Chapter 4

Power

闕志達

台灣大學電機系

Power Dissipation

Dynamic power

- Switching power
  - Charging and discharging of load capacitance
- Short-circuit power
  - Current when PMOS/NMOS turn partially on simultaneously
- Static power
  - Subthreshold current in MOS
  - Leakage current through reverse-biased junction
  - Gate leakage
  - Current in ratioed logic

Power Dissipation Sources

- $P_{total} = P_{dynamic} + P_{static}$
- Dynamic power: $P_{dynamic} = P_{switching} + P_{shortcircuit}$
  - Switching load capacitances
  - Short-circuit current
- Static power: $P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention})V_{DD}$
  - Subthreshold leakage
  - Gate leakage
  - Junction leakage
  - Contention current
- Refer to Chapter 2 for three leakage scenarios.

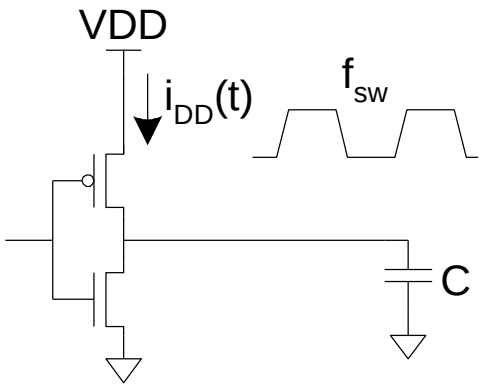
Switching Power

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

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

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

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



Activity Factor in Switching Power

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where α = activity factor
  - α is the probability of node transitions from 0 to 1.
  - If the signal is a clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = ½$

- The average switching power (power used to charge and discharge load capacitance of a CMOS gate) for a square-wave input having a repetition period of $t = 1/f$ and activity factor α is given by

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

Switching Power Example

- 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 65 nm process ($\lambda = 25$nm)
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)
- Estimate switching power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current.

Solution

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

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

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

Switching Power Reduction

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

Activity Factor Estimation

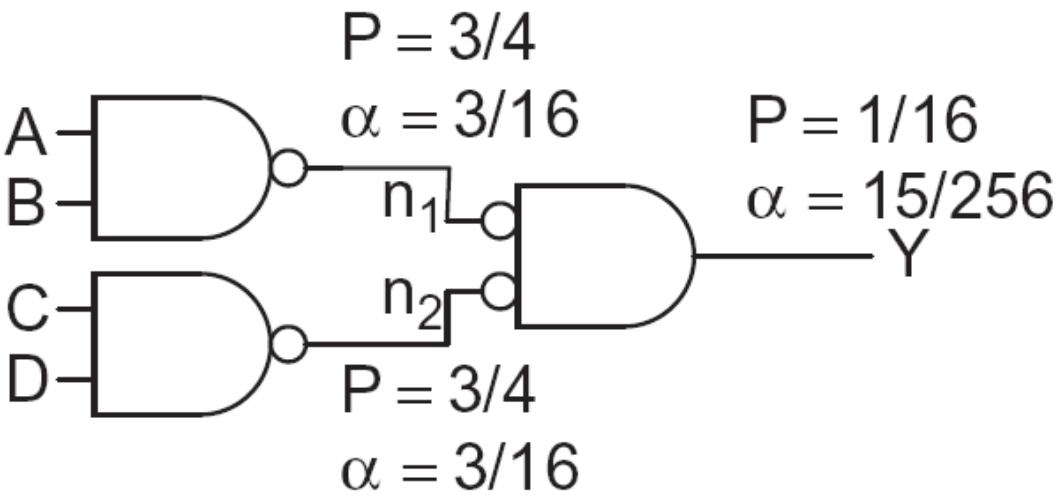
- Let P_i = Prob(node i = 1)
  - $\overline{P}_i = 1-P_i$
- $\alpha_i = P_i * \overline{P}_i$
- Completely random data has P = 0.5 and α = 0.25
- Data is often not completely random
  - e.g. upper bits of 64-bit words representing bank account balances are usually 0
- Data propagating through ANDs and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

Switching 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	$P_A \overline{P}_B + \overline{P}_A P_B$

Example

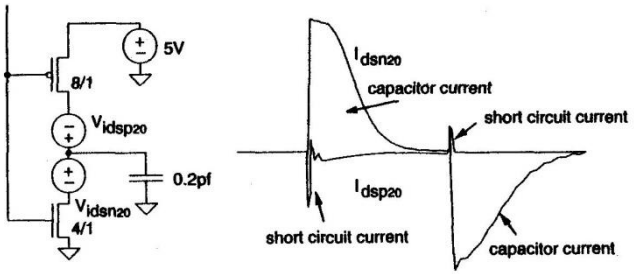
- A 4-input AND is built out of two levels of gates
- Estimate the activity factor at each node if the inputs have P = 0.5



Short-Circuit Power Dissipation

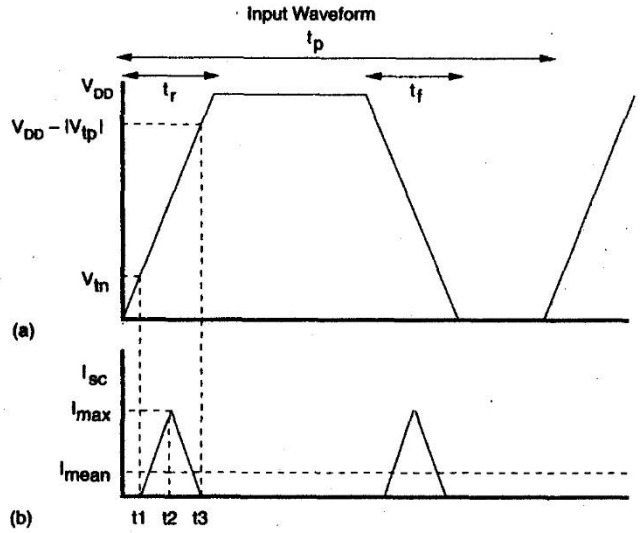
- During transition from '0' to '1' or from '1' to '0', both NMOS and PMOS are on for a short period of time. This results in a short circuit current from V_{DD} to ground and thus some short-circuit power dissipation, P_{sc}. As the load capacitance grows larger, the short circuit current becomes comparatively smaller than the charge or discharge currents. Also longer input rise time and fall time increases the short circuit current.
- The short circuit power dissipation is given by

$$P_{sc} = I_{mean} V_{DD}$$



Short-Circuit Power Dissipation

- < 10% of dynamic power if rise/fall times are comparable for input and output
- Slow input rise time and fall time can result in significant short circuit power dissipation.
- Thus it is good practice to keep all edges fast if power dissipation is a concern.
- We will generally ignore this.



Static Power

- Static power is consumed even when chip is quiescent.
  - Leakage draws power from nominally OFF devices
- Some other logic, e.g. Pseudo-NMOS (introduced in a later chapter) has constant current between V_{DD} and GND.
  - Ratioed circuits burn power in fight between ON transistors
- The total static power dissipation is

$$P_{static} = I_{static} V_{DD}$$

- In 0.13um technology, the static power very often becomes comparable to dynamic power.
- Even worse in 28nm/14nm/7nm technology.

Static Power Example

- Revisit power estimation for 1 billion transistor chip
- Estimate static power consumption
  - Subthreshold leakage
    - Normal V_t: 100 nA/μm
    - 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

Solution

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

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

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

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

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

50% transistors are OFF

50% transistors are ON

15% of the dynamic power

p.6

- 1 billion transistor chip
  - 50M logic transistors
    - Average width: 12 λ
    - Activity factor =
  - 9
    - Activity factor = 0.02
  - 1.0 V 65 nm process (λ = 25nm)
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)

Subthreshold leaka

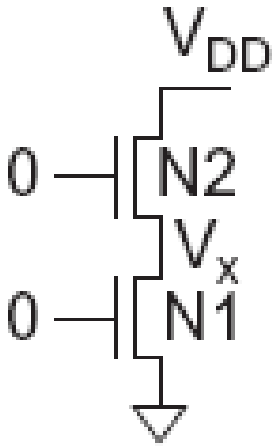
- Normal V_t: 100 nA/μm
- High V_t: 10 nA/μm
- High Vt used in all memories and in 95% of logic gates

Gate leakage 5 nA/μm

p.15

Stack Effect

- Series OFF transistors have less leakage
  - $V_x > 0$, so N2 has negative V_{gs}
  - Leakage through 2-stack reduces ~10x
  - Leakage through 3-stack reduces further

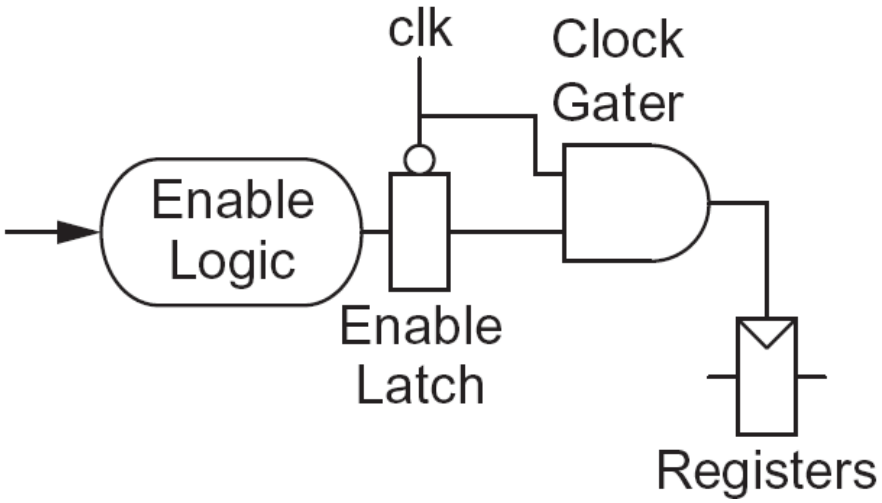


Low-Power Design

- Dynamic Power Saving
  - Turn off unused modules (α). Called clock gating.
  - Use just enough speed for computation
  - Dynamic voltage frequency scaling (DVFS); running the processor at 2/3 speed and a lower V_{DD} may save 70% power.
  - Avoid slow-rising/falling waveforms in large-current circuits (I/O).
- Static Power saving
  - Use body effect to raise V_t in the idle transistors or lower V_t of the active transistors. Called back-bias technique.
  - Lower or turn off V_{DD} of the idle modules. Called Vdd gating, also called power gating.

Clock Gating

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

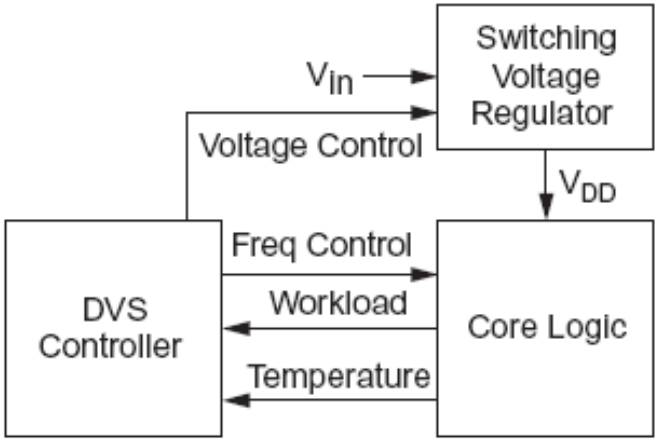


Capacitance

- 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
  - 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



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*
  - Decrease V_b (See Chapter 2 for body effect)
    - *Reverse body bias* in sleep → higher V_{sb}, so higher V_t
    - Or forward body bias in active mode → lower V_{sb}

Gate Leakage

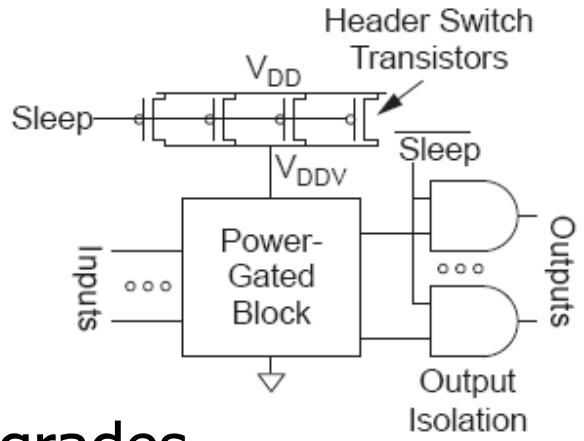
- Extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches subthreshold leakage at 65 nm and below in some processes
- An order of magnitude less for PMOS than NMOS
- Control leakage in the process using $t_{ox} > 10.5$ Å
  - High-k gate dielectrics help
  - Some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

Junction Leakage

- From reverse-biased p-n junctions
  - Between diffusion and substrate or well
- Ordinary diode leakage is negligible

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