Introduction to CMOS VLSI Design

Chapter 5
Power

Outline

- Power and Energy
- Dynamic Power — công suất khi làm việc
- Static Power — công suất khi không hoạt động
  vd: tivi tắt nhưng vẫn cấp nguồn

Power and Energy

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

tức thời

- Instantaneous Power: $P(t) =$ I(t) * V(t)

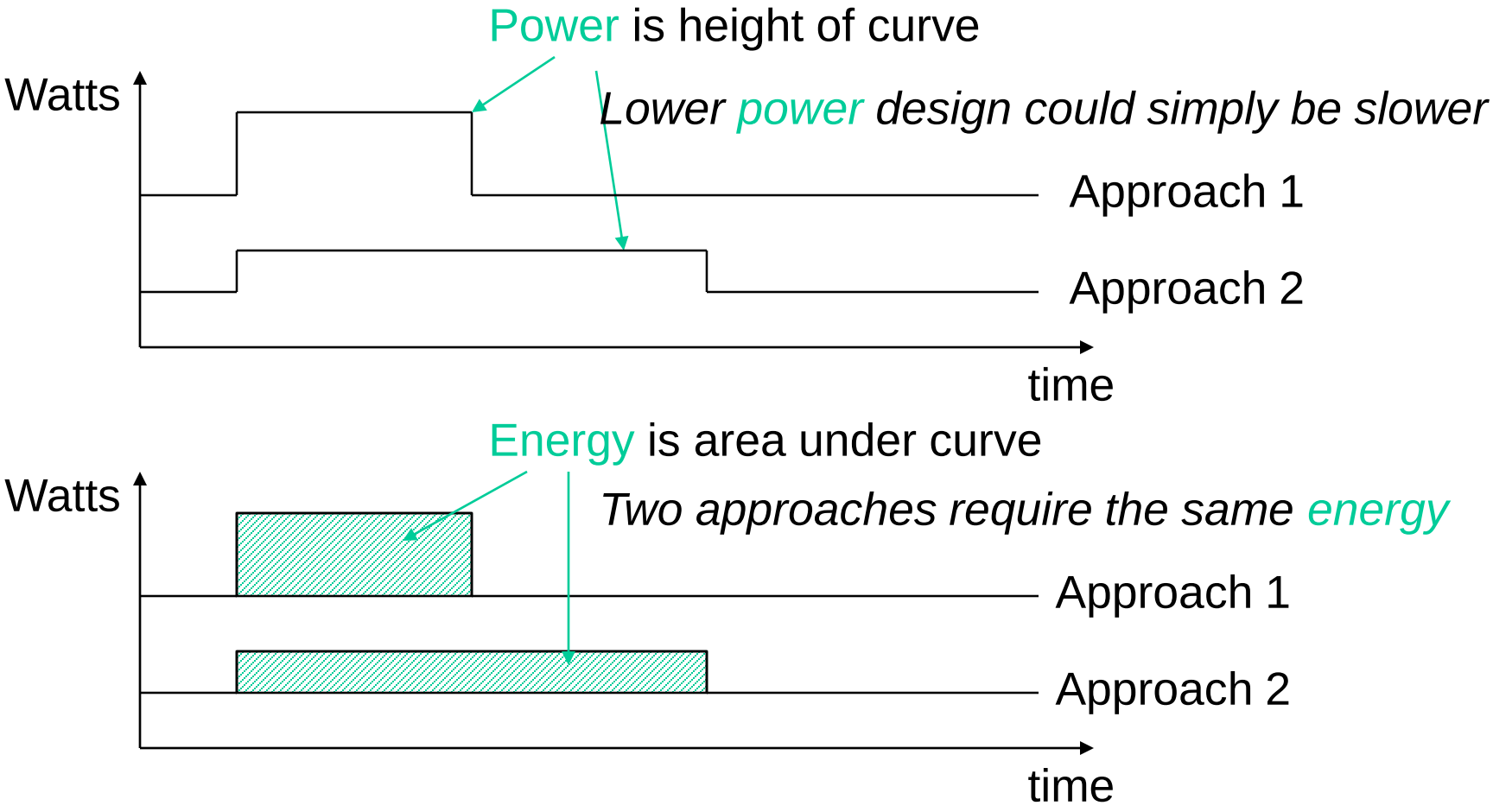
tổng công suất theo thời gian liên tục

- Energy: $E = \int_0^T P(t)\,dt$

tỉ số giữa năng lượng và thời gian

- Average Power: $P_{avg} = \frac{E}{T} = \frac{1}{T}\int_0^T P(t)\,dt$

Power versus Energy



tiêu tán

Power Dissipation Sources

tổng công suất tiêu thụ

- $P_{total} = P_{dynamic} + P_{static}$
- Dynamic power: $P_{dynamic} = P_{switching} + P_{shortcircuit}$
  - Switching load capacitances
  - Short-circuit (crowbar) current
- Static power: $P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention})V_{DD}$

ngưỡng dưới

  - Subthreshold leakage
  - Gate leakage

mối nối

  - Junction leakage
  - Contention current

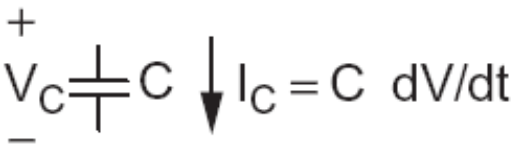
Power in Circuit Elements

$$P_{VDD}(t) = I_{DD}(t) V_{DD}$$

$$P_R(t) = \frac{V_R^2(t)}{R} = I_R^2(t) R$$

$$E_C = \int_0^{\infty} I(t) V(t)\, dt = \int_0^{\infty} C \frac{dV}{dt} V(t)\, dt$$

$$= C \int_0^{V_C} V(t)\, dV = \tfrac{1}{2} C V_C^2$$



Charging a Capacitor

lên mức 1

- When the gate output rises
  - Energy stored in capacitor is

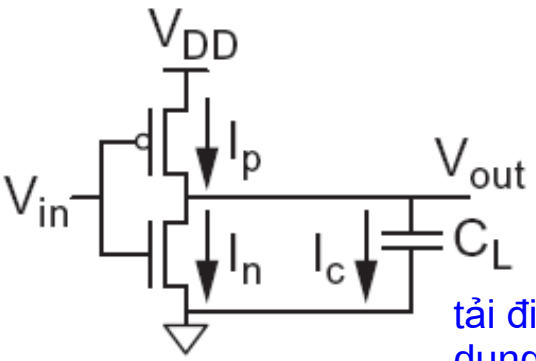
$$E_C = \frac{1}{2} C_L V_{DD}^2$$

  - But energy drawn from the supply is

$$E_{VDD} = \int_0^\infty I(t) V_{DD} dt = \int_0^\infty C_L \frac{dV}{dt} V_{DD} dt$$

$$= C_L V_{DD} \int_0^{V_{DD}} dV = C_L V_{DD}^2$$

  - Half the energy from V_{DD} is dissipated in the pMOS transistor as heat, other half stored in capacitor



tải điện dung CL

- When the gate output falls

dạng sóng đồ thị đổ

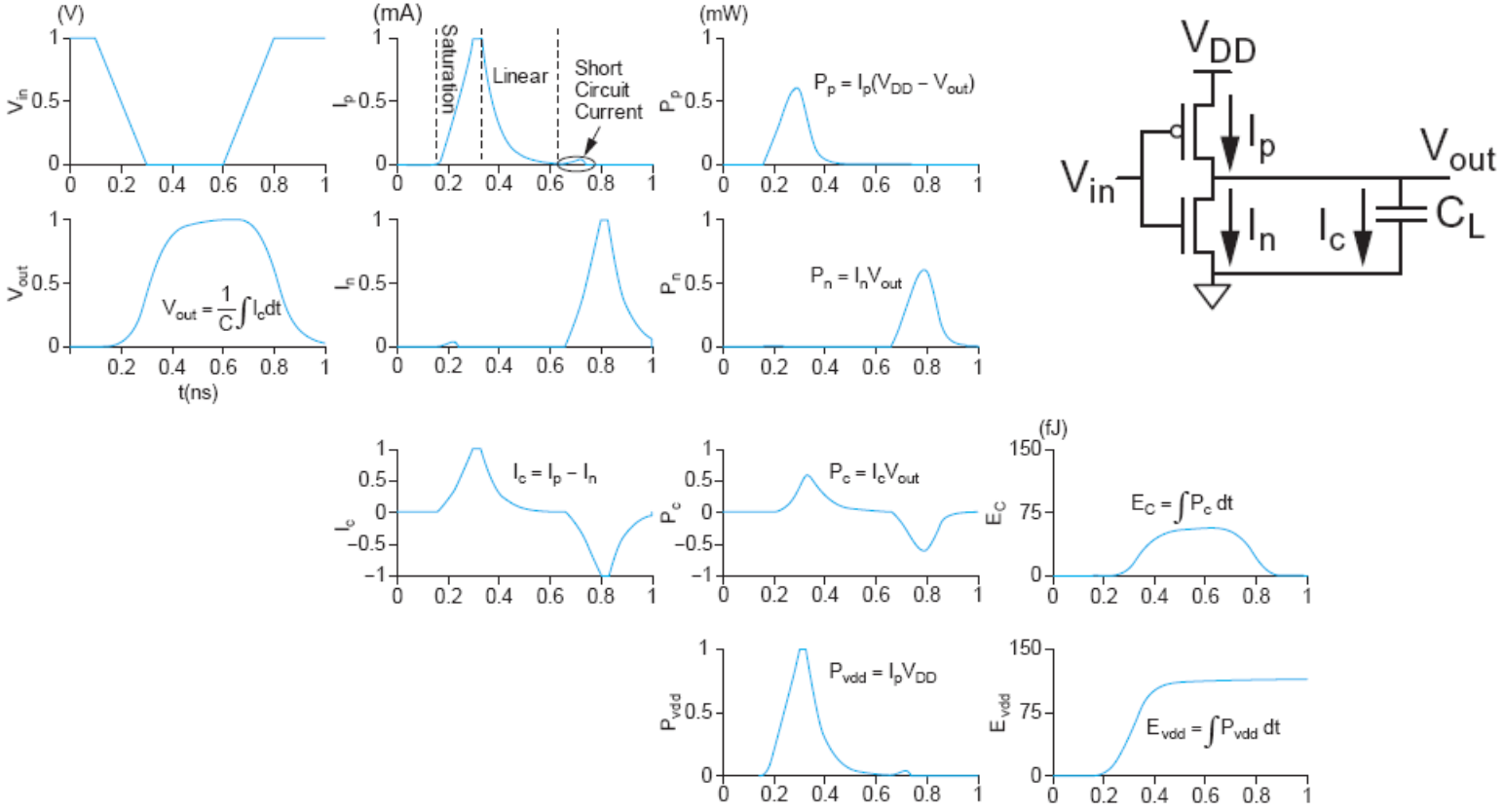
  - Energy in capacitor is dumped to GND
  - Dissipated as heat in the nMOS transistor

dưới dạng nhiệt

Switching Waveforms

10^-15 10^9

- Example: V_{DD} = 1.0 V, C_L = 150 fF, f = 1 GHz



Switching Power

$$P_{\text{switching}} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}\,dt$$

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

$$= \frac{V_{DD}}{T}\left[Tf_{\text{sw}}CV_{DD}\right]$$

$$= CV_{DD}{}^2 f_{\text{sw}}$$

Over T, $T * f_{sw}$ times cap C charges

$$\int_0^T iDD(t)dt$$

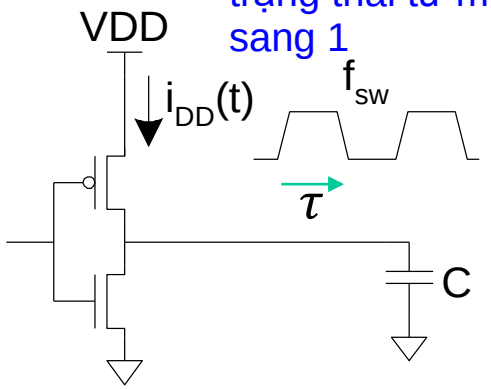
$$= \int_0^\tau C\frac{dV}{dt}dt * (T * f_{sw})$$

$$= \int_0^{VDD} CdV * (T * f_{sw})$$

fsw: the gate switches at some average freq

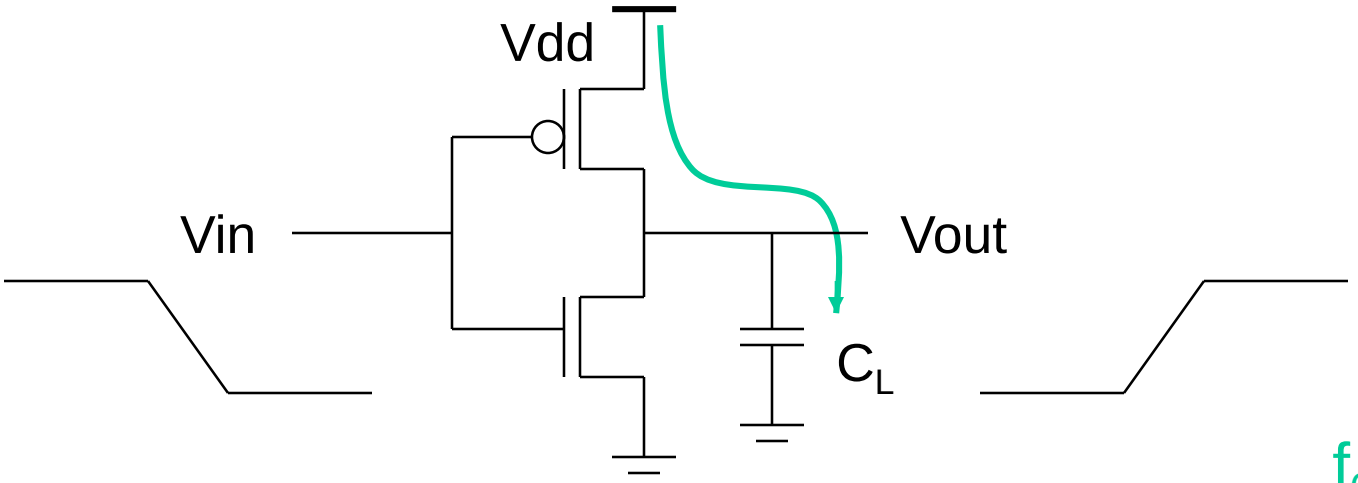
Over T, T*fsw times gate switches

tần suất chuyển mức trạng thái từ mức 0 sang 1



điện dung ngõ vào

Dynamic Power



ko khả thi

Energy/transition = $C_L * V_{DD}^2 * P_{0\to1}$

$f_{0\to1}$

tần số hoạt động của mạch

giảm dc

P_{dyn} = Energy/transition * f = $C_L * V_{DD}^2$ * $P_{0\to1} * f$

xác suất mạch chuyển đổi trạng thái từ logic 0 sang 1

Data dependent - a function of switching activity!

Activity Factor

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where α = activity factor (prob. that Circuit 0→1)
  - If the signal is a clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = ½$

- Dynamic power:

$$P_{\text{switching}} = \alpha C V_{DD}^{\ 2} f$$

Reducing Dynamic Power

Capacitance:
Function of fan-out, wire length, transistor sizes

điện dung kí sinh, layout, đi dây, kích thước transistor, giảm quy trình

Supply Voltage:
Has been dropping with successive generations

$$P_{dyn} = C_L V_{DD}^2 P_{0\rightarrow1} f$$

Activity factor:
How often, on average, do wires switch?

phụ thuộc chuỗi dữ liệu ngõ vào

Clock frequency:
Increasing…

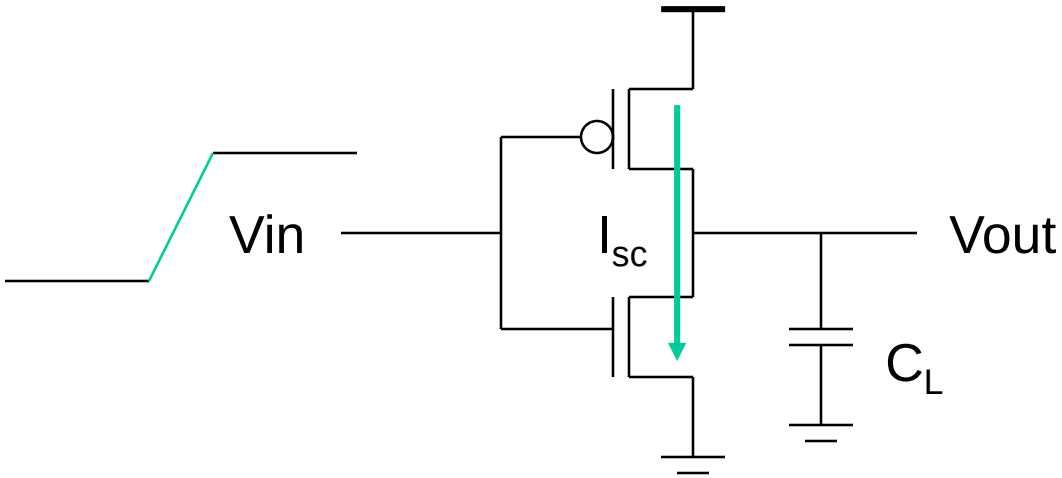
Low V-Swing Signaling

- Driver power
- Receiver power
- Latching receivers

Short Circuit (Crowbar) Current

- When transistors switch, both nMOS and pMOS networks may be momentarily ON at once ngay lập tức
- Leads to a blip of “short circuit” current.
- < 10% of dynamic power if rise/fall times are comparable for input and output
- We will generally ignore this component.
  It is included in circuit simulation

Short Circuit Power



hữu hạn

Finite slope of the input signal causes a direct current path between V_{DD} and GND for a short period of time during switching when both the NMOS and PMOS transistors are conducting.

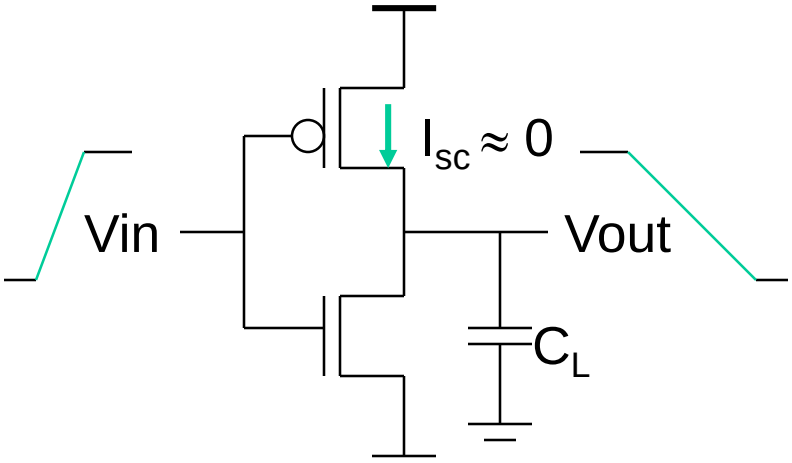
Short Circuit Currents

$$E_{sc} = t_{sc} V_{DD} I_{peak} P_{0\rightarrow1}$$

$$P_{sc} = t_{sc} V_{DD} I_{peak} f_{0\rightarrow1}$$

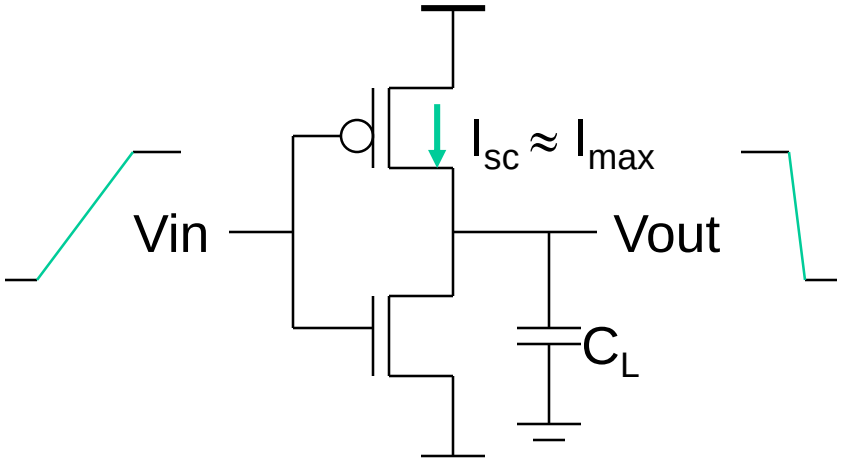
- Duration and slope of the input signal, t_{sc}
- I_{peak} determined by
  - the saturation current of the P and N transistors which depend on their sizes, process technology, temperature, etc.
  - strong function of the ratio between input and output slopes
    - a function of C_L

Impact of C_L on P_{sc} in next Stage



Large capacitive load

Output fall time significantly larger than input rise time.



Small capacitive load

Output fall time substantially smaller than the input rise time.

Dynamic Power Example 5.1

$\lambda = \frac{\text{process}}{2} = \frac{50\text{nm}}{2} = 25\text{nm}$

- 1 billion transistor chip
  - 50M logic transistors
    - Average width: 12 λ
    - Activity factor = 0.1
  - 950M memory transistors
    - Average width: 4 λ
    - Activity factor = 0.02
  - 1.0 V 50 nm process
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)
- Estimate dynamic power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current.

Solution

Clogic = 50*10^6 * W * (Cg + Cd) Cmem = 950*10^6 * W * (Cg + Cd)

$$C_{\text{logic}} = \left(50\times10^{6}\right)\left(12\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8\,fF/\mu m\right) = 27\ \text{nF}$$

$$C_{\text{mem}} = \left(950\times10^{6}\right)\left(4\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8\,fF/\mu m\right) = 171\ \text{nF}$$

$$P_{\text{dynamic}} = \left[0.1C_{\text{logic}} + 0.02C_{\text{mem}}\right]\left(1.0\right)^{2}\left(1.0\ \text{GHz}\right) = 6.1\ \text{W}$$

Pdyn = Pdyn(logic) + Pdyn(mem) = $\alpha_l \cdot V^2 \cdot C_l \cdot f + \alpha_m \cdot V^2 \cdot C_m \cdot f$

Dynamic Power Reduction

- $P_{\text{switching}} = \alpha C V_{DD}^{\ 2} f$
- Try to minimize:
  - Activity factor
  - Capacitance
  - Supply voltage
  - Voltage swing
  - Frequency

Activity Factor Estimation

- Let P_i = Prob(node i = 1)
  - $\bar{P}_i = 1-P_i$
- $\alpha_i = \bar{P}_i * P_i$
- Completely random data has P = 0.5 and α = 0.25
- Data is often not completely random
  - e.g. upper bits of 64-bit words representing bank account balances are usually 0
- Data propagating through ANDs and ORs has lower activity factor
  - Depends on design, but typically $\alpha \approx 0.1$

xác suất

Gates and Probability

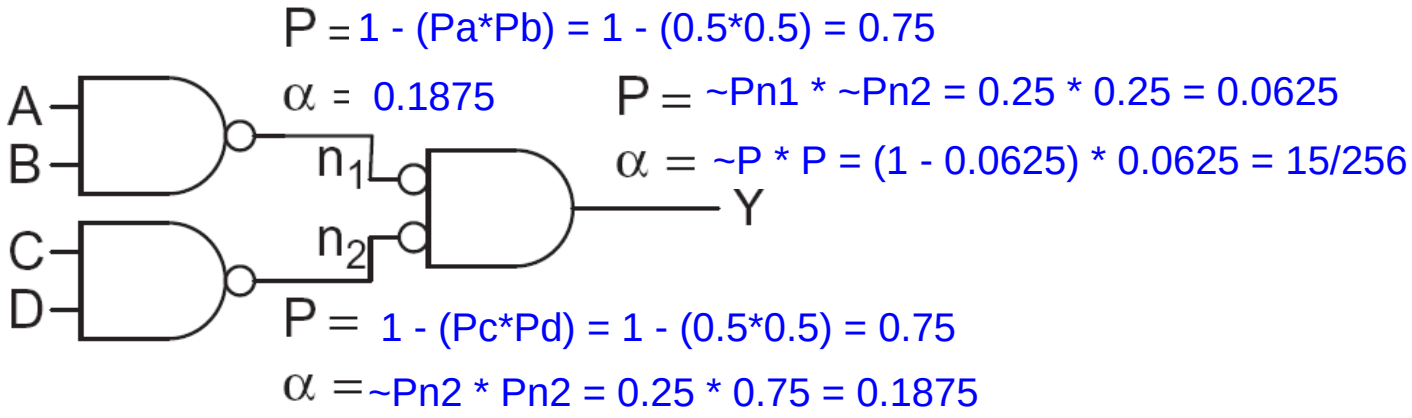
Gate	P_Y
AND2	$P_A P_B$
AND3	$P_A P_B P_C$
OR2	$1 - \overline{P}_A \overline{P}_B$
NAND2	$1 - P_A P_B$
NOR2	$\overline{P}_A \overline{P}_B$
XOR2	$P_A \overline{P}_B + \overline{P}_A P_B$

PA: prob. that input A = 1, PB: prob. that input B = 1

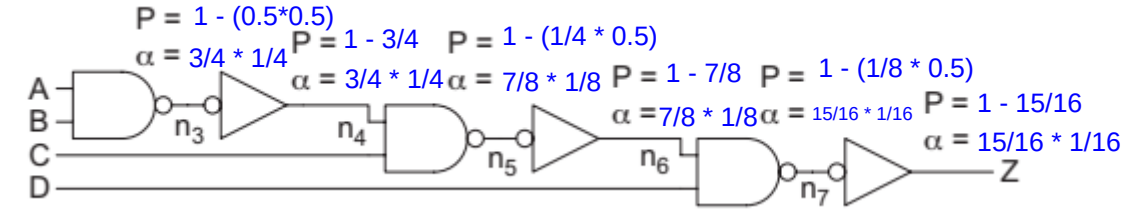
$\alpha_Y = \overline{P_Y} P_Y$ where $\overline{P_Y} = 1 - P_Y$

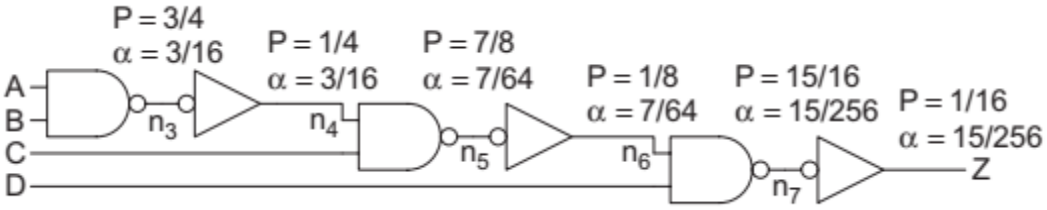
Example 5.2 (a)

- A 4-input AND is built out of two levels of gates
- Estimate the activity factor at each node if the inputs have P = 0.5 and inputs are uncorrelated with each other and in time:



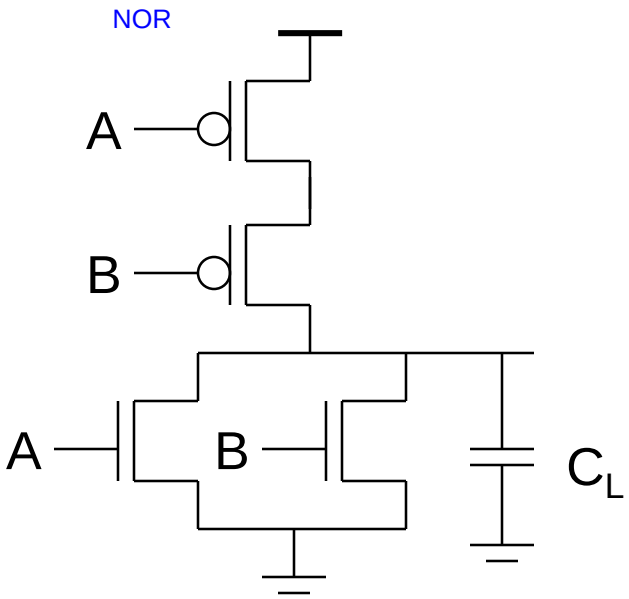
Example 5.2 (b)

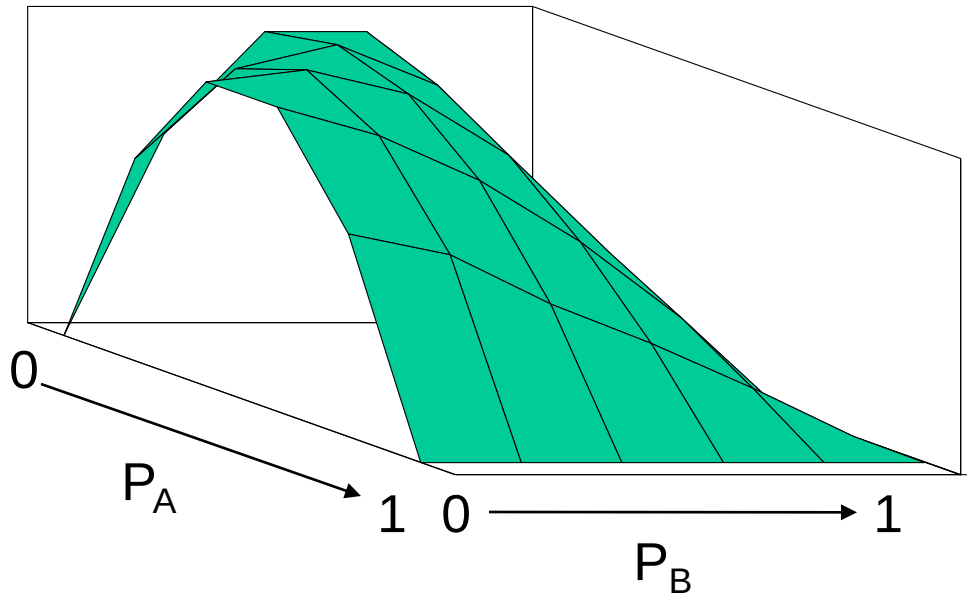

P = 1 - (0.5*0.5)
α = 3/4 * 1/4
P = 1 - 3/4
α = 3/4 * 1/4
P = 1 - (1/4 * 0.5)
α = 7/8 * 1/8
P = 1 - 7/8
α = 7/8 * 1/8
P = 1 - (1/8 * 0.5)
α = 15/16 * 1/16
P = 1 - 15/16
α = 15/16 * 1/16
A
B
C
D
n3
n4
n5
n6
n7
Z


P = 3/4
α = 3/16
P = 1/4
α = 3/16
P = 7/8
α = 7/64
P = 1/8
α = 7/64
P = 15/16
α = 15/256
P = 1/16
α = 15/256
A
B
C
D
n3
n4
n5
n6
n7
Z

Transition Probabilities

- Switching activity is a strong function of the input signal statistics
  - P_A and P_B are the probabilities that inputs A and B are one

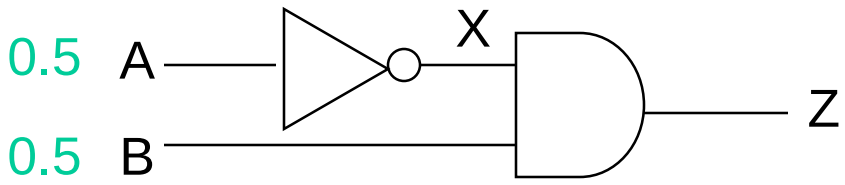




$$P_{0\rightarrow1} = P_0 \times P_1 = (1-(1-P_A)(1-P_B))*(1-P_A)(1-P_B)$$

Transition Probabilities

	$P_{0\rightarrow1} = P_{out=0} \times P_{out=1}$
NOR	$(1 - (1 - P_A)(1 - P_B)) \times (1 - P_A)(1 - P_B)$
OR	$(1 - P_A)(1 - P_B) \times (1 - (1 - P_A)(1 - P_B))$
NAND	$P_A P_B \times (1 - P_A P_B)$
AND	$(1 - P_A P_B) \times P_A P_B$
XOR	$(1 - (P_A + P_B - 2P_A P_B)) \times (P_A + P_B - 2P_A P_B)$



For X: $P_{0\rightarrow1} = P_0 \times P_1 = (1-P_A)\,P_A = 0.5 \times 0.5 = 0.25$

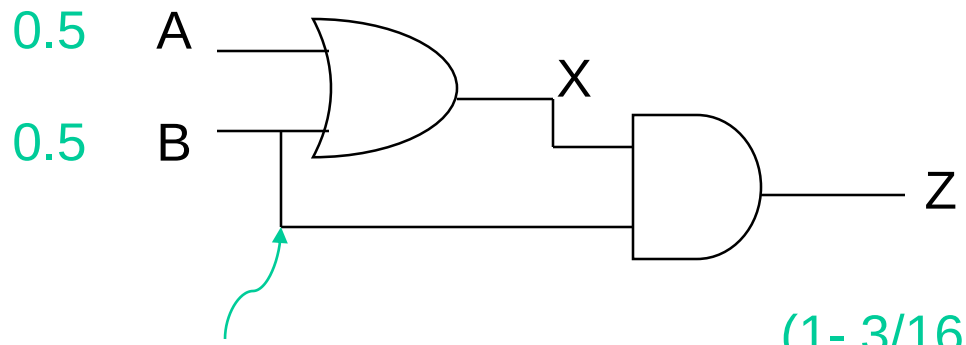
For Z: $P_{0\rightarrow1} = P_0 \times P_1 = (1-P_X P_B)\,P_X P_B = (1 - (0.5 \times 0.5)) \times (0.5 \times 0.5) = 3/16$

mối tương quan

Inter-signal Correlations

- Determining switching activity is complicated by the fact that signals exhibit correlation in space and time
  - reconvergent fan-out

(1-0.5)(1-0.5)x(1-(1-0.5)(1-0.5)) = 3/16



(1- 3/16 x 0.5) x (3/16 x 0.5) = 0.085

Reconvergent

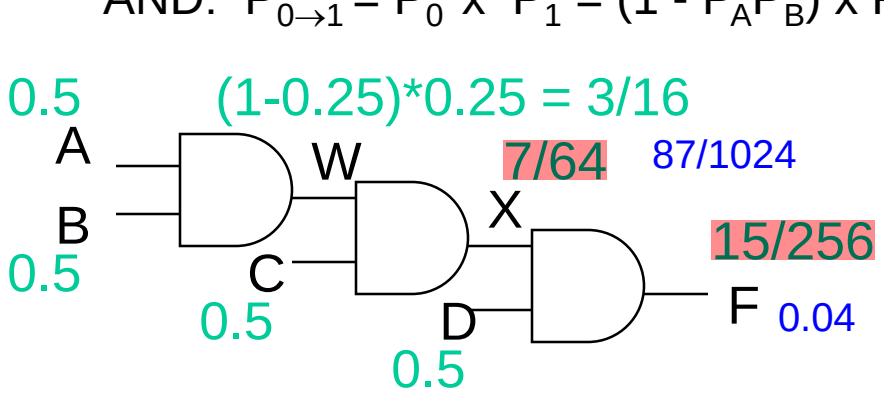
is P(Z=1) = P(B=1) & P(A=1 | B=1) ? What is Z?

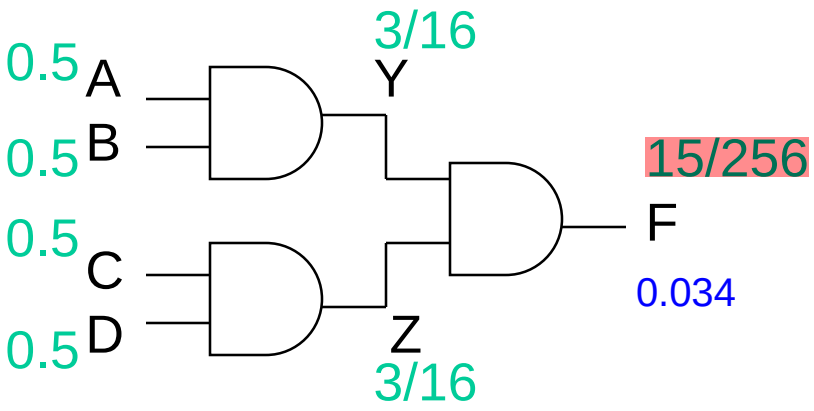
- Have to use conditional probabilities

Logic Restructuring

- Logic restructuring: changing the topology of a logic network to reduce transitions

AND: $P_{0\to1} = P_0 \times P_1 = (1 - P_A P_B) \times P_A P_B$



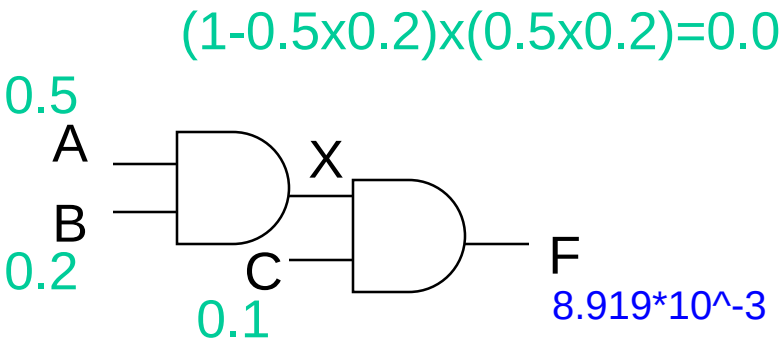


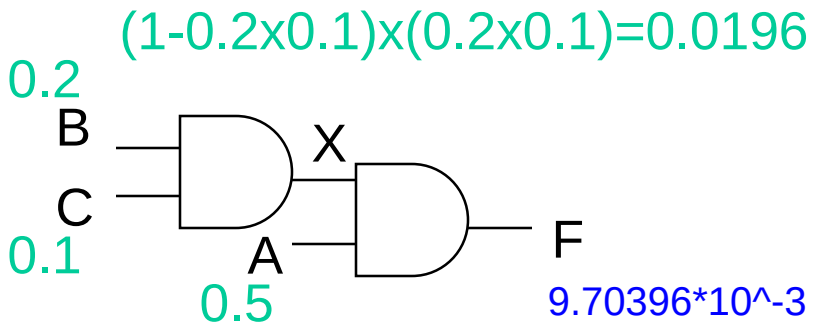
Chain implementation has a lower overall switching activity than the tree implementation for random inputs

Ignores glitching effects

trục trặc

Input Ordering



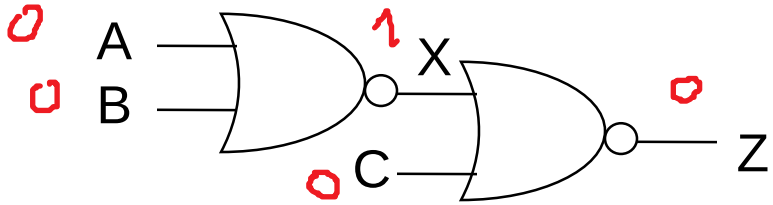


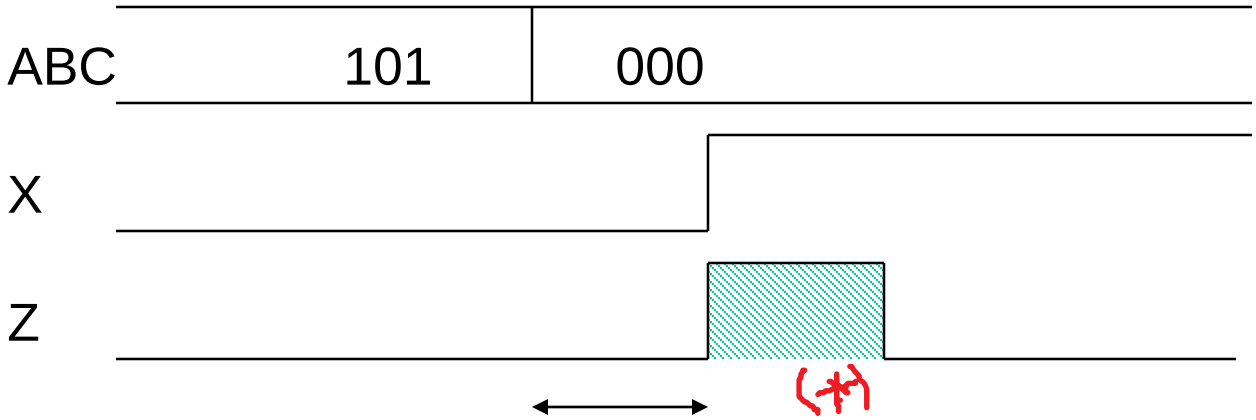
Beneficial to postpone the introduction of signals with a high transition rate (signals with signal probability close to 0.5)

Glitching

giả mạo

- Gates have a nonzero propagation delay resulting in spurious transitions or glitches (dynamic hazards) hiểm nguy
  - glitch: node exhibits multiple transitions in a single cycle before settling to the correct logic value

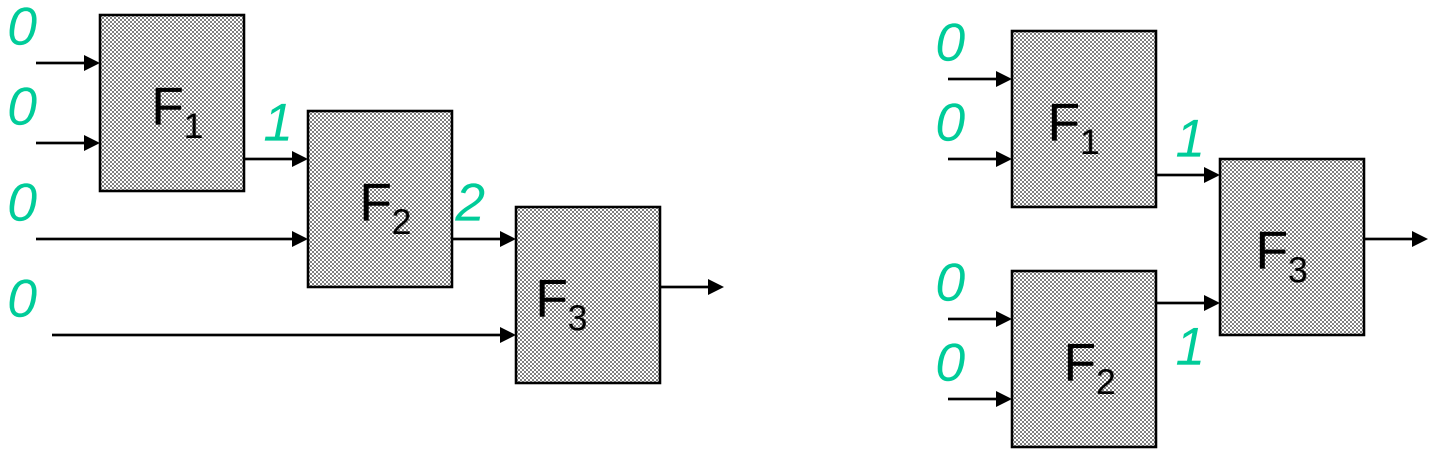




Unit Delay

Balanced Delay Paths

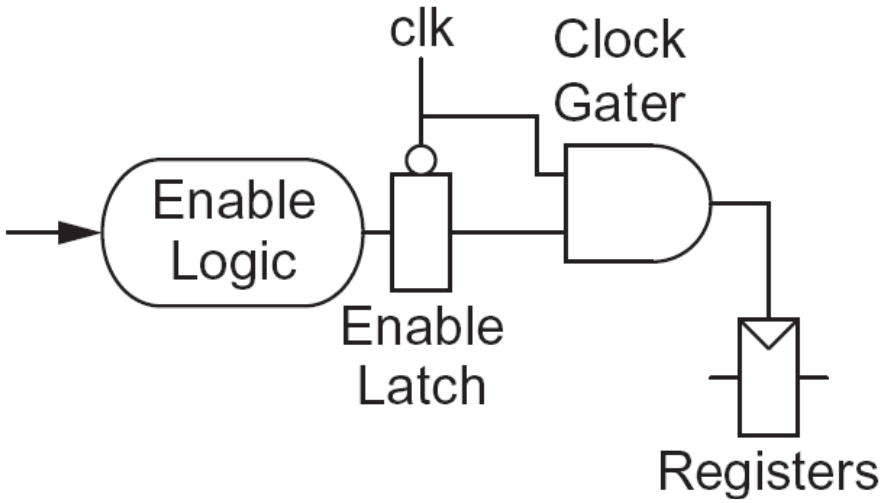
- Glitching is due to a mismatch in the path lengths in the logic network; if all input signals of a gate change simultaneously, no glitching occurs



So equalize the lengths of timing paths through logic

Clock Gating

- The best way to reduce the activity is to turn off the clock to registers in unused blocks
  - Saves clock activity ($\alpha = 1$)
  - Eliminates all switching activity in the block
  - Requires determining if block will be used

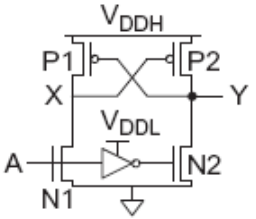


Capacitance

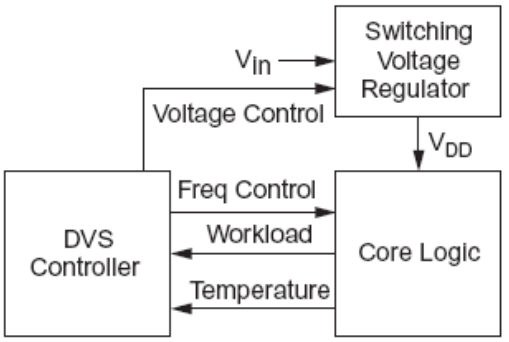
- Gate capacitance
  - Fewer stages of logic
  - Small gate sizes
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other
  - Drive long wires with inverters or buffers rather than complex gates

Voltage / Frequency

- Run each block at the lowest possible voltage and frequency that meets performance requirements
- Voltage Domains miền
  - Provide separate supplies to different blocks
  - Level converters required when crossing from low to high V_{DD} domains



- Dynamic Voltage Scaling
  - Adjust V_{DD} and f according to workload



Multiple V_{DD} Domains

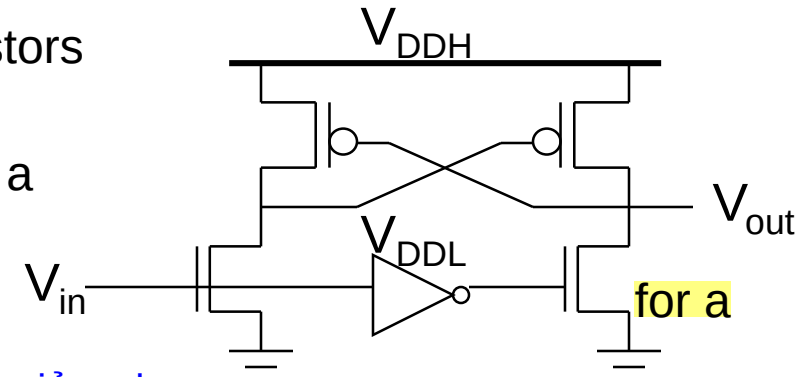
- IO / core
  - 2.5V or 3.3V IO to interface with chips using existing standards
  - 1.2V to 0.8V core for small geometry transistors
  - Need 2 thicknesses of gate oxides
- Still lower power, low speed circuits in the core
- How to get signals between V_{DD} domains?

hình học

- Multiple core supplies not as popular as multiple V_T masks
  - High V_T for low power (low leakage, low crowbar)
  - Low V_T for high speed (high current)

Multiple V_{DD}

- How many V_{DD}? – Two is becoming common
  - Many chips already have two supplies (one for core and one for I/O)
- When combining multiple supplies, level converters are required whenever a module at the lower supply drives a gate at the higher supply (step-up)
  - If a gate supplied with V_{DDL} drives a gate at V_{DDH}, the PMOS never turns off
    - The cross-coupled PMOS transistors do the level conversion
    - The NMOS transistor operate on a reduced supply
  - Level converters are not needed step-down change in voltage



giảm nhẹ

for a

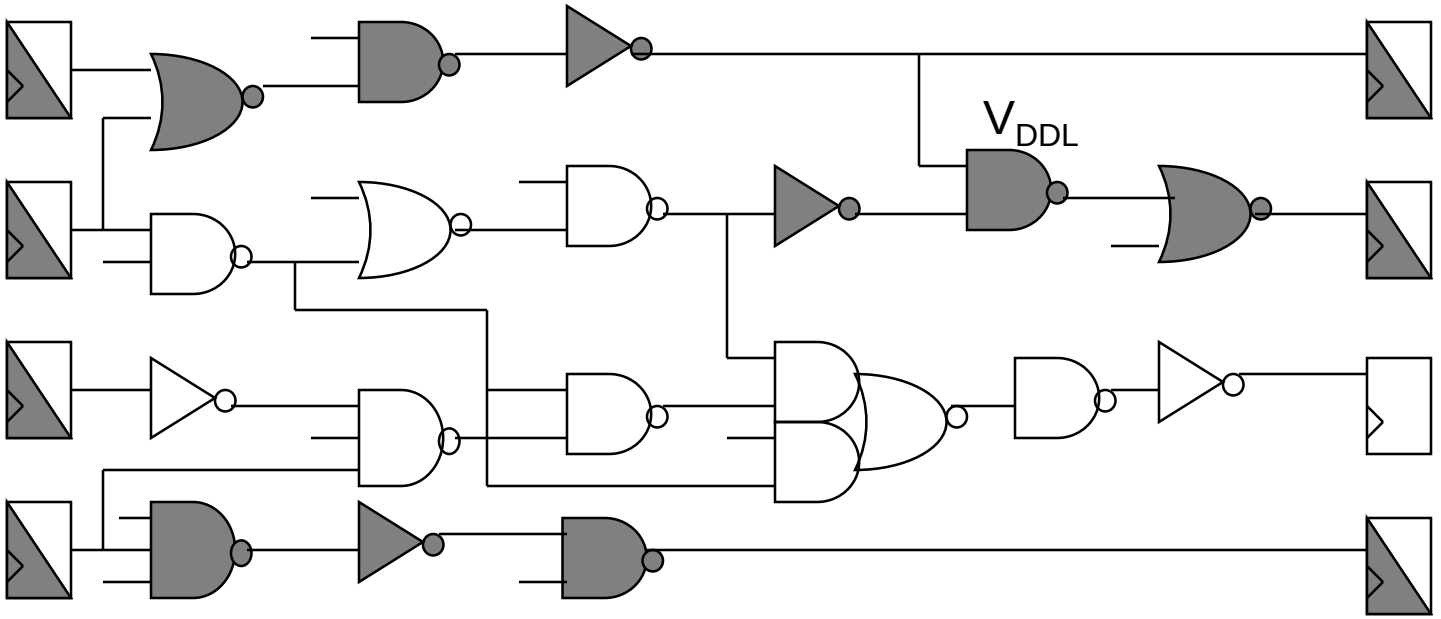
  - Overhead of level converters can be mitigated by doing conversions at register boundaries and embedding the level conversion inside the flipflop

ranh giới

Dual-Supplies

quan trọng

- Minimum energy consumption is achieved if all logic paths are critical (have the same delay)
- Clustered voltage-scaling (thành cụm)
  - Each path starts with V_{DDH} and switches to V_{DDL} (gray logic gates) when delay slack (chùng xuống) is available
  - Level conversion is done in the flipflops at the end of the paths



Static Power

- Static power is consumed even when chip is quiescent.
  - Leakage draws power from nominally OFF devices
  - Ratioed circuits burn power in fight between ON transistors

ngay cả

không hoạt động

trên danh nghĩa

tỷ lệ

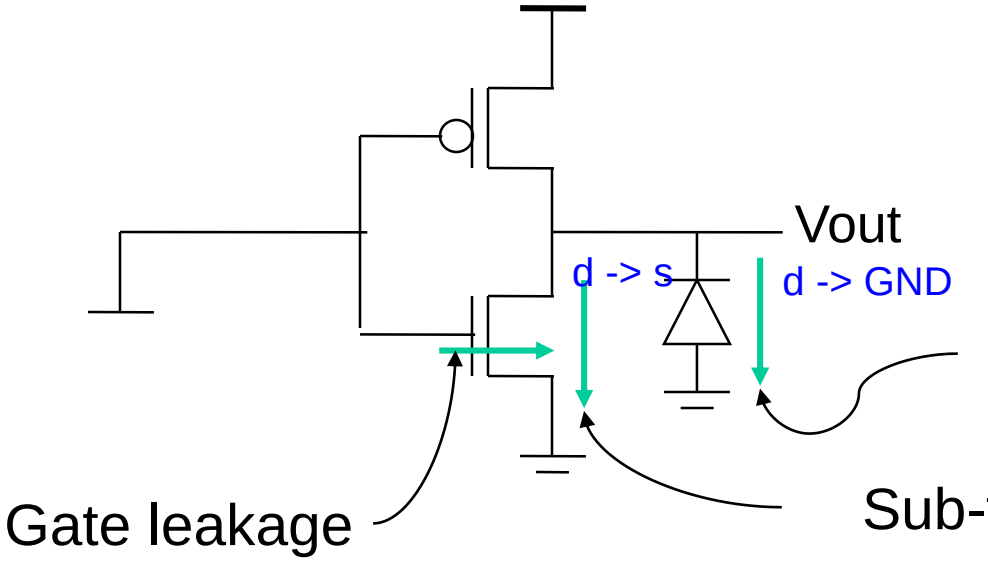
về trạng thái standby đen thui, nhưng vẫn còn đèn sáng bên trong vì vẫn được cấp nguồn, có những dòng rò đi qua linh kiện điện tử, vẫn có tiêu thụ công suất nhất định gọi là công suất tĩnh

những cực của transistor có những dòng rò

công suất rò

Leakage Power

tổng công suất rò gọi là công suất tĩnh



$V_{DD}\ I_{leakage}$

giao lộ

Drain junction leakage

Gate leakage

Sub-threshold current

cốt yếu

Sub-threshold current is the dominant factor.

slide 41

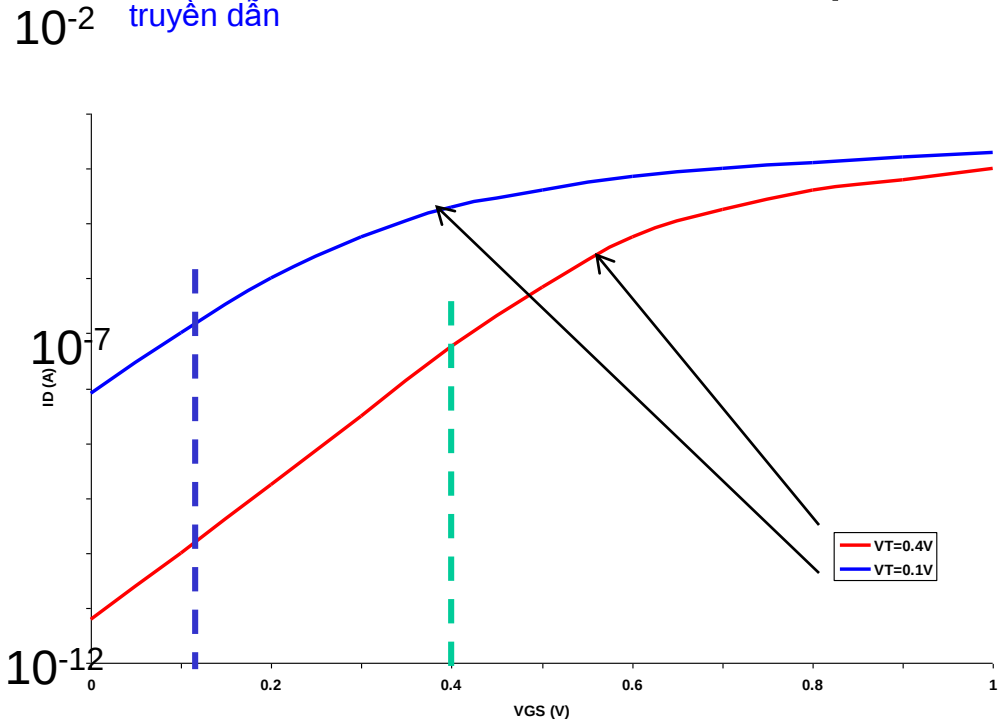
All increase exponentially with temperature!

Leakage and V_T

- Continued scaling of supply voltage and the subsequent scaling of threshold voltage will make subthreshold conduction a dominant component of power dissipation.

tiêu tán

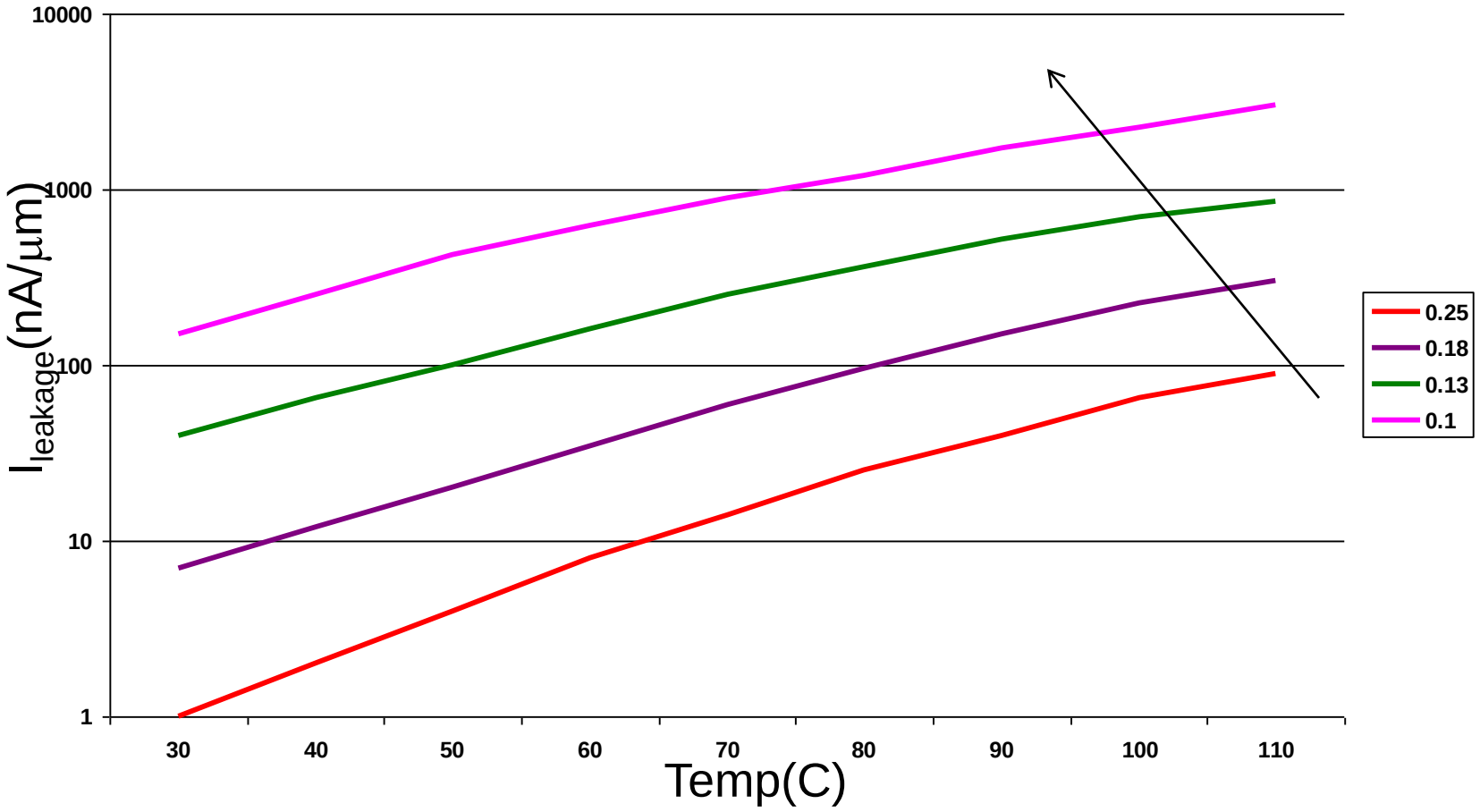
truyền dẫn



- Each 255mV increase in V_T gives 3 orders of magnitude reduction in leakage (but adversely affects performance)

xấu

Leakage Current Increase



Static Power Example

- Revisit power estimation for 1 billion transistor chip
- Estimate static power consumption
  - Subthreshold leakage
    - Normal V_t: 100 nA/μm 5%
    - High V_t: 10 nA/μm
    - High Vt used in all memories and in 95% of logic gates
  - Gate leakage 5 nA/μm
  - Junction leakage negligible không đáng kể

Solution

slide 18

5%

$$W_{\text{normal-V}_t} = \left(50\times10^6\right)\left(12\lambda\right)\left(0.025\mu\text{m}/\lambda\right)\left(0.05\right) = 0.75\times10^6\ \mu\text{m}$$

$$W_{\text{high-V}_t} = \left[\left(50\times10^6\right)\left(12\lambda\right)\left(0.95\right)+\left(950\times10^6\right)\left(4\lambda\right)\right]\left(0.025\mu\text{m}/\lambda\right) = 109.25\times10^6\ \mu\text{m}$$

$$I_{sub} = \left[W_{\text{normal-V}_t}\times100\ \text{nA}/\mu\text{m}+W_{\text{high-V}_t}\times10\ \text{nA}/\mu\text{m}\right]/2 = 584\ \text{mA}$$

$$I_{gate} = \left[\left(W_{\text{normal-V}_t}+W_{\text{high-V}_t}\right)\times5\ \text{nA}/\mu\text{m}\right]/2 = 275\ \text{mA}$$

$$\text{P}_{static} = \left(584\ \text{mA}+275\ \text{mA}\right)\left(1.0\ \text{V}\right) = 859\ \text{mW}$$

14% of 6.1W dynamic power

What if 90% idle?

Subthreshold Leakage

- For $V_{ds} > 50$ mV

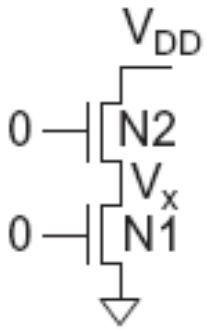
$$I_{sub} \approx I_{off} 10^{\frac{V_{gs} + \eta(V_{ds} - V_{DD}) - k_\gamma V_{sb}}{S}}$$

- I_{off} = leakage at $V_{gs} = 0$, $V_{ds} = V_{DD}$

Typical values in 65 nm
I_{off} = 100 nA/μm @ V_t = 0.3 V
I_{off} = 10 nA/μm @ V_t = 0.4 V
I_{off} = 1 nA/μm @ V_t = 0.5 V
η = 0.1
k_γ = 0.1
S = 100 mV/decade

Stack Effect

- Series OFF transistors have less leakage
  - $V_x > 0$, so N2 has negative V_{gs}



$$I_{sub} = \underbrace{I_{off} 10^{\frac{\eta(V_x - V_{DD})}{S}}}_{N2} = \underbrace{I_{off} 10^{\frac{-V_x + \eta((V_{DD} - V_x) - V_{DD}) - k_\gamma V_x}{S}}}_{N1}$$

$$V_x = \frac{\eta V_{DD}}{1 + 2\eta + k_\gamma}$$

$$I_{sub} = I_{off} 10^{\frac{-\eta V_{DD}\left(\frac{1+\eta+k_\gamma}{1+2\eta+k_\gamma}\right)}{S}} \approx I_{off} 10^{\frac{-\eta V_{DD}}{S}}$$

  - Leakage through 2-stack reduces ~10x
  - Leakage through 3-stack reduces further

Leakage Control

- Leakage and delay trade off
  - Aim for low leakage in sleep and low delay in active mode
- To reduce leakage:
  - Increase V_t: *multiple V_t*
    - Use low V_t only in critical circuits
  - Increase V_s: *stack effect*
    - *Input vector control* in sleep
  - Decrease V_b

đảo ngược khuynh hướng

    - *Reverse body bias* in sleep
    - Or forward body bias in active mode

Gate Leakage

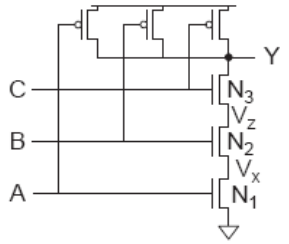
- Extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches subthreshold leakage at 65 nm and below in some processes
- An order of magnitude less for pMOS than nMOS
- Control leakage in the process using $t_{ox} > 10.5$ Å
  - High-k gate dielectrics help
  - Some processes provide multiple t_{ox}
    - e.g. thicker oxide for 3.3 V I/O transistors
- Control leakage in circuits by limiting V_{DD}

NAND3 Leakage Example

- 100 nm process

I_{gn} = 6.3 nA I_{gp} = 0

I_{offn} = 5.63 nA I_{offp} = 9.3 nA



Input State (ABC)	I_{sub}	I_{gate}	I_{total}	V_x	V_z
000	0.4	0	0.4	stack effect	stack effect
001	0.7	0	0.7	stack effect	$V_{DD} - V_t$
010	0	1.3	1.3	intermediate	intermediate
011	3.8	0	10.1	$V_{DD} - V_t$	$V_{DD} - V_t$
100	0.7	6.3	7.0	0	stack effect
101	3.8	6.3	10.1	0	$V_{DD} - V_t$
110	5.6	12.6	18.2	0	0
111	28	18.9	46.9	0	0

trung gian

Data from [Lee03]

Junction Leakage

- From reverse-biased p-n junctions
  - Between diffusion and substrate or well
- Ordinary diode leakage is negligible
- Band-to-band tunneling (BTBT) can be significant
  - Especially in high-V_t transistors where other leakage is small
  - Worst at $V_{db} = V_{DD}$

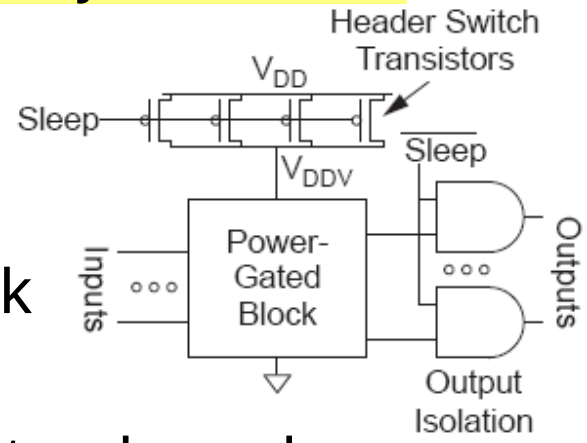
gây ra

làm trầm trọng thêm

- Gate-induced drain leakage (GIDL) exacerbates
  - Worst for $V_{gd} = -V_{DD}$ (or more negative)

Power Gating

- Turn OFF power to blocks when they are idle to save leakage
  - Use virtual V_{DD} (V_{DDV})
  - Gate outputs to prevent invalid logic levels to next block



- Voltage drop across sleep transistor degrades performance during normal operation
  - Size the transistor wide enough to minimize impact
- Switching wide sleep transistor costs dynamic power
  - Only justified when circuit sleeps long enough