ECE260B Winter 22

Power and Interconnect

Prof. Mingu Kang

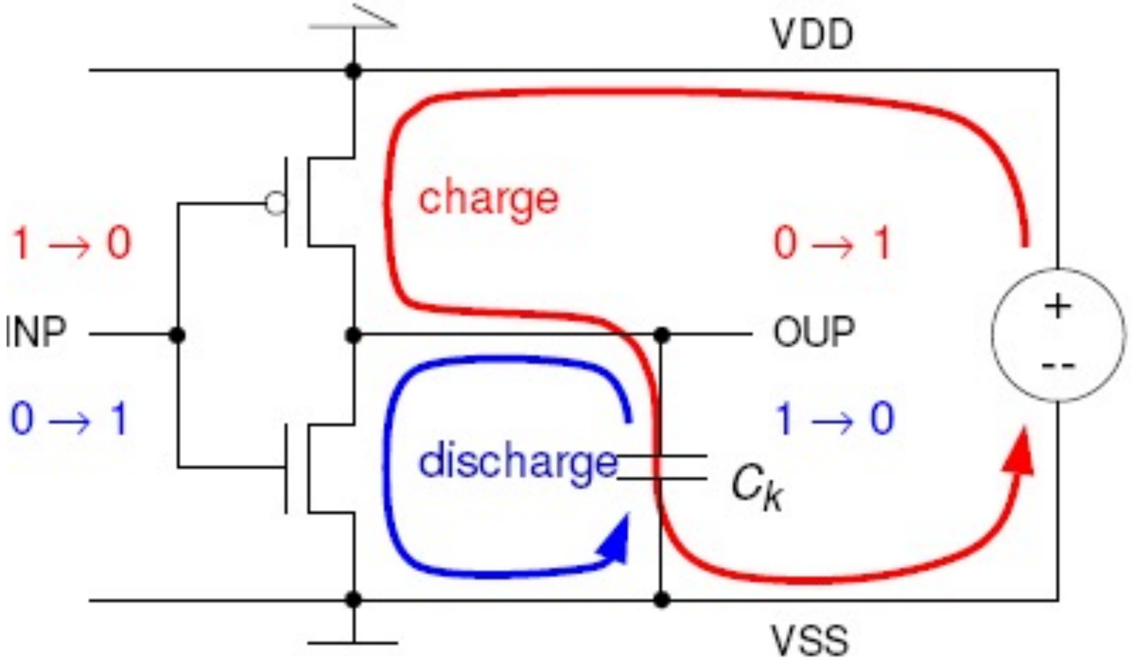
UCSD Computer Engineering

Power

Power Consumption

- As transistor counts and clock frequencies have increased, power consumption has skyrocketed
- Dynamic power dissipation due to
  - Charging and discharging of load capacitances
  - “short-circuit” current while both PMOS and NMOS networks are partially ON
- Static power dissipation due to
  - Subthreshold conduction through OFF transistors
  - Tunneling current through gate oxide

Dynamic Power (during output transition from 0 -> 1)



- Energy delivered from power supply = $C_L V_{DD}^2$
- Energy stored in capacitor is ½ $C_L V_{DD}^2$
- Energy dissipated in the PMOS register (by heat) is ½ $C_L V_{DD}^2$
- No power dissipation from V_{DD} during 1 ->0 transition

Proof

Energy stored in the capacitor

$$E = \int_0^\infty v(t)i(t)dt = \int_0^\infty V\left[1 - e^{\frac{-t}{RC}}\right]\frac{V}{R}e^{\frac{-t}{RC}}dt$$

$$E = \frac{1}{2}CV^2$$

Energy dissipated in the register

$$E = R\int_0^\infty i^2(t)dt = R\frac{V^2}{R^2}\int_0^\infty e^{\frac{-2t}{RC}}dt$$

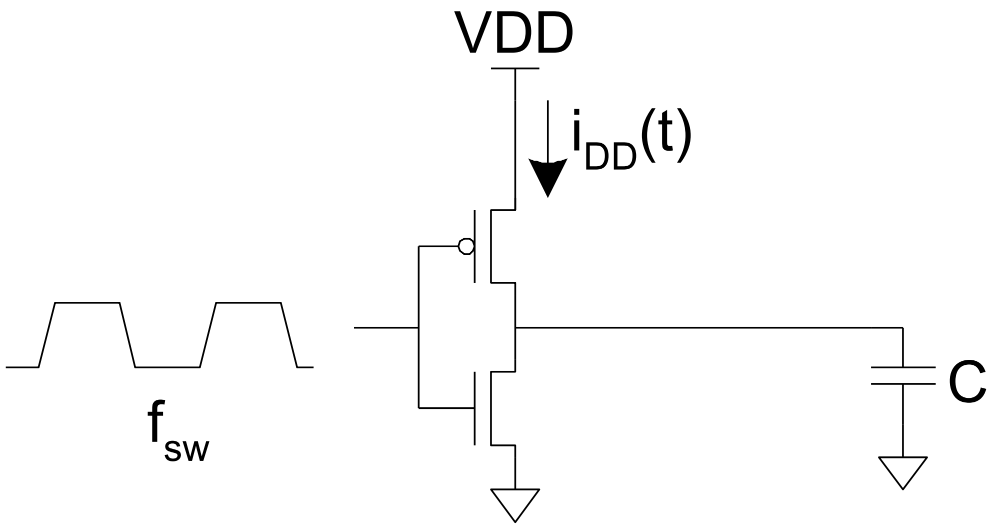
$$E = \frac{1}{2}CV^2$$

$$E_{trans} = \int_0^\infty Vi(t)dt = \int_0^\infty \frac{V^2}{R}e^{\left(\frac{-t}{RC}\right)}dt$$

$$E_{trans} = CV^2$$

Dynamic Power Dissipation with Switching Activity

$$
\begin{aligned}
P_{\text{dynamic}} &= \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}\,dt \\
&= \frac{V_{DD}}{T}\int_0^T i_{DD}(t)\,dt \\
&= \frac{V_{DD}}{T}\left[\overset{\text{(\#cycles)}}{Tf_{\text{sw}}}\,\underset{\text{(charge per cycle)}}{CV_{DD}}\right] \\
&= CV_{DD}{}^{2} f_{\text{sw}}
\end{aligned}
$$

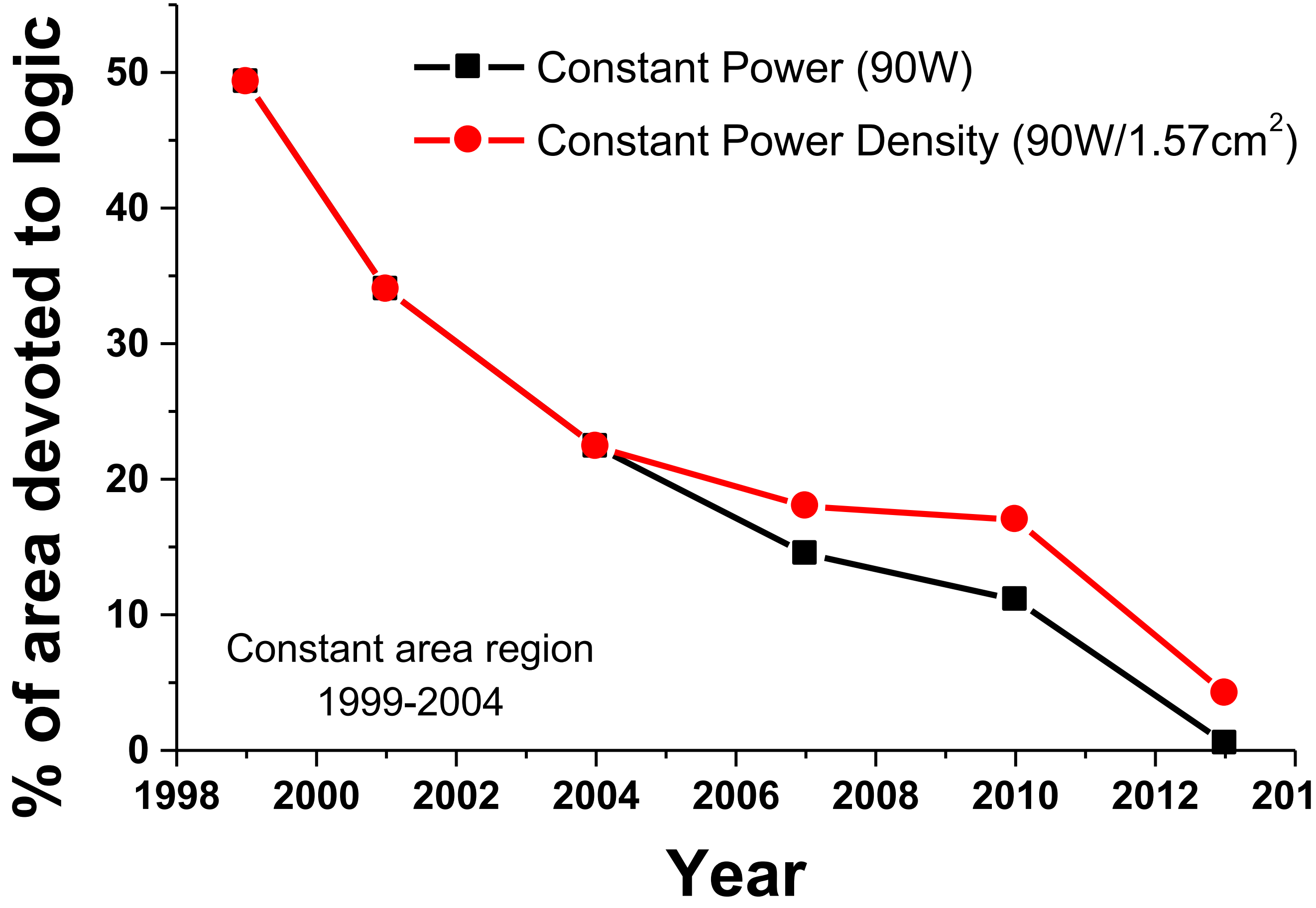


- This repeats T*f_{sw} times over an interval of T
- Power dissipation = $C_L V_{DD}^2 f_{sw}$

Activity Factor

- Suppose the system clock frequency = f_{CLK}
- Let $f_{sw} = \alpha_{0\to1} f_{CLK}$, where $\alpha_{0\to1}$ = activity factor
  - If the signal is a clk, $\alpha_{0\to1}$ is 1
  - If the signal switches once per cycle, $\alpha_{0\to1}$ is 0.5
- Depends on design, but typically $\alpha_{0\to1} = 0.1$ → now perhaps 0.03 – 0.05
- Dynamic power: $P_{dyn} = \alpha_{0\to1} C_L V_{DD}^2 f_{CLK}$
- **Activity factors are decreasing: Why?**
  - (Cf. “Dark Silicon”)

“Dark Silicon” Analysis in 2001 ITRS



- Portion of (switched) logic content at any given moment is approaching to zero due to power limits.
- Unfortunately, resource utilization is also decreasing

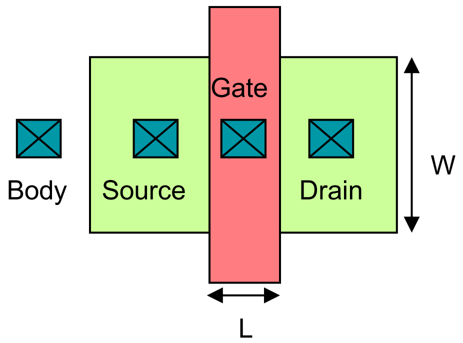
Short Circuit Power

- When transistors switch, both *n*MOS and *p*MOS networks may be momentarily ON at once
- Leads to “short circuit” current
- < 10% of dynamic power if rise/fall times are comparable for input and output
  - This is <u>one</u> reason why you have transition time limits in your “electrical rule checks” (ERCs)

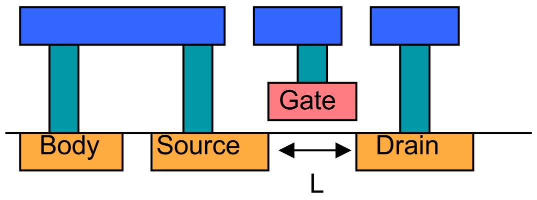
Basic Concepts of Power Optimization

- Goal
  - Maximize power reduction **under a given timing requirement**
- There is a tradeoff between power and delay
  - To reduce power, we need to sacrifice speed
  - **However, since not all timing paths are timing critical, we can use surplus timing to reduce power**
- Main idea
  - Critical timing path: Use faster (higher drive, higher power) cells
  - Non-critical timing path: Use slower (lower drive, lower power) cells

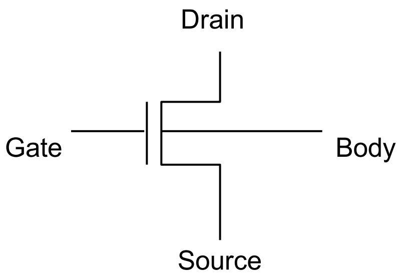
Power Optimization Knob in MOS



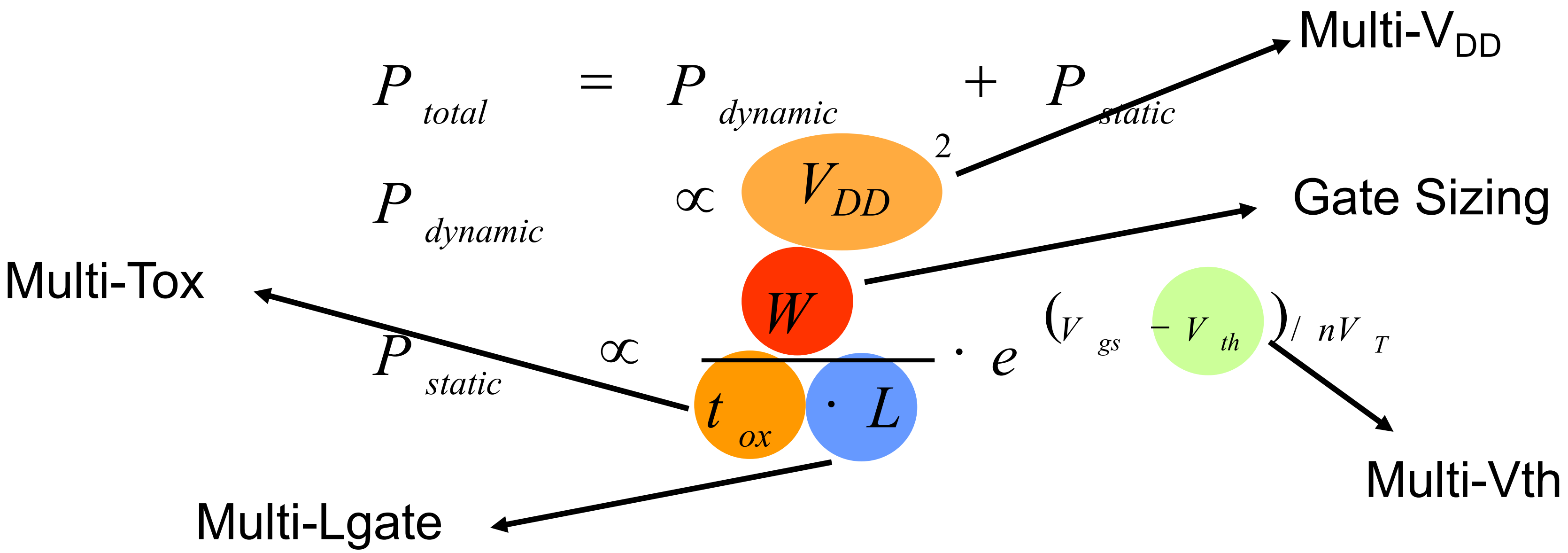
Top View



Side View



Schematic



$$P_{total} = P_{dynamic} + P_{static}$$

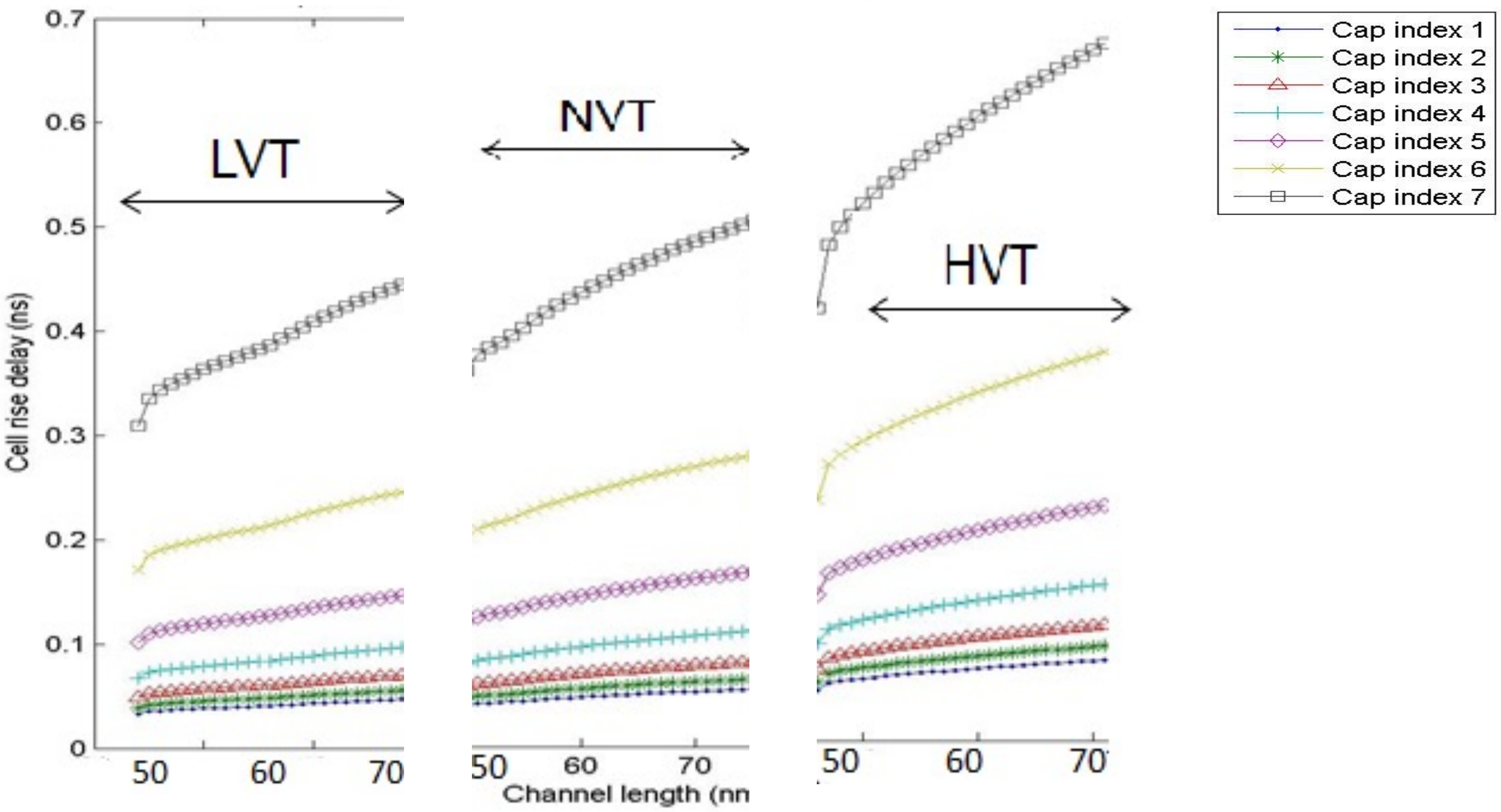
$$P_{dynamic} \propto V_{DD}^{2}$$

$$P_{static} \propto \frac{W}{t_{ox} \cdot L} \cdot e^{(V_{gs} - V_{th})/nV_{T}}$$

Pros and Cons of Low-Power Techniques

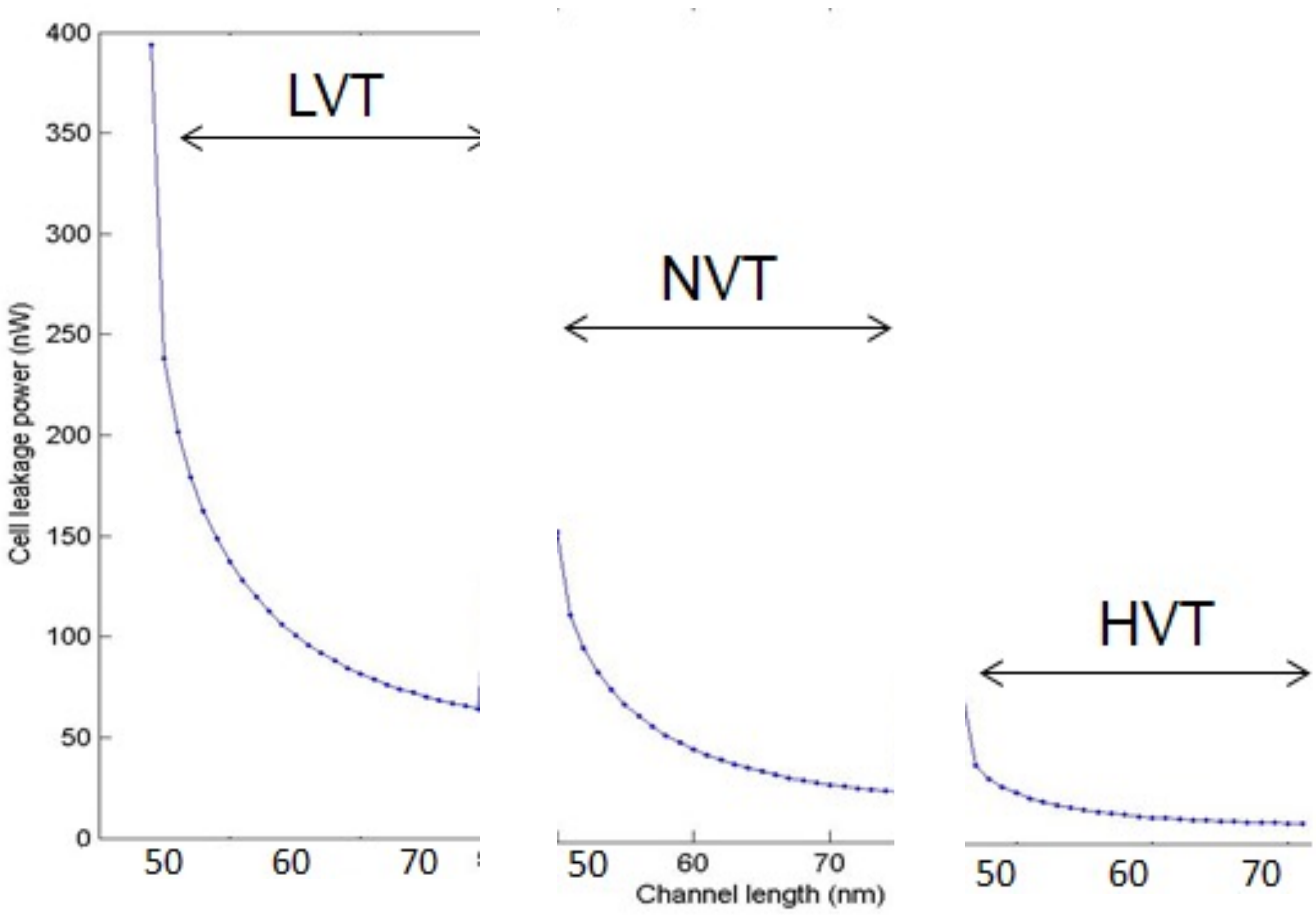
	Pros	Cons
Multi-V_{DD}	▪ Most effective to reduce power (especially, dynamic power)	▪ Additional area for regulators, power networks and level-shifters are required ▪ Difficult to control voltage of individual cells
Multi-Vth	▪ Easy to make cell variants (just change doping)	▪ Additional masks and manufacturing steps are required ▪ Can increase dynamic power
Multi-Lgate	▪ No additional mask or process are required ▪ Many types of cell variants can be applicable → fine-grain control	▪ Additional cell layouts for all the variants are required ▪ Can increase dynamic power

Delay vs. Gate Length in 65 nm Process



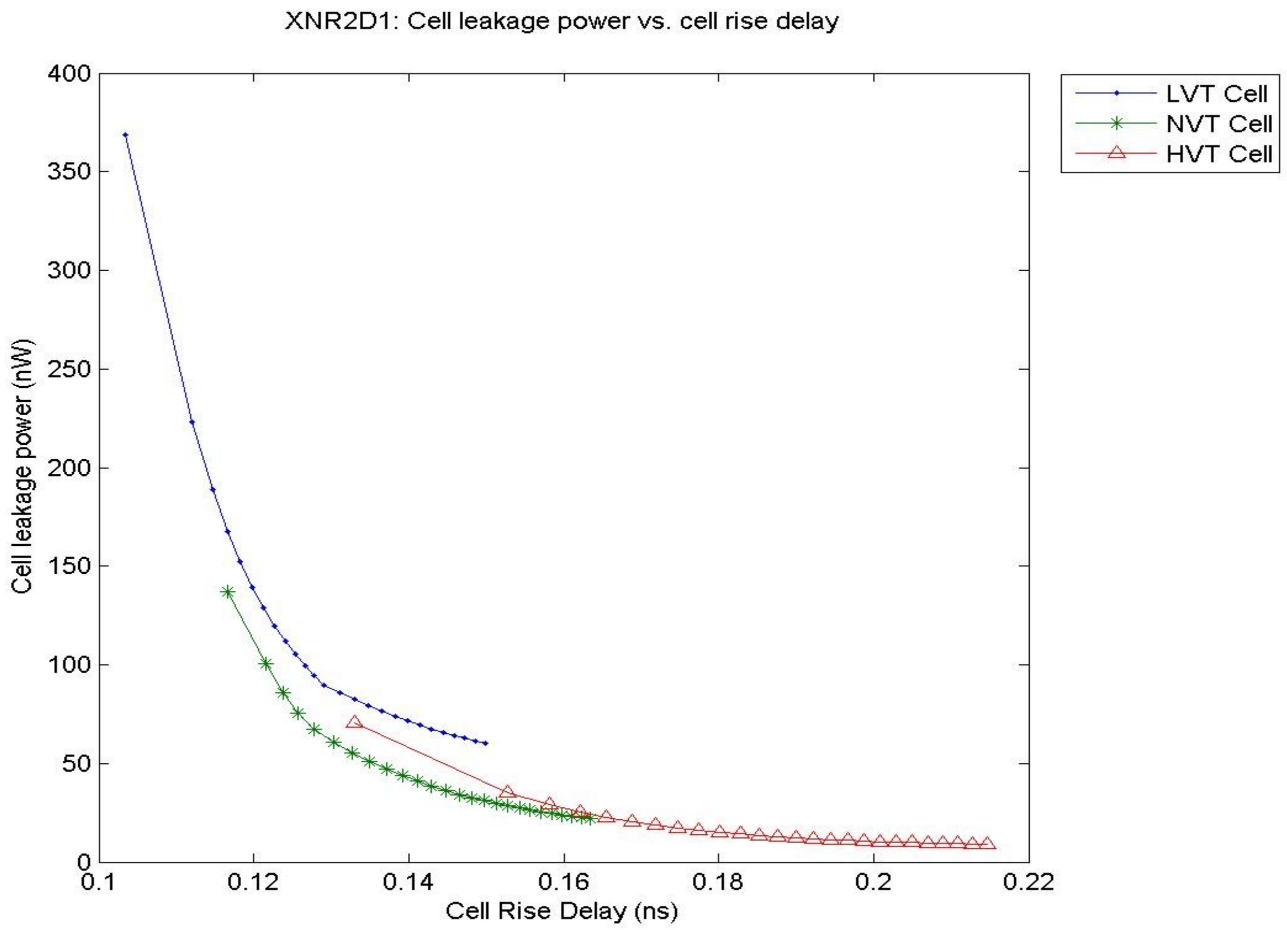
- delay increases ~linearly with increasing channel length of the device.

Leakage vs. Gate Length in 65 nm Process



- leakage decreases exponentially with increasing channel length of the device.

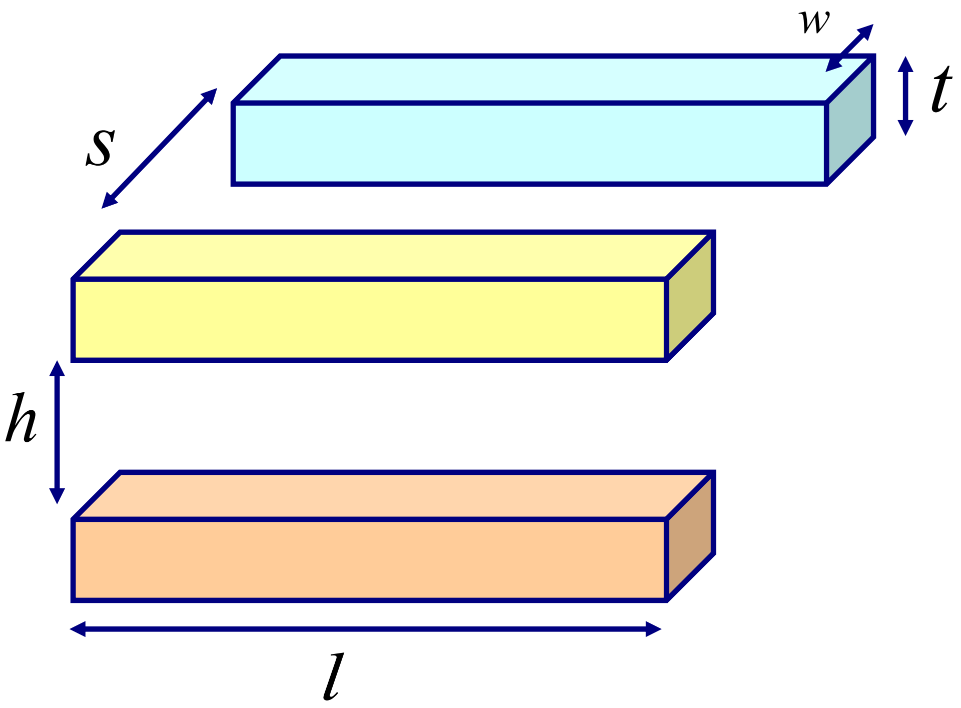
Leakage vs. Delay Length in 65 nm Process



- leakage and delay are inversely proportional.

Interconnect

Interconnect Dimensions



w: width of interconnect

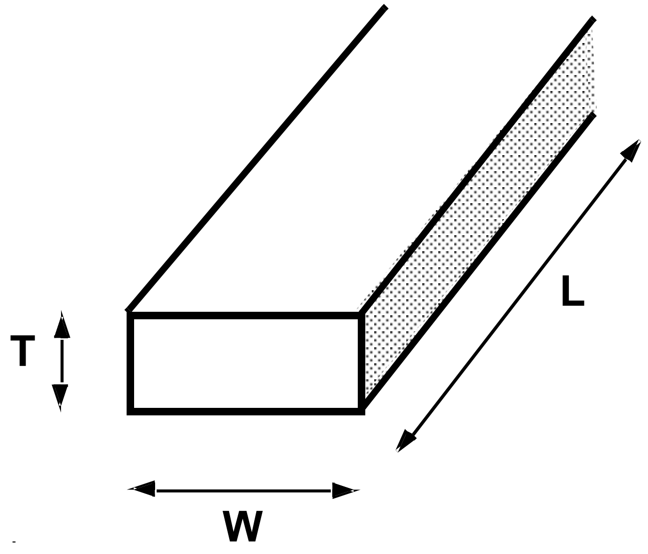
s: spacing between interconnects on same layer

h: dielectric thickness (spacing between interconnects in two vertically adjacent layers)

l: length of interconnect

t: thickness of interconnect

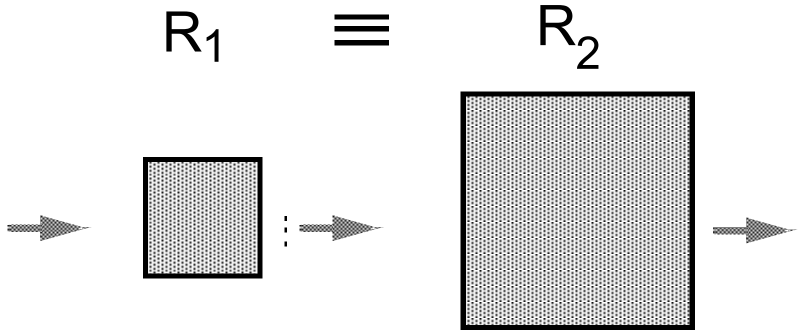
Resistance & Sheet Resistance



$$R = \frac{\rho}{T}\frac{L}{W}$$

(controlled by designer)

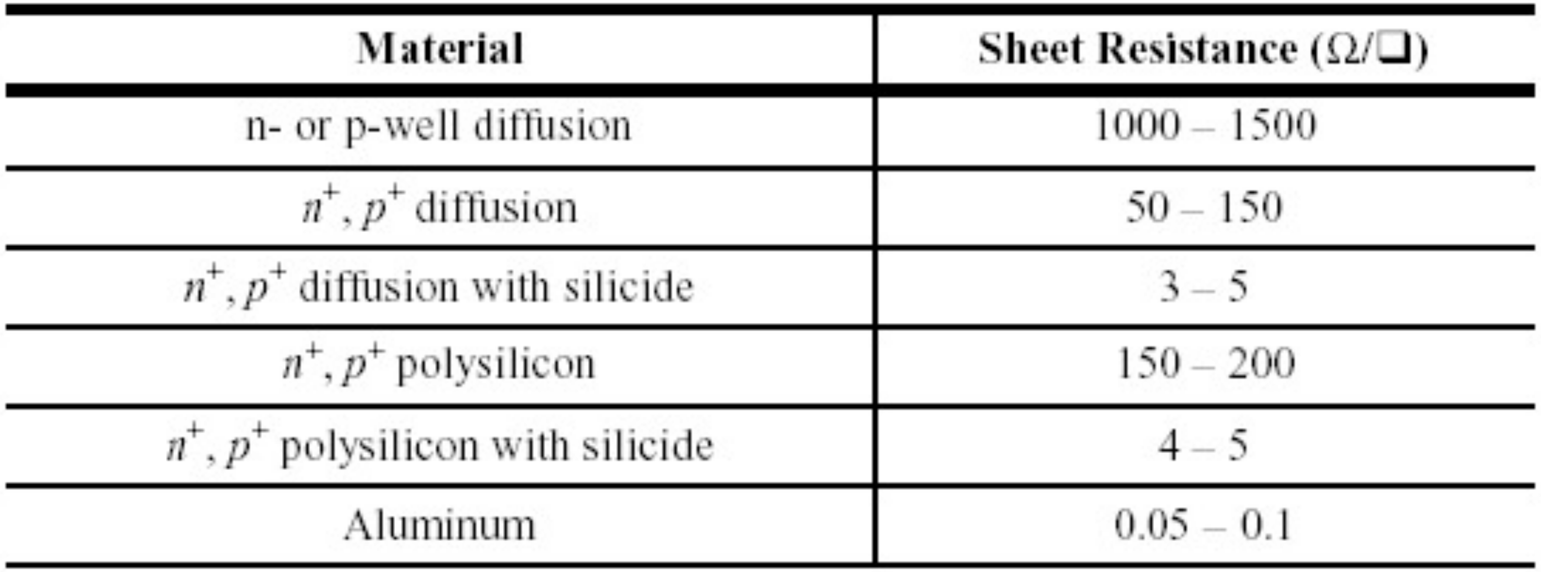
Sheet Resistance
$R_\square$ (intrinsic to process)



Top view showing W x L

- Sheet resistance: resistance when W = L
- Resistance between R_1 vs. R_2 is same

Interconnect Resistance



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

- Resistance scales badly

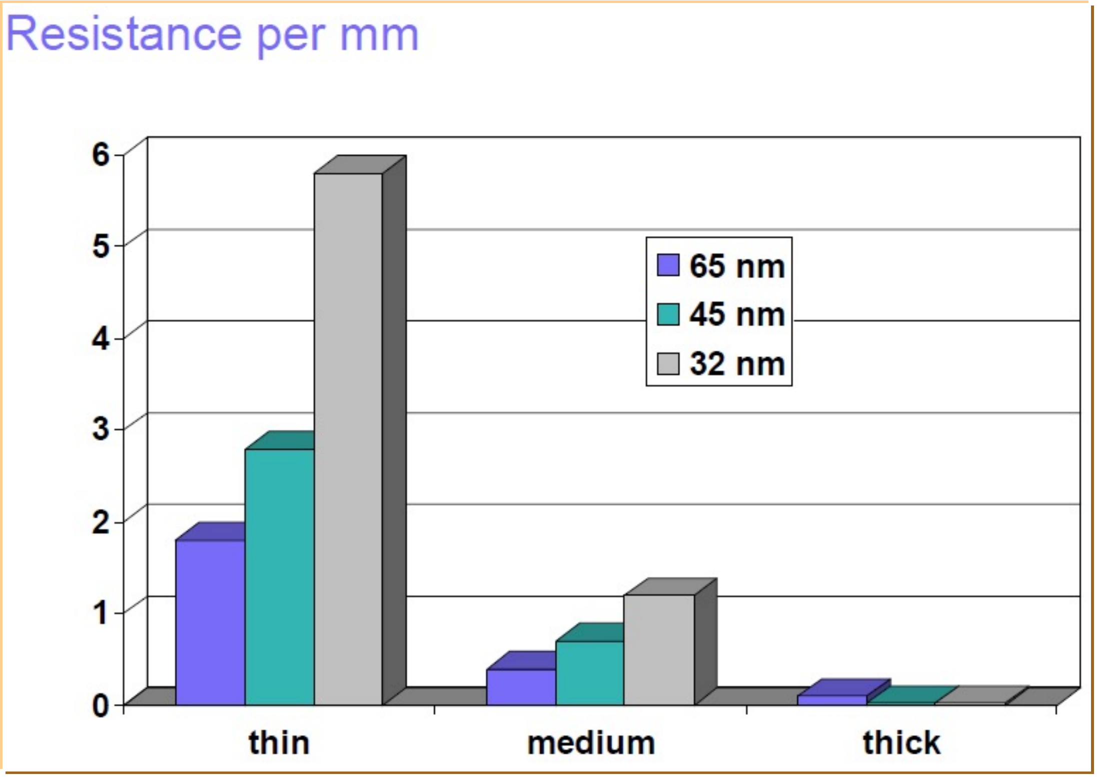
  - True scaling would reduce width and thickness by S each node → roughly the case, since aspect ratio (AR) is constant

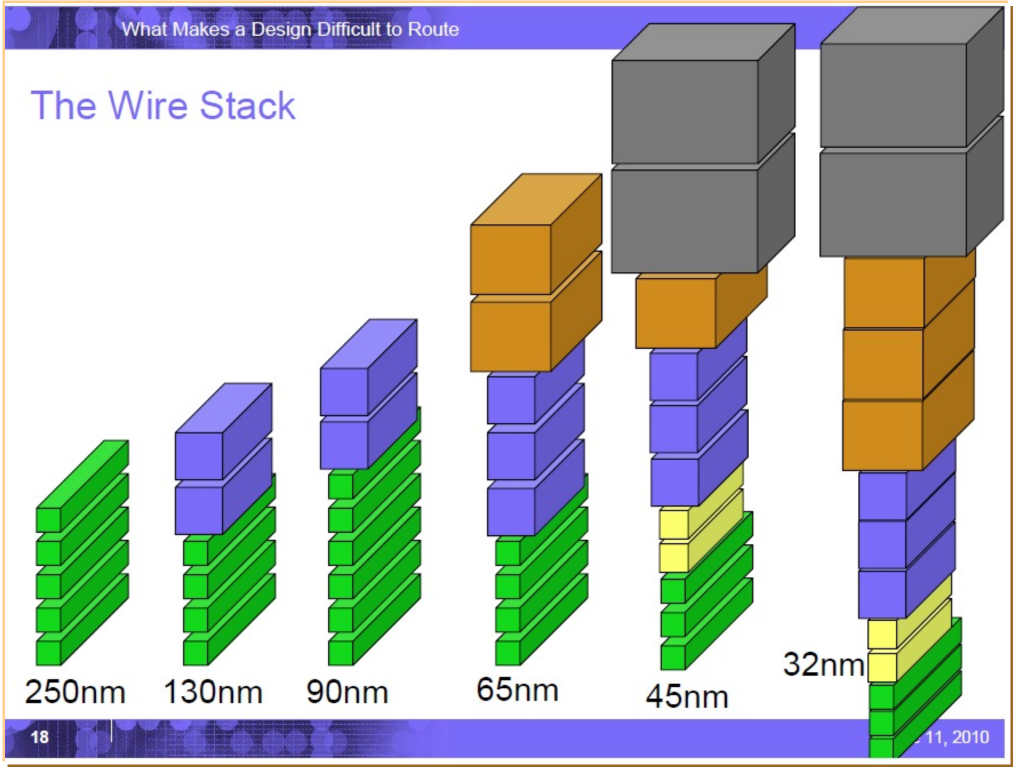
  → R ~ S^2 for a fixed line length and material

- **Reverse scaling** → global wires get relatively slower with respect to (faster) devices

  - Skin effect: At very large dimensions and at higher frequencies, current crowds to edges (outer layers) of conductors → R increases more

Trends in Interconnect


Resistance per mm


What Makes a Design Difficult to Route

The Wire Stack

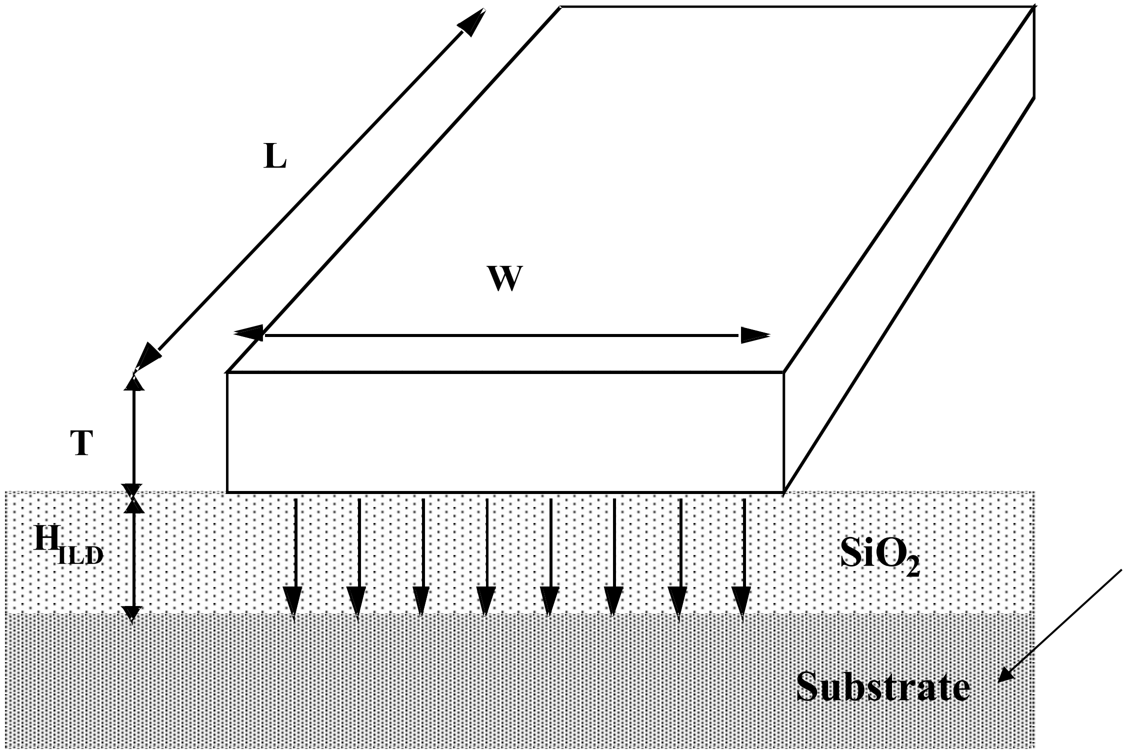
18

11, 2010

Parallel Capacitance Plate Model

ILD = interlevel (or, interlayer) dielectric

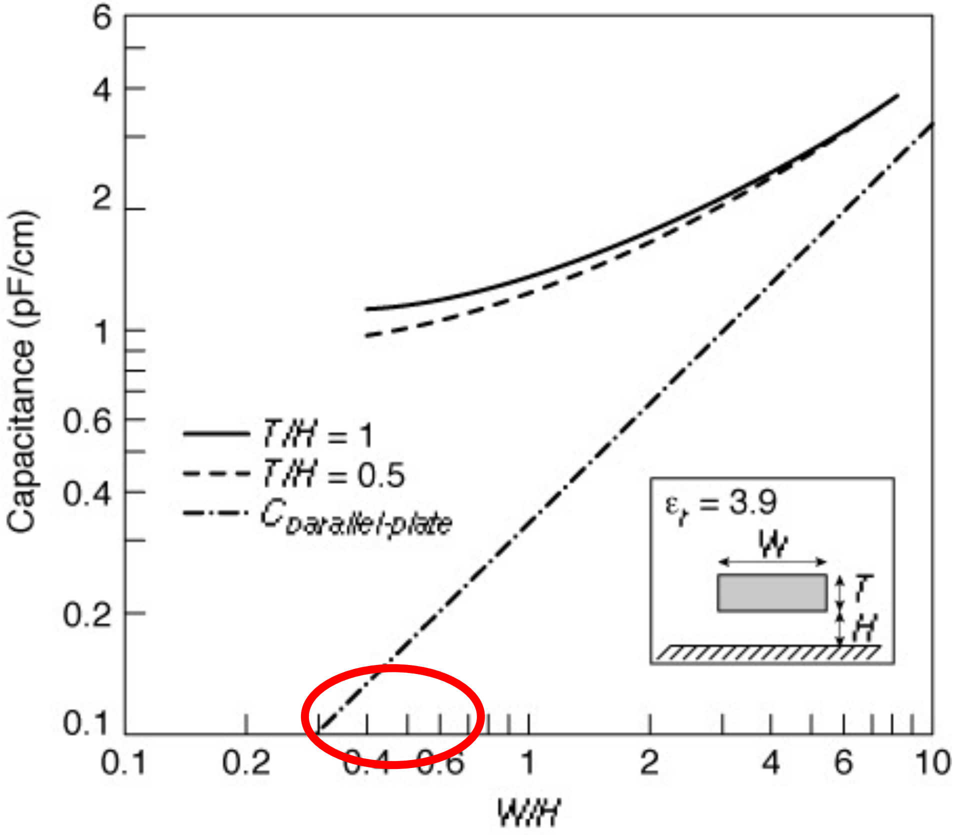
$H_{ILD} \equiv t_{ox}$



Bottom plate of cap can be either substrate or another metal layer

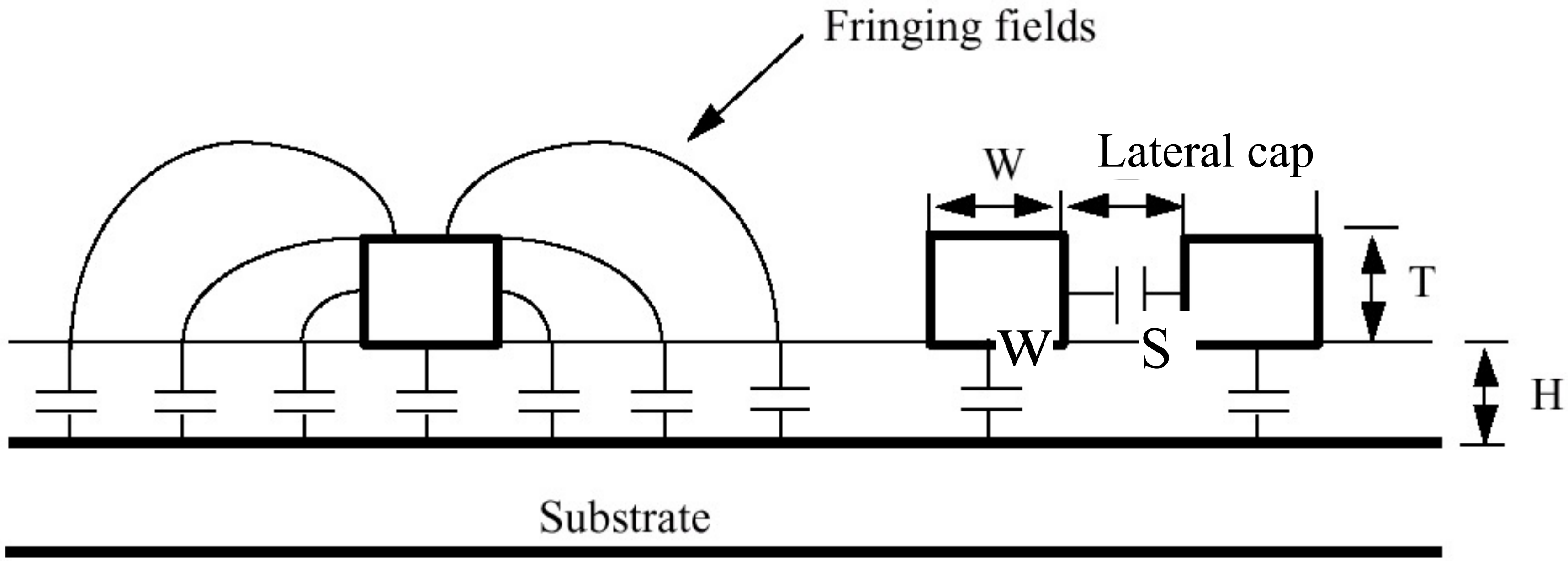
$C_{int} = e_{ox} * (W*L / t_{ox})$

Capacitance Values for Different Configurations



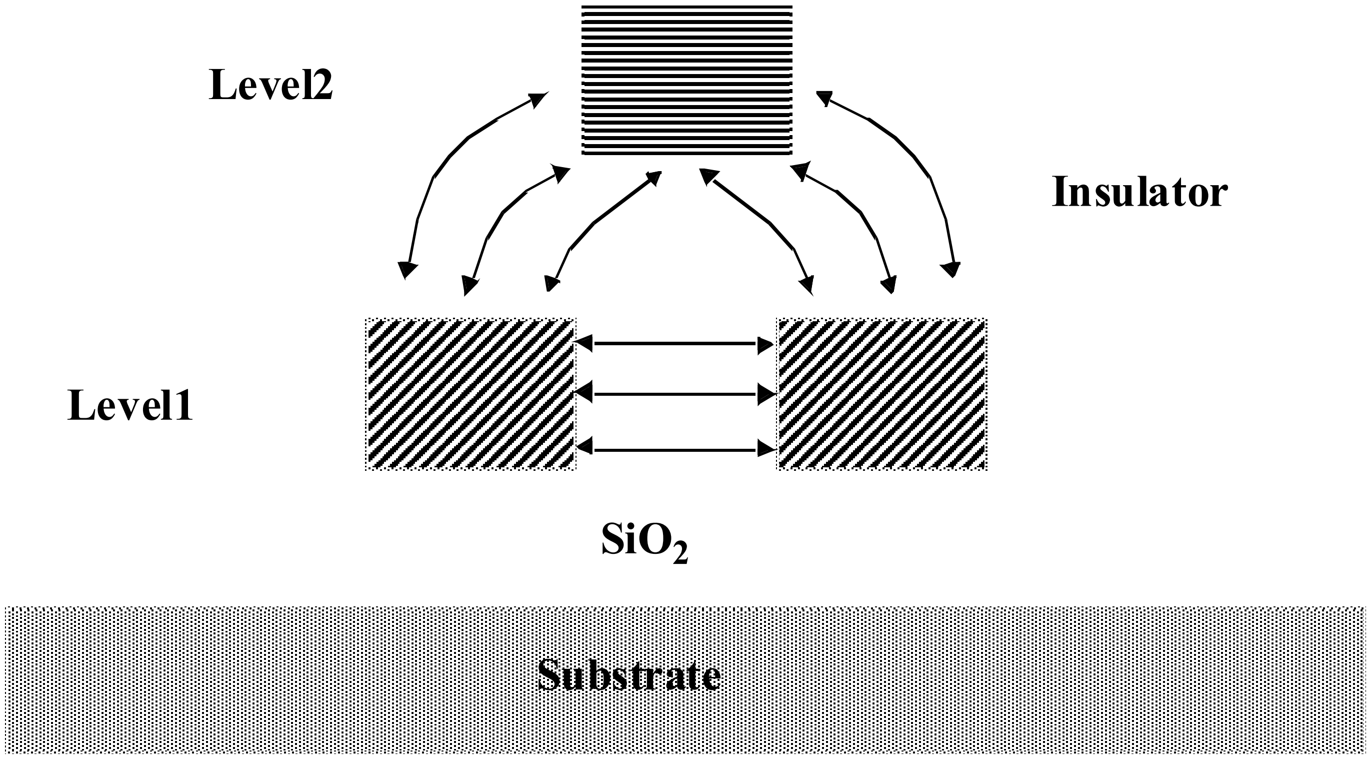
- Parallel-plate model substantially underestimates capacitance as W drops below order of H
  - More lateral / fringing capacitance elements, esp. relative to ground (substrate) coupling

Capacitance Values for Different Configurations



- Line dimensions: W, S (space), T, H

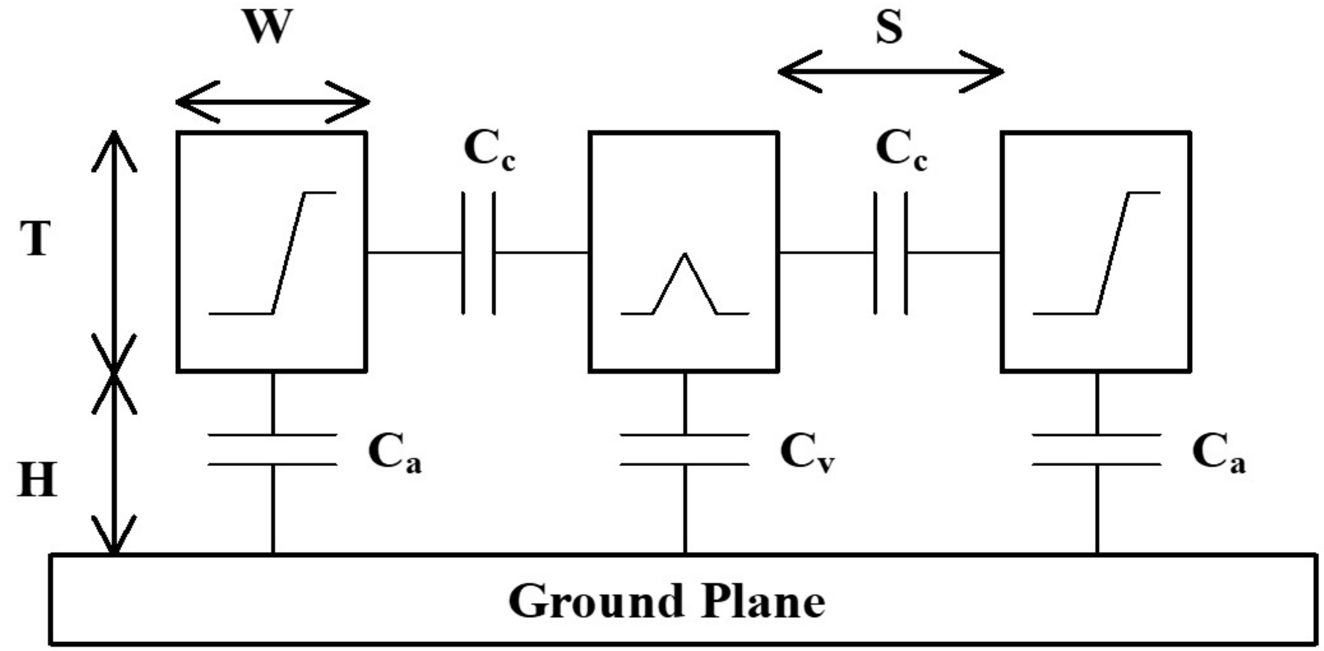
Inter-wire (Coupling) Capacitance



- Coupling effects among neighboring wires

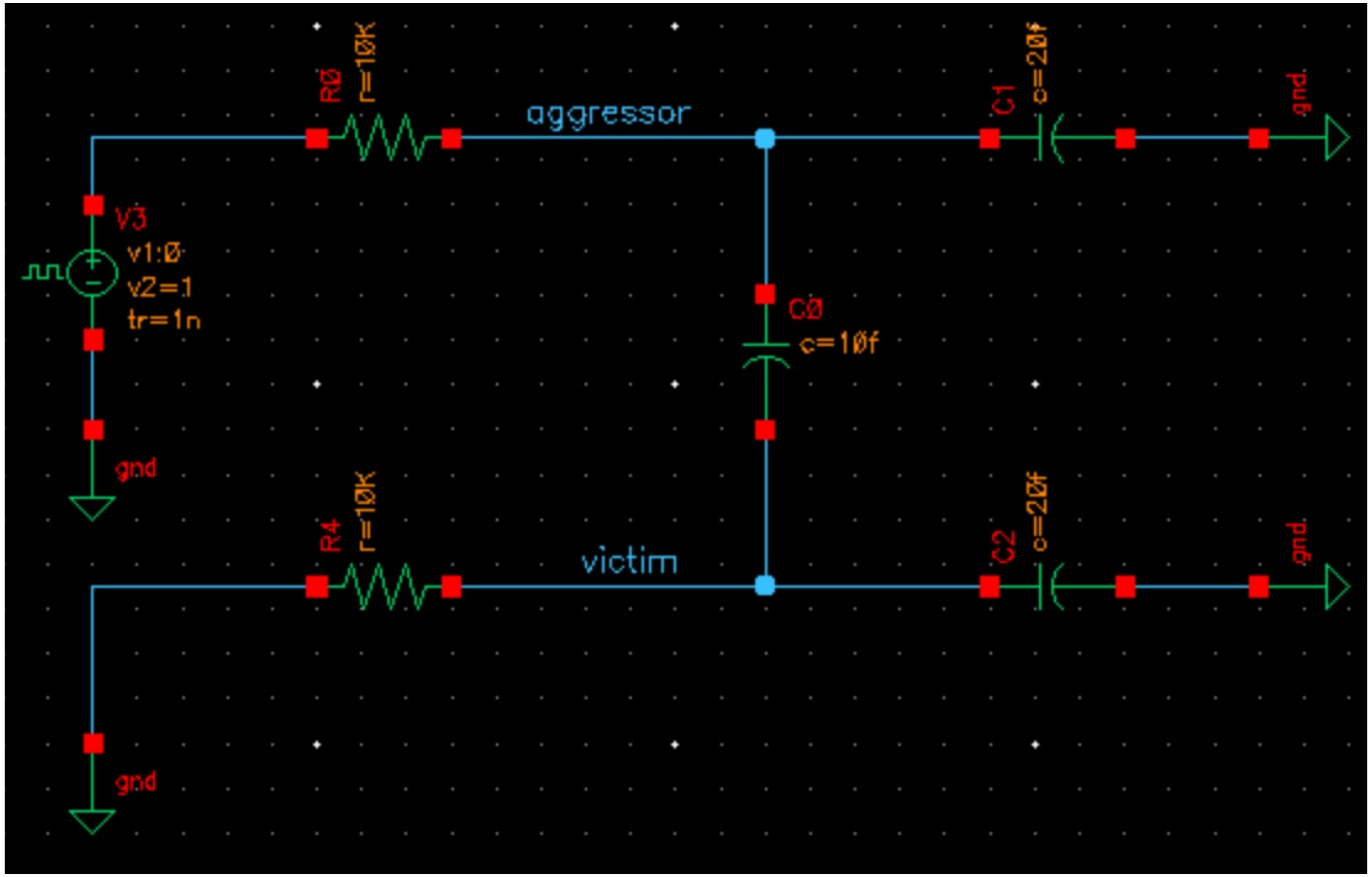
  - Includes cross-over / cross-under wires on other layers

Coupling Noise



- Cross-section: victim (v) and aggressors (a)
- Interwire capacitance allows neighboring wires to interact
- Charge injected across C_c results in temporary (in static logic) glitch in voltage at the victim

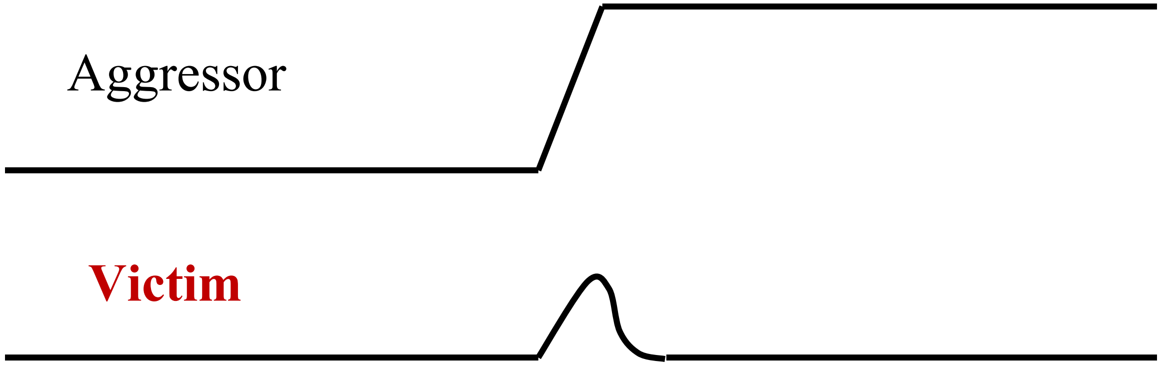
Spectre Simulation (Demo): Victim – Aggressor



- Rising slope with 100p / 1ns
- Simulations with C0 = 5 / 10 fF

Crosstalk from Capacitive Coupling

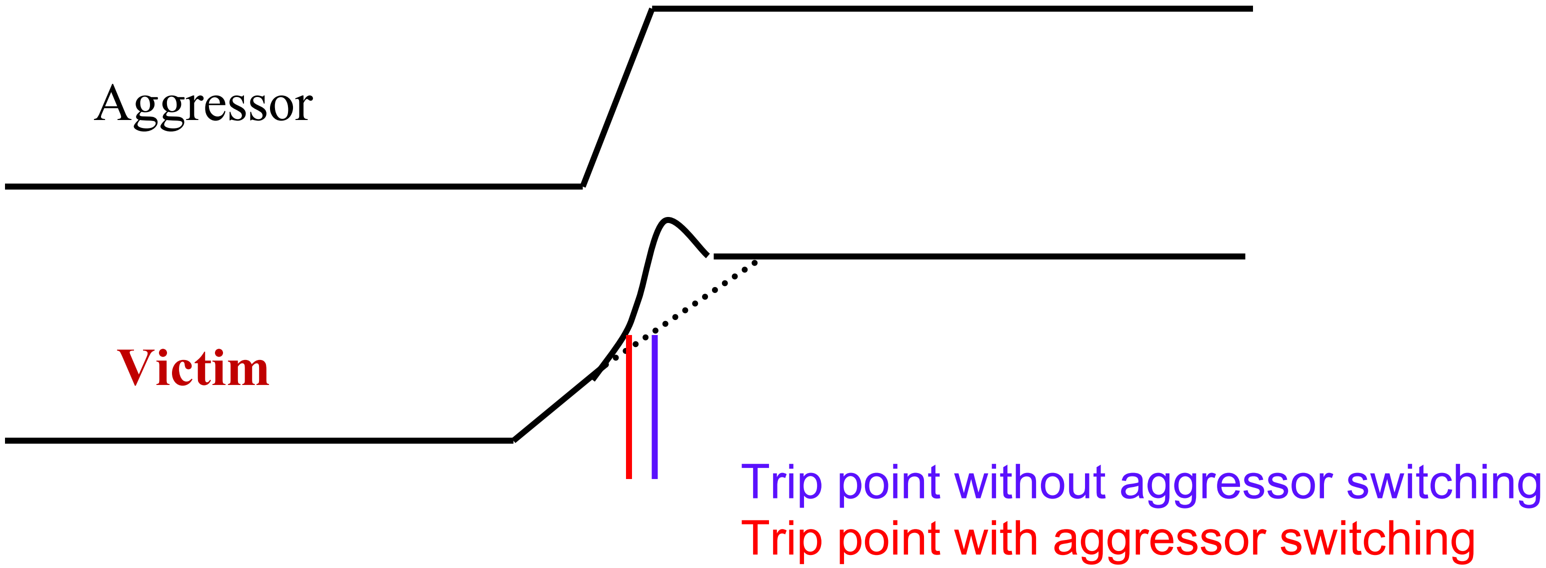
- Glitches caused by capacitive coupling between wires
  - An “aggressor” wire switches
  - A “**victim**” wire is charged or discharged by the coupling capacitance
- This is bad if:
  - The victim is a clock or asynchronous reset
  - The victim is a signal whose value is being latched at that moment



Slide courtesy of Paul Rodman, ReShape

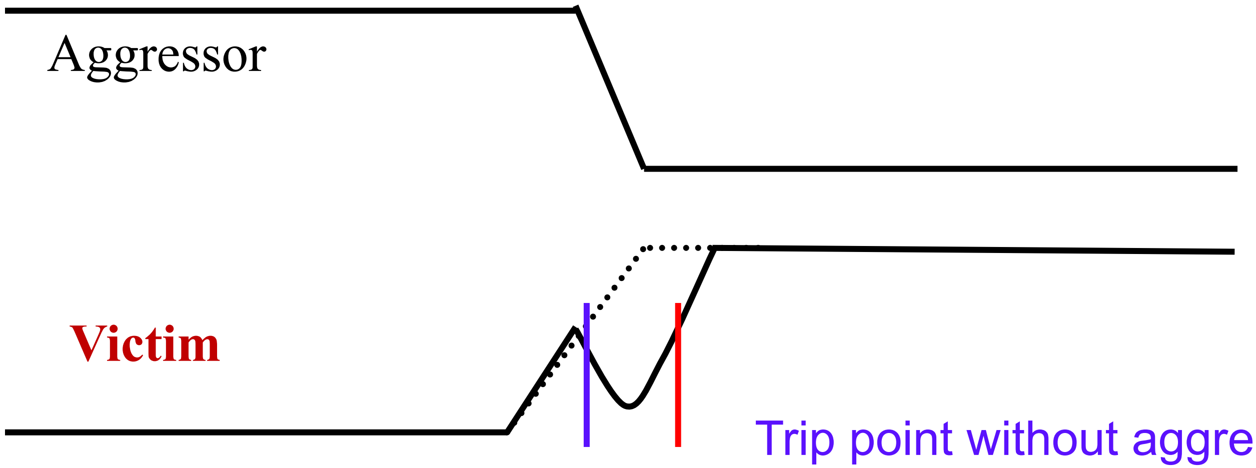
Crosstalk: Timing Pull-In

- A switching **victim** is sped up by a coupled aggressor that is switching in the same direction
- This is bad if your path now violates “hold time” (minimum path delay constraint) checks
- can be fixed by adding delay elements to your path


Aggressor
Victim
Trip point without aggressor switching
Trip point with aggressor switching

Crosstalk: Timing Push-out

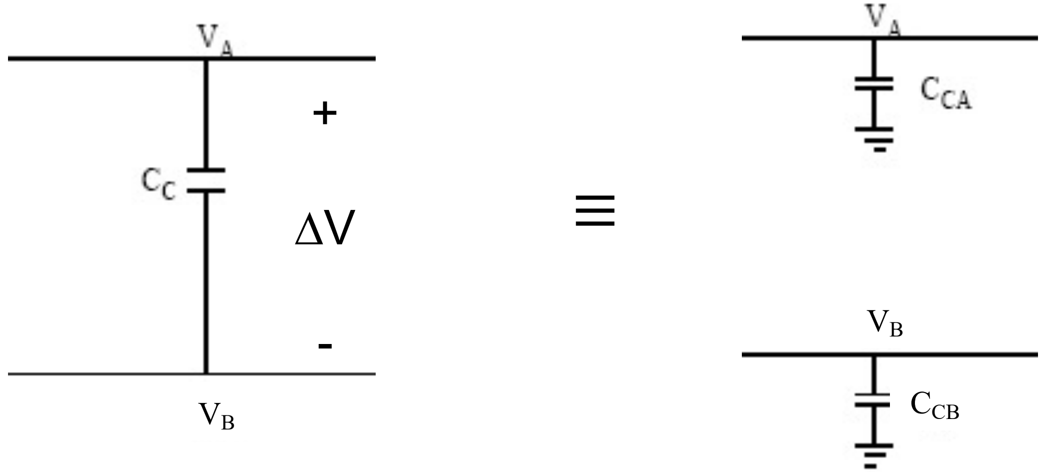
- A switching **victim** is slowed down by a coupled aggressor that is switching in the opposite direction
- This is bad if your path now violates setup time checks
- can be fixed by spacing the wires, using stronger drivers, …



Trip point without aggressor switching

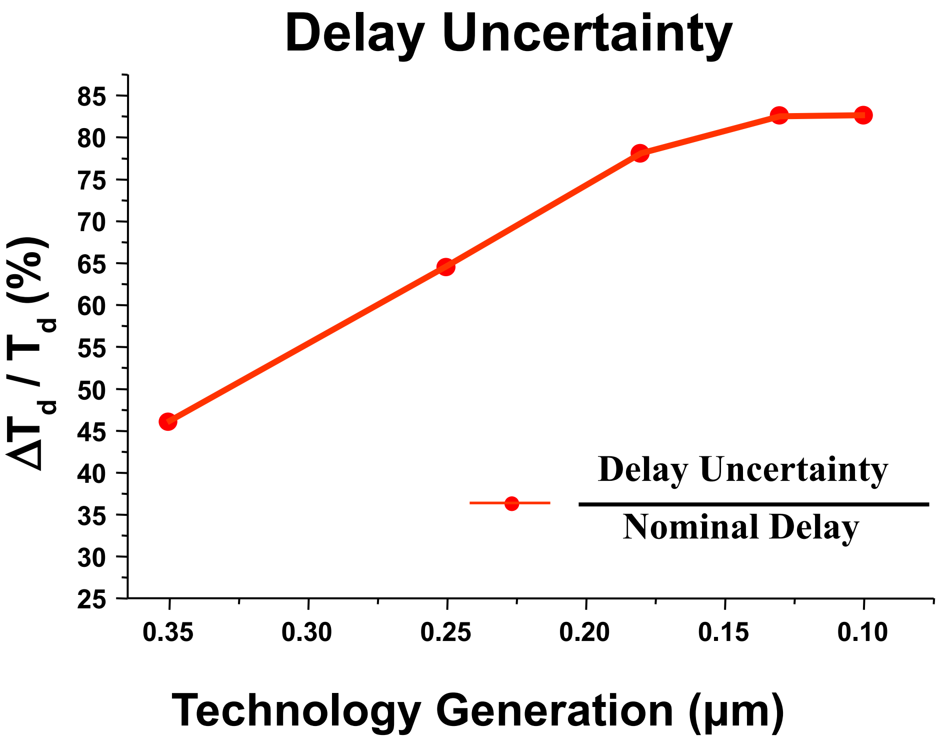
Trip point with aggressor switching

Miller Coupling Effect



- “A” switches but “B” does not: $\Delta V = V_{DD}$. A node “A” sees cap to node B is C_C.
  - “Miller Coupling Factor” (MCF) = 1. Thus $C_{C,eff} = C_C$.
- “A” and “B” switch in same direction: no voltage change: $\Delta V = 0$, C_c is effectively absent
  - MCF = 0, Thus $C_{C,eff} = 0$.
- “A” and “B” switch in opposite directions: voltage change $\Delta V = 2V_{DD}$, twice as much charge is required, C_c is effectively doubled
  - MCF = 2, Thus $C_{C,eff} = 2C_C$.

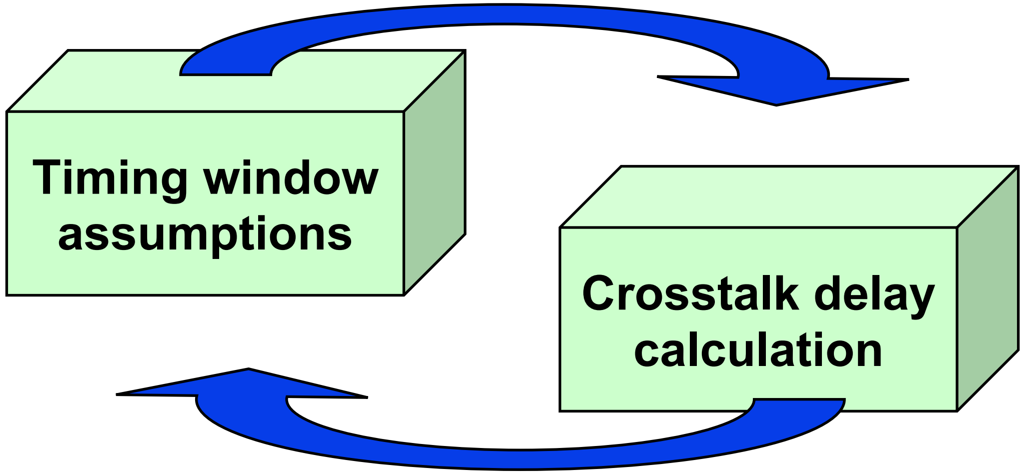
Timing Uncertainty over Technology Scaling



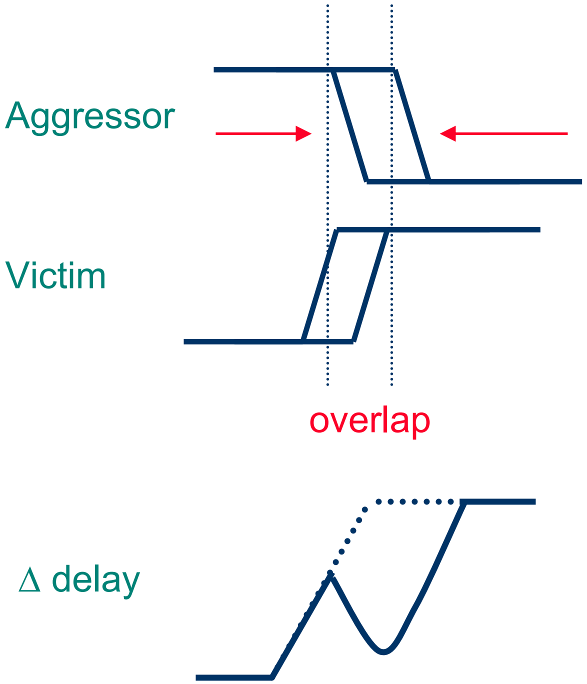
- Relatively greater coupling noise due to the reduced line space
- Tighter timing budgets to achieve fast circuit speed

Calculation Flow

- Timing window overlaps enable crosstalk delay variation
- Chicken-egg situation: delay vs. crosstalk
- Iteration
  - Starting with the assumption that all timing windows are overlapped (pessimistic about the unknowns)
  - Refine calculation by reducing pessimism

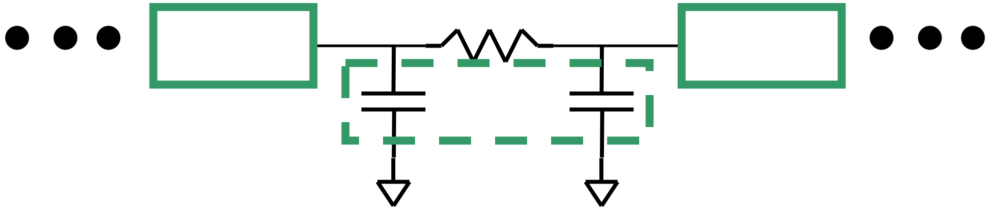

Timing window assumptions

Crosstalk delay calculation



Interconnect Modeling

- Model in SPICE by using R and C.
  - Π Model is usually used, where resistance and capacitance of an interconnect is distributed
  - Distributed model uses N segments
    - More accurate but can have computational cost
    - Number of nodes blows up
  - Lumped model uses 1 segment of Π
    - Adequate only for local (short, point to point) nets

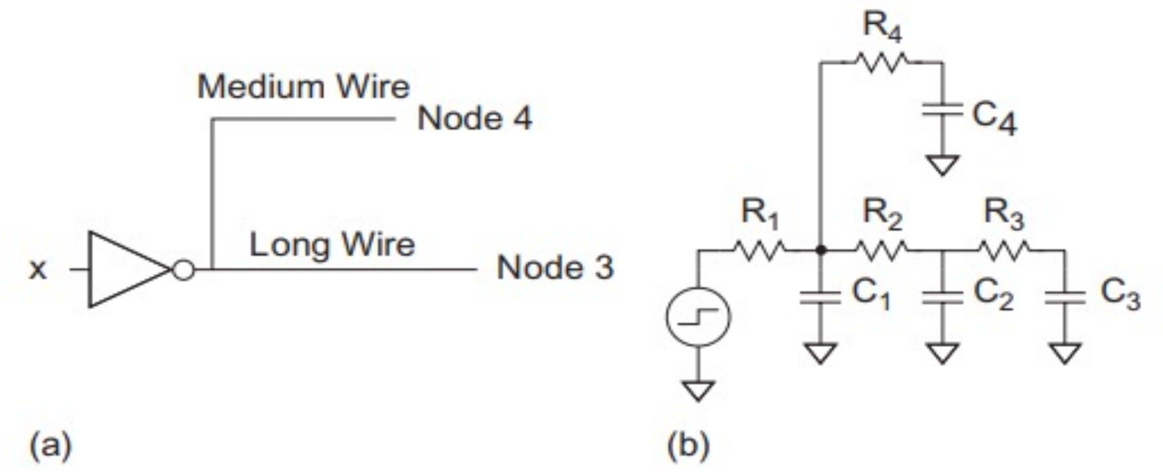


Distributed model of a single wire using multiple Π segments

Elmore Delay Model

$$\boldsymbol{t_{pd}(k) = 0.69\,\Sigma_{\mathrm{i}}\mathrm{R}_{\mathrm{is}}\mathrm{C}_{\mathrm{i}}}$$

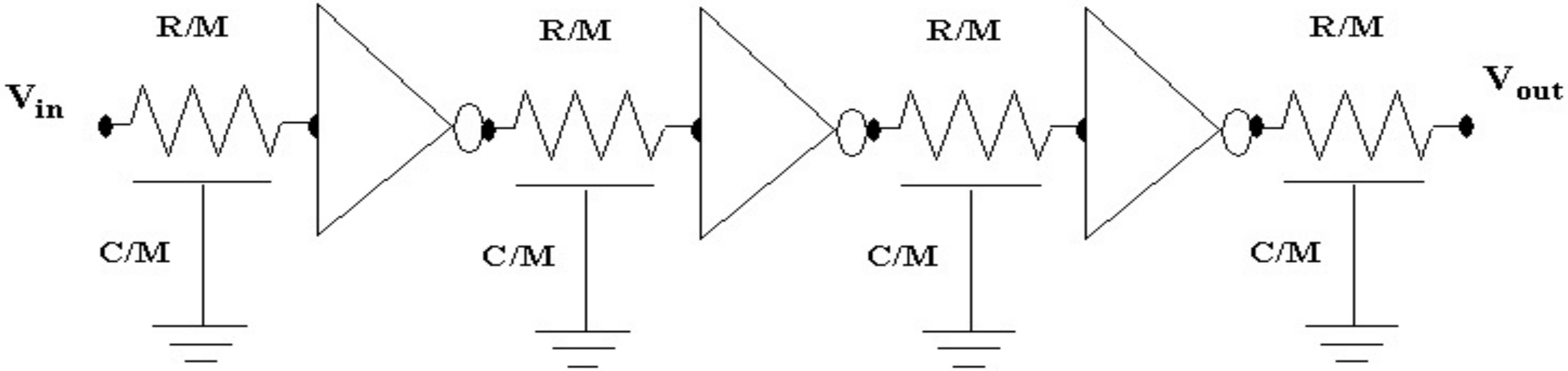
- C_i is the capacitance at node i
- R_{is} is the total on-path (= on the shared path to s_k) resistance from the source s_0 to target node i



$$T_{D_3} = R_1C_1 + \left(R_1 + R_2\right)C_2 + \left(R_1 + R_2 + R_3\right)C_3 + R_1C_4$$

$$T_{D_4} = R_1C_1 + R_1\left(C_2 + C_3\right) + \left(R_1 + R_4\right)C_4$$

Reducing RC Delay With Repeaters

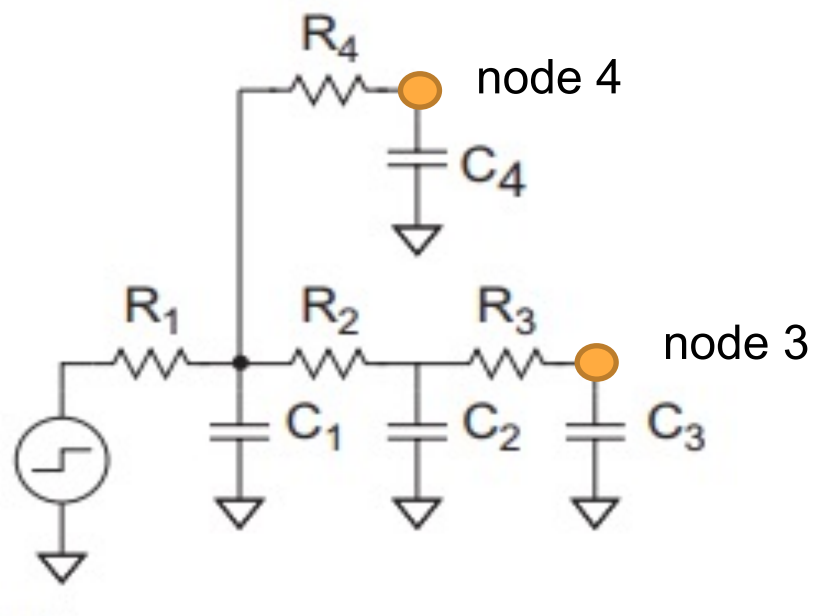


- Repeater

  - strong driver (usually inverter or pair of inverters for non-inversion)

  - placed along a long RC line to “break up” the line and reduce delay

HW: Elmore Delay



- Perform SPICE simulation with following parameters

  - Draw schematic

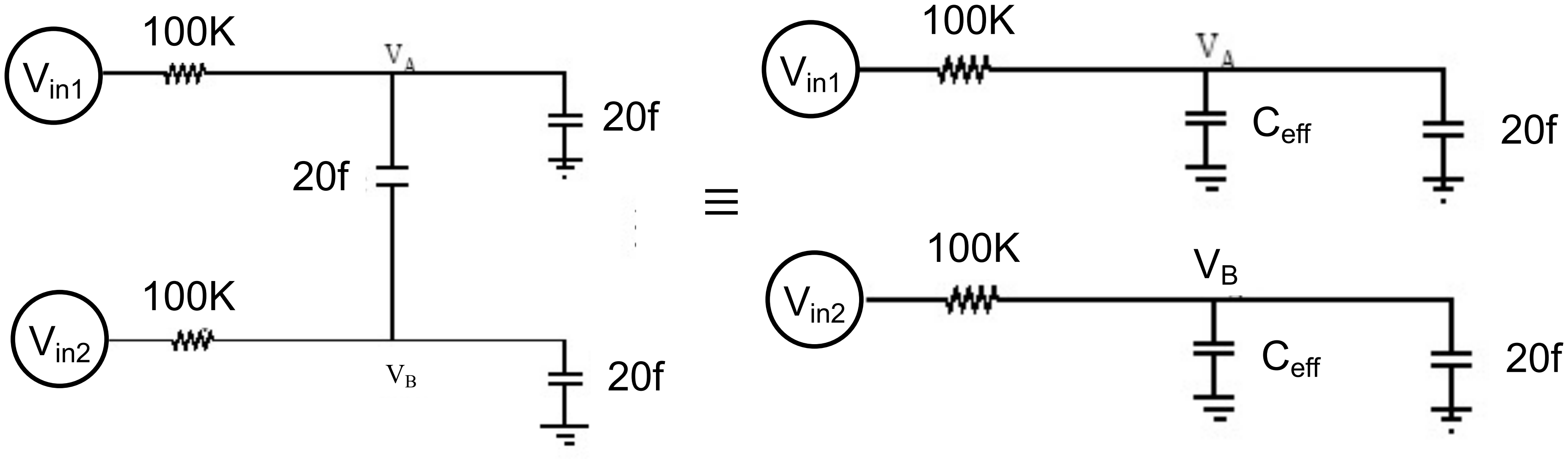
  - R3=30K while all the other R = 10K

  - C1=C2=C3=C4 = 20fF

  - Measure 50%-50% delay from the input to node 3 & 4

  - Compare with your theoretic calculation

Miller Coupling Effect



- Perform simulation for the following three cases

1) “A” switches but “B” does not:

2) “A” and “B” switch in same direction

3) “A” and “B” switch in opposite directions

What will be the C_{eff} value for above three cases ?

Simulate left and right cases to prove your calculated value (by comparing the delay).

(right bottom does not need to be simulated)