Power Outline

❑ Power and Energy

❑ Dynamic Power

❑ Static Power

Power and Energy

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

❑ Instantaneous Power: $P(t) = I(t)V(t)$

❑ Energy: $E = \int_0^T P(t)dt$

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

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

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

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

  - – But energy drawn from the supply is
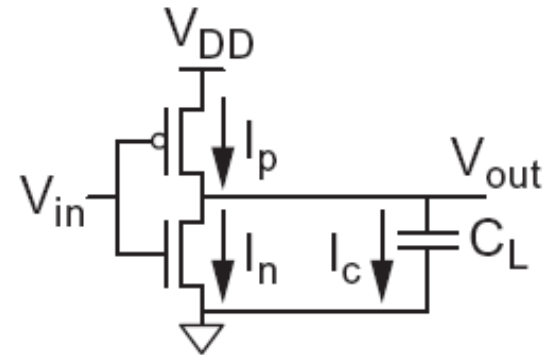
$$E_{VDD} = \int_0^\infty I(t) V_{DD} dt = \int_0^\infty C_L \frac{dV}{dt} V_{DD} dt$$

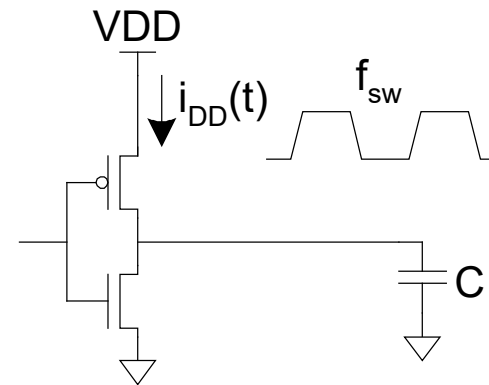$$= C_L V_{DD} \int_0^{V_{DD}} dV = C_L V_{DD}^2$$

  - – Half the energy from V_{DD} is dissipated in the pMOS transistor as heat, other half stored in capacitor
- ❑ When the gate output falls
  - – Energy in capacitor is dumped to GND
  - – Dissipated as heat in the nMOS transistor

Switching Power

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

$$= \frac{V_{DD}}{T} \int_0^T C \frac{dV}{dt} dt$$

$$= \frac{V_{DD}}{T} \int_0^{V_{DD}} C dV$$

$$= C \frac{V_{DD}^2}{T} = C V_{DD}^2 f_{sw}$$

VDD

$i_{DD}(t)$ f_{sw}

C

f_{sw}—average switching frequency

Activity Factor

❑ Suppose the system clock frequency = f

❑ Let $f_{sw} = \alpha f$, where α = activity factor
  – If the signal is a clock, $\alpha = 1$
  – If the signal switches once per cycle, $\alpha = ½$

❑ Dynamic power:

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

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

❑ We will generally ignore this component

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 is the power consumption when the circuit is off. Will be discussed later.

Dynamic 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 dynamic power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current.

Solution

$$C_{\text{logic}} = \left(50 \times 10^6\right)\left(12\lambda\right)\left(0.025\,\mu m / \lambda\right)\left(1.8\,fF / \mu m\right) = 27 \text{ nF}$$

$$C_{\text{mem}} = \left(950 \times 10^6\right)\left(4\lambda\right)\left(0.025\,\mu m / \lambda\right)\left(1.8\,fF / \mu m\right) = 171 \text{ nF}$$

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

Dynamic Power Reduction

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

$\square$ Try to minimize:

- Activity factor
- Capacitance
- Supply voltage
- Frequency

Activity Factor Estimation

❑ Let P_i = Probability for node i = 1
  – $\bar{P}_i$ = 1-P_i (node i = 0)

❑ $\alpha_i = \bar{P}_i * P_i$

❑ Completely random data has P = 0.5, then 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_γ
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
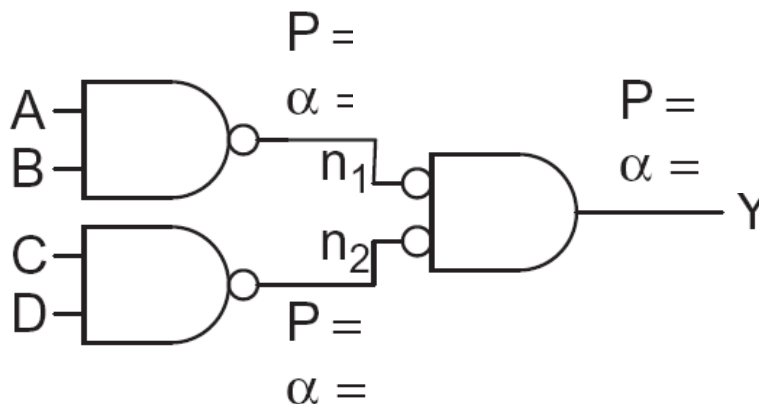
Clock Gating

❑ 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
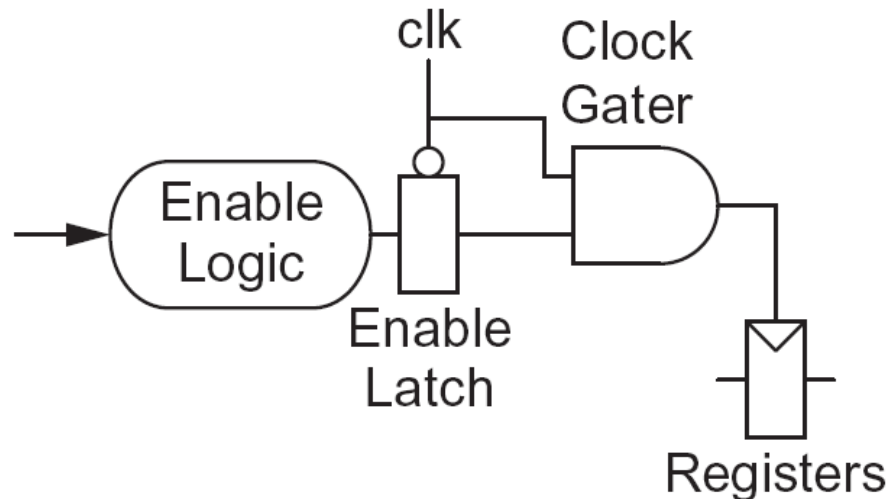
Capacitance

❑ Gate and diffusion capacitance

– Fewer stages of logic

– Small gate sizes

– Good floorplanning with Euler path technique to reduce diffusion breaks

❑ 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

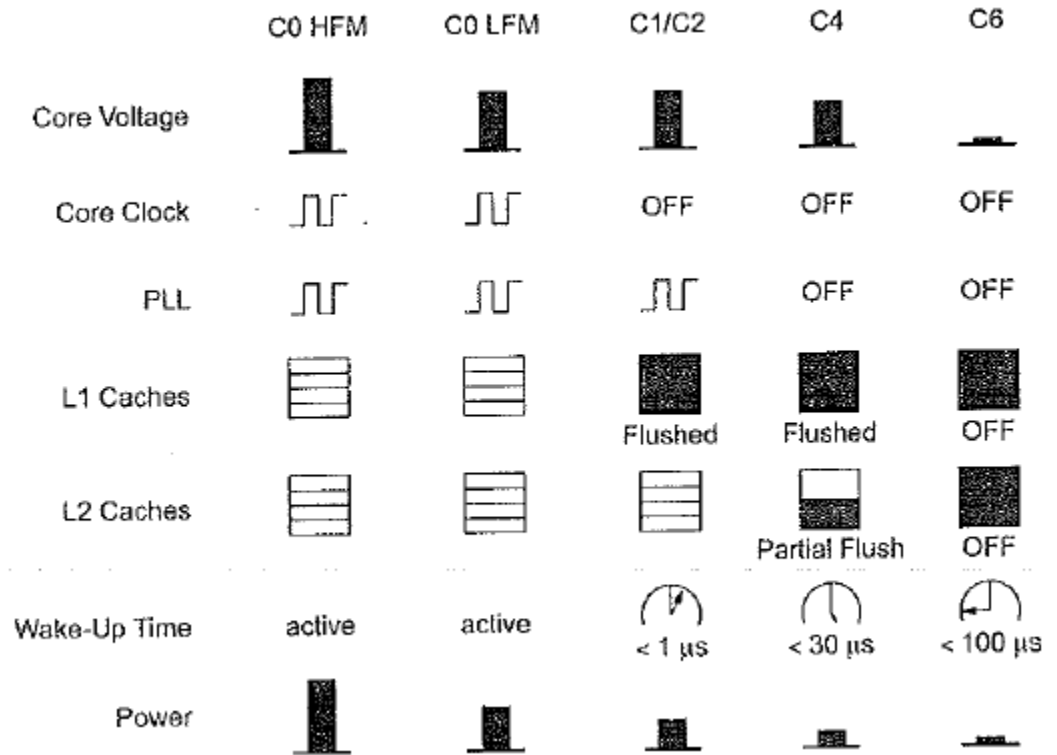   – Level converters required when crossing from low to high V_{DD} domains
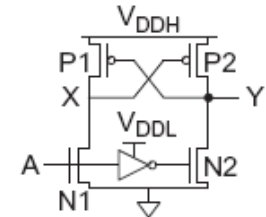
❑ Dynamic Voltage Scaling

    – Adjust V_{DD} and f according to workload
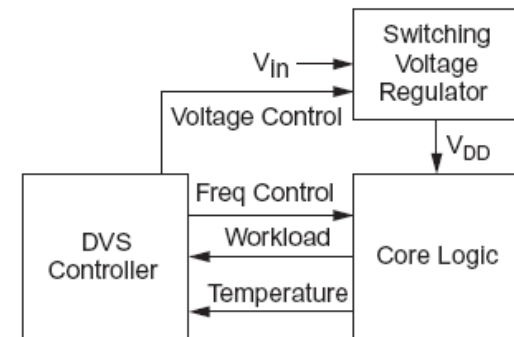
Power Management

❑ Power Management

- Turn off portions of the chip when they are not active

- Multiple frequency (2GHz-600MHz)

- Multiple core voltage (1.0V-0.75V)

	C0 HFM	C0 LFM	C1/C2	C4	C6
Core Voltage					
Core Clock			OFF	OFF	OFF
PLL				OFF	OFF
L1 Caches			Flushed	Flushed	OFF
L2 Caches				Partial Flush	OFF
Wake-Up Time	active	active	< 1 µs	< 30 µs	< 100 µs
Power					

2W P_{avg}=220mW   80mW

Intel Atom processor power management mode

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

❑ Static power: $P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention})V_{DD}$

– Subthreshold leakage

– Gate leakage

– Junction leakage

– Contention current

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

$$W_{\text{normal-V}_t} = \left(50 \times 10^6\right)\left(12\lambda\right)\left(0.025\mu\text{m}/\lambda\right)\left(0.05\right) = 0.75 \times 10^6 \ \mu\text{m}$$

$$W_{\text{high-V}_t} = \left[\left(50 \times 10^6\right)\left(12\lambda\right)\left(0.95\right) + \left(950 \times 10^6\right)\left(4\lambda\right)\right]\left(0.025\mu\text{m}/\lambda\right) = 109.25 \times 10^6 \ \mu\text{m}$$

$$I_{sub} = \left[W_{\text{normal-V}_t} \times 100 \text{ nA}/\mu\text{m} + W_{\text{high-V}_t} \times 10 \text{ nA}/\mu\text{m}\right]/2 = 584 \text{ mA}$$

$$I_{gate} = \left[\left(W_{\text{normal-V}_t} + W_{\text{high-V}_t}\right) \times 5 \text{ nA}/\mu\text{m}\right]/2 = 275 \text{ mA}$$

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

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
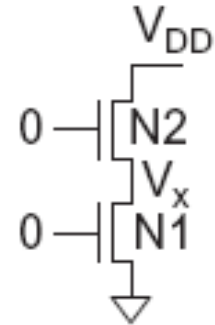η = 0.1
k_γ = 0.1
S = 100 mV/decade

η: DIBL (Drain Induced Barrier Lowering) coefficient;

$K\gamma$: body effect coefficient;

S: subthreshold slope

Stack Effect

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

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

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

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

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

  – 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
    $$I_{sub} \approx I_{off} 10^{\frac{V_{gs} + \eta(V_{ds} - V_{DD}) - k_\gamma V_{sb}}{S}}$$
    • *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
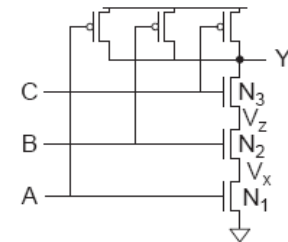
| Input State (ABC)	I_{sub}	I_{gate}	I_{total}	V_x	V_z
000	0.4	0	0.4	stack effect	stack effect
001	0.7	0	0.7	stack effect	$V_{DD} - V_t$
010	0	1.3	1.3	intermediate	intermediate
011	3.8	0	10.1	$V_{DD} - V_t$	$V_{DD} - V_t$
100	0.7	6.3	7.0	0	stack effect
101	3.8	6.3	10.1	0	$V_{DD} - V_t$
110	5.6	12.6	18.2	0	0
111	28	18.9	46.9	0	0

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 $V_{db} = V_{DD}$

❑ 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
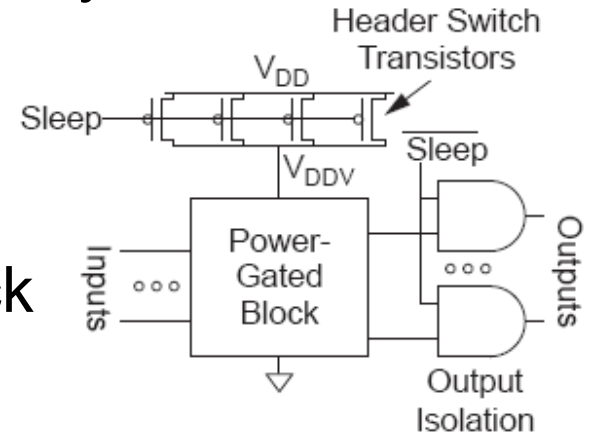
❑ Voltage drop across sleep transistor degrades performance during normal operation
  – Size the transistor wide enough to minimize impact

❑ Switching wide sleep transistor costs dynamic power
  – Only justified when circuit sleeps long enough

Energy Delay Product

❑ Popular metric for measuring the performance

  – Power-Delay-Product (PDP) @ frequency f

  – Energy-Delay-Product (EDP)

   • Estimating the gate delay considering the supply voltage and threshold voltage change

$$\tau = k \frac{CV_{dd}}{\left(V_{dd} - V_t\right)^{\alpha}}$$

   » α: velocity saturation index (1-2)

   » k: process parameter

   • Energy drawn through Vdd, $E_{VDD} = C_L V_{DD}^2$

$$EDP = k \frac{C^2 V_{dd}^{\,3}}{\left(V_{dd} - V_t\right)^{\alpha}}$$

Minimum Energy Delay Product

❑ Find the optimal Vdd for minimum EDP-- differentiating EDP with respect to Vdd $EDP = k \dfrac{C^2 V_{dd}^{\ 3}}{\left(V_{dd} - V_t\right)^{\alpha}}$

$$V_{dd-opt} = \frac{3}{3-\alpha} V_t$$

❑ For a typical α, $\quad V_{dd\text{-}opt} \approx 2V_t$

Summary

❑ Dynamic Power

❑ Static Power