\ k'_n\ V_{DSATn}\ )$$

Impacts of NMOS/PMOS Ratio



$\beta = (W/L_p)/(W/L_n)$

β of 2.4 (= 31 kΩ/13 kΩ) gives symmetrical response

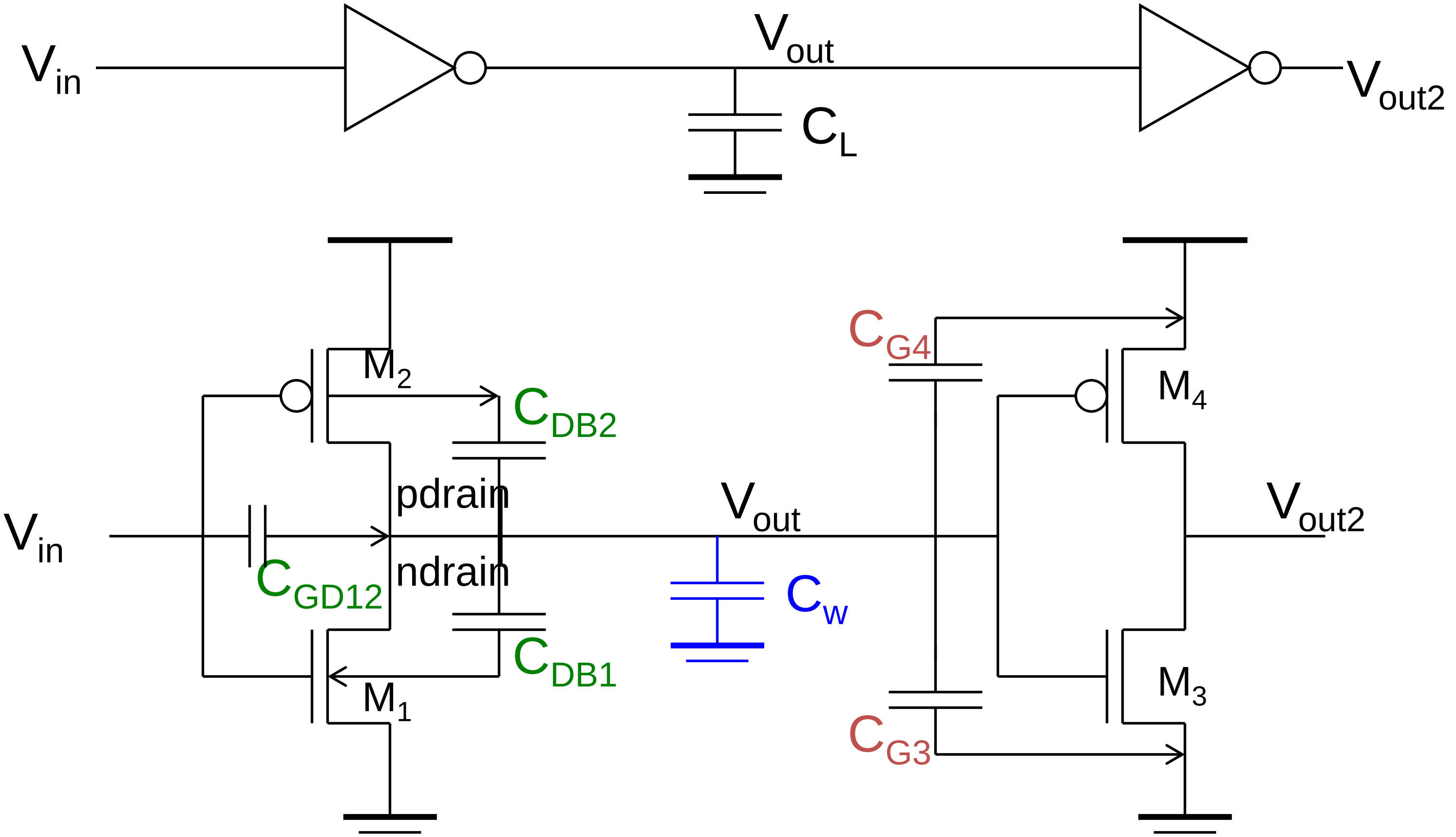
β of 1.6 to 1.9 gives optimal performance

Calibrating Delays

- **Step RC delay model is a good first-order approximation**
- **Accuracy can be improved by including:**
  - Slope effects
  - Non-linear capacitive loading
  - Signal arrival times
  - Wire models

Sources of Capacitance

- intrinsic MOS transistor capacitances
- extrinsic MOS transistor (fanout) capacitances
- wiring (interconnect) capacitance



MOS Capacitances

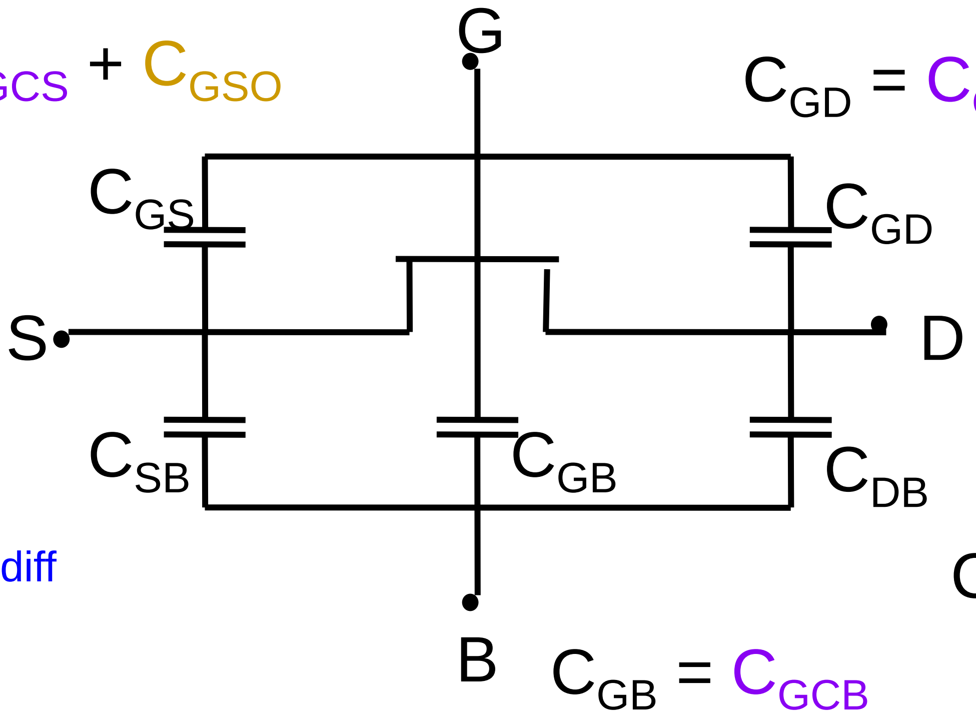
- **Gate capacitance**
  - Non-linear channel capacitance
  - Linear overlap, fringing capacitances
  - Miller effect on overlap capacitance
- **Non-linear drain diffusion capacitance**
  - PN junction
- **Wiring capacitances**
  - Linear

Intrinsic MOS Capacitances

- Structure capacitances
- Channel capacitances
- Diffusion capacitances from the depletion regions of the reverse-biased *pn*-junctions

$C_{GS} = C_{GCS} + C_{GSO}$

$C_{GD} = C_{GCD} + C_{GDO}$



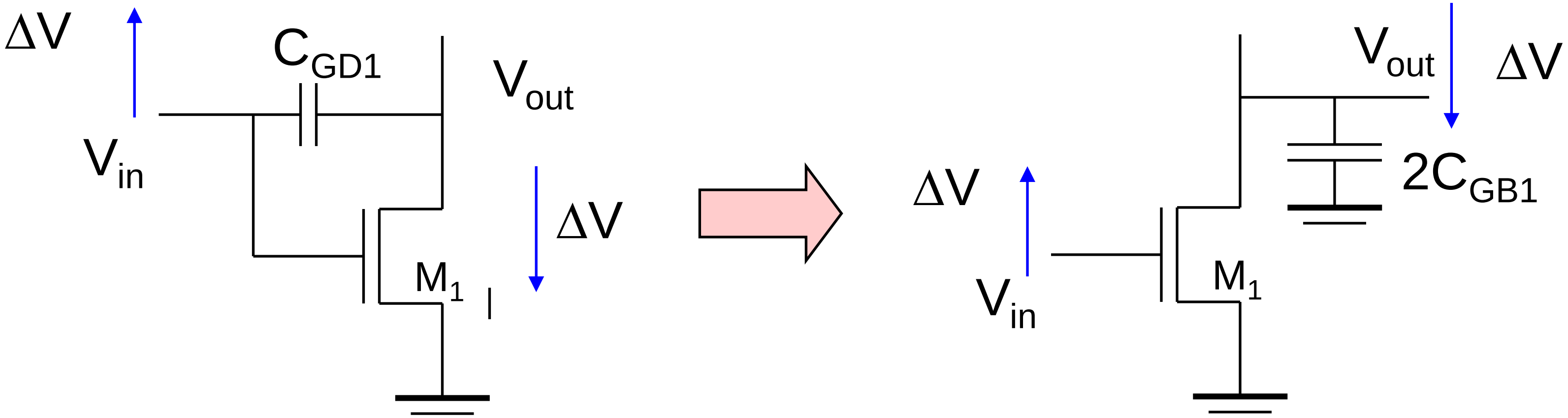
$C_{SB} = C_{Sdiff}$

$C_{DB} = C_{Ddiff}$

$C_{GB} = C_{GCB}$

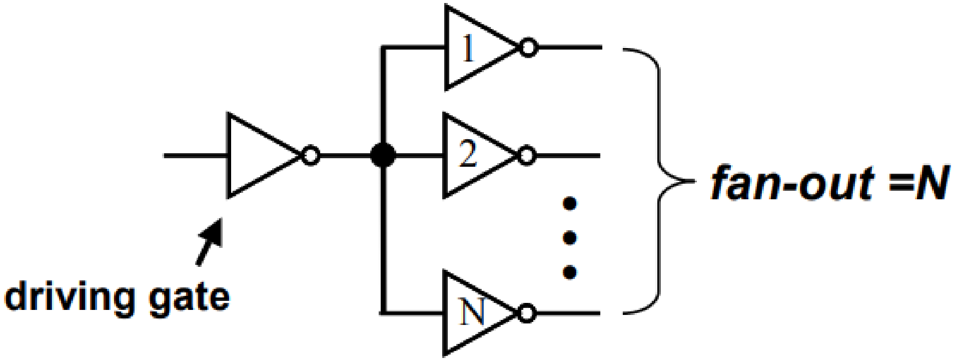
Gate-Drain Capacitance: The Miller Effect

- M1 and M2 are either in cut-off or in saturation.
- The floating gate-drain capacitor is replaced by a capacitance-to-ground (gate-bulk capacitor).
- Miller Effect: A capacitor experiencing identical but opposite voltage swings at both its terminals can be replaced by a capacitor to ground whose value is two times the original value.



Fan-Out of a Cell (gate)

- Typically, the output of a logic gate is connected to the input(s) of one or more logic gates
- The fan-out is the number of gates that are connected to the output of the driving gate



- Fanout leads to increased capacitive load on the driving gate, and therefore longer propagation delay
  - The input capacitances of the driven gates sum, and must be charged through the equivalent resistance of the driver

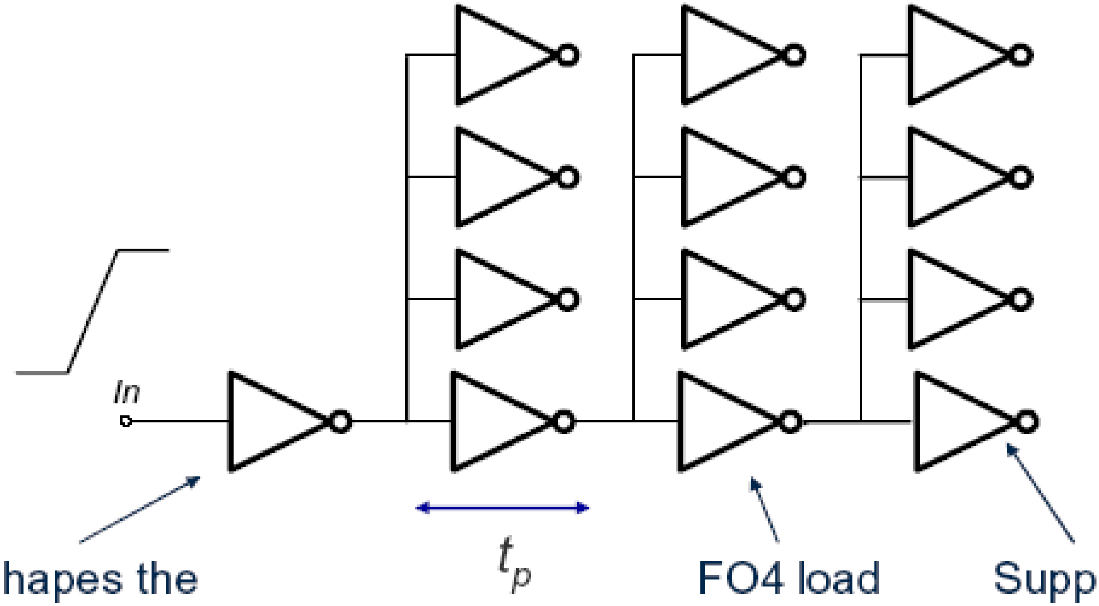
Extrinsic (Fan-Out) Capacitance

- The extrinsic, or fan-out, capacitance is the total gate capacitance of the loading gates M3 and M4.

$$\begin{aligned} C_{fan\text{-}out} &= C_{gate}\,(NMOS) + C_{gate}\,(PMOS) \\ &= (C_{GSOn} + C_{GDOn} + W_n L_n C_{ox}) + (C_{GSOp} + C_{GDOp} + W_p L_p C_{ox}) \end{aligned}$$

- Simplification of the actual situation
  - Assumes all the components of C_{gate} are between V_{out} and GND (or V_{DD})
  - Assumes the channel capacitances of the loading gates are constant

FO4 Inverter Delay



Shapes the input slope to FO4

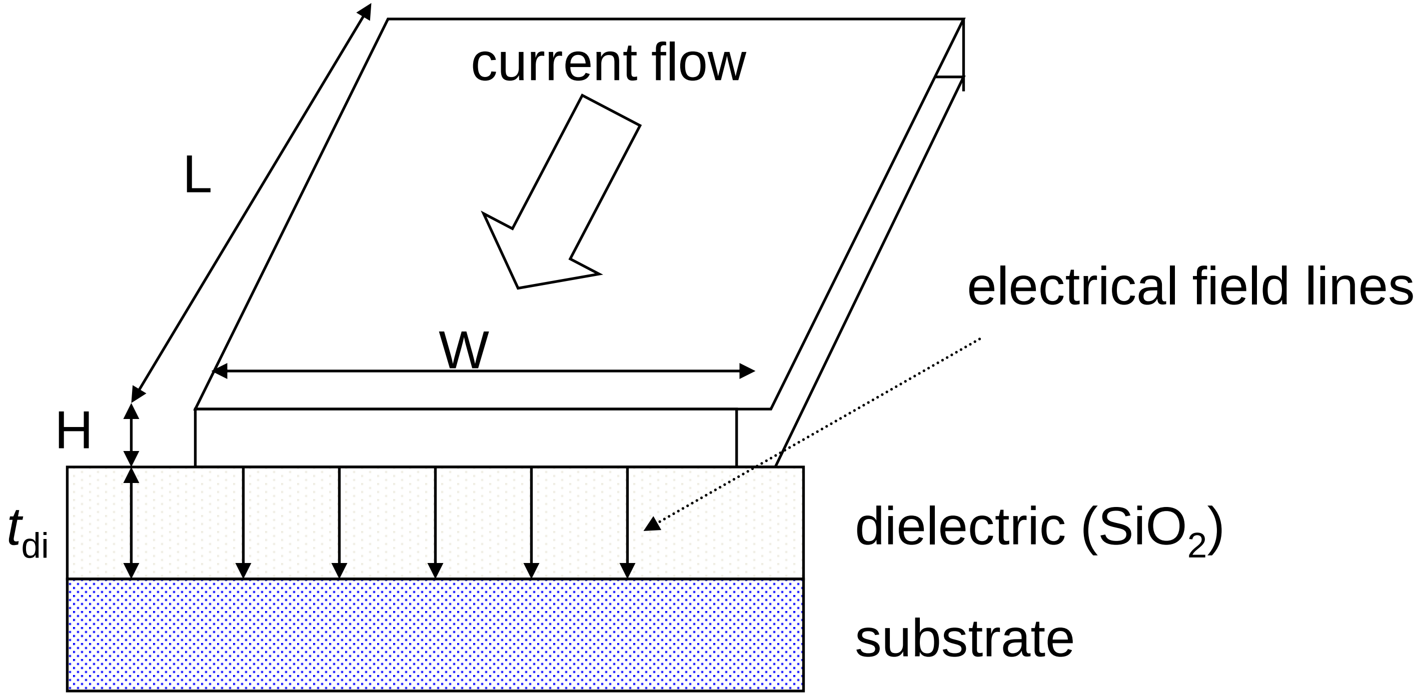
FO4 load

Suppresses Miller kickback

Wiring Capacitance

- The wiring capacitance depends upon the length and width of the connecting wires and is a function of the fan-out from the driving gate and the number of fan-out gates.
- Wiring capacitance is growing in importance with the scaling of technology.

Parallel Plate Wiring Capacitance



current flow

electrical field lines

dielectric (SiO_2)

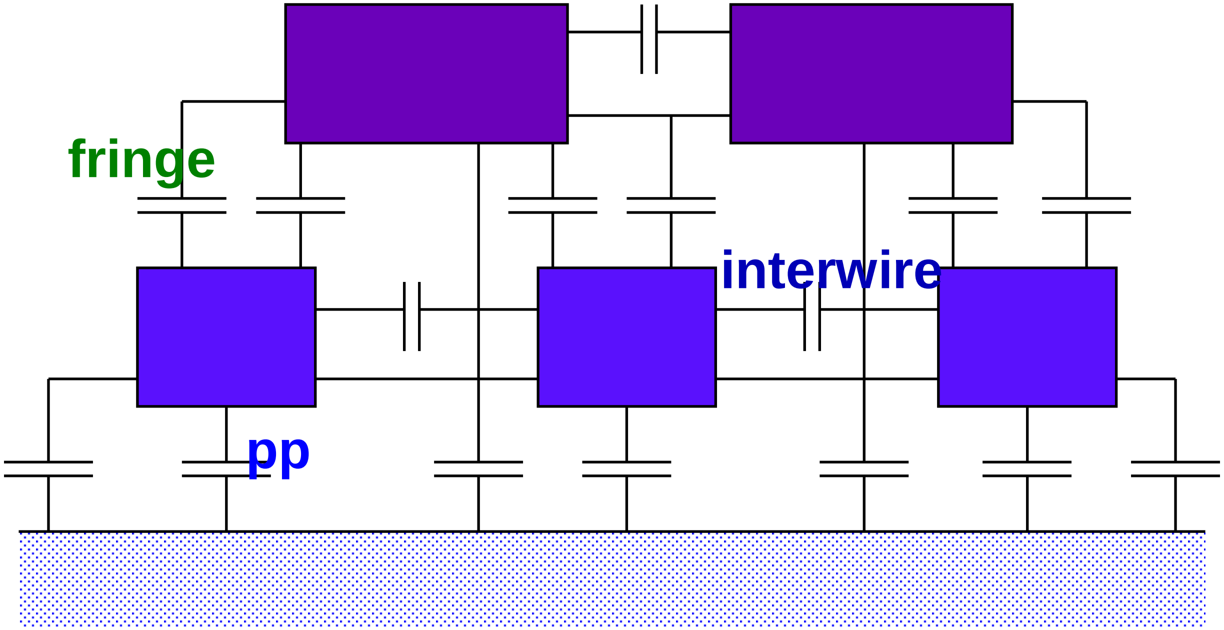
substrate

permittivity constant (SiO_2= 3.9)

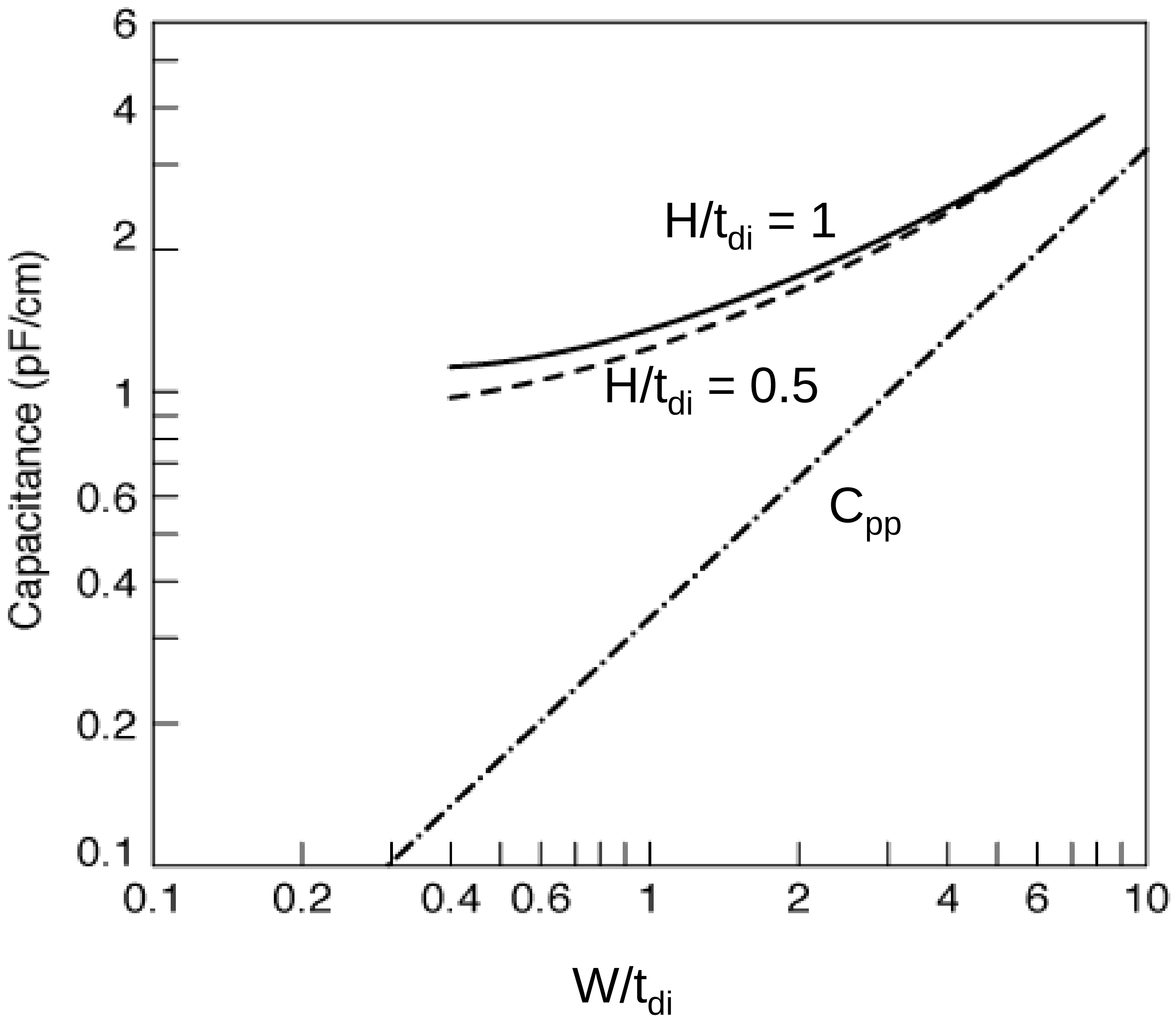
$$C_{pp} = (\varepsilon_{di}/t_{di})\ WL$$

Sources of Interwire Capacitance

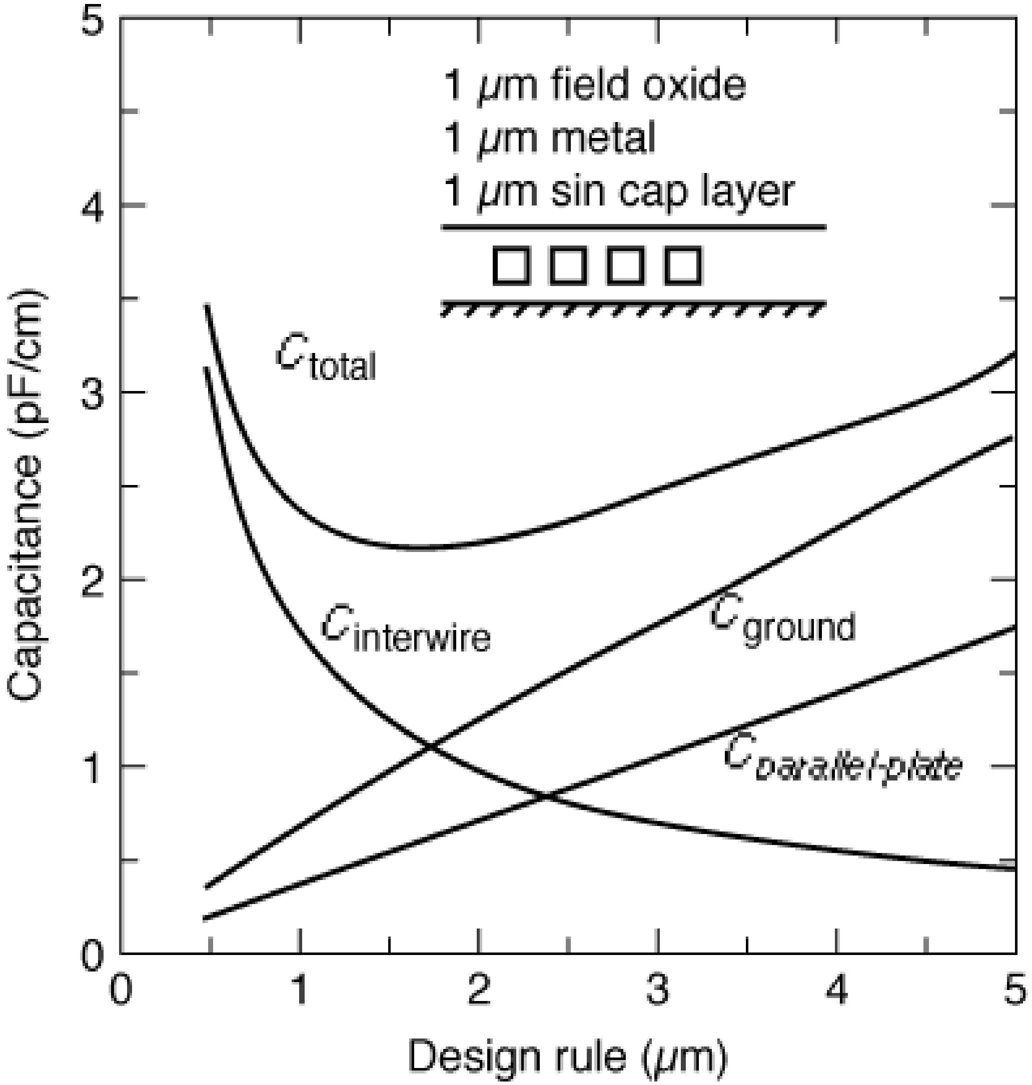
$$
\begin{aligned}
C_{wire} &= C_{pp} + C_{fringe} + C_{interwire} \\
&= (\varepsilon_{di}/t_{di})WL \\
&\quad + (2\pi\varepsilon_{di})/\log(t_{di}/H) \\
&\quad + (\varepsilon_{di}/t_{di})HL
\end{aligned}
$$



Impact of Fringe Capacitance



Impact of Interwire Capacitance

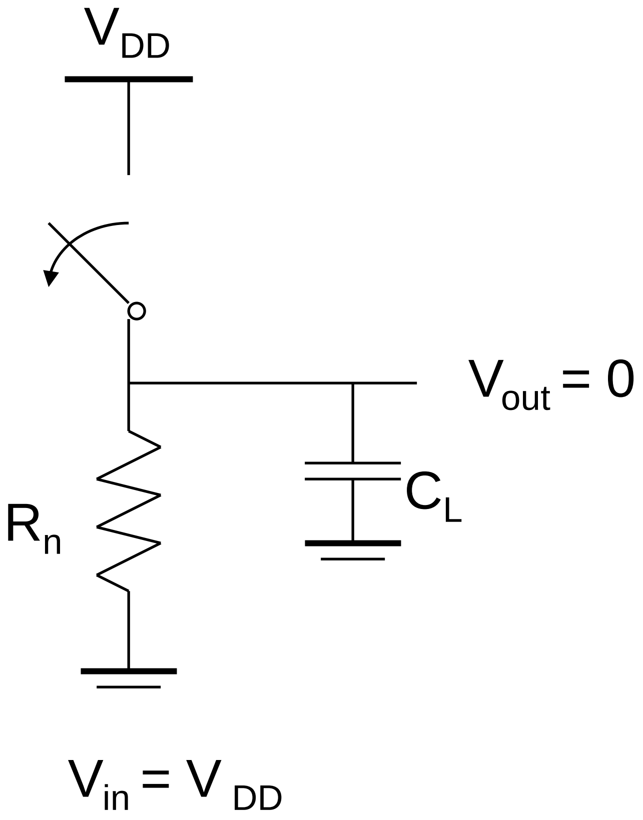


Wiring Insights

- For W/H < 1.5, the fringe component dominates the parallel-plate component. Fringing capacitance can increase the overall capacitance by a factor of 10 or more.
- When W/H < 1.75 interwire capacitance starts to dominate
- Interwire capacitance is more pronounced for wires in the higher interconnect layers (further from the substrate)
- Rules of thumb
  - Never run wires in diffusion
  - Use poly only for short runs
  - Shorter wires – lower R and C
  - Thinner wires – lower C but higher R
- Wire delay nearly proportional to L^2

CMOS Inverter : Dynamic

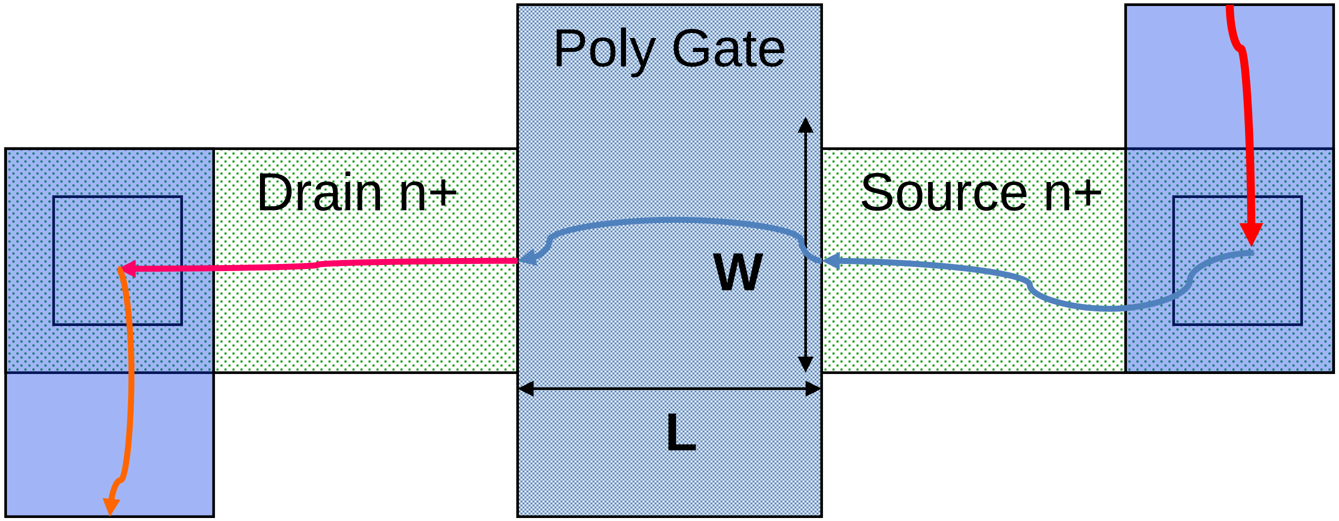
- Transient, or dynamic, response determines the maximum speed at which a device can be operated.



$$t_{pHL} = f(R_n, C_L)$$

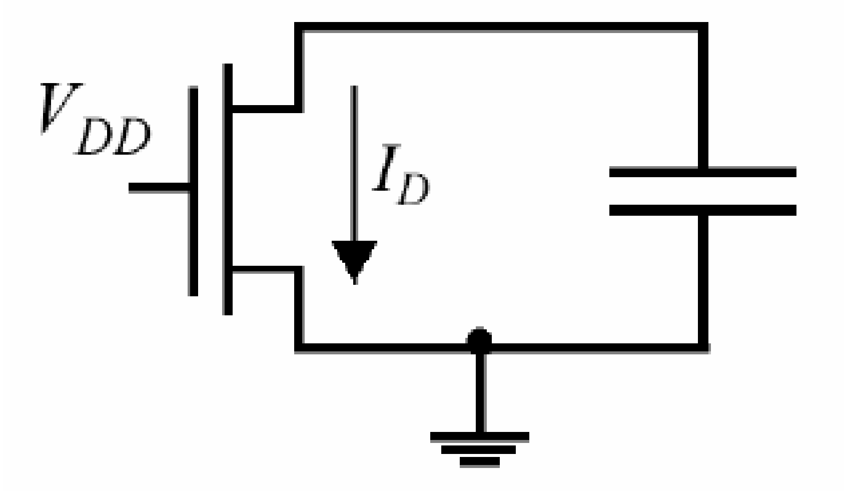
Sources of Resistance

- MOS structure resistance - R_{on}
- Source and drain resistance
- Contact (via) resistance
- Wiring resistance



MOS Transistor as a Switch

Discharging a capacitor



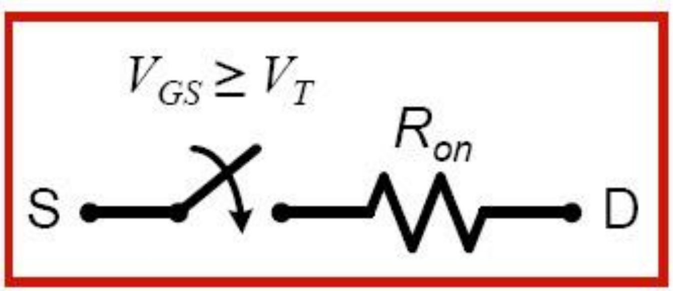
- Can solve:

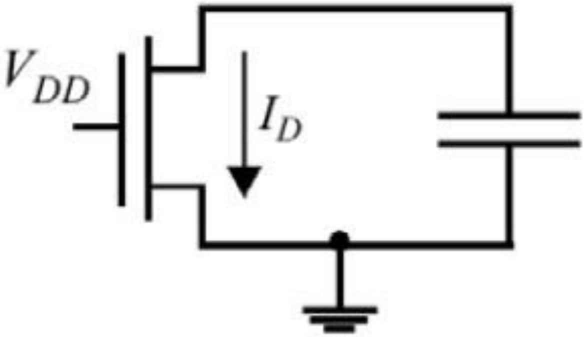
$$i_D = i_D(v_{DS})$$

$$i_D = C\frac{dV_{DS}}{dt}$$

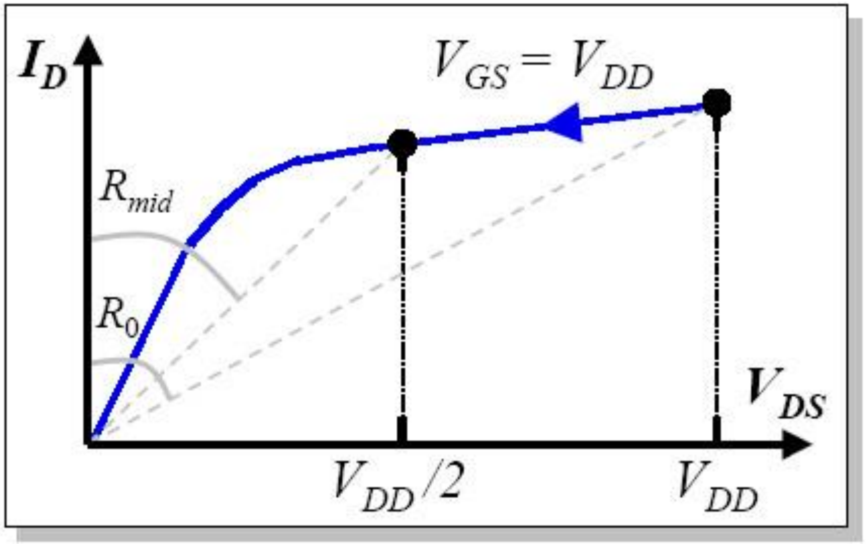
- Prefer using equivalent resistances

Equivalent MOS Resistance





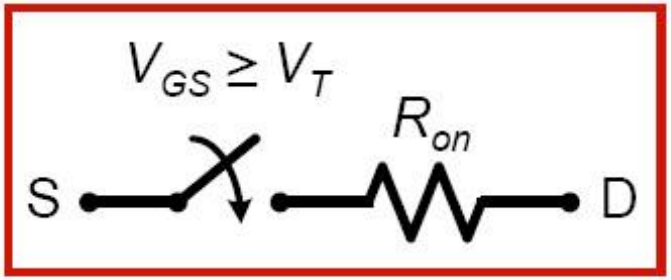
Traversed path

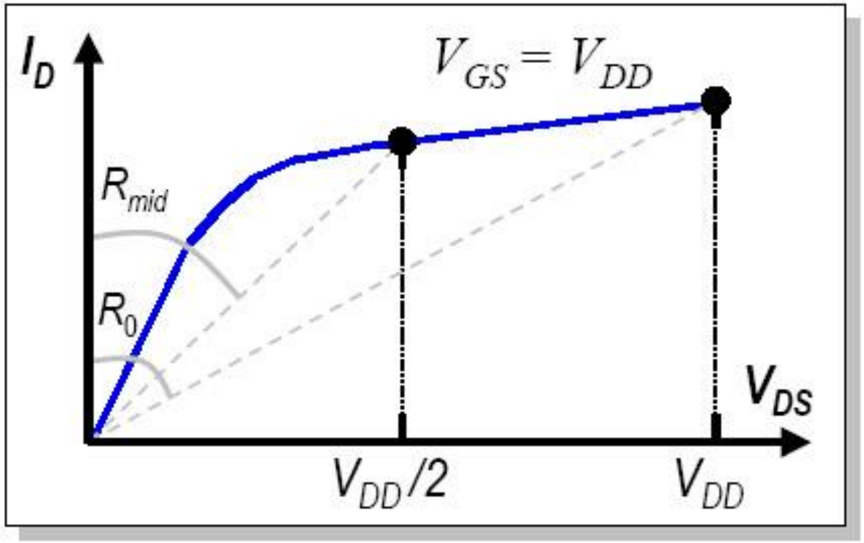


$$R_{eq} = avg(R_{on}(t))\Big|_{t=t_1}^{t_2} = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} R_{on}(t)\cdot dt = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \frac{V_{DS}(t)}{I_D(t)}\cdot dt$$

$$R_{eq} \approx \frac{1}{2}\cdot\left(R_{on}(t_1) + R_{on}(t_2)\right)$$

Equivalent MOS Resistance





$$R_{eq} = \frac{1}{2}\cdot(R_0 + R_{mid})$$

$$R_{eq} = \frac{1}{2}\cdot\left(\frac{V_{DD}}{I_{DSAT}\cdot(1+\lambda\cdot V_{DD})} + \frac{V_{DD}/2}{I_{DSAT}\cdot(1+\lambda\cdot V_{DD}/2)}\right)$$

$$\boxed{R_{eq} \approx \frac{3}{4}\cdot\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{5}{6}\cdot\lambda\cdot V_{DD}\right)}$$

Approximate MOS Resistance

Solving the integral:

$$R_{eq} = \frac{1}{-V_{DD}/2} \int_{V_{DD}}^{V_{DD}/2} \frac{V}{I_{DSAT}(1+\lambda V)} dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1 - \frac{7}{9}\lambda V_{DD}\right)$$

with $I_{DSAT} = k'\frac{W}{L}\left((V_{DD} - V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2}\right)$

Averaging resistances:

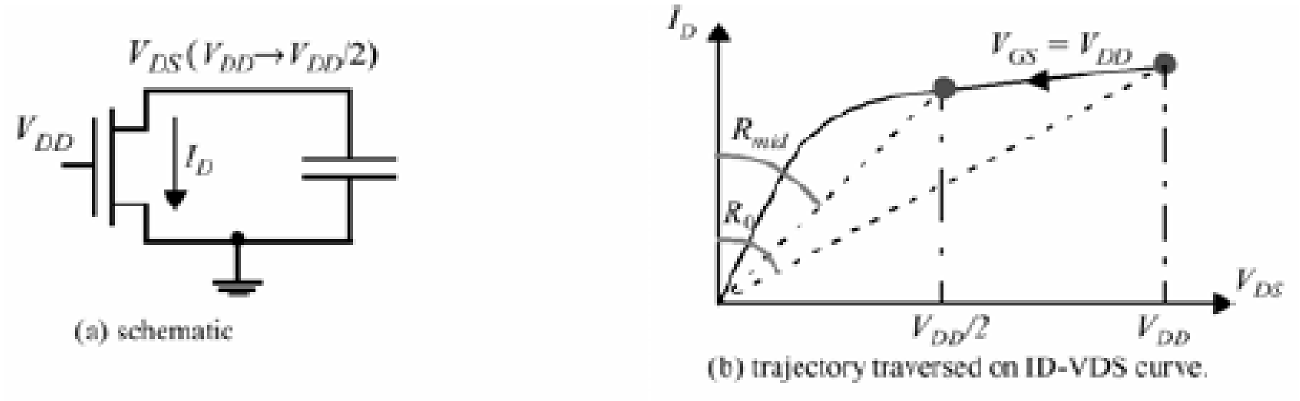
$$R_{eq} = \frac{1}{2}\left(\frac{V_{DD}}{I_{DSAT}(1+\lambda V_{DD})} + \frac{V_{DD}/2}{I_{DSAT}(1+\lambda V_{DD}/2)}\right) \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1 - \frac{5}{6}\lambda V_{DD}\right)$$

Often just: $\boxed{R_{eq} \approx \frac{3}{4} \cdot \frac{V_{DD}}{I_{DSAT}}}$

CMOS Performance

Propagation delay: $t_{pHL} = (\ln 2) R_{eqn} C_L \quad t_{pLH} = (\ln 2) R_{eqp} C_L$

ln2 = 0.7



(a) schematic

(b) trajectory traversed on ID-VDS curve.

Short channel

$$R_{eq} \neq f(V_{DD})$$

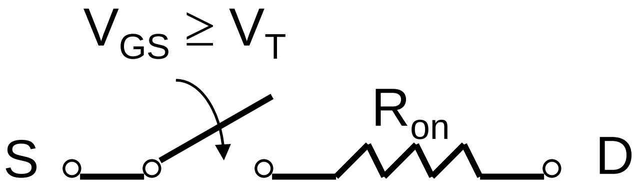
Long channel

$$R_{eq} \propto \frac{1}{V_{DD}}$$

for $V_{DD} >> V_T$

MOS Structure Resistance

- The simplest model assumes the transistor is a switch with an infinite “off” resistance and a finite “on” resistance R_{on}

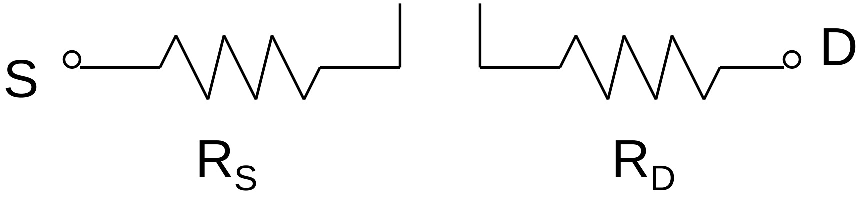


- However R_{on} is nonlinear, so use instead the average value of the resistances, R_{eq}, at the end-points of the transition (V_{DD} and $V_{DD}/2$)

$$R_{eq} = \tfrac{1}{2}\left(R_{on}(t_1) + R_{on}(t_2)\right)$$

$$R_{eq} = \tfrac{3}{4}\, V_{DD}/I_{DSAT}\,(1 - 5/6\ \lambda V_{DD})$$

Source and Drain Resistance



$$R_{S,D} = (L_{S,D}/W)R$$

where $L_{S,D}$ is the length of the source or drain diffusion
R is the sheet resistance of the source or drain diffusion (20 to 100 Ω/)

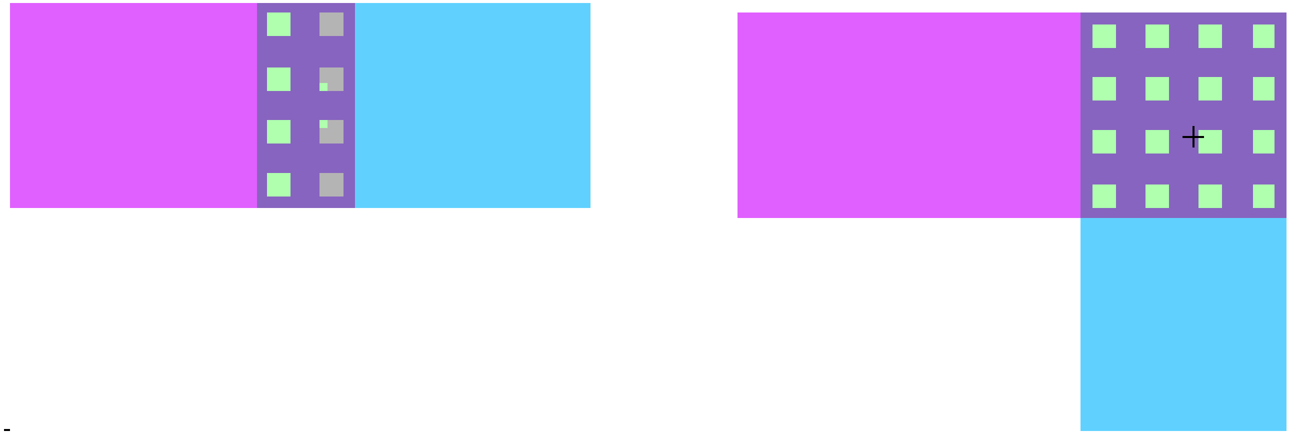
- More pronounced with scaling since junctions are shallower
- With silicidation R is reduced to the range 1 to 4 Ω/

Contact Resistance

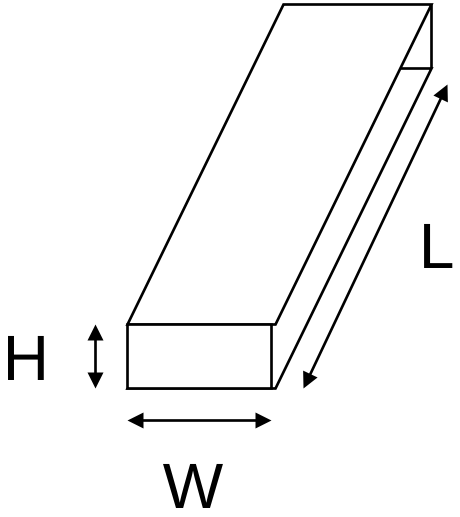
- Transitions between routing layers (contacts through via's) add extra resistance to a wire
  - keep signals wires on a single layer whenever possible
  - avoid excess contacts
  - reduce contact resistance by making vias larger (beware of current crowding that puts a practical limit on the size of vias) or by using multiple minimum-size vias to make the contact
- Typical contact resistances, R_C, (minimum-size)
  - 5 to 20 Ω for metal or poly to n+, p+ diffusion and metal to poly
  - 1 to 5 Ω for metal to metal contacts
- More pronounced with scaling since contact openings are smaller

Contacts Resistance

- Use many contacts for lower R
  - Many small contacts for current crowding around periphery



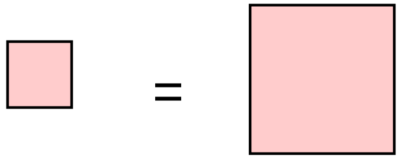
Wire Resistance



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

Sheet Resistance R

$R_{1□} = R_{2□}$

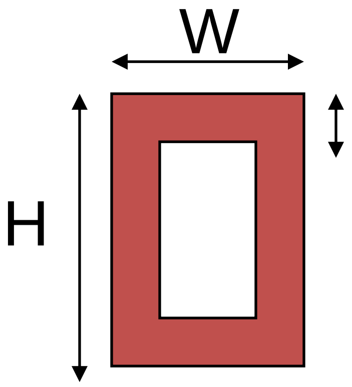


Material	ρ(Ω-m)
Silver (Ag)	1.6 x 10^{-8}
Copper (Cu)	1.7 x 10^{-8}
Gold (Au)	2.2 x 10^{-8}
Aluminum (Al)	2.7 x 10^{-8}
Tungsten (W)	5.5 x 10^{-8}

Material	Sheet Res. (Ω/)
n, p well diffusion	1000 to 1500
n+, p+ diffusion	50 to 150
n+, p+ diffusion with silicide	3 to 5
polysilicon	150 to 200
polysilicon with silicide	4 to 5
Aluminum	0.05 to 0.1

Skin Effect

- At high frequency, currents tend to flow primarily on the surface of a conductor with the current density falling off exponentially with depth into the wire



$\delta = \sqrt{(\rho/(\pi f \mu))}$

where f is frequency

$\mu = 4\pi \times 10^{-7}$ H/m

$\delta = 2.6\ \mu$m

for Al at 1 GHz

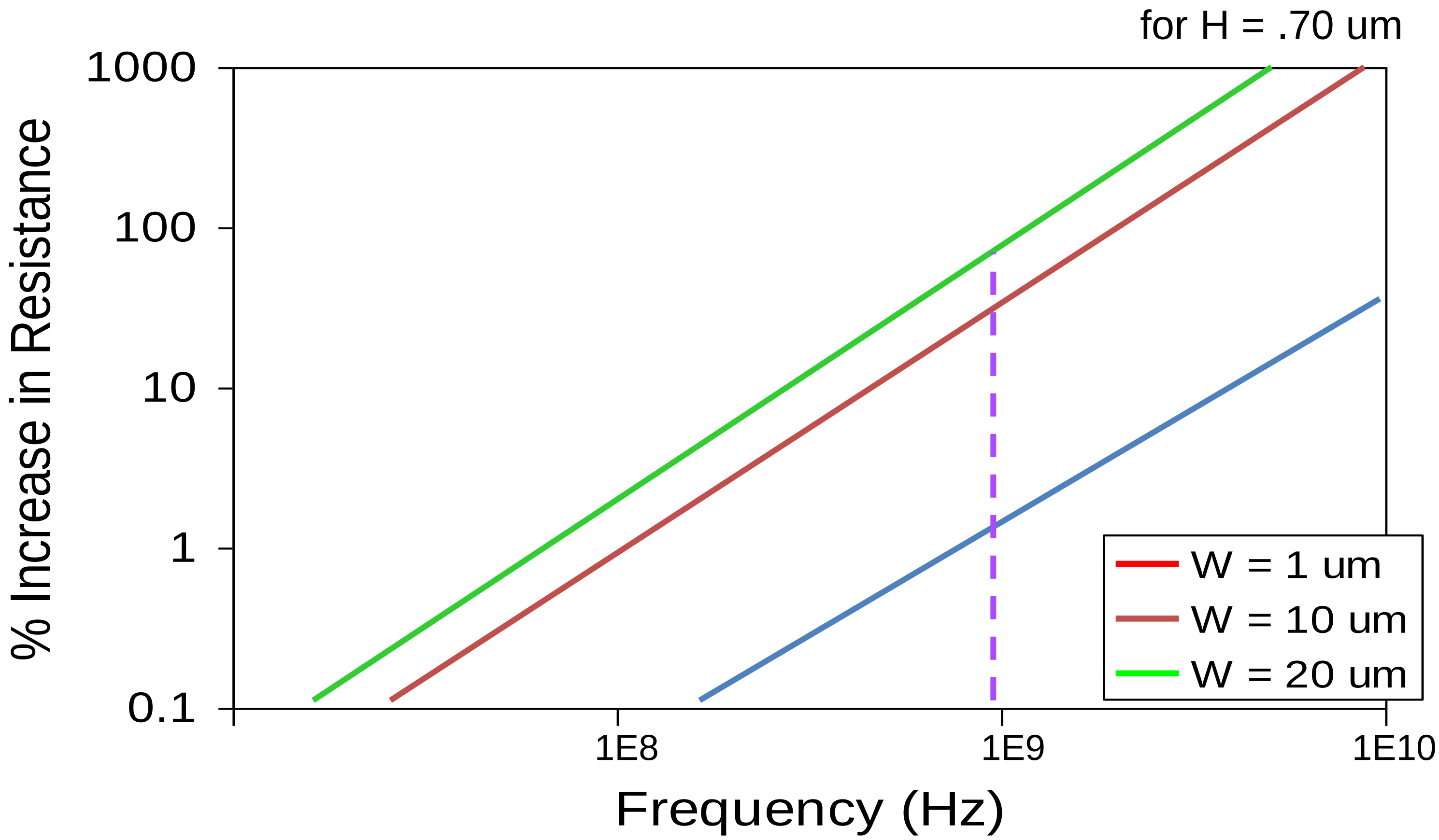
so the overall cross section is ~ $2(W+H)\delta$

- The onset of skin effect is at f_s - where the skin depth is equal to half the largest dimension of the wire.

$$f_s = 4\ \rho / (\pi\ \mu\ (\max(W,H))^2)$$

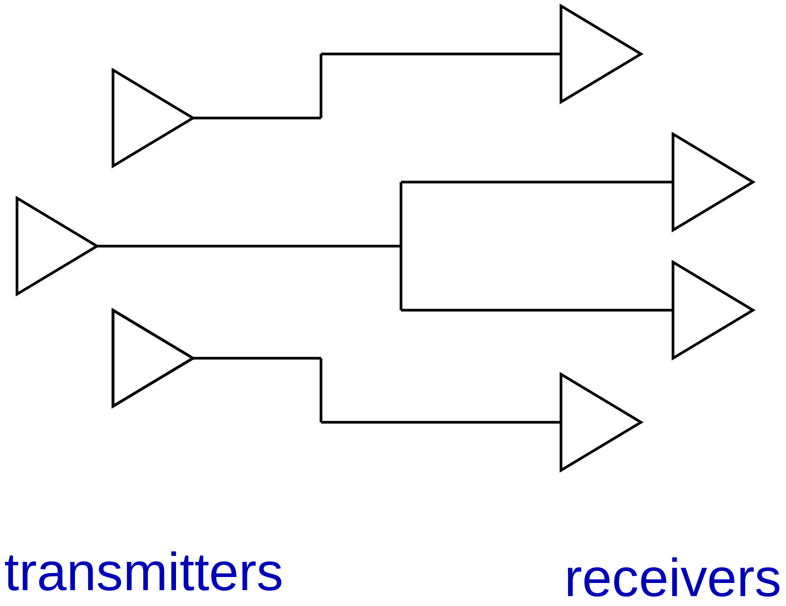
- An issue for high frequency, wide (tall) wires (i.e., clocks!)

Skin Effect for Different W’s



- A 30% increase in resistance is observe for 20 μm Al wires at 1 GHz (versus only a 1% increase for 1 μm wires)

The Wire

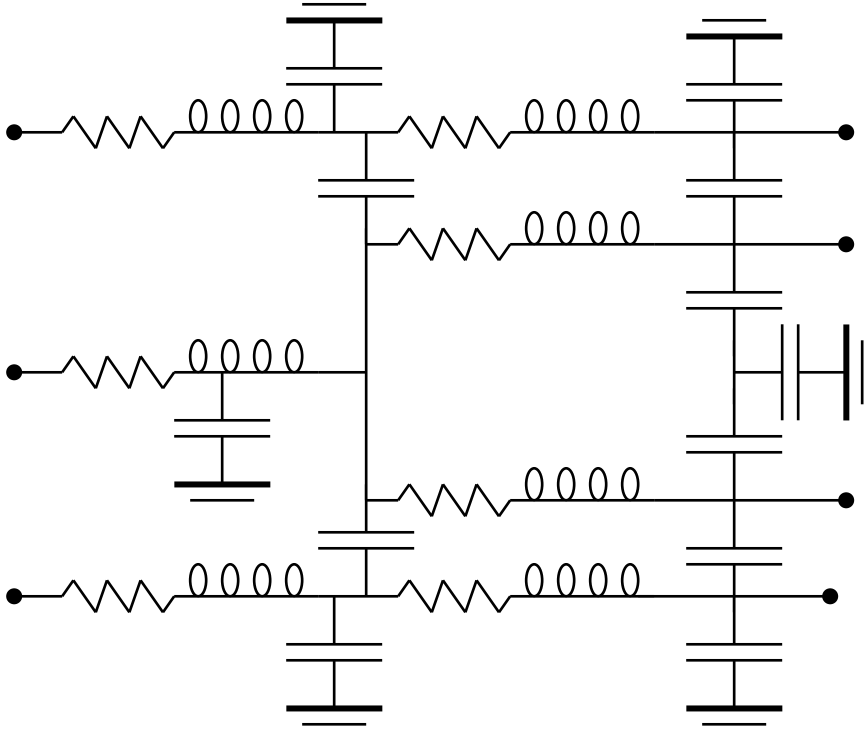


schematic

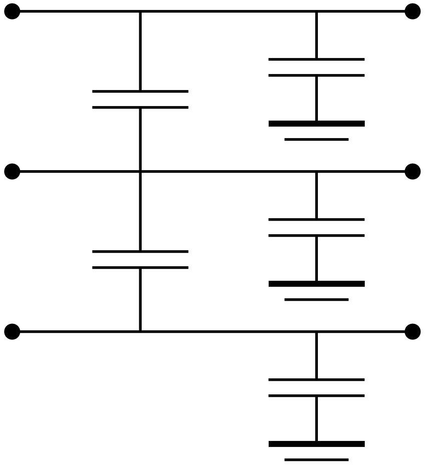
physical

Wire Models

- Interconnect parasitics (capacitance, resistance, and inductance)
  - reduce reliability
  - affect performance and power consumption



All-inclusive (C,R,I) model



Capacitance-only

Parasitic Simplifications

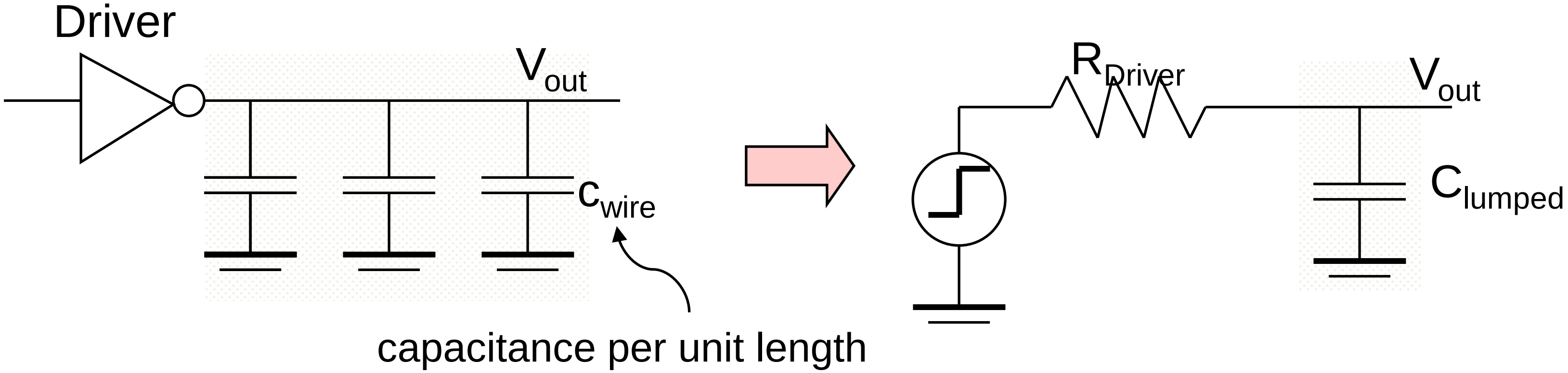
- Inductive effects can be ignored
  - if the resistance of the wire is substantial enough (as is the case for long metal wires with small cross section)
  - if the rise and fall times of the applied signals are slow enough

- When the wire is short, or the cross-section is large, or the interconnect material has low resistivity, a capacitance only model can be used

- When the separation between neighboring wires is large, or when the wires run together for only a short distance, interwire capacitance can be ignored and all the parasitic capacitance can be modeled as capacitance to ground

Wire Delay Models

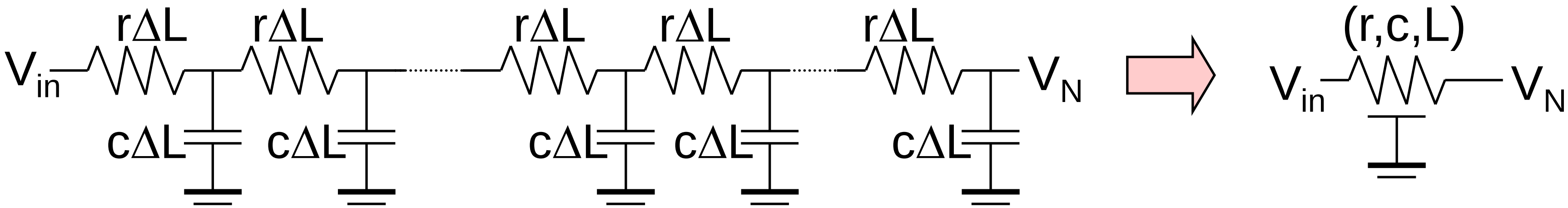
- Ideal wire
  - same voltage is present at every segment of the wire at every point in time - at equi-potential
  - only holds for *very short* wires, i.e., interconnects between *very* nearest neighbor gates
- Lumped C model
  - when only a single parasitic component (C, R, or L) is dominant the different fractions are lumped into a single circuit element
    - When the resistive component is small and the switching frequency is low to medium, can consider only C; the wire itself does not introduce any delay; the only impact on performance comes from wire capacitance



- good for short wires; pessimistic and inaccurate for long wires

Lumped/Distributed Delay Models

- Lumped RC model
  - total wire resistance is lumped into a single R and total capacitance into a single C
  - good for short wires; pessimistic and inaccurate for long wires
- Distributed RC model
  - circuit parasitics are distributed along the length, L, of the wire
    - c and r are the capacitance and resistance per unit length

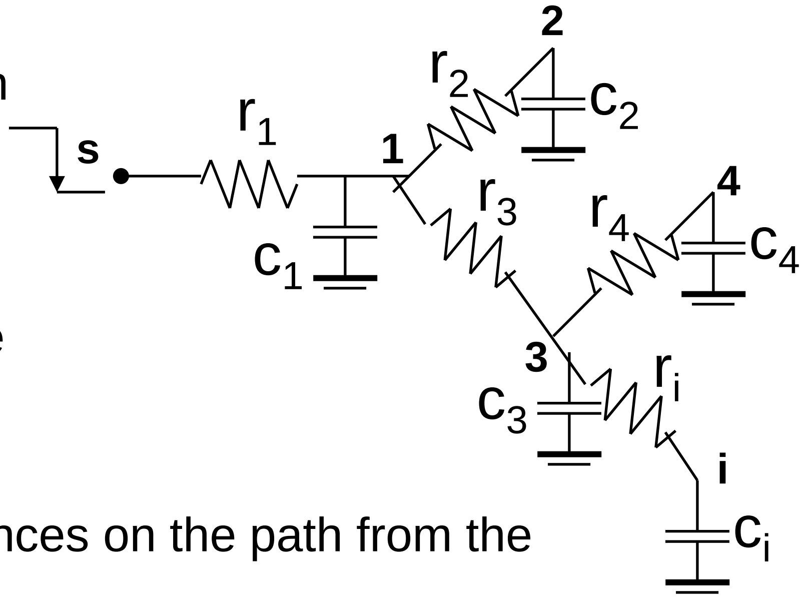


- Delay is determined using the Elmore delay equation

$$\tau_{Di} = \sum_{k=1}^{N} c_k r_{ik}$$

RC Tree Definitions

- RC tree characteristics
  - A unique resistive path exists between the source node and any node of the network
    - ◆Single input (source) node, s
    - ◆All capacitors are between a node and GND
    - ◆No resistive loops



  - Path resistance (sum of the resistances on the path from the input node to node *i*)

$$r_{ii} = \sum_{j=1}^{i} r_j \Rightarrow (r_j \in [\text{path}(s \rightarrow i)]$$

  - Shared path resistance (resistance shared along the paths from the input node to nodes *i* and *k*)

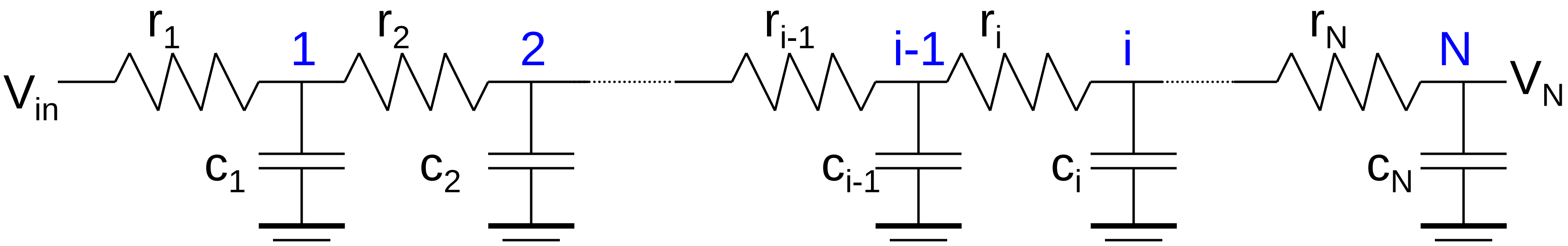
$$r_{ik} = \sum_{j=1}^{N} r_j \Rightarrow (r_j \in [\text{path}(s \rightarrow i) \cap \text{path}(s \rightarrow k)])$$

- A typical wire is a chain network with (simplified) Elmore delay of

$$\tau_{DN} = \sum_{i=1}^{N} c_i r_{ii}$$

Chain Network Elmore Delay

$\tau_{D1}=c_1r_1$ $\tau_{D2}=c_1r_1+c_2(r_1+r_2)$



$\tau_{Di}=c_1r_1+c_2(r_1+r_2)+\ldots+c_i(r_1+r_2+\ldots+r_i)$

Elmore delay equation $$\tau_{DN}=\sum c_i r_{ii}=\sum^{N} c_i \sum^{i} r_j$$

$\tau_{Di}=c_1r_{eq}+2c_2r_{eq}+3c_3r_{eq}+\ldots+ic_ir_{eq}$

Distributed RC Model for Simple Wires

- A length L RC wire can be modeled by N segments of length L/N
  - The resistance and capacitance of each segment are given by r L/N and c L/N

$\tau_{DN} = (L/N)^2(cr+2cr+\ldots+Ncr) = (crL^2)\,(N(N+1))/(2N^2) = CR((N+1)/(2N))$

where R (= rL) and C (= cL) are the total lumped resistance and capacitance of the wire

- For large N $\boxed{\tau_{DN} = RC/2 = rcL^2/2}$
  - Delay of a wire is a quadratic function of its length, L
  - The delay is 1/2 of that predicted (by the lumped model)

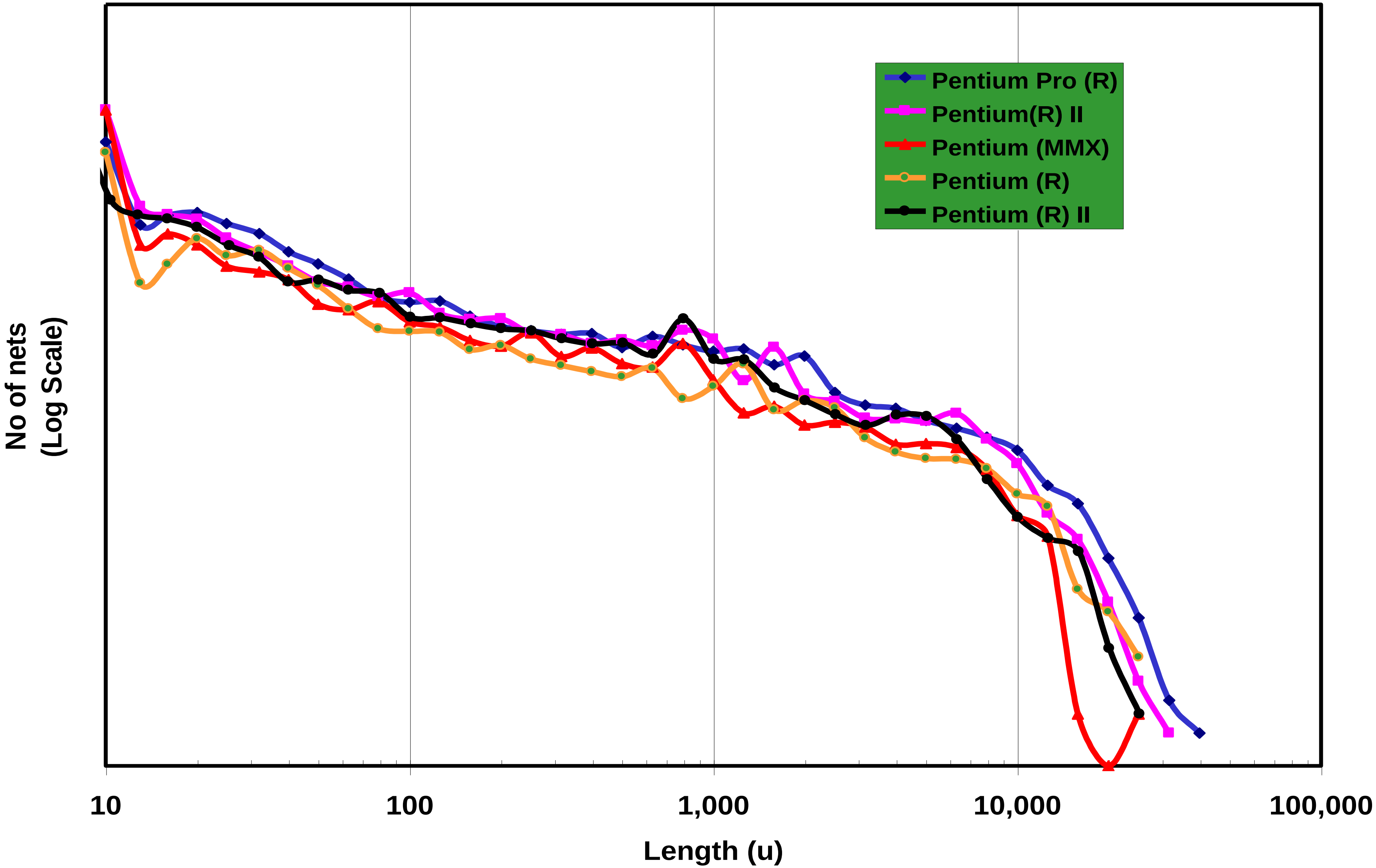
Step Response Points

Voltage Range	Lumped RC	Distributed RC
0 → 50% (t_p)	0.69 RC	0.38 RC
0 → 63% (τ)	RC	0.5 RC
10% → 90% (t_r)	2.2 RC	0.9 RC
0 → 90%	2.3 RC	1.0 RC

Time to reach the 50% point is $t = \ln(2)\tau = 0.69\tau$

Time to reach the 90% point is $t = \ln(9)\tau = 2.2\tau$

Nature of Interconnect

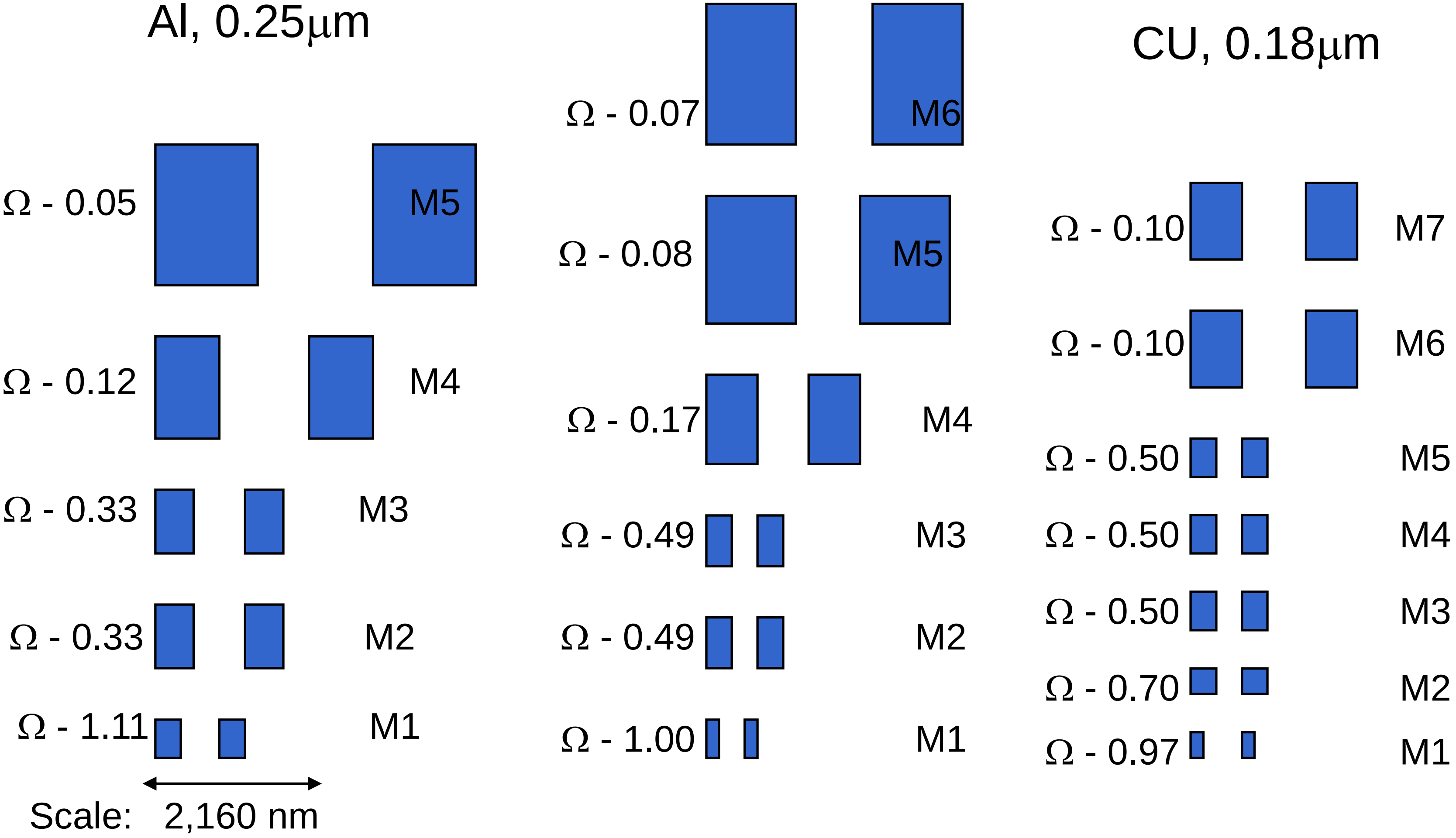


Wire Spacing

Intel P856.5
Al, 0.25µm

Intel P858
Al, 0.18µm

IBM CMOS-8S
CU, 0.18µm



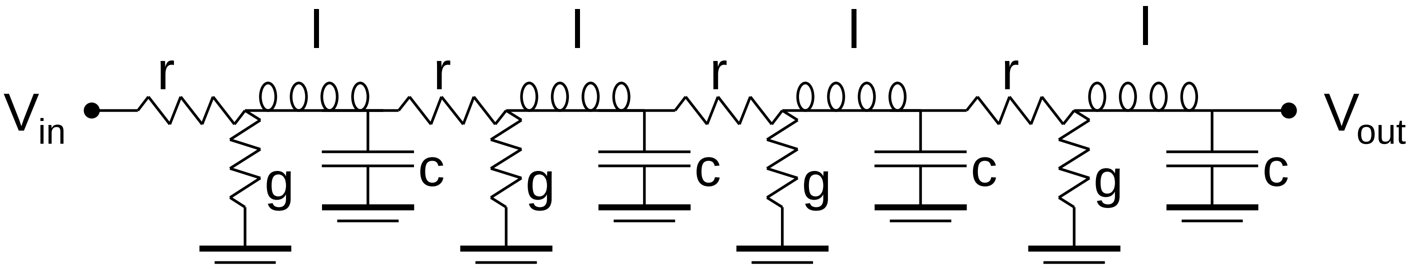
Intel P856.5:
Ω - 0.05 M5
Ω - 0.12 M4
Ω - 0.33 M3
Ω - 0.33 M2
Ω - 1.11 M1
Scale: 2,160 nm

Intel P858:
Ω - 0.07 M6
Ω - 0.08 M5
Ω - 0.17 M4
Ω - 0.49 M3
Ω - 0.49 M2
Ω - 1.00 M1

IBM CMOS-8S:
Ω - 0.10 M7
Ω - 0.10 M6
Ω - 0.50 M5
Ω - 0.50 M4
Ω - 0.50 M3
Ω - 0.70 M2
Ω - 0.97 M1

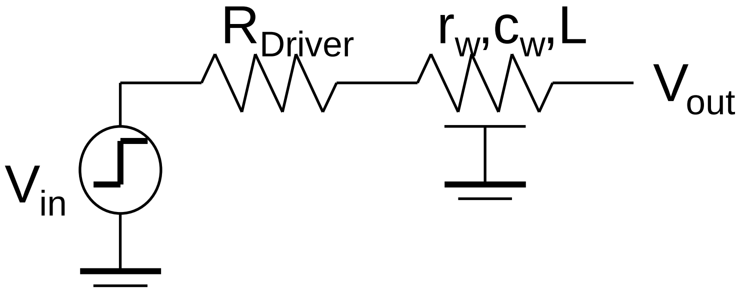
Inductance of Wires

- When the rise and fall times of the signal become comparable to the time of flight of the signal waveform across the line, then the inductance of the wire starts to dominate the delay behavior



- Must consider wire transmission line effects
  - ◆Signal propagates over the wire as a wave (rather than diffusing as in rc only models)
    - ➢ Signal propagates by alternately transferring energy from
      capacitive to inductive modes

Delay of MOS + Wire



- Total propagation delay consider driver and wire

$$\tau_D = R_{Driver}C_w + (R_wC_w)/2 = R_{Driver}C_w + 0.5 r_w c_w L^2$$

and $t_p = 0.69\ R_{Driver}C_w + 0.38\ R_wC_w$

where $R_w = r_wL$ and $C_w = c_wL$

- The delay introduced by wire resistance becomes dominant when

  $(R_wC_w)/2 \geq R_{Driver}\ C_W$ (when $L \geq 2R_{Driver}/R_w$)

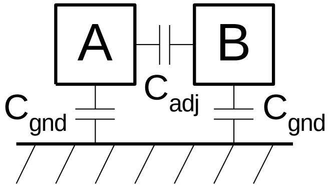
  - For an $R_{Driver} = 1\ \text{k}\Omega$ driving an 1 μm wide Al1 wire, L_{crit} is 2.67 cm

Crosstalk

- A capacitor does not like to change its voltage instantaneously.
- A wire has high capacitance to its neighbor.
  - When the neighbor switches from 1-> 0 or 0->1, the wire tends to switch too.
  - Called capacitive *coupling* or *crosstalk*.
- Crosstalk effects
  - Noise on nonswitching wires
  - Increased delay on switching wires

Crosstalk Delay

- Assume layers above and below on average are quiet
  - Second terminal of capacitor can be ignored
  - Model as $C_{gnd} = C_{top} + C_{bot}$
- Effective C_{adj} depends on behavior of neighbors
  - *Miller effect*

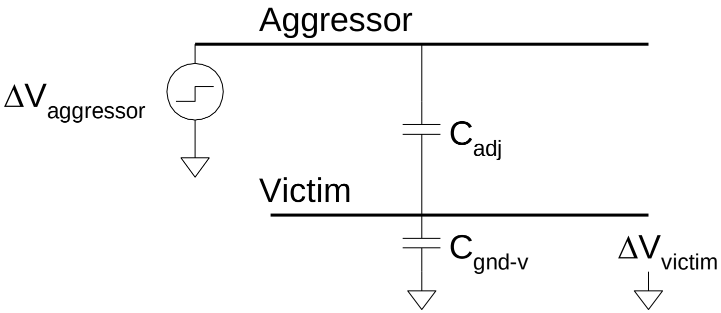


B	**ΔV**	$\mathbf{C_{eff(A)}}$	**MCF**
Constant	V_{DD}	$C_{gnd} + C_{adj}$	1
Switching with A	0	C_{gnd}	0
Switching opposite A	$2V_{DD}$	$C_{gnd} + 2\ C_{adj}$	2

Crosstalk Noise

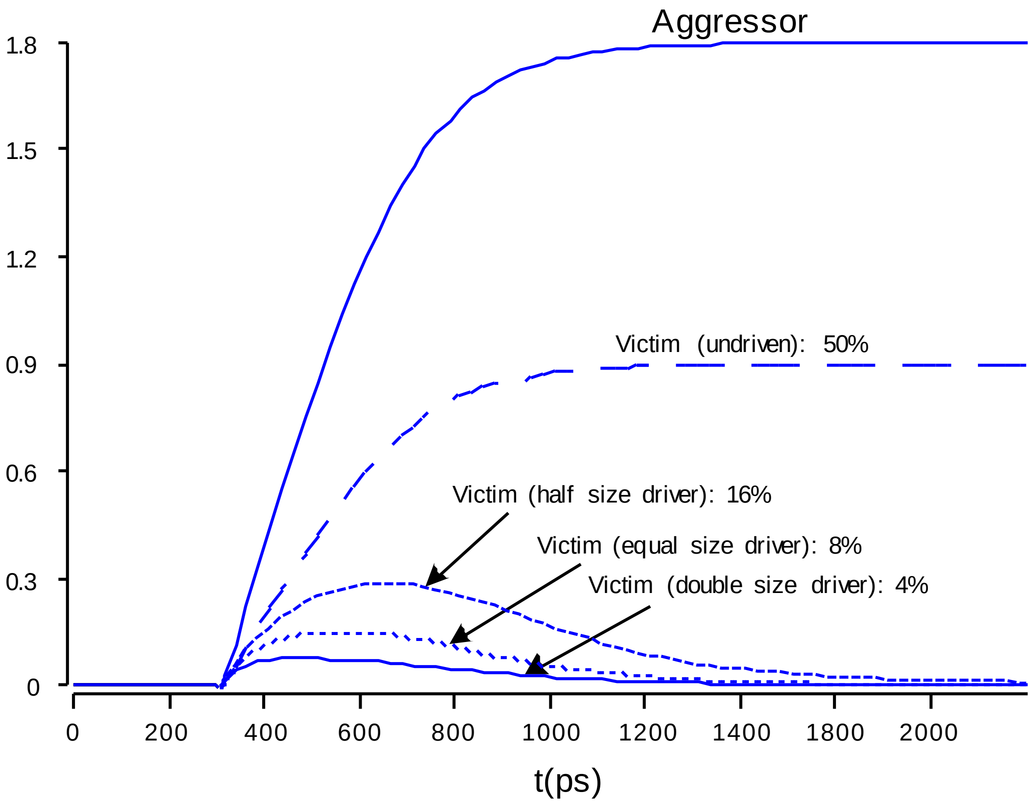
- Crosstalk causes noise on nonswitching wires
- If victim is floating:
  - model as capacitive voltage divider

$$\Delta V_{victim} = \frac{C_{adj}}{C_{gnd-v} + C_{adj}} \Delta V_{aggressor}$$



Coupling Waveforms

- Simulated coupling for $C_{adj} = C_{victim}$

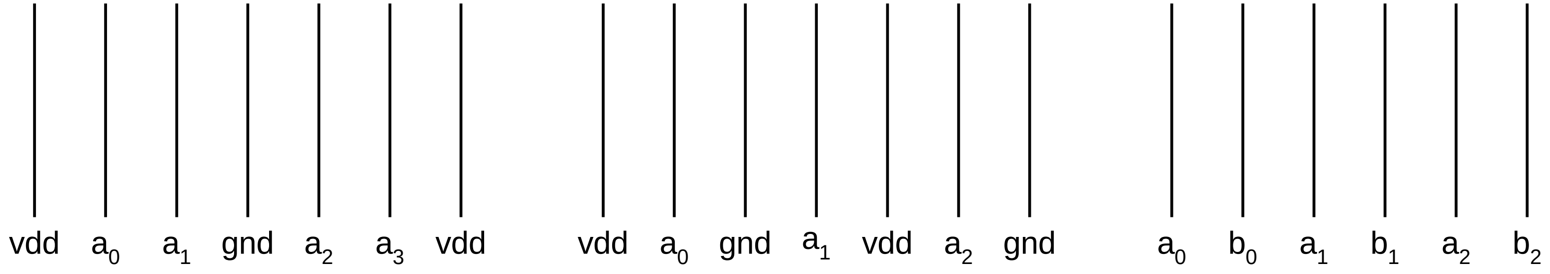


Noise Implications

- *So what* if we have noise?
- If the noise is less than the noise margin, nothing happens
- Static CMOS logic will eventually settle to correct output even if disturbed by large noise spikes
  - But glitches cause extra delay
  - Also cause extra power from false transitions
- Dynamic logic never recovers from glitches
- Memories and other sensitive circuits also can produce the wrong answer

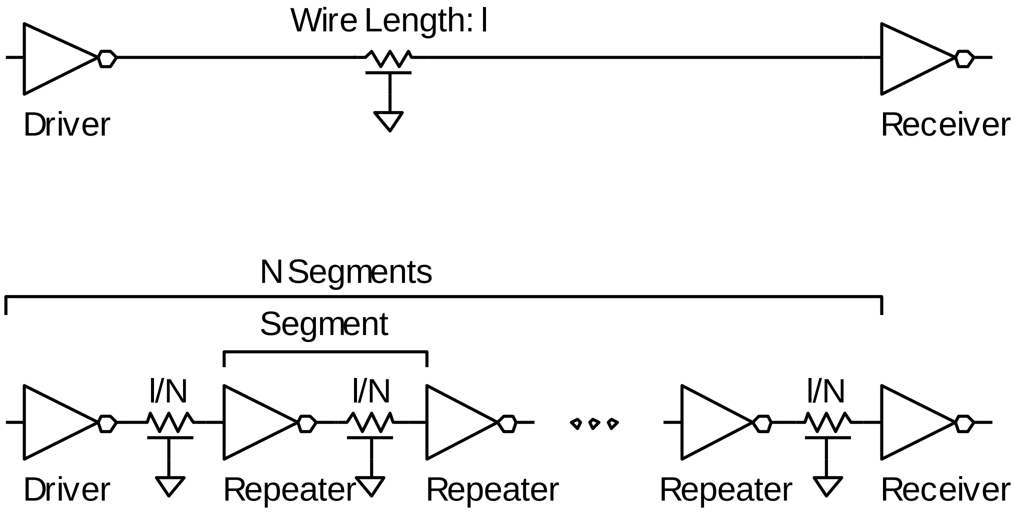
Wire Engineering

- Goal: achieve delay, area, power goals with acceptable noise
- Degrees of freedom:
  - Width
  - Spacing
  - Layer
  - Shielding



Repeaters

- R and C are proportional to *l*
- RC delay is proportional to l^2
  - Unacceptably great for long wires
- Break long wires into N shorter segments
  - Drive each one with an inverter or buffer


Wire Length: l
Driver
Receiver
N Segments
Segment
l/N
l/N
l/N
Driver
Repeater
Repeater
Repeater
Receiver