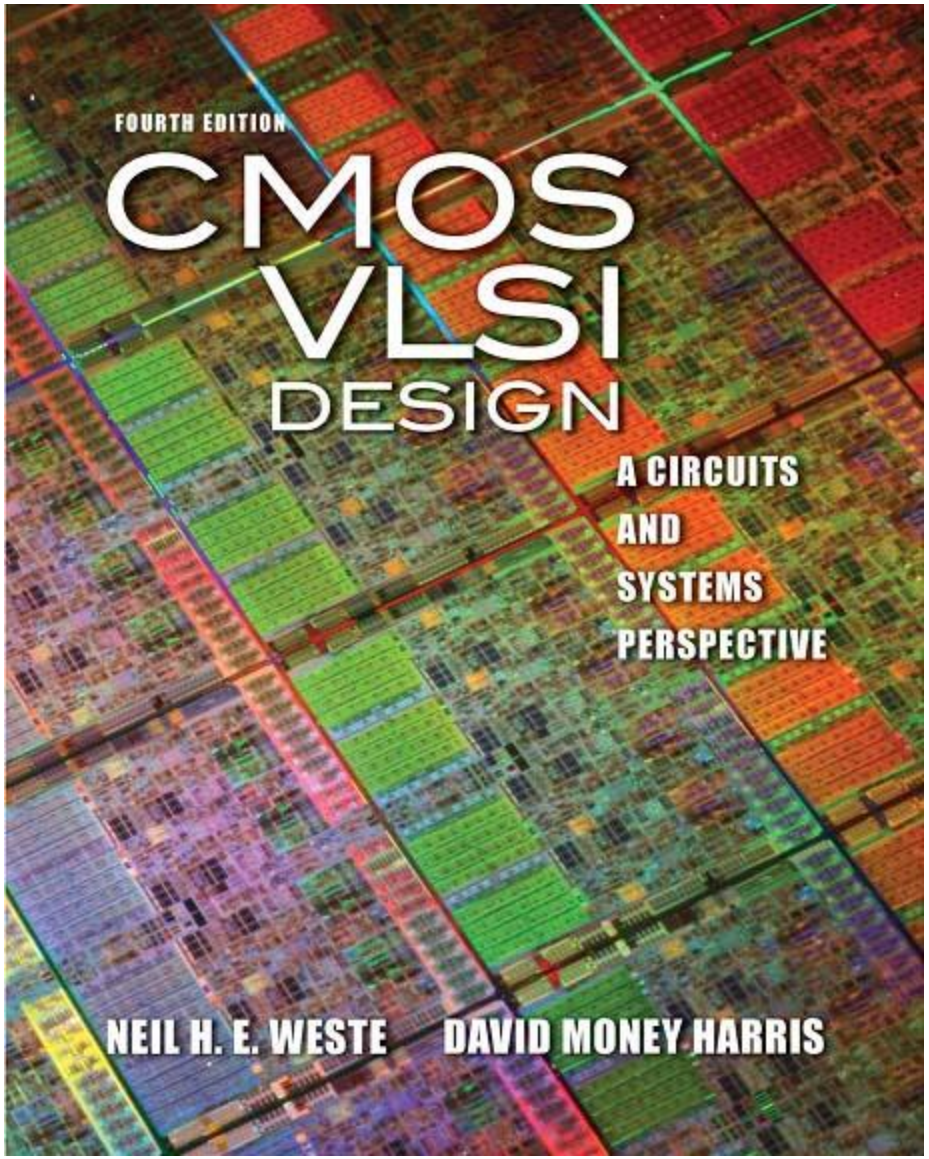


Lecture 7: Power

Outline

- Power and Energy
- Dynamic Power
- Static Power

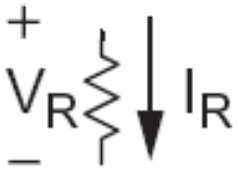
Power and Energy

- Power is drawn from a voltage source attached to the V_{DD} pin(s) of a chip.
- Instantaneous Power: $P(t) =$
- Energy: $E =$
- Average Power: $P_{\text{avg}} =$

Power in Circuit Elements

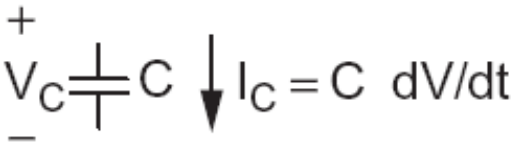
$$P_{VDD}(t) = I_{DD}(t) V_{DD}$$

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

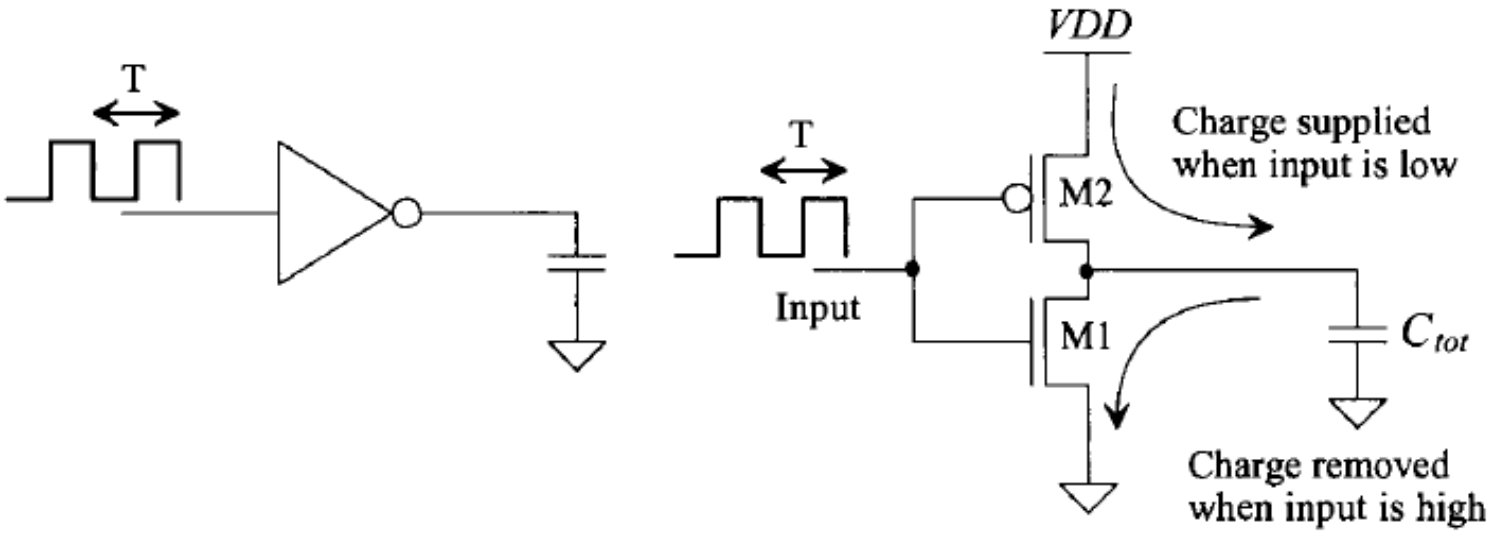


$$E_C = \int_0^\infty I(t) V(t)\, dt = \int_0^\infty C \frac{dV}{dt} V(t)\, dt$$

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



Dynamic power dissipation



Charging a Capacitor

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

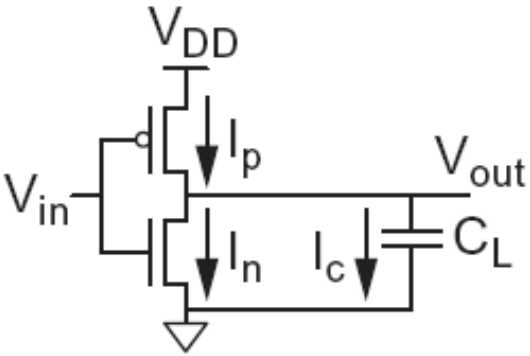
$$E_C = \tfrac{1}{2} C_L V_{DD}^2$$

  - But energy drawn from the supply is

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

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

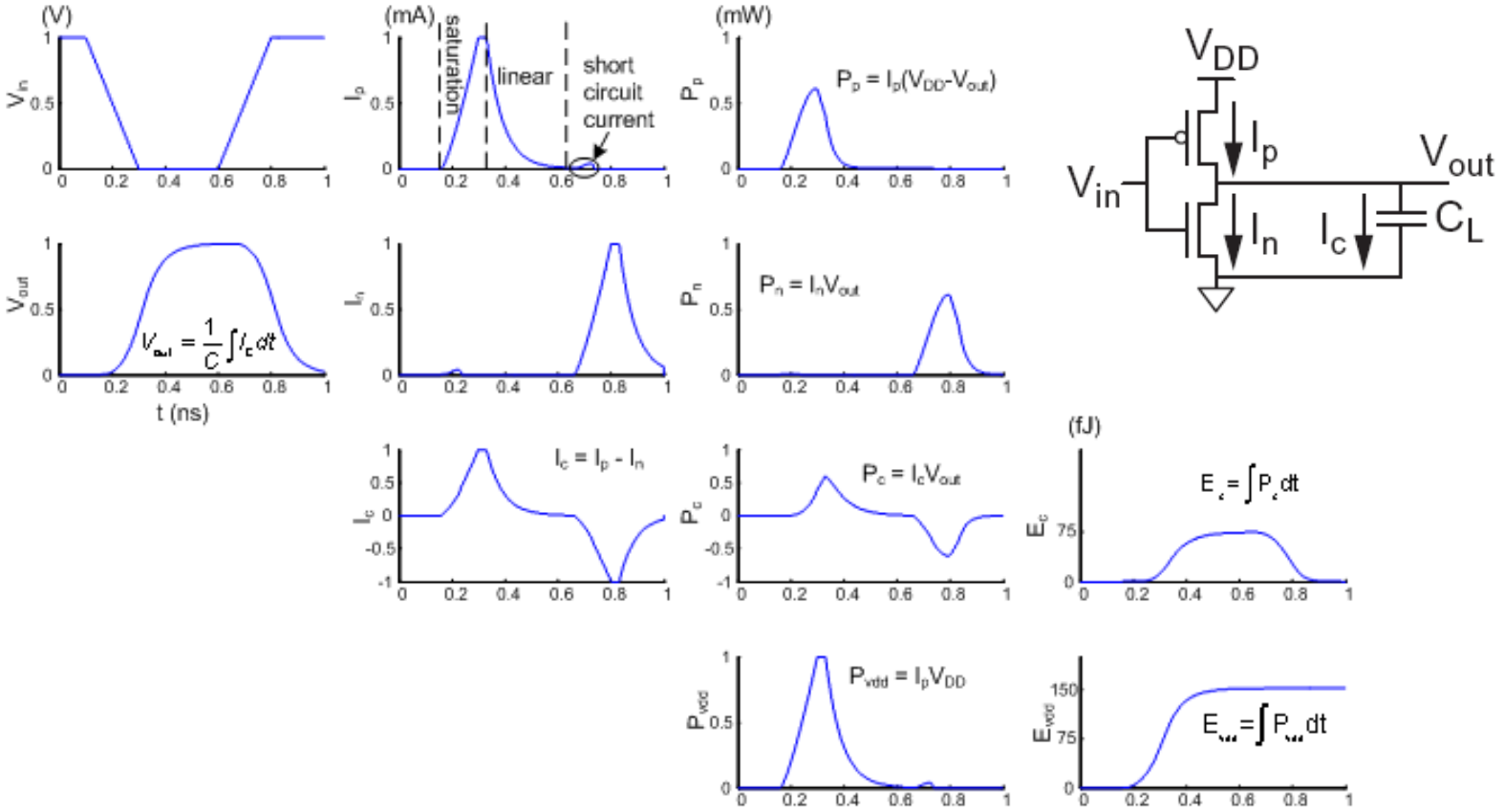
  - Half the energy from V_{DD} is dissipated in the pMOS transistor as heat, other half stored in capacitor



- When the gate output falls
  - Energy in capacitor is dumped to GND
  - Dissipated as heat in the nMOS transistor

Switching Waveforms

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



Switching Power

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

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

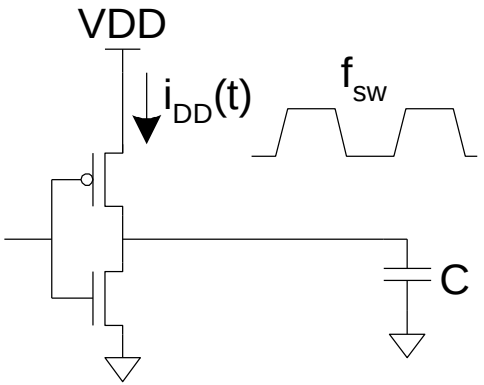
$$= \frac{V_{DD}}{T}\left[Tf_{\text{sw}}CV_{DD}\right]$$

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

$$I_{avg} = \frac{Q}{T} = \frac{V_{DD}C}{T}$$

$$P_{avg} = V_{DD}\cdot I_{avg} = \frac{C\cdot V_{DD}^2}{T}$$

$$= CV_{DD}^2 f_{SW}$$



Activity Factor

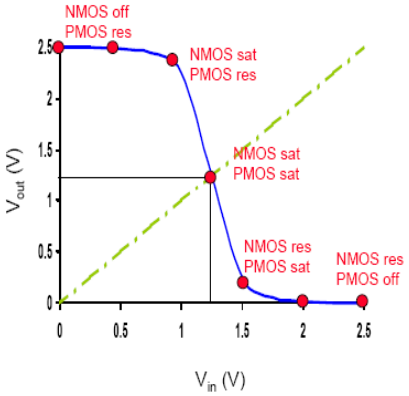
- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where α = activity factor
  - If the signal is a clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = ½$

- 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

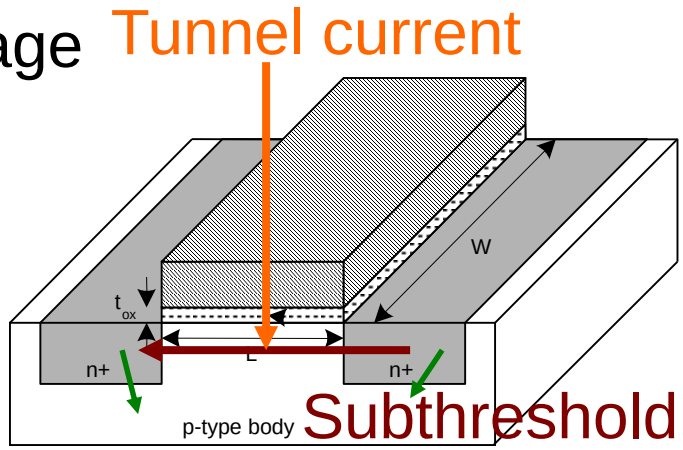


Power Dissipation Sources

- $P_{total} = P_{dynamic} + P_{static}$
- Dynamic power: $P_{dynamic} = P_{switching} + P_{shortcircuit}$
  - Switching load capacitances
  - Short-circuit current

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

- Static power: $P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention})V_{DD}$
  - Subthreshold leakage
  - Gate leakage
  - Junction leakage
  - Contention current in rationed circuits



Power equations in CMOS

Power Equations in CMOS

$$P = \alpha f C_L V_{DD}^2 + V_{DD} I_{peak} (P_{0\rightarrow1} + P_{1\rightarrow0}) + V_{DD} I_{leak}$$

Dynamic power	Short-circuit power	Leakage power
(≈ 40 - 70% today and decreasing relatively)	(≈ 10 % today and decreasing absolutely)	(≈ 20 – 50 % today and increasing)

Dynamic Power Example

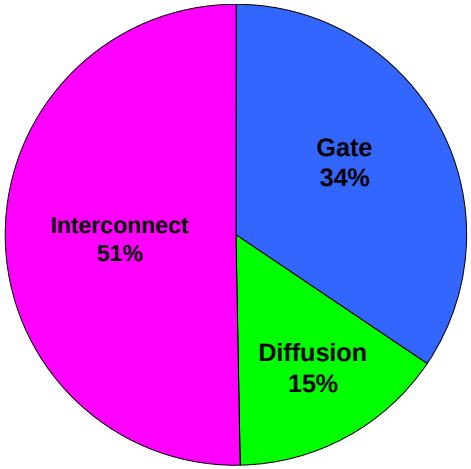
- 1 billion transistor chip $\rightarrow R \approx \frac{1}{\beta(V_{gs}-V_t)} = \frac{1}{\mu C_{ox}} \frac{L}{W} \frac{1}{(V_{gs}-Vt)}$
  - 50M logic transistors
    - Average width: 12 λ
    - Activity factor = 0.1
  - 950M memory transistors
    - Average width: 4 λ
    - Activity factor = 0.02
  - 1.0 V 65 nm process
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)
- Estimate dynamic power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current.

Solution

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

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

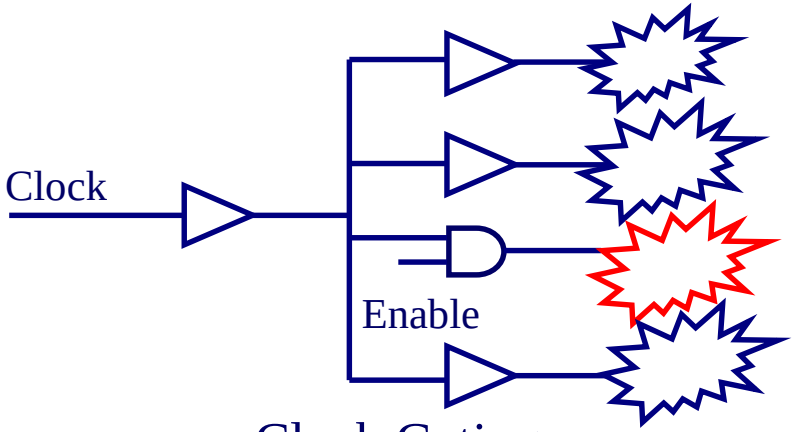
$$P_{\text{dynamic}} = \left[0.1C_{\text{logic}} + 0.02C_{\text{mem}}\right]\left(1.0\right)^{2}\left(1.0\text{ GHz}\right) = 6.1\text{ W}$$



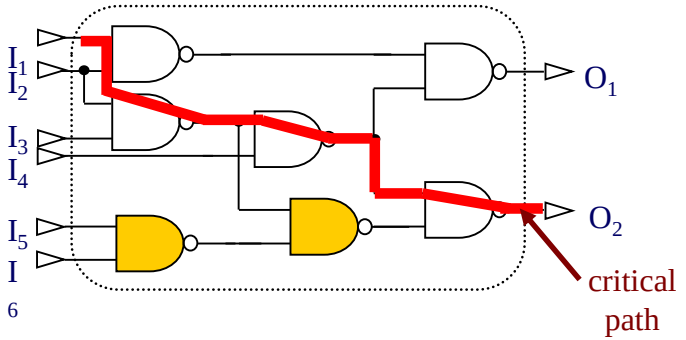
Total dynamic Power [source: Intel'03]

Dynamic Power Reduction

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



Clock Gating



only reduce supply voltage of non critical gates

Activity Factor Estimation

- Let P_i = Prob(node i = 1)
  - $\overline{P}_i = 1 - P_i$
- $\alpha_i = \overline{P}_i * P_i$
- 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 $\alpha \approx 0.1$

Define P_i to be the probability that node i is 1. $\overline{P}_i = 1 - P_i$ is the probability that node i is 0. α_i, the activity factor of node i, is the probability that the node is 0 on one cycle and 1 on the next. If the probability is uncorrelated from cycle to cycle,

Switching Probability

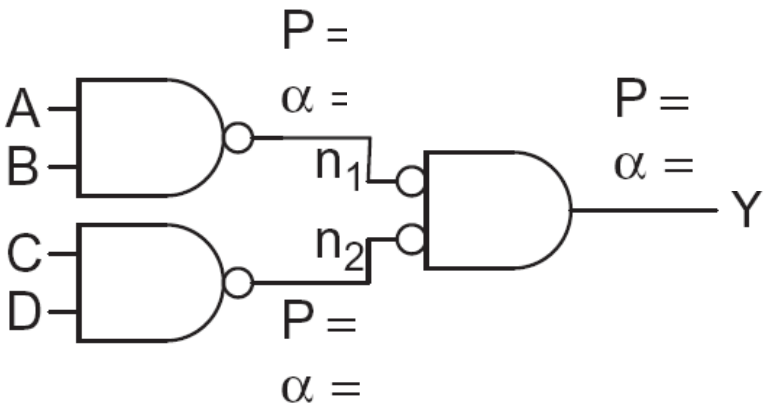
Gate	P_Y
AND2	$P_A P_B$
AND3	$P_A P_B P_C$
OR2	$1 - \overline{P}_A \overline{P}_B$
NAND2	$1 - P_A P_B$
NOR2	$\overline{P}_A \overline{P}_B$
XOR2	$P_A \overline{P}_B + \overline{P}_A P_B$

$$P(\overline{AB}) = \overline{P(AB)}$$
$$= 1 - P_A P_B$$

$$P(AB) = P_A P_B$$
$$P(ABC) = P_A P_B P_C$$
$$P(A+B) = P(\overline{\overline{A+B}})$$
$$= \overline{P(\overline{A}\,\overline{B})} = 1 - \overline{P}_A \overline{P}_B$$
$$P(\overline{A+B}) = P(\overline{A}\,\overline{B})$$
$$= \overline{P}_A \overline{P}_B$$
$$P(A\overline{B} + \overline{A}B) =$$
$$P(A\overline{B}) + P(\overline{A}B)$$
$$= P_A \overline{P}_B + \overline{P}_A P_B$$

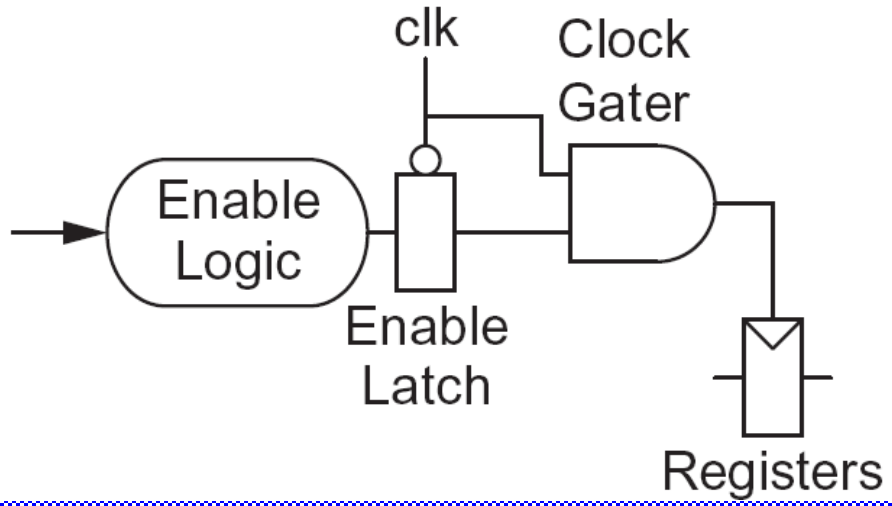
Example

- 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

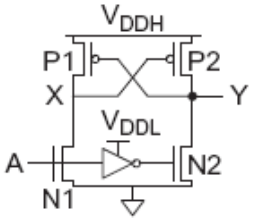


Capacitance

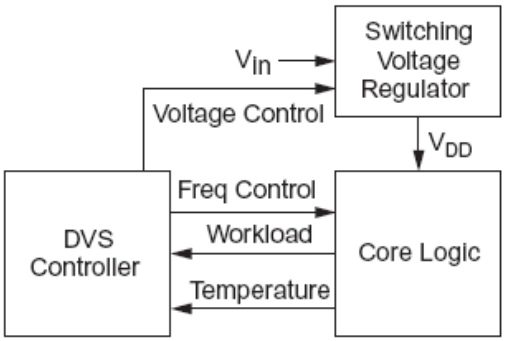
- Gate capacitance
  - Fewer stages of logic
  - Small gate sizes
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other
  - Drive long wires with inverters or buffers rather than complex gates

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



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

$$I_{ds} = I_{ds0} e^{\frac{V_{gs} - V_t}{n v_T}} \left[1 - e^{\frac{-V_{ds}}{v_T}} \right]$$

$$v_T = \frac{kT}{q}$$

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 V_t used in all memories and in 95% of logic gates
  - Gate leakage 5 nA/μm
  - Junction leakage negligible

Solution

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

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

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

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

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

Subthreshold Leakage

- For $V_{ds} > 50$ mV

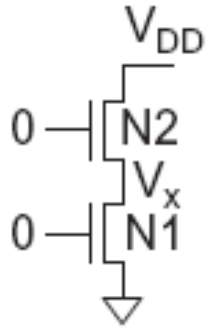
$$I_{sub} \approx I_{off} 10^{\frac{V_{gs} + \eta(V_{ds} - V_{DD}) - k_\gamma V_{sb}}{S}}$$

- 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

Stack Effect

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



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

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

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

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

$\eta = 0.1$

V_{DD}=1.0 V

S = 100 mV/decade

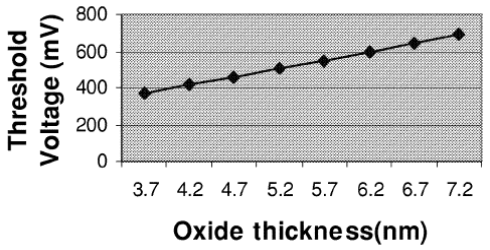
  - Leakage through 2-stack reduces ~10x
  - Leakage through 3-stack reduces further

Threshold current control

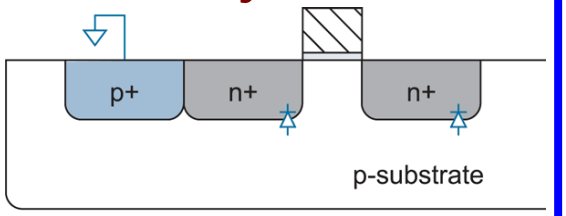
- Low V_{th} devices switch faster, and are therefore useful on critical delay paths to minimize clock periods $R=\frac{1}{\mu C_{ox}}\frac{L}{W}\frac{1}{(V_{gs}-Vt)}$
- The penalty is that low V_{th} devices have substantially higher static leakage power
- High V_{th} devices are used on non-critical paths to reduce static leakage power without incurring a delay penalty
- Typical high V_{th} devices reduce static leakage by 10 times compared with low V_{th} devices
- One method of creating devices with multiple threshold voltages is to apply different bias voltages (V_b) to the base or bulk terminal of the transistors.
- Other methods involve adjusting the gate oxide thickness, gate oxide dielectric constant (material type), or dopant concentration in the channel region beneath the gate oxide

Leakage Control

- Leakage and delay trade off γ = *body effect coefficient*
  - Aim for low leakage in sleep and low delay in active mode $V_t = V_{t0} + \gamma\left(\sqrt{\phi_s + V_{sb}} - \sqrt{\phi_s}\right)$
- 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
    - *Reverse body bias* in sleep(increase V_t)
    - Or forward body bias in active mode(decrease V_t) ϕ_s = *surface potential* at threshold



2. Body Bias



Gate Leakage

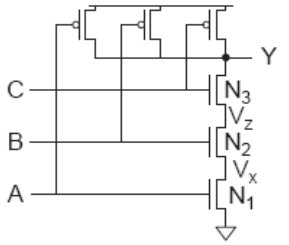
- Extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches subthreshold leakage at 65 nm and below in some processes
- An order of magnitude less for pMOS than nMOS
- Control leakage in the process using $t_{ox} > 10.5$ Å
  - High-k gate dielectrics help
  - Some 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

  Gate leakage: I_{gn} = 6.3 nA I_{gp} = 0

  Sub leakage: I_{offn} = 5.63 nA I_{offp} = 9.3 nA

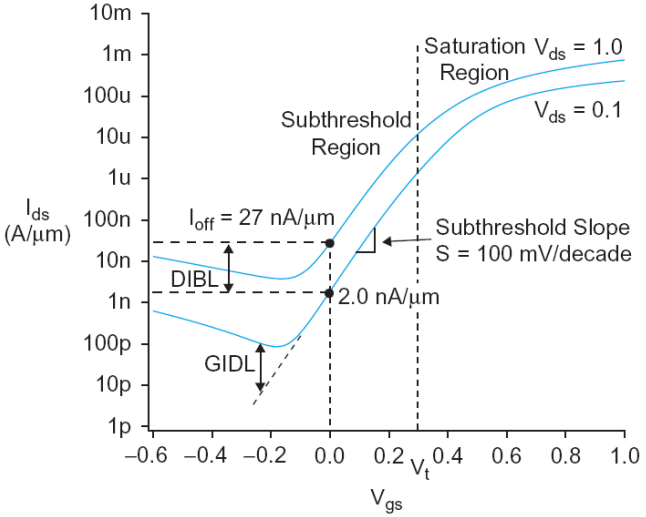


Input State (ABC)	I_{sub}	I_{gate}	I_{total}	V_x	V_z
000	0.4	0	0.4	stack effect	stack effect
001	0.7	0	0.7	stack effect	$V_{DD} - V_t$
010	0.7	1.3	2.0	intermediate	intermediate
011	3.8	0	3.8	$V_{DD} - V_t$	$V_{DD} - V_t$
100	0.7	6.3	7.0	0	stack effect
101	3.8	6.3	10.1	0	$V_{DD} - V_t$
110	5.6	12.6	18.2	0	0
111	28	18.9	46.9	0	0

Data from [Lee03]

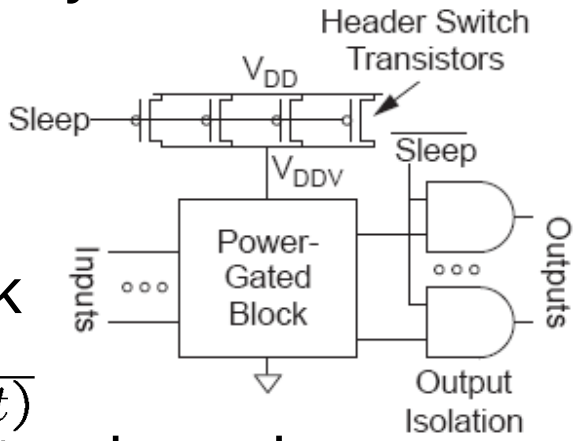
Junction Leakage

- 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 le
  - W
- Gate (GIDL) exacerbates
  - W 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



$$\rightarrow R \approx \frac{1}{\beta(V_{gs}-V_t)} = \frac{1}{\mu C_{ox}} \frac{L}{W} \frac{1}{(V_{gs}-Vt)}$$

- Voltage drop across sleep transistor degrades performance during normal operation
  - Size the transistor wide enough to minimize impact $P_{\text{switching}} = \alpha C V_{DD}^{\ 2} f$
- Switching wide sleep transistor costs dynamic power
  - Only justified when circuit sleeps long enough