EE3230 Lecture 5: Circuit Characterization and Performance Estimation II

Ping-Hsuan Hsieh (謝秉璇)

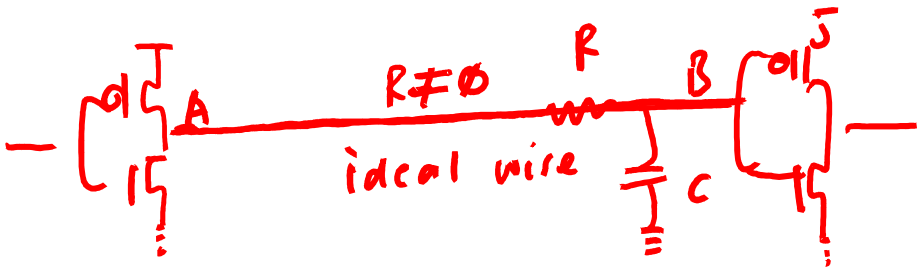
Delta Building R908

EXT 42590

phsieh@ee.nthu.edu.tw

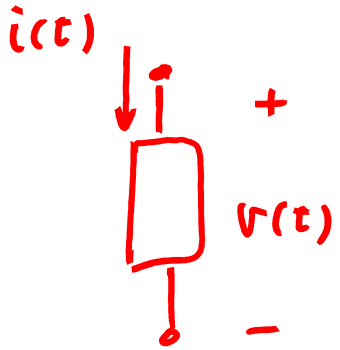
Outline

- Delay estimation
- Logical effort and transistor sizing
- **Power dissipation** (consumption) Ch5
- Interconnect Ch6
- Wire engineering
- Design margin
- Reliability
- Scaling

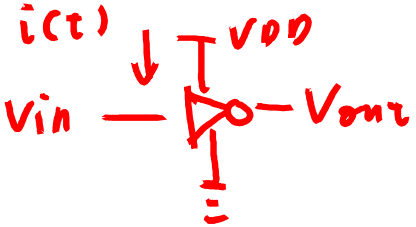


Power and Energy

- Power is drawn from a voltage source attached to the V_{DD} pin(s) of a chip



- Instantaneous Power: $p(t) = v(t) \cdot i(t) = V_{DD} \cdot i(t)$



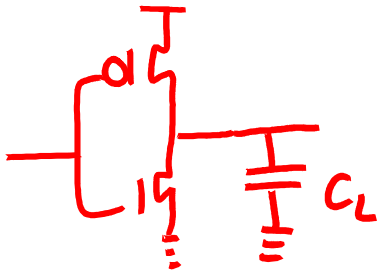
- Energy: $E = \int_0^T p(t)\,dt = V_{DD} \cdot \int_0^T i(t)\,dt$

- Average Power: $P_{avg} = \frac{E}{T} = \frac{1}{T} V_{DD} \int_0^T i(t)\,dt = V_{DD} \cdot \bar{I}$

Static and Dynamic Power Dissipation

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

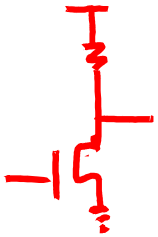
diff modes: active, stand-by, sleep



- Static power dissipation
  - Sub-threshold conduction through OFF transistors
  - Tunneling current through gate oxide
  - Leakage through reverse-biased diodes
  - Contention current in ratioed circuits

SOI 較佳 Bd leakage performance

leakage



non-zero static current

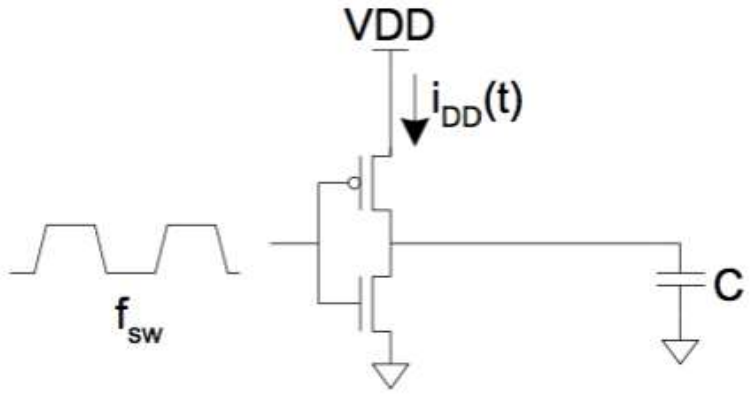
- Dynamic power dissipation
  - Charging and discharging of load capacitance
  - "Short-circuit" current while both PMOS and NMOS networks are partially ON

Dynamic Power (I)

- Dynamic power is required to charge and discharge load capacitances when transistors switch
- One cycle involves a rising and falling output.
- On rising output, charge $Q = CV_{DD}$ is required
- On falling output, charge is dumped to GND
- This repeats f_{sw} times per second

$\Rightarrow E = CV_{DD}^2$

$$\bar{P} = \frac{CV_{DD}^2}{T_{dur}} \propto CV_{DD}^2 \cdot f_{switching}$$



Dynamic Power (II) 充電.

$$P_{dynamic} = \frac{1}{T}\left(\int_0^T V_{DD}\, i(t)\, dt\right)$$ energy

instant power

從supply的角度看進去module

$$= \frac{V_{DD}}{T}\cdot\int_0^T i(t)\,dt$$

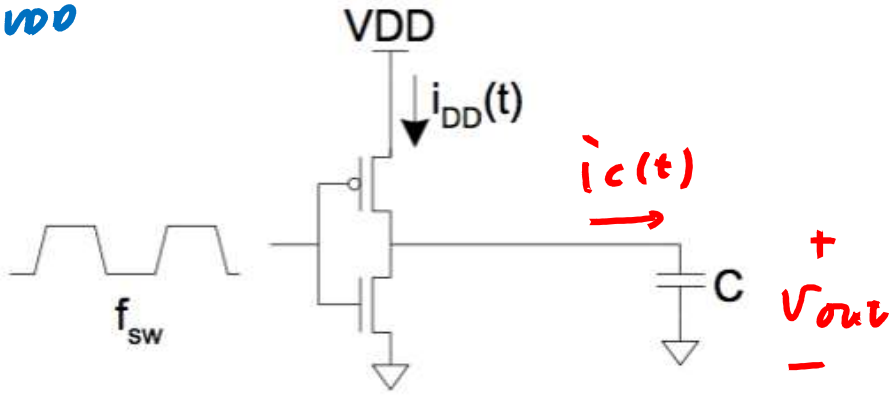
$$= \frac{V_{DD}}{T}\cdot C\cdot V_{DD}$$ energy = E_{VDD}

$$= C V_{DD}^2 \cdot f$$

$$E_C = \int_0^\infty i(t)\, v_{out}(t)\, dt$$

$$= \int_0^\infty C\,\frac{dv_{out}(t)}{dt}\cdot v_{out}(t)\, dt$$

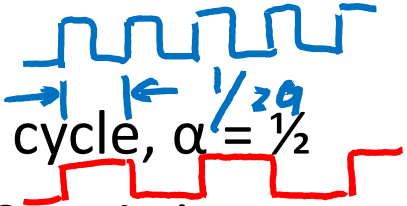
$$= \frac{1}{2} C V_{DD}^2$$



Activity Factor

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where **α = activity factor**
  - If the signal is a clock, α = 1
  - If the signal switches once per cycle, α = ½
  - **Dynamic gates:** Switch either 0 or 1 times per cycle, a = ½
  - **Static gates:** Depends on design, typically α = 0.1

@ 2GHz



CMOS random signal, switch with a probability of 0.5 → α=0.25

empirical data

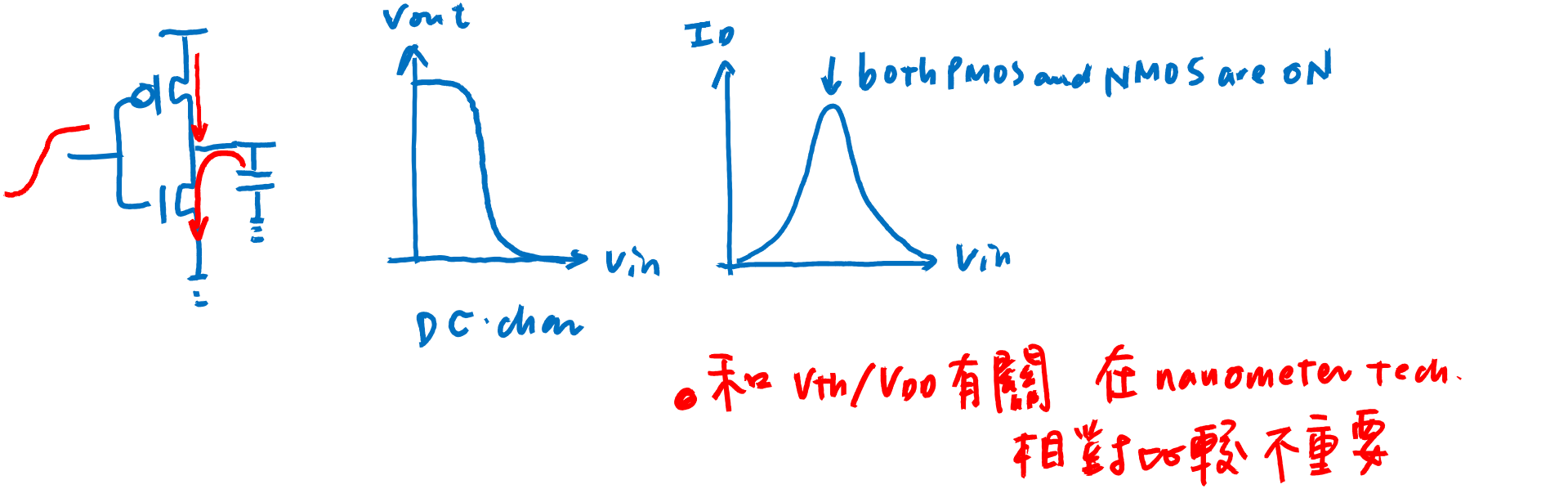
- Dynamic power:

$$P_{dynamic} = \alpha C V_{DD}^2 f_{CK}$$

Short-Circuit Current

- When transistors switch, both NMOS and PMOS networks may be momentarily ON at once
- Leads to a blip of **short-circuit current**
- < 10% of dynamic power if rise/fall times are comparable for input and output (well-controlled)

特別是受 input slope 影響



- 和 Vth/VDD 有關 在 nanometer tech. 相對比較不重要

Example (1)

- 200M transistor chip System
  - 20M logic transistors
    - Average width: 12 λ
  - 180M memory transistors
    - Average width: 4 λ
  - 1.2-V 100-nm process (λ = 0.5* feature size = 50nm) min channel length
  - C_g = 2 *f*F/μm

$$\bar{P} = \alpha \cdot C V_{DD}^2 \cdot f_{sw}$$

Dynamic Power Consumption (II)

- Static CMOS logic gates: activity factor = 0.1
- Memory arrays: activity factor = 0.05 (many banks and partially activated at a time!)
- Estimate dynamic power consumption per MHz.
  - (Neglect wire capacitance and short-circuit current)

$$C_{logic} = (20 \cdot 10^6)(12 \cdot 50\,nm)\left(\frac{2\,fF}{\mu m}\right) = 24\,nF$$

$$C_{mem} = (180 \cdot 10^6)(4 \cdot 50\,nm)\left(\frac{2\,fF}{\mu m}\right) = 7.2\,nF$$

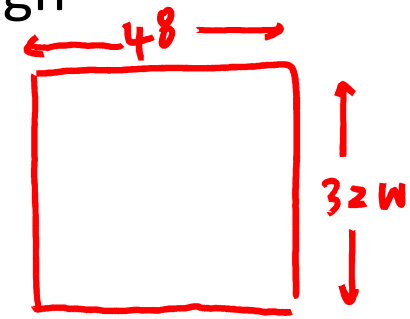
$$\bar{P} = \frac{0.1 \cdot 24\,nF \cdot 1.2^2 \cdot 10^6 + 0.05 \cdot 7.2\,nF \cdot 1.2^2 \cdot 10^6}{10^9} = 0.6\,W$$

Static Power Consumption

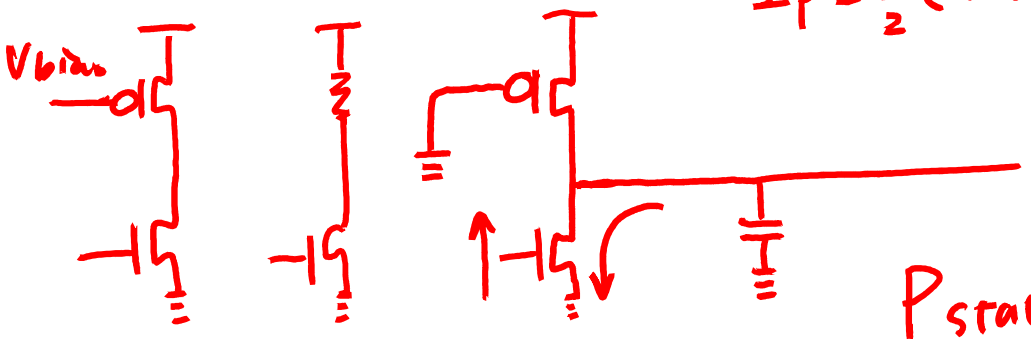
- Static power is consumed even when chip is quiescent.
  - (1) – Ratioed circuits burn power in fight with ON transistors
  - (2) – Leakage draws power from nominally OFF devices

Ratioed Example

- The chip contains a 32 word x 48 bit ROM
  - Uses 1:32 pseudo-nMOS decoder and bit-line pull-ups
  - In average, one wordline and 24 bitlines are high
- Find static power drawn by the ROM
  - $\beta = 75\ \mu A/V^2$, $V_{DD} = 1.8\ V$
  - $V_{tp} = -0.4V$



$I_p = \frac{\beta}{2}(V_{DD} - |V_{th1}|)^2 = 13\mu A$



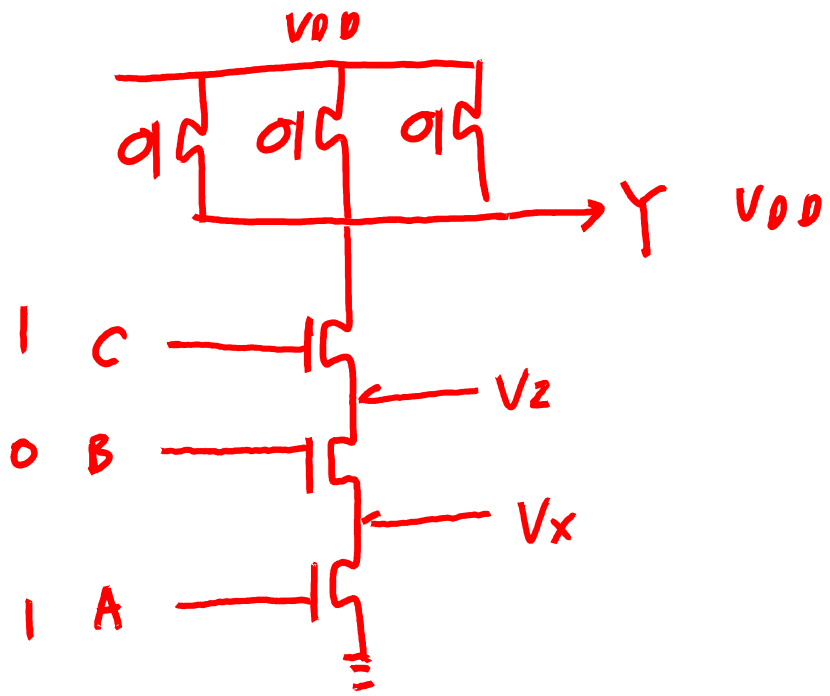
$P_{static} = (31+24)\,13\mu A \cdot 1.8\,V$

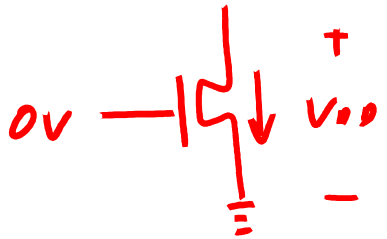
$= 1.2mW$ 和C無關

NAND3

leakage 與 data dependent

- 15-A oxide with 65-nm channel length
- (1) Gate leakage through ON NMOS is 6.3 nA ($V_{GS}=V_{DD}$) (ignore PMOS gate leakage)
- (2) Subthreshold leakage through NMOS with $V_{DS}=V_{DD}$ is 5.63 nA, and is 9.3 nA for PMOS





NAND3

- Subthreshold leakage I_{SUB}
- 000 Stacking
- 110
- 011
- 111

ABC	ISUB	IGATE	Vx	Vz
000	0.4		Stack	Stack
001				
010	0	1.3		
011	3.8		Vdd - Vth	Vdd - Vth
100				
101				
110	5.6		0	0
111	28		0	0

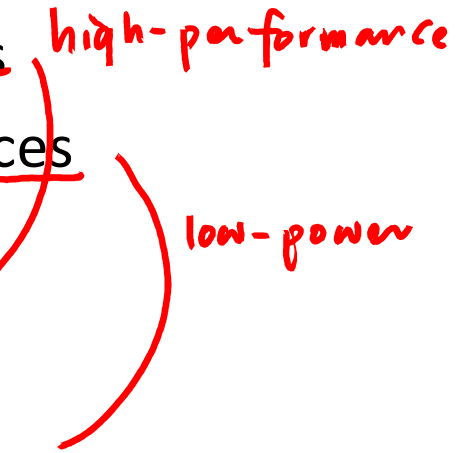
NAND3

- Gate leakage I_{GATE}
- 000
- 001
- 010
- 101

ABC	ISUB	IGATE	Vx	Vz
000		0		
001		~0		$V_{DD}-V_{th}$
010	0	1.3	intermediate	intermediate
011				
100				
101		6.3	0	$V_{DD}-V_{th}$
110				
111				

Leakage Example (I)

- The process has two threshold voltages and two oxide thicknesses.
- Subthreshold leakage:
  - 20 nA/μm for low V_{th} devices
  - 0.02 nA/μm for high V_{th} devices
- Gate leakage:
  - 3 nA/μm for thin oxide
  - 0.002 nA/μm for thick oxide
- Memories use low-leakage transistors everywhere
- Gates use low-leakage transistors on 80% of logic



Leakage Example (II)

- Estimate static power:
  - High leakage (performance): $(20 \cdot 10^6)(20\%)(12.50\text{nm}) = 2.4 \times 10^6\ \mu\text{m}$
  - Low leakage: $(20 \cdot 10^6)(80\%)(12.50\text{nm}) + (180 \cdot 10^6)(4.50\text{nm}) = 45.6 \cdot 10^6\ \mu\text{m}$

$$I_{static} = 2.4 \cdot 10^6\,(20\text{nA} + 3\text{nA})/2 + 45.6 \cdot 10^6\,(0.02\text{nA} + 0.002\text{nA})/2$$

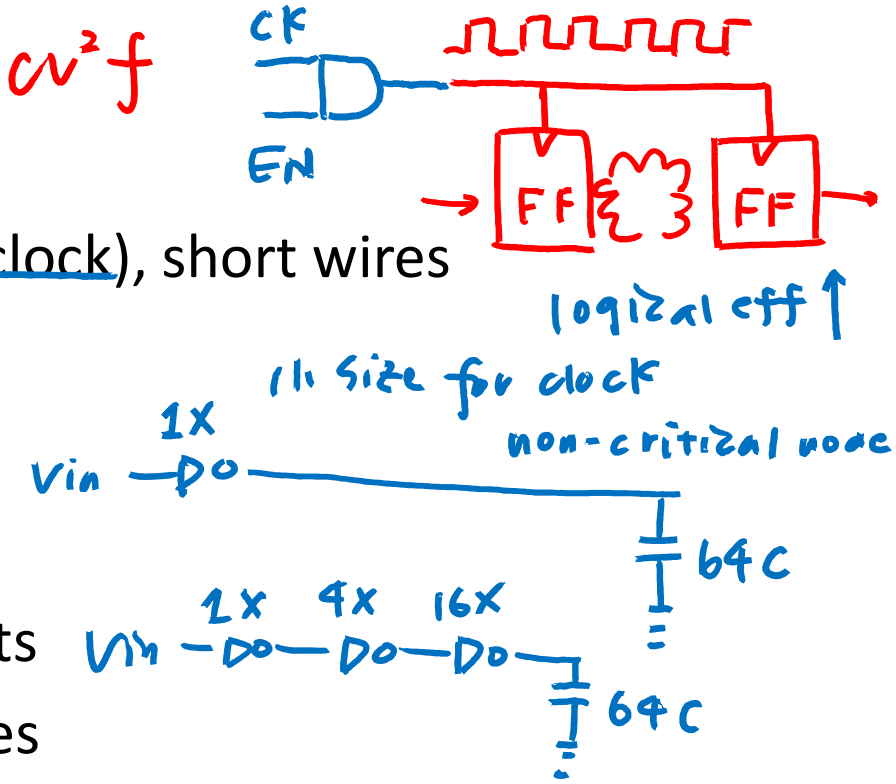
$$= 27.6 + 0.5016 = 28.1016\ \text{mA}$$

- Withow leakage devices, P_{static} = 552 mW!!

Low Power Design

- To rduce dynamic power
  - α: clock gating, sleep mode
  - C: small transistors (esp. on clock), short wires
  - V_{DD}: lowest suitable voltage
  - f: lowest suitable frequency
- To reduce static power
  - Selectively use ratioed circuits
  - Selectively use low V_{th} devices
  - Leakage reduction:

    stacked devices, body bias, low temperature

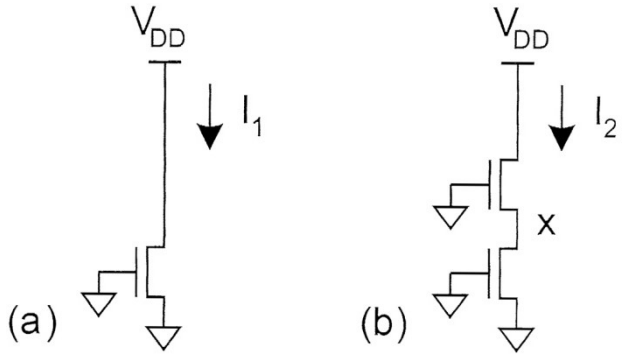


DVFS

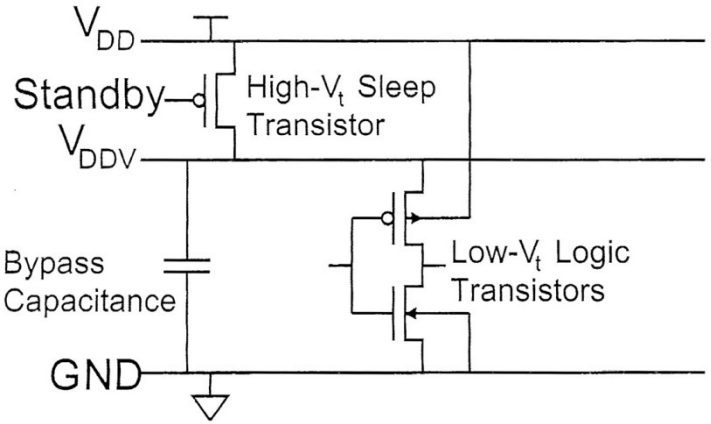
dynamic voltage-frequency scaling

Reduce Static Power

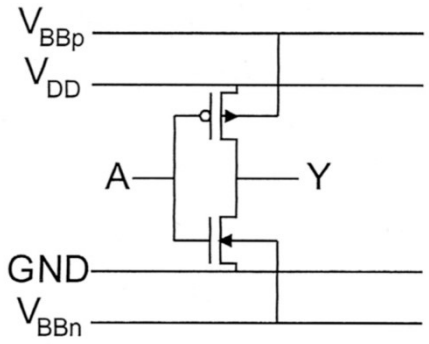
- Leakage stack effect

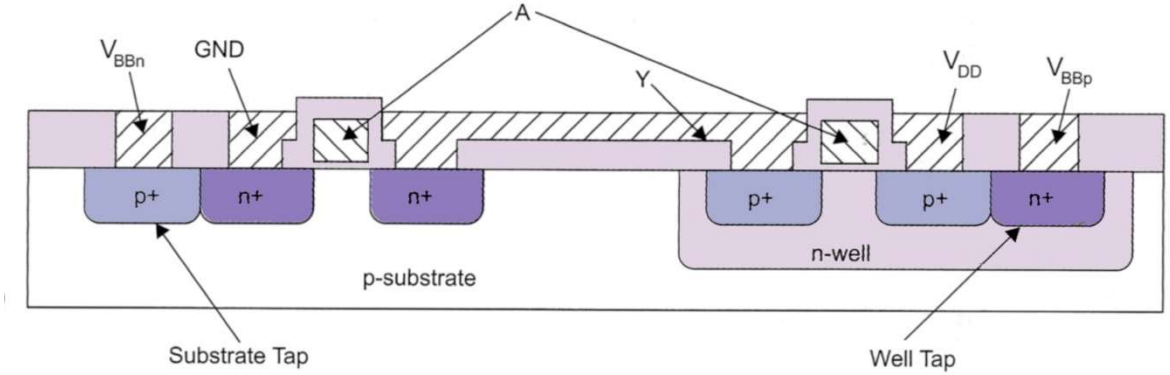


- MTCMOS: multiple threshold CMOS



- Body bias





Outline

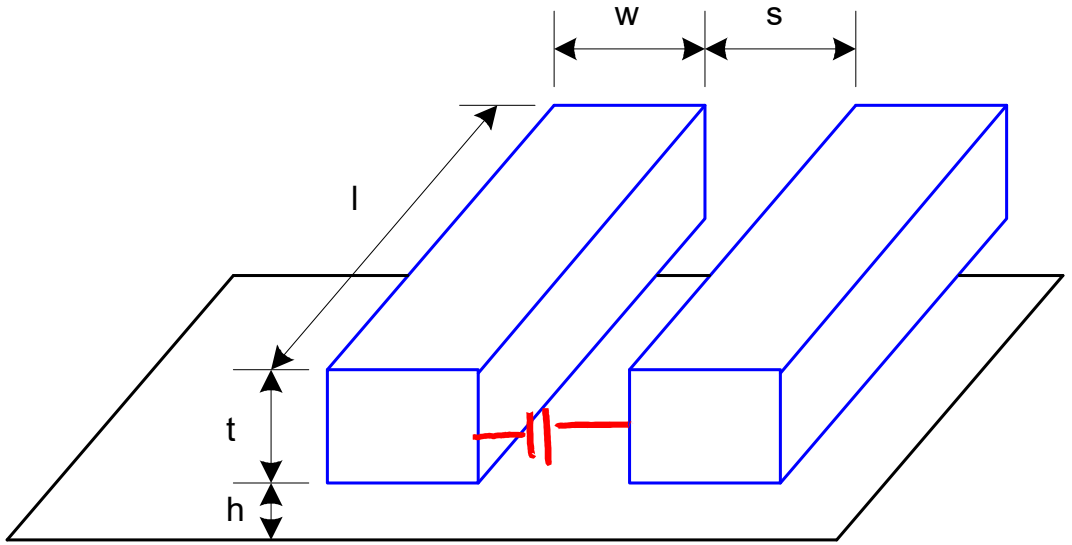
- Delay estimation
- Logical effort and transistor sizing
- Power dissipation
- **Interconnect**
- Wire engineering
- Design margin
- Reliability
- Scaling

Interconnect

- Chips are mostly made of wires called *interconnect*
  - In stick diagrams, wires determine size
  - Transistors are little things under the wires
  - Many layers of wires
- Wires are as important as transistors
  - Speed RC
  - Power C
  - Noise (coupling)
- Alternating layers run orthogonally

Wire Geometry

- Pitch = w + s
- Aspect ratio: AR = t/w
  - Old processes had AR << 1
  - Modern processes have AR ≈ 2 ↑ taller
    - Pack in many skinny wires



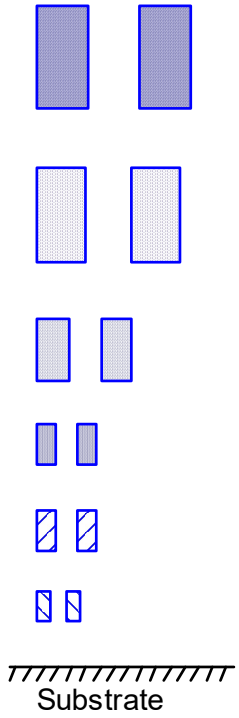
Layer Stack

- AMI 0.6-µm process has 3 metal layers
- Modern processes use 6-10+ metal layers CMP
- Example: Intel 180 nm process
- M1: thin, narrow (< 3λ)
  - High density cells
- M2-M4: thicker
  - For longer wires
- M5-M6: thickest
  - For V_{DD}, GND, clk

Layer	T (nm)	W (nm)	S (nm)	AR
6	1720	860	860	2.0
	1000			
5	1600	800	800	2.0
	1000			
4	1080	540	540	2.0
	700			
3	700	320	320	2.2
	700			
2	700	320	320	2.2
	700			
1	480	250	250	1.9
	800			

BEOL

Front-end of line FEOL

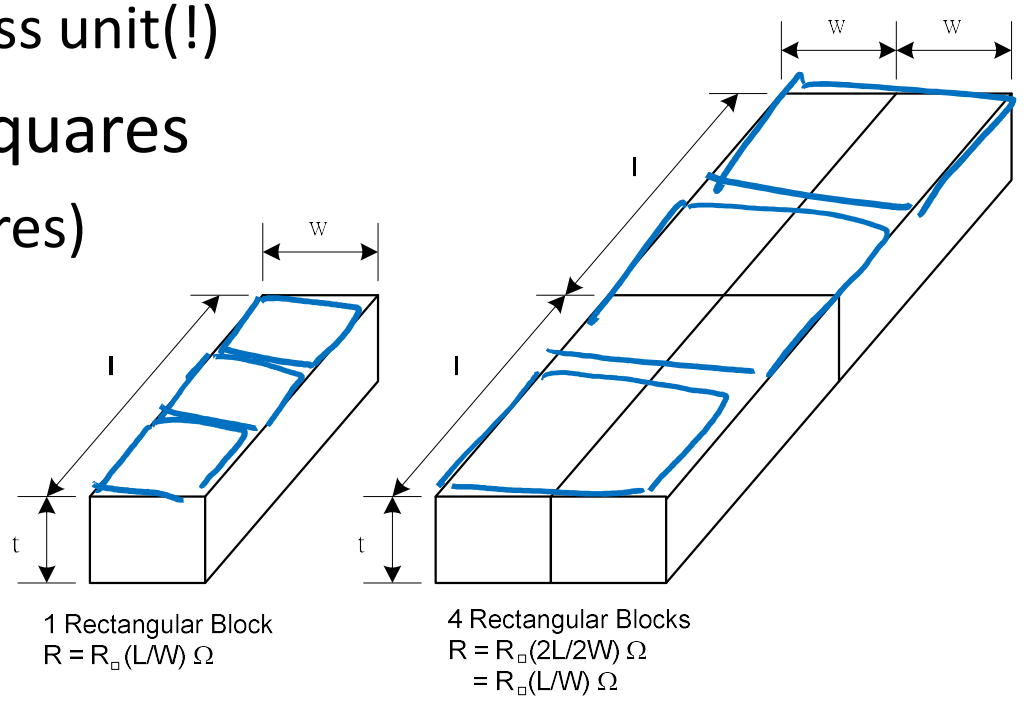


Wire Resistance

- ρ = *resistivity* (Ω*m)

$$R = \rho \cdot \frac{l}{w \cdot t} = \frac{\rho}{t} \cdot \frac{l}{w}$$

- $R_{\square}$ = *sheet resistance* (Ω/□)
  - □ is a dimensionless unit(!)
- Count number of squares
  - $R = R_{\square} * (\#\ \text{of squares})$



1 Rectangular Block
$R = R_{\square}(L/W)\ \Omega$

4 Rectangular Blocks
$R = R_{\square}(2L/2W)\ \Omega$
$= R_{\square}(L/W)\ \Omega$

Choice of Metals

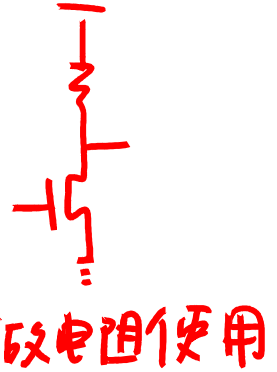
- Until 180 nm, most wires were aluminum
- Modern processes often use copper
  - Cu atoms diffuse into silicon and damage FETs
  - Must be surrounded by a diffusion barrier

Metal	Bulk resistivity (μΩ*cm)
Silver (Ag)	1.6
Copper (Cu)	1.7
Gold (Au)	2.2
Aluminum (Al)	2.8
Tungsten (W)	5.3
Molybdenum (Mo)	5.3

Sheet Resistance

- Typical sheet resistances in 180-nm process

Layer	Sheet Resistance (Ω/□)
Diffusion (silicided)	3-10
Diffusion (no silicide)	50-200
Polysilicon (silicided)	3-10
Polysilicon (no silicide)	50-400
Metal1	0.08
Metal2	0.05
Metal3	0.05
Metal4	0.03
Metal5	0.02
Metal6	0.02

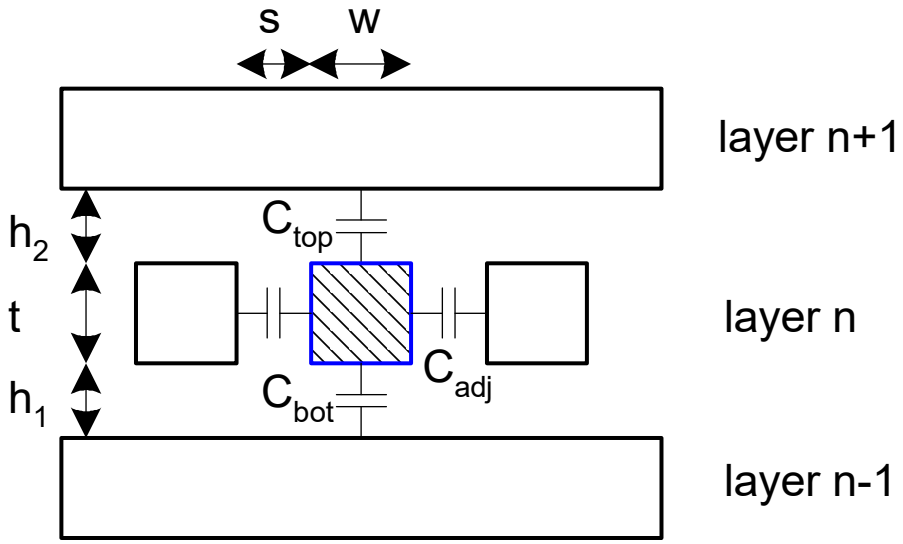


Contact Resistance

- Contacts and vias also have 2-20 Ω
- Use many contacts for lower R
  - Many small contacts for current crowding around periphery

Wire Capacitance

- Wire has capacitance per unit length
  - To neighbors
  - To layers above and below
- $C_{total} = C_{top} + C_{bot} + 2C_{adj}$

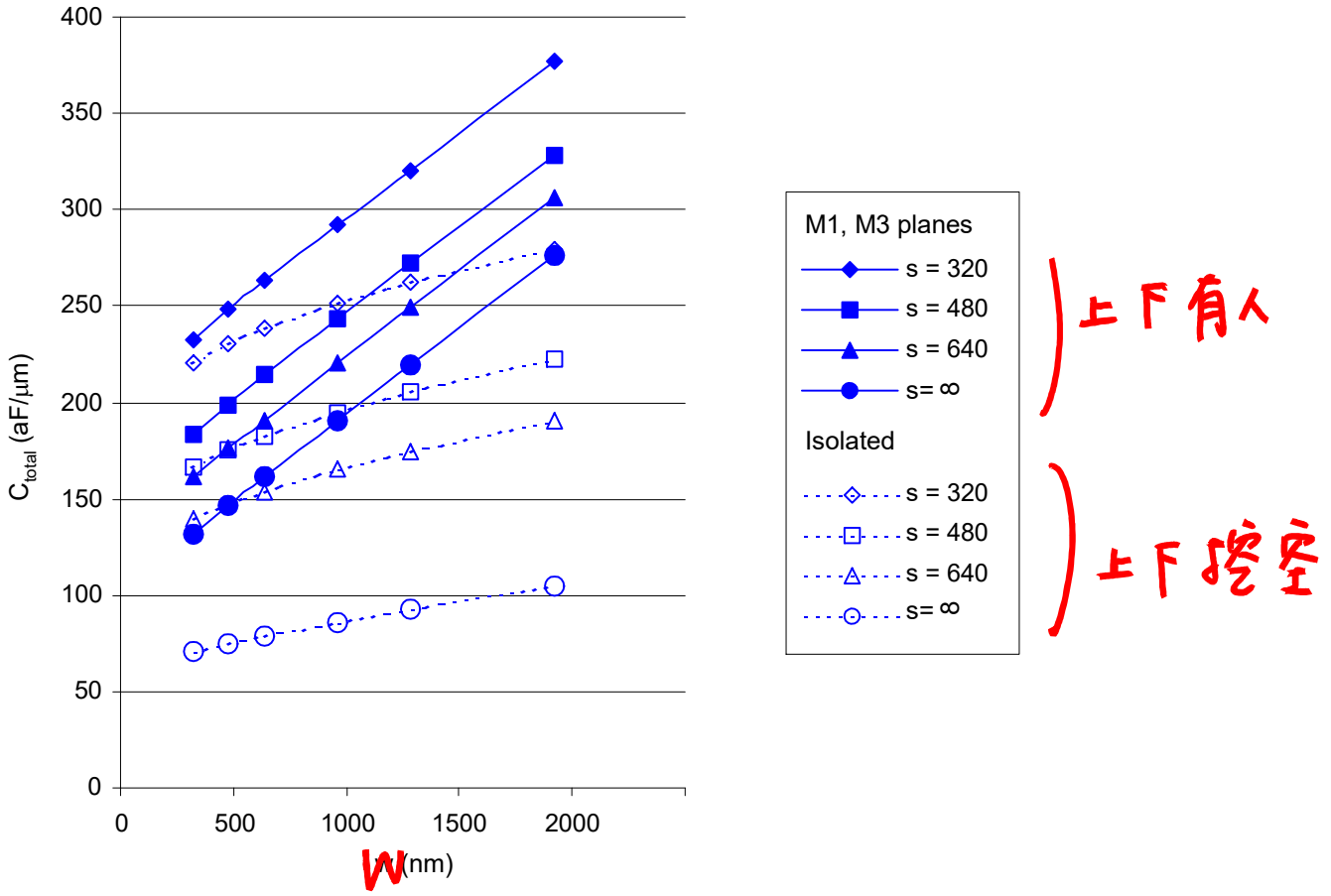


Capacitance Trend

- Parallel plate equation: $C = \varepsilon A/d$
  - Wires are not parallel plates, but obey trends
  - Increasing area (W, t) increases capacitance
  - Increasing distance (s, h) decreases capacitance
- Dielectric constant
  - $\varepsilon = k\varepsilon_0$
  - $\varepsilon_0 = 8.85 \times 10^{-14}$ F/cm
  - $k = 3.9$ for SiO_2
- Processes are starting to use low-k dielectrics
  - $k \approx 3$ (or less) as dielectrics use air pockets

M2 Capacitance Data

- Typical wires have ~ 0.2 *f*F/µm
  - Compare to 2 *f*F/µm for gate capacitance

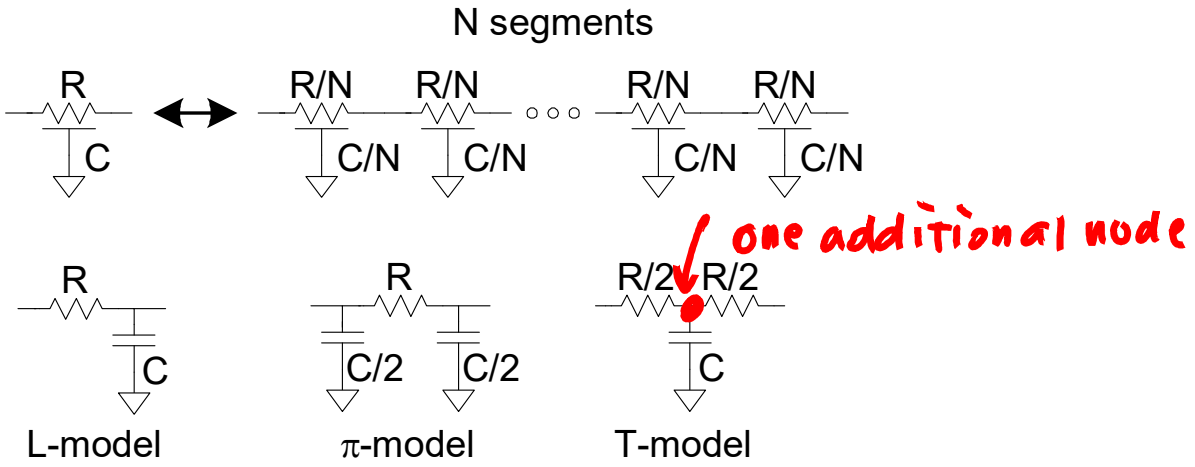


Diffusion and Polysilicon

- Diffusion capacitance is very high (about 2 *f*F/μm)
  - Comparable to gate capacitance
  - Diffusion also has high resistance
  - Avoid using diffusion *runners* for wires!
- Polysilicon has lower C but high R
  - Use for transistor gates
  - Occasionally for very short wires between gates

Lumped Element Models

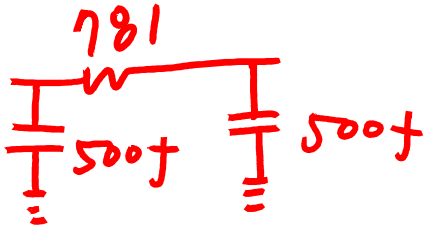
- Wires are a distributed system
  - Approximate with lumped element models

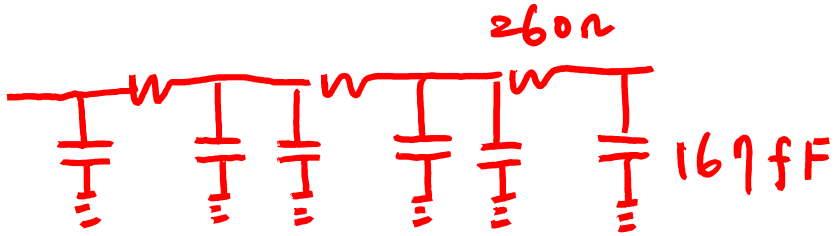


- 3-segment π-model is accurate to 3% in simulation
- L-model needs 100 segments for same accuracy!
- Use single segment π-model for Elmore delay

Example

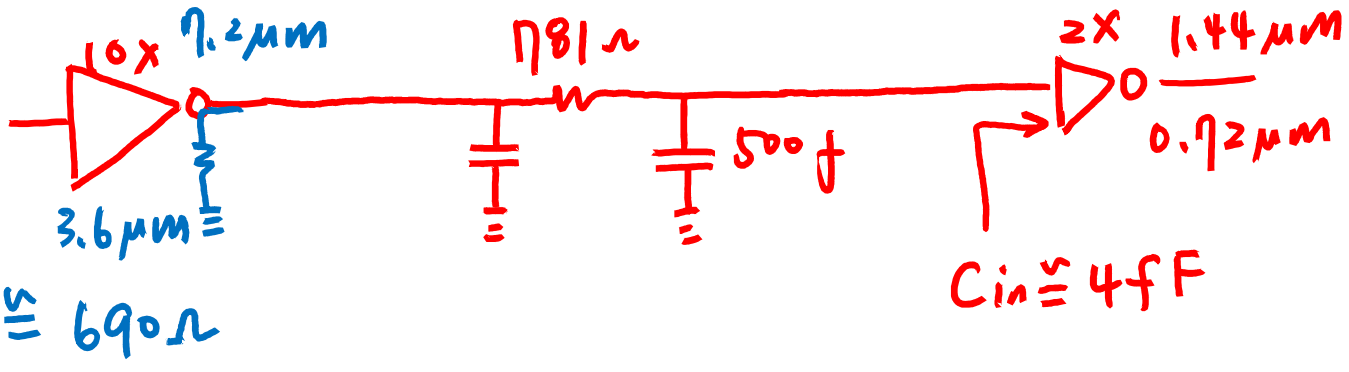
- M2 wire in 180-nm process
  - 5-mm long
  - 0.32-μm wide ≈ W_{min}
- Construct a 3-segment π-model
  - $R_{\square}$ = 0.05 Ω/□ → R = 781 Ω
  - $C_{permicron}$ = 0.2 *f*F/μm → C = 1 pF





Wire RC Delay

- Estimate the delay of a 10x inverter driving a 2x inverter at the end of the 5-mm wire from previous example
  - Effective R = 2.5 kΩ/μm for gates, C = 2 *f*F/μm
  - Unit inverter: 4λ = 0.36 μm nMOS, 8λ = 0.72 μm pMOS

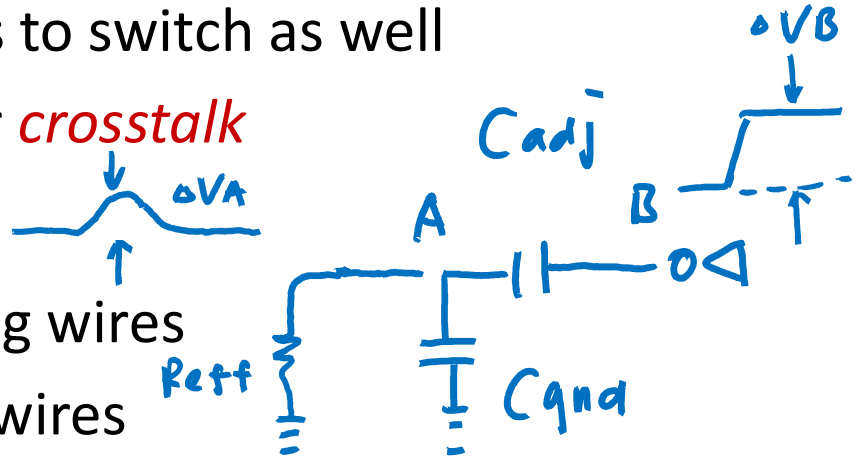


$$t_{paf} = 690 \cdot 500f + (690 + 981)(504f) \cong 1.1ns$$

$$F04 \cong 5RC = 110ps$$

Crosstalk coupling noise

- Capacitor do not change voltage instantaneously
- A wire has high capacitance to its neighbor
  - When the neighbor (aggressor) switches from 1→0 or 0→1, the wire (victim) tends to switch as well
  - Called *capacitive coupling* or *crosstalk*
- Impacts
  - Cause noise on non-switching wires
  - Increase delay on switching wires



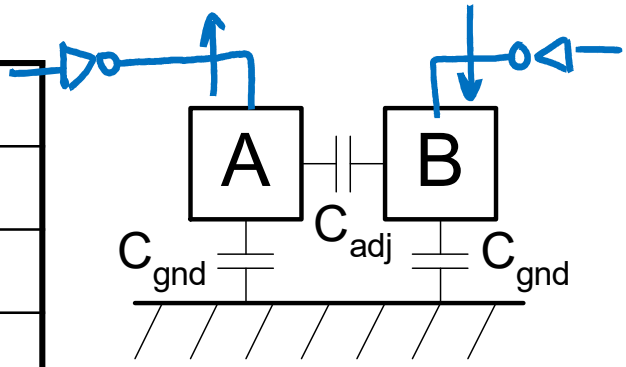
$$\Delta V_A = \Delta V_B \frac{C_{adj}}{C_{gnd} + C_{adj}}$$

Crosstalk Delay

- Assume layers above and below in average are quiet
  - Second terminal of capacitor can be ignored
  - Modeled as $C_{gnd} = C_{top} + C_{bot}$
- Effective C_{adj} depends on behavior of neighbors
  - *Miller Coupling Factor (MCF)*

跨在 Cadj 身上的 0V

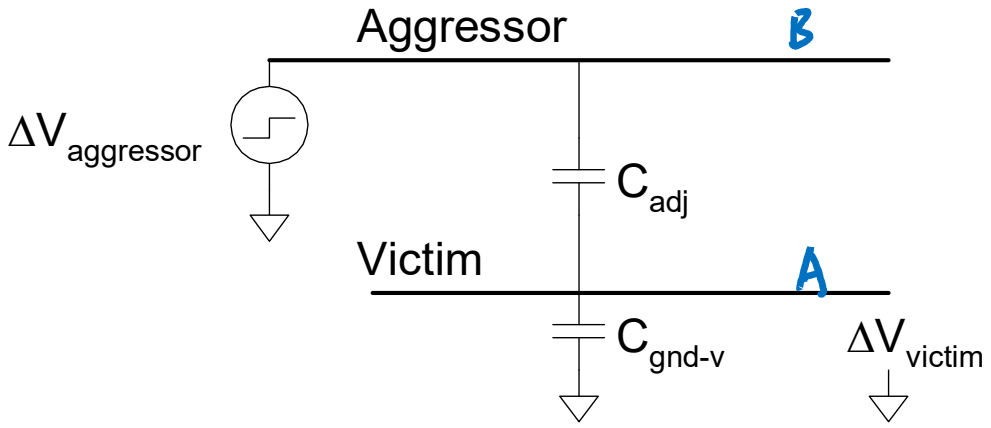
B	ΔV	$C_{eff(A)}$	MCF
Constant	V_{DD}	$C_{gnd} + C_{adj}$	1
Switching with A	0	C_{gnd}	0
Switching opposite A	$2V_{DD}$	$C_{gnd} + 2\,C_{adj}$	2



Crosstalk Noise (Floating Victims)

- Crosstalk causes noise on non-switching wires
- If victim is floating
  - modeled as capacitive voltage divider

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

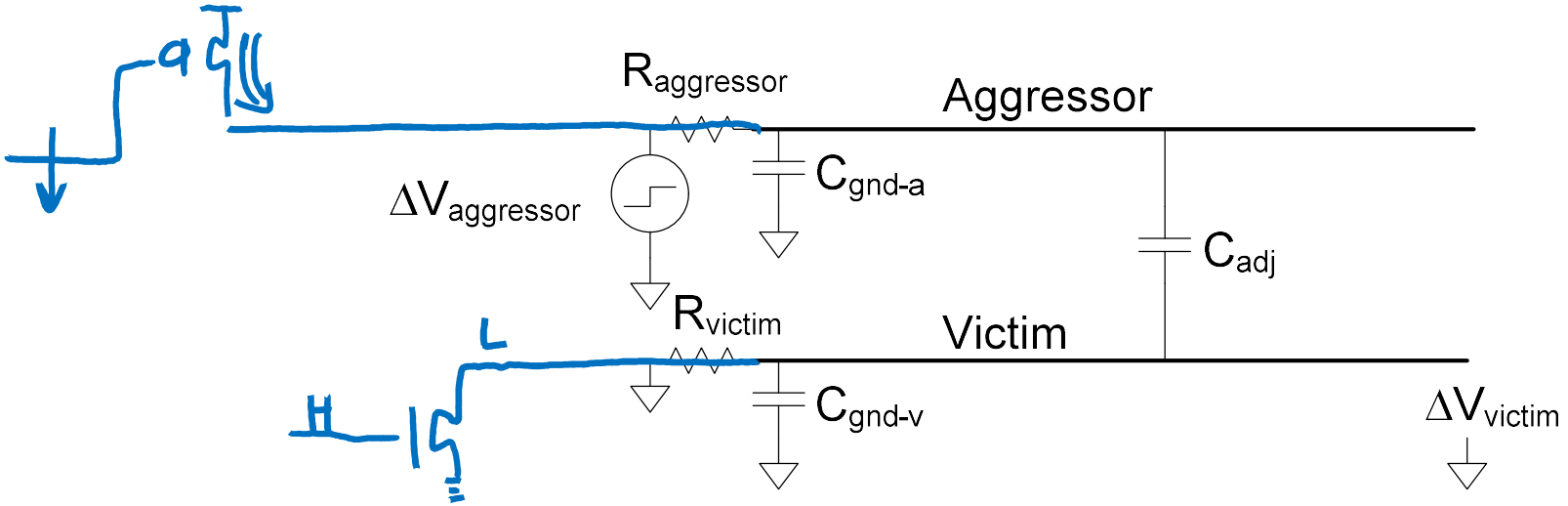


Crosstalk Noise (Driven Victims)

- Usually victim is driven by a gate that fights noise
  - Noise depends on relative resistances
  - Assume victim driver in linear region and aggressor driver in saturation (considering inverter operation)
  - With equal sizes, $R_{aggressor}$ = 2-4 x R_{victim}

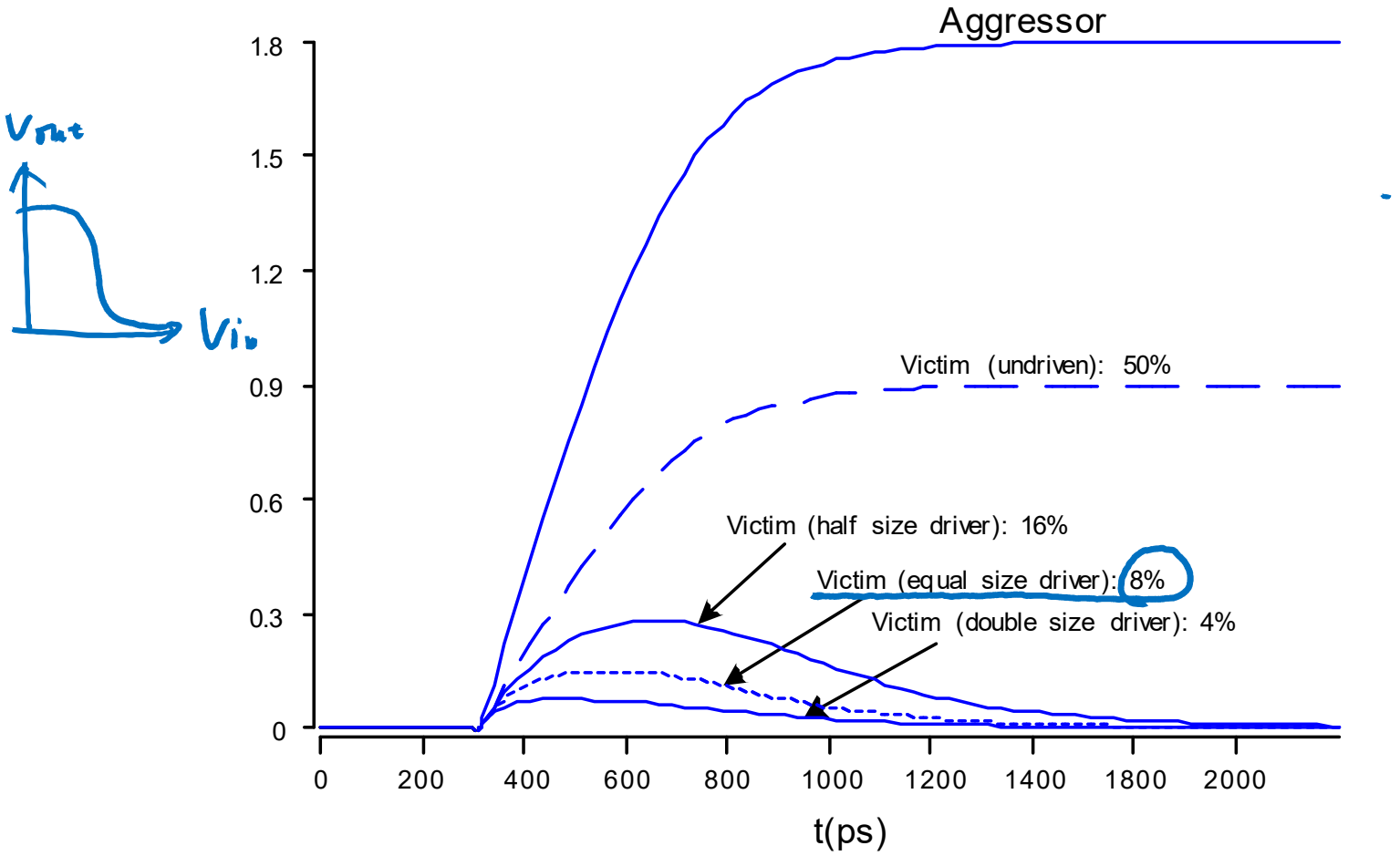
$\delta V_c = \delta V_a \frac{C_{adj}}{C_{gnd} + C_{adj}} \frac{1}{1+k}$

$K = \frac{\tau_{aggressor}}{\tau_{victim}}$ ↓ aggressor 弱



Coupling Waveforms

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



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 (floating node)
- Memories and other sensitive circuits also can produce wrong outputs

Outline

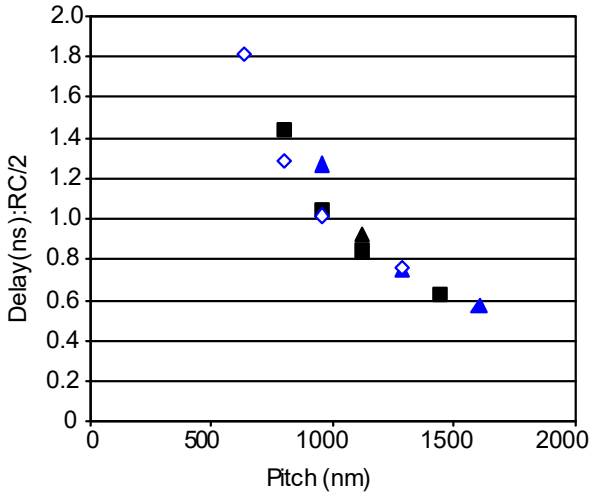
- Delay estimation
- Logical effort and transistor sizing
- Power dissipation
- Interconnect
- **Wire engineering** ←
- Design margin
- Reliability
- Scaling

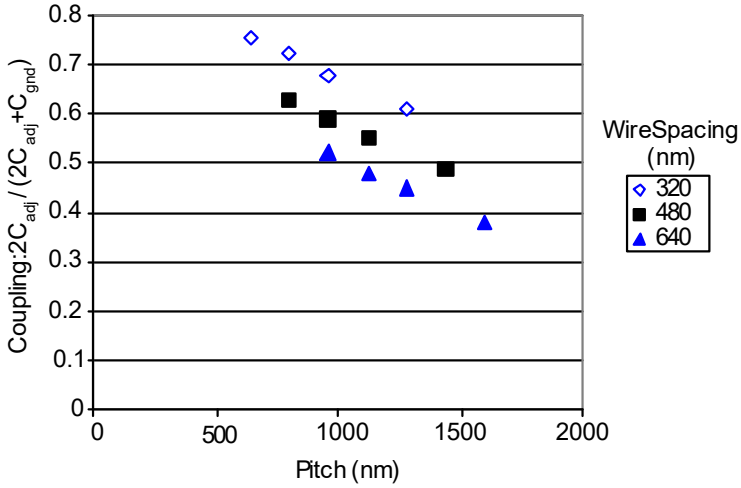
Wire Engineering

- Goal: to achieve delay, area, and power goals with acceptable noise
- Degrees of freedom:

Wire Engineering

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

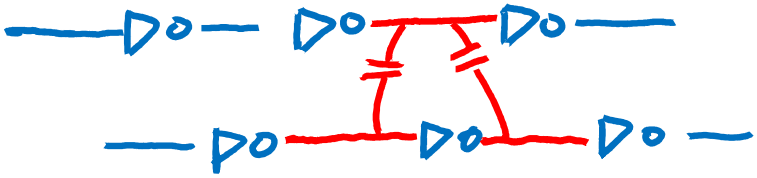


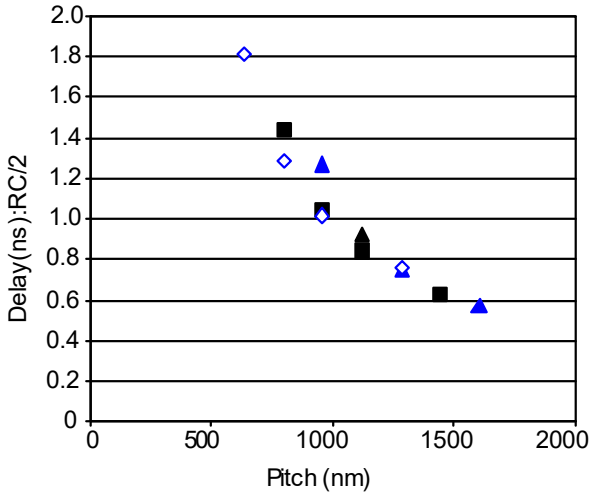


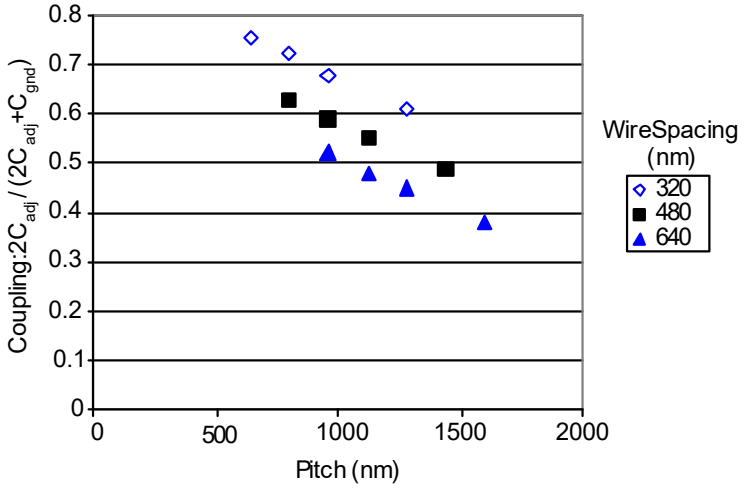
Wire Engineering

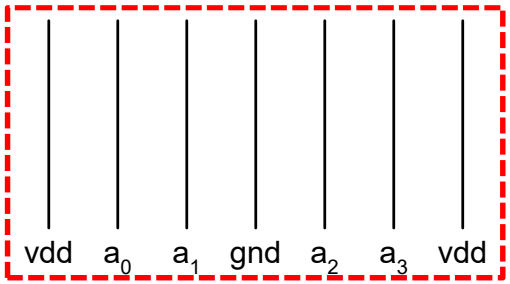
Staggered repeater

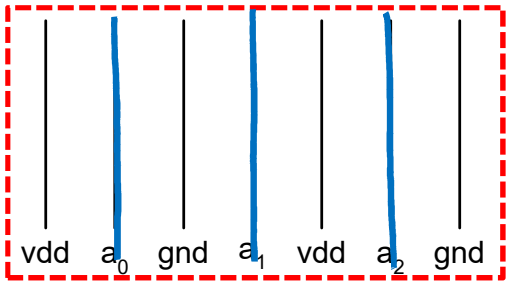
- Goal: to achieve delay, area, and power goals with acceptable noise
- Degrees of freedom:
  - Width
  - Spacing
  - Layer
  - Shielding

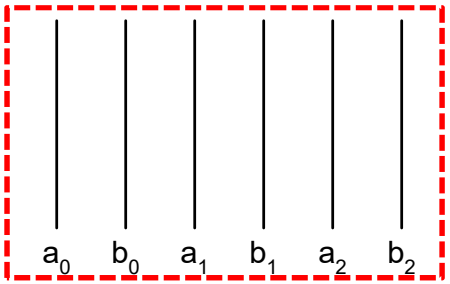






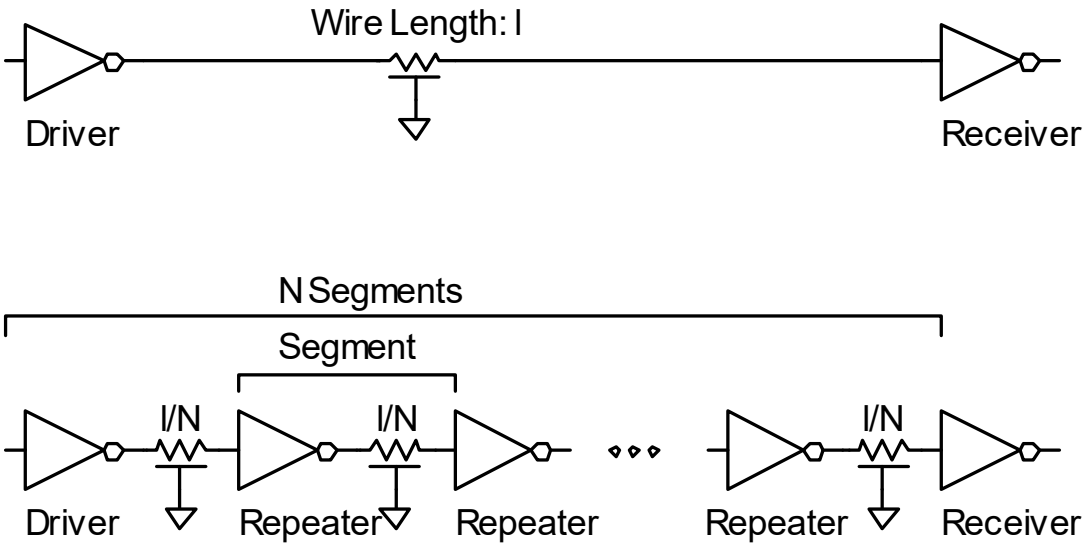






Repeaters

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



Repeater Design

1. How many repeaters should we use?
2. How large should each one be?

- Equivalent Circuit
  - Wire length l
    - Wire Capacitance $C_w \cdot l$, Resistance $R_w \cdot l$

cap/length

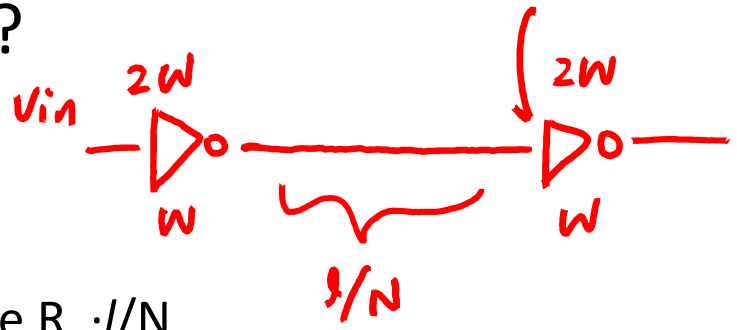
res/length

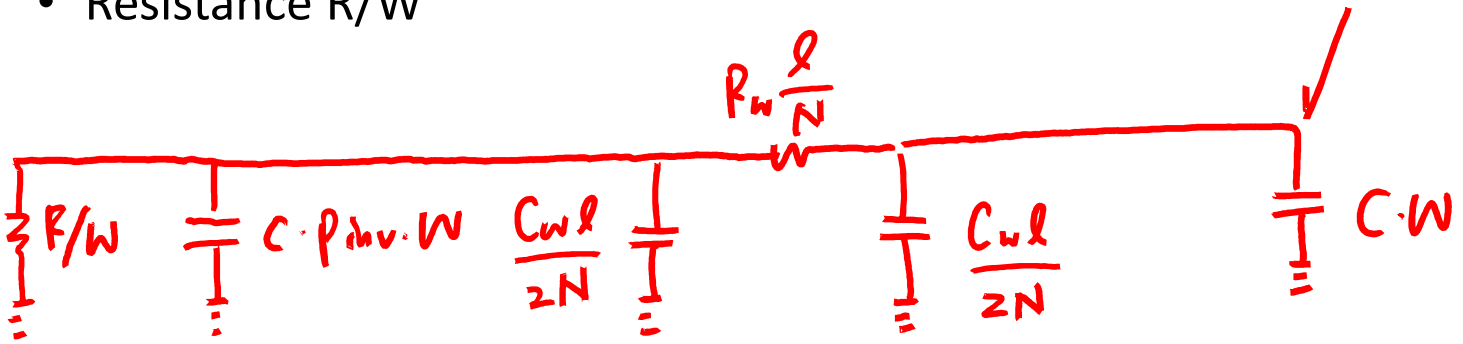
  - Inverter size W (nMOS = W, pMOS = 2W)
    - Gate capacitance $C \cdot W$, parasitic capacitance $C \cdot p_{inv} \cdot W$
    - Resistance R/W

(drain)

Repeater Design

- How many repeaters should we use?
- How large should each one be?
- Equivalent Circuit
  - Wire length *l/N*
    - Wire Capacitance $C_w \cdot l/N$, Resistance $R_w \cdot l/N$
  - Inverter size W (nMOS = W, pMOS = 2W)
    - Gate capacitance $C \cdot W$, parasitic capacitance $C \cdot p_{inv} \cdot W$
    - Resistance R/W





Repeater Results

* Elmore Delay

$$t_{pd} = N\left(\frac{R}{W}\left(C_w\frac{l}{N} + CW(1+p_{inv})\right) + R_w\frac{l}{N}\left(\frac{C_w l}{2N} + CW\right)\right)$$

$$= f(N, W)$$

- Differentiate wrt W and N and find out optimum

$$\frac{l}{N} = \sqrt{\frac{2RC(1+p_{inv})}{R_w C_w}}$$

with $FO4 = 5RC$ (implied $p_{inv}=1$)

with folded transistor $p_{inv}=0.5$

$$0.77\sqrt{\frac{FO4}{R_w C_w}}$$

$$\frac{t_{pd}}{l} = \left(2+\sqrt{2(1+p_{inv})}\right)\sqrt{RCR_wC_w} \leq 1.67\sqrt{FO4\,R_wC_w}$$

$$W = \sqrt{\frac{R \cdot C_w}{R_w C}}$$

$$\frac{E}{\ell} = C_w \left(1 + \sqrt{\frac{1+P_{inv}}{2}}\right) \cdot V_{DD}^2$$

$$\cong 1.8\eta \cdot C_w V_{DD}^2$$

≒38% power compared to no repeaters

★ in addition to min(tpd), 也可以 min(EDP) energy-delay product

→ 0.8mm 左右一個 W=11μm 的 inverter

$tpd/\ell = 47\,ps/mm$ $E/\ell = 0.26\,pJ/mm$

Example

- 65-nm technology, middle routing layer with 2x width, spacing, and height.
  - Rw=200 ohm/mm, Cw= 0.2 pF/mm
  - FO4=15 ps

$$\frac{t_{pd}}{\ell} = 1.69\sqrt{15p \cdot 200 \cdot 0.2p} \cong 41\ ps/mm$$

$$\frac{E}{\ell} = 0.4\ pJ/mm$$

$$\frac{\ell}{N} = 0.45mm$$ 尺一個 W=18μm (pmos=36μm) 的 inverter

in addition to min(t_{pd})

也可以 min (EDP) energy-delay product

可得 $\frac{t_{pd}}{\ell} = 47\ ps/mm$ 而 $\frac{E}{\ell} = 0.26\ pJ/mm$

合是 0.8 mm 才放一個 $W = 11\ \mu m$ 的 inverter