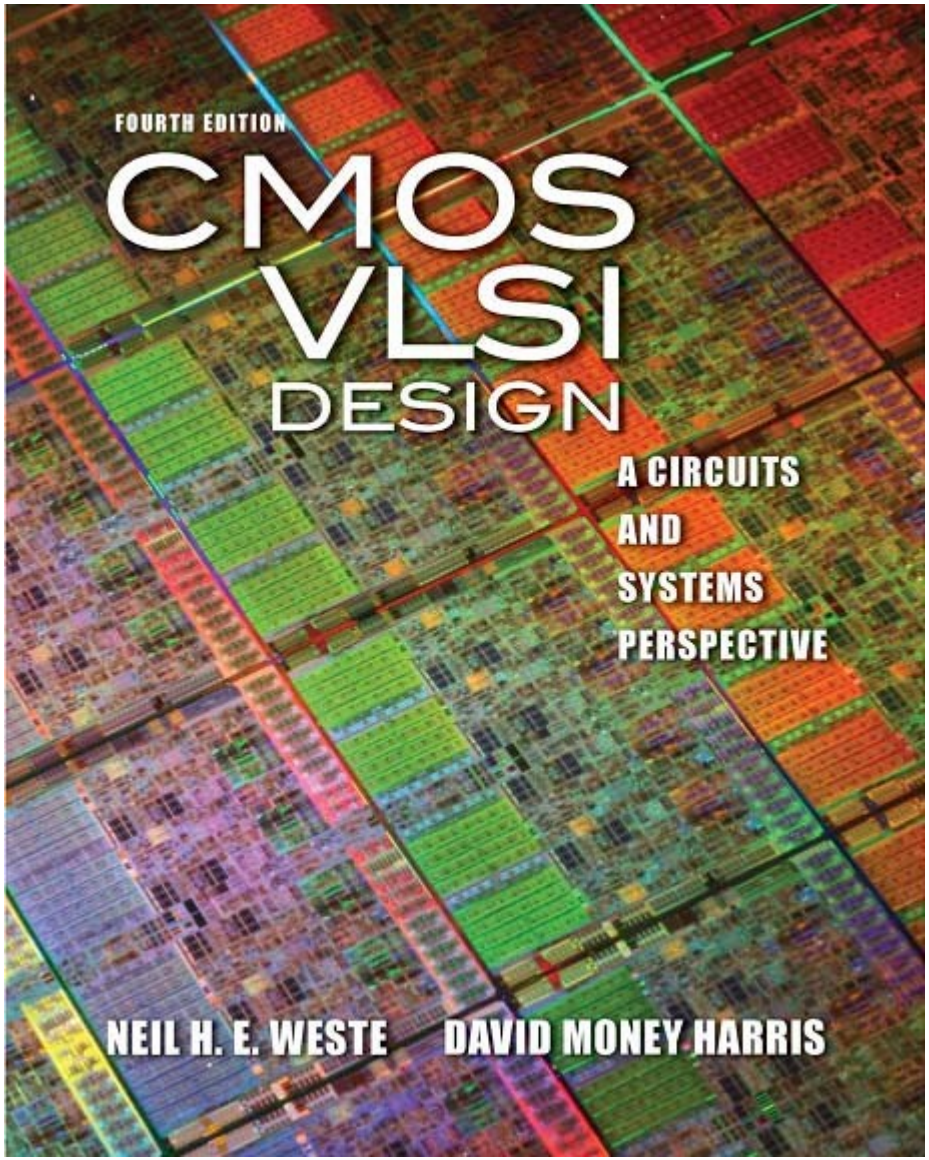


Lecture 7: Power

Outline

- Power and Energy
- Dynamic Power
- Static Power

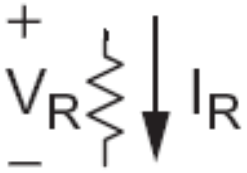
Power and Energy

- 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

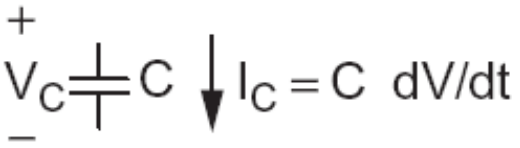
$$P_{VDD}(t) = I_{DD}(t) V_{DD}$$

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



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

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



Charging a Capacitor

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

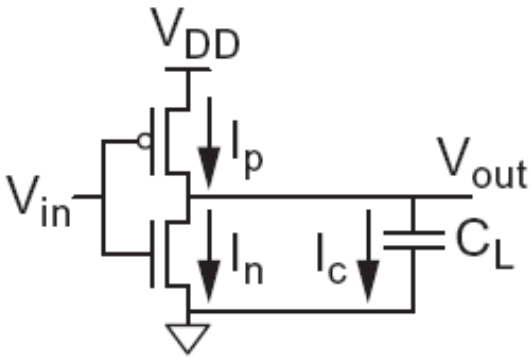
$$E_C = \frac{1}{2} C_L V_{DD}^2$$

  - But energy drawn from the supply is

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

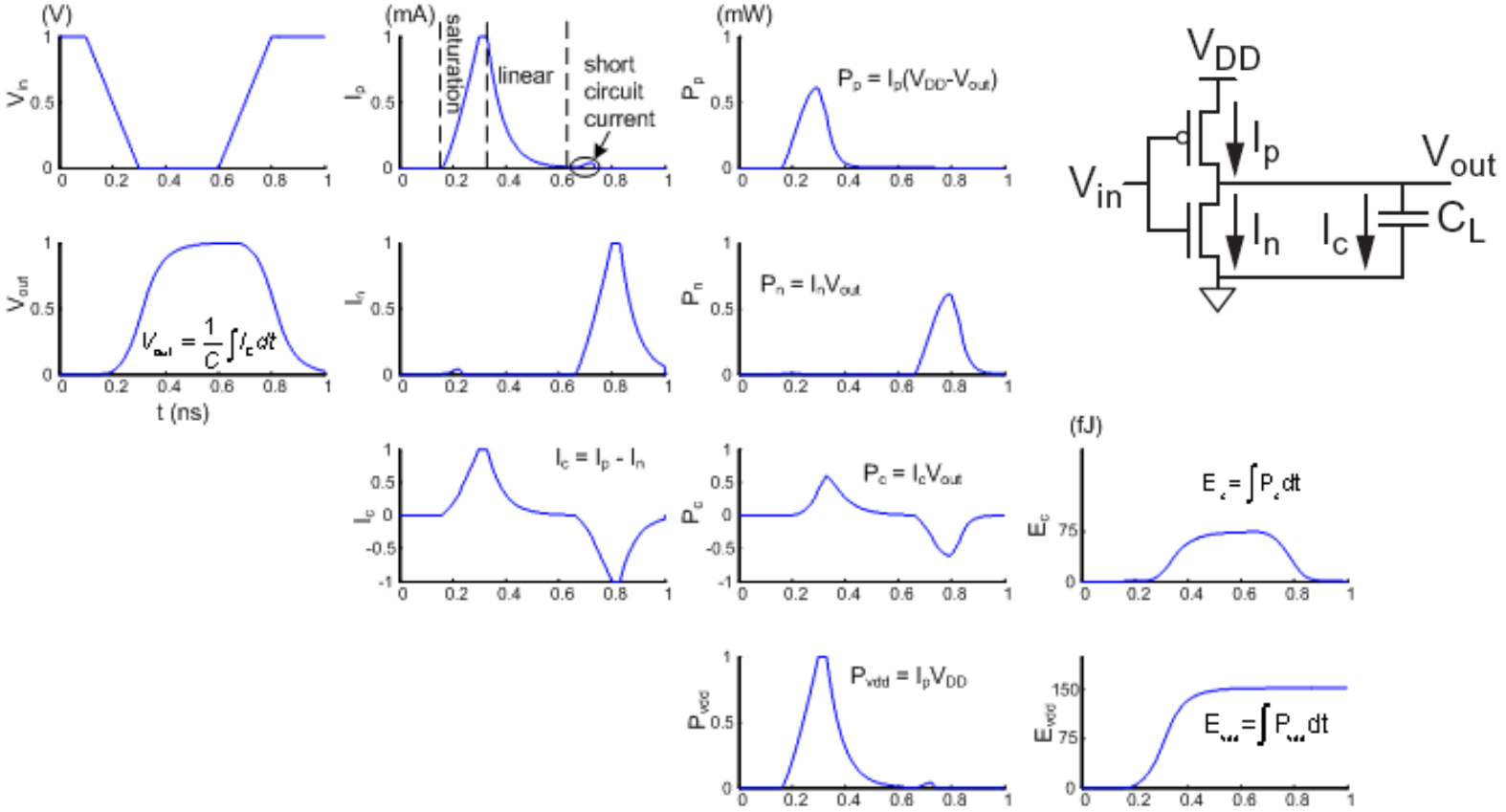
$$= C_L V_{DD} \int_0^{V_{DD}} dV = C_L V_{DD}^2$$

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



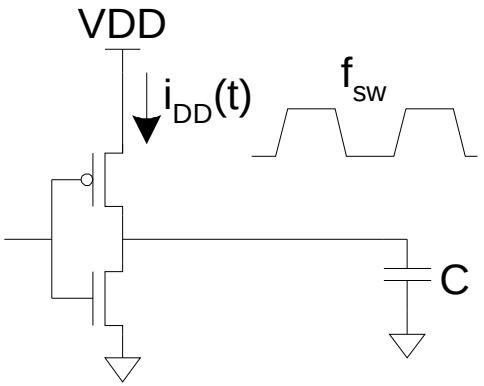
Switching Power

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

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

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

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



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

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
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)
- Estimate dynamic power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current.

Solution

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

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

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

Dynamic Power Reduction

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

Activity Factor Estimation

Activity factor of a node is the probability that it switches from 0 to 1.

- Define $\overline{P}_i$ to be the probability that node i is 1.
  - $\overline{P}_i = 1-P_i$ is the probability that node i is 0.
- $\alpha_i = P_i * \overline{P}_i$; the activity factor of node i, is the probability that the node is 0 on one cycle and 1 on the next
- Completely random data has P = 0.5 and $\alpha = 0.25$
- Data propagating through ANDs and ORs has lower activity factor
  - Depends on design, but typically $\alpha \approx 0.1$

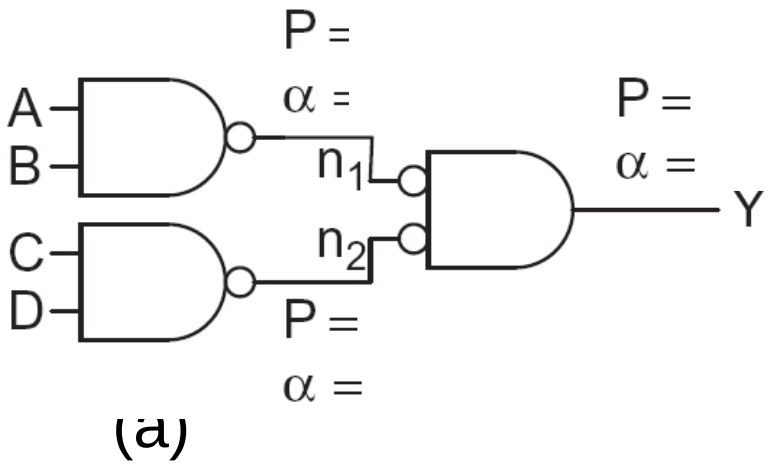
Switching Probability

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

Determine the Activity Factors at each node:

- 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

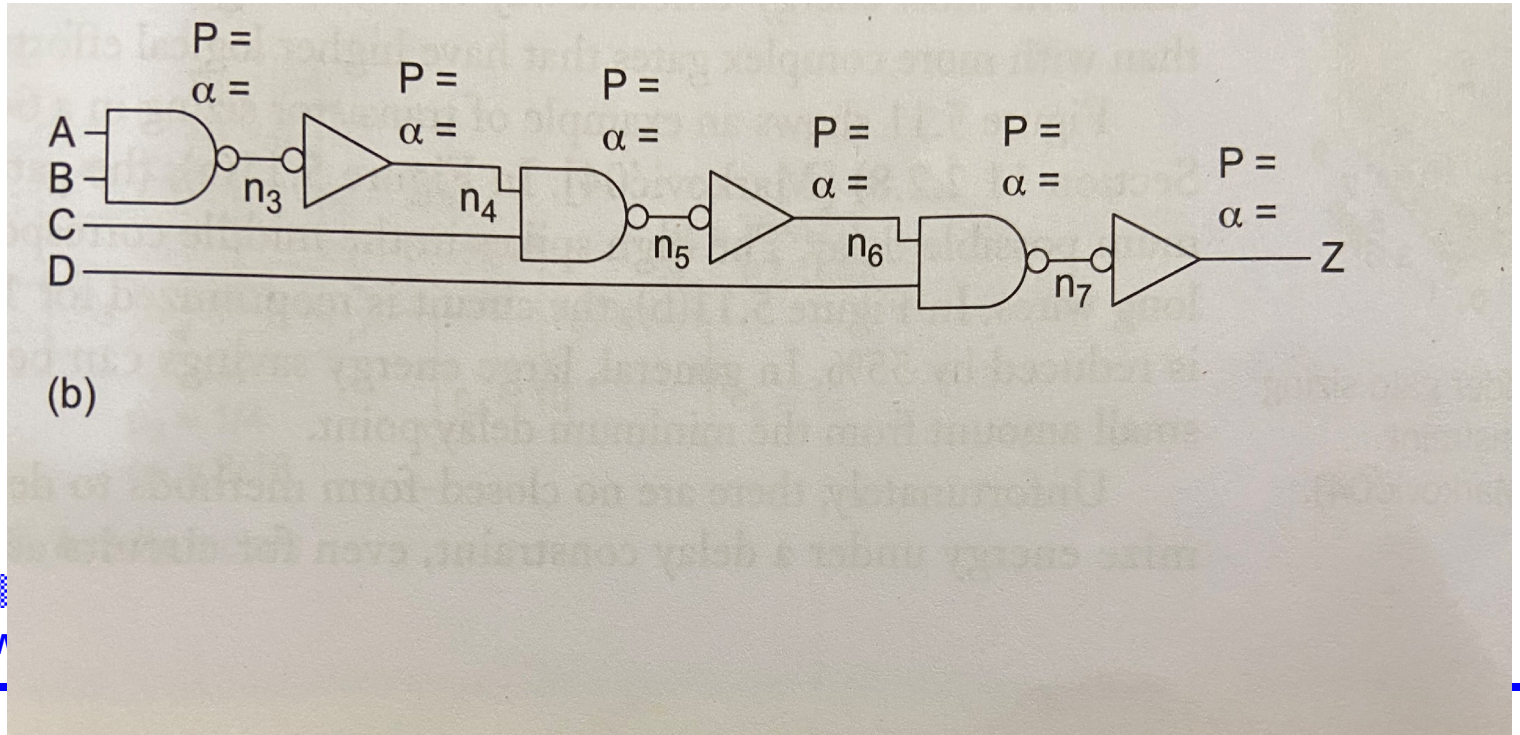


(a)

Example

Determine the Activity Factors at each node:

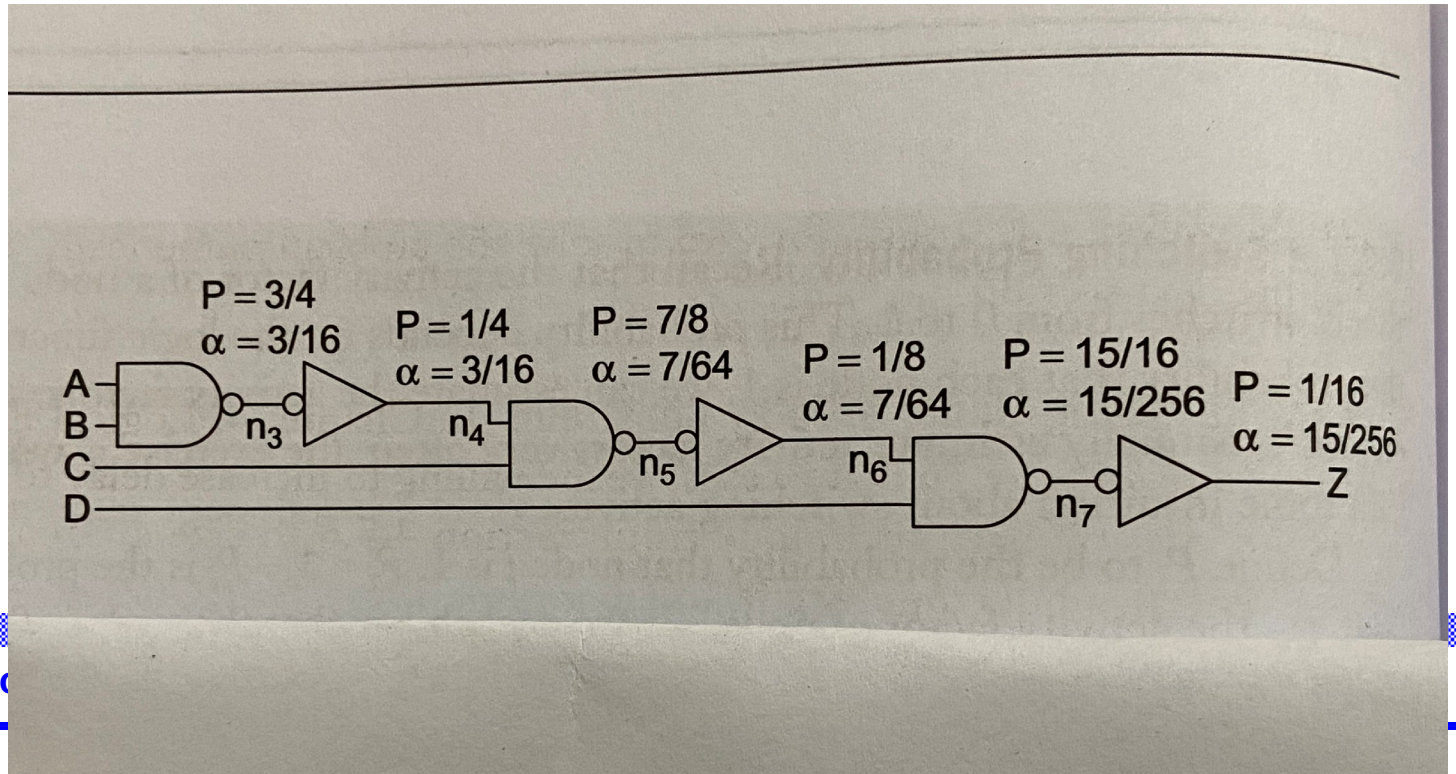
- A chain of gates
- Estimate the activity factor at each node if the inputs have P = 0.5



Example

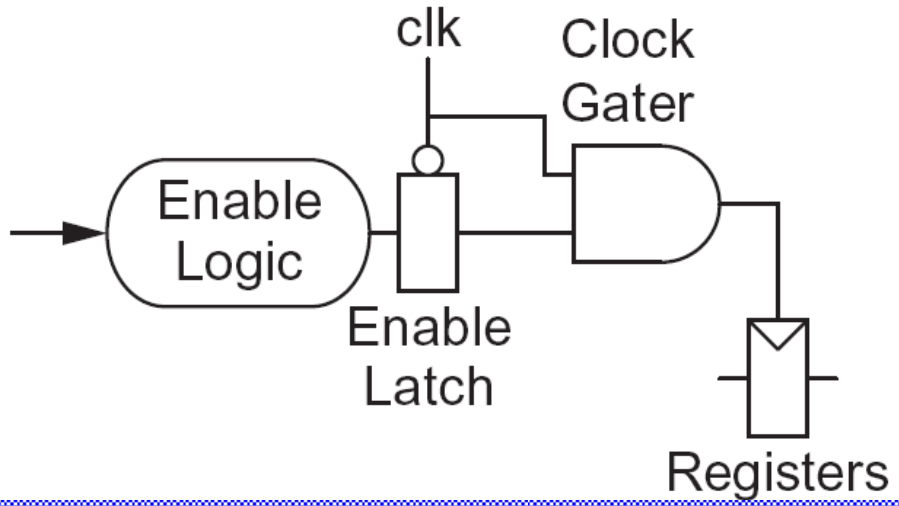
Determine the Activity Factors at each node:

- A chain 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

  Device-Switching capacitance is reduced by

  –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

Gate Sizing under a Delay Constraint

Consider a model to compute the energy of a circuit. If a unit inverter has gate capacitance 3C, then a gate with logical effort g, parasitic delay p, and drive x has gx times as much gate capacitance and px times as much diffusion capacitance. The energy of the entire circuit:

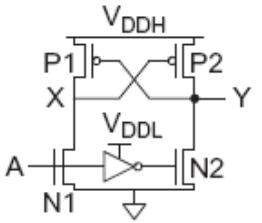
Energy = $3CV^2_{DD} \sum_{i€nodes} \alpha_i (C_{wire}/3c + p_i x_i + \sum_{j€fanout(i)} g_j x_j)$

$= \sum_{i€nodes} \alpha_i (c_i + p_i x_i + \sum_{j€fanout(i)} g_j x_j)$

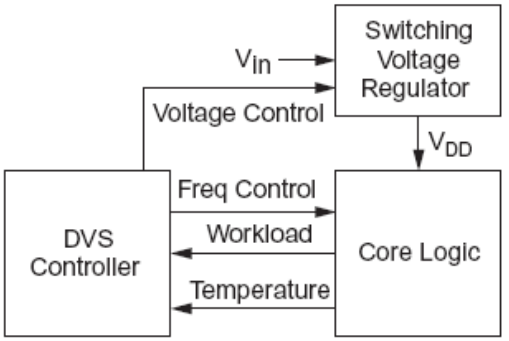
$= \sum_{i€nodes} \alpha_i x_i d_i)$

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 not switching.
  - Leakage draws power from nominally OFF devices
  - Leakage power was of concern primarily during sleep mode and it was negligible compared to dynamic power.
  - In nanometer processes with low threshold voltages and thin gate oxides, leakage can account for as much as a third of total active power.

Static Power Sources

- Static power arises from
  - Subthreshold leakage
  - Gate leakage
  - Junction leakage
  - Contention current

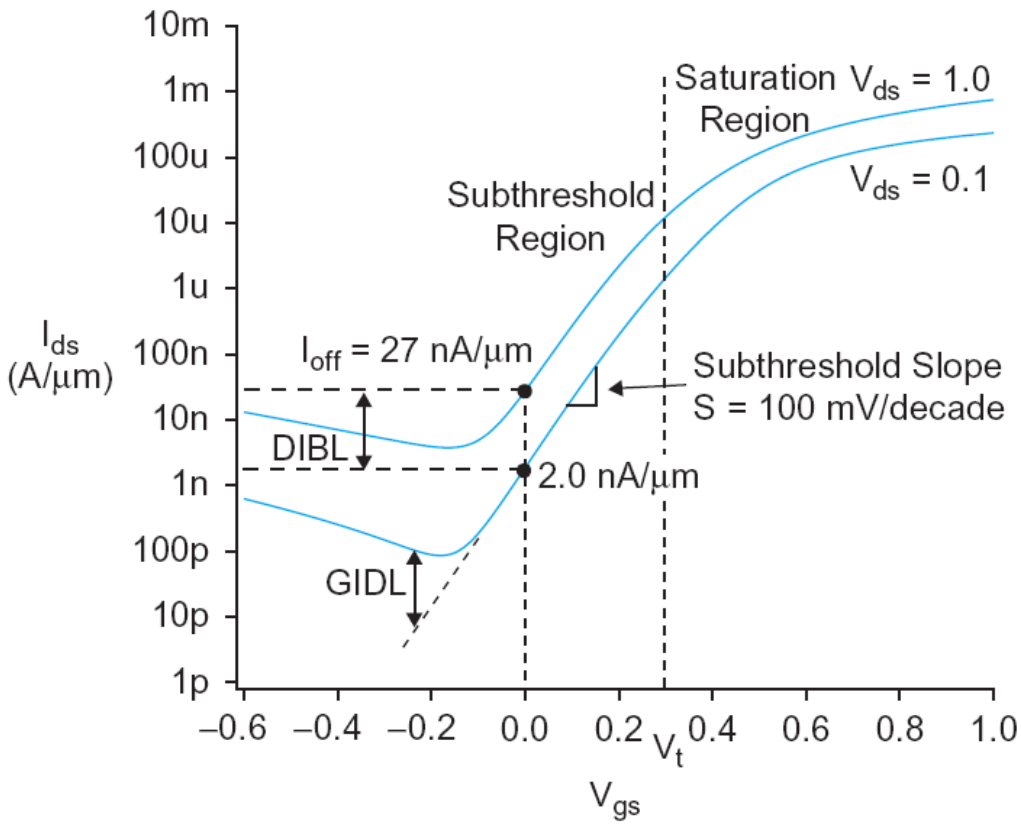
Static power: $P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention})V_{DD}$

Leakage Sources

- Subthreshold conduction
  - Transistors can't abruptly turn ON or OFF
  - Dominant source in contemporary transistors
- Gate leakage
  - Tunneling through ultrathin gate dielectric
- Junction leakage
  - Reverse-biased PN junction diode current

Leakage

- What about current in cutoff?
  - Current doesn't go to 0 in cutoff

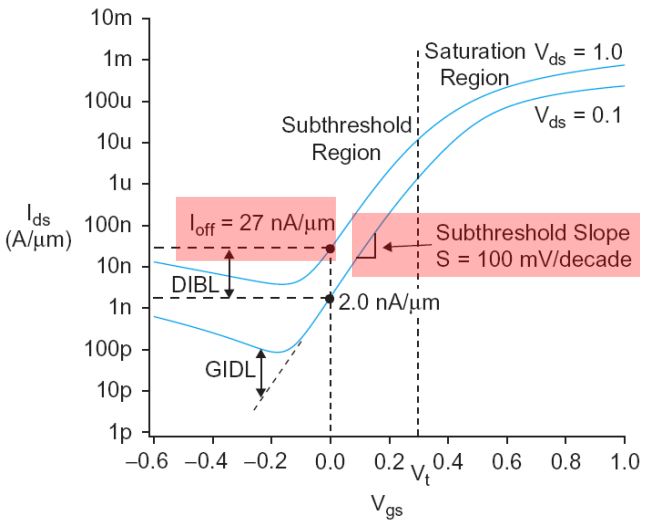


Subthreshold Leakage

- Subthreshold leakage exponential with V_{gs}

$$I_{ds} = I_{ds0} e^{\frac{V_{gs} - V_{t0} + \eta V_{ds} - k_\gamma V_{sb}}{n v_T}} \left(1 - e^{\frac{-V_{ds}}{v_T}}\right)$$

- n is process dependent
  - typically 1.3-1.7
- Rewrite relative to I_{off} on log scale



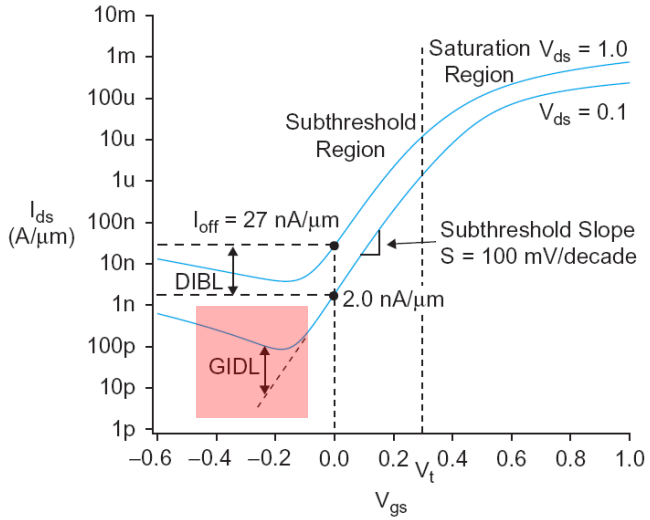
$$I_{ds} = I_{off} 10^{\frac{V_{gs} + \eta(V_{ds} - V_{dd}) - k\gamma V_{sb}}{S}} \left(1 - e^{\frac{-V_{ds}}{v_t}}\right)$$

$$S = \left[\frac{d(\log_{10} I_{ds})}{dV_{gs}}\right]^{-1} = n v_T \ln 10$$

- S ≈ 100 mV/decade @ room temperature
- DIBL – Drain Induced Barrier Lowering, GIDL - Gate-Induced Drain Leakage

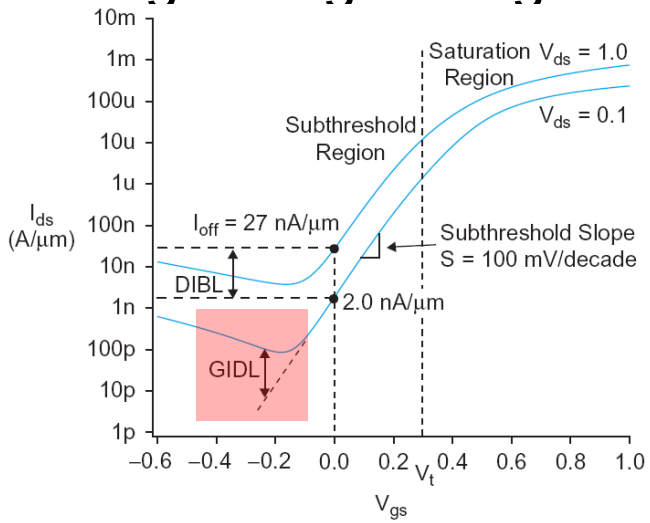
Drain Induced Barrier Lowering

- V_{ds} creates an electric field that affects the threshold voltage $V_t = V_{t0} – \eta V_{ds}$, η is DIBL coefficient, typical value is 0.1.
- DIBL causes I_{ds} to increase with V_{ds} in saturation.



Gate-Induced Drain Leakage

- Occurs at overlap between gate and drain
  - Most pronounced when drain is at V_{DD}, gate is at a negative voltage
  - Thwarts efforts to reduce subthreshold leakage using a negative gate voltage



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

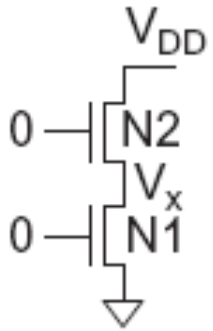
S = 100 mV/decade

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

Stack Effect

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



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

$N1$ $N2$

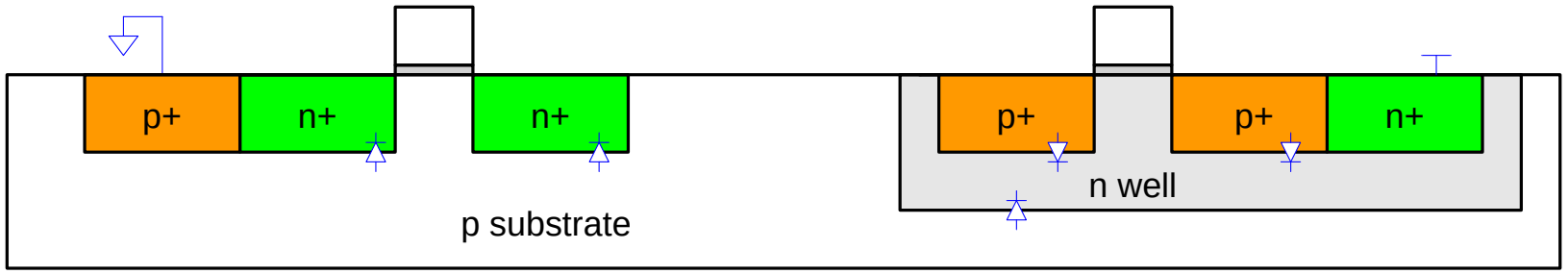
$$V_x = \frac{\eta V_{DD}}{1 + 2\eta + k_\gamma}$$

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

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

Junction Leakage

- Reverse-biased p-n junctions have some leakage
  - Ordinary diode leakage
  - Band-to-band tunneling (BTBT)
  - Gate-induced drain leakage (GIDL)



Diode Leakage

- Reverse-biased p-n junctions have some leakage

$$I_D = I_S \left(e^{\frac{V_D}{v_T}} - 1 \right)$$

- At any significant negative diode voltage, $I_D = -I_s$
- I_s depends on doping levels
  - And area and perimeter of diffusion regions
  - Typically < 1 fA/μm^2 (negligible)

Band-to-Band Tunneling

- Tunneling across heavily doped p-n junctions
  - Especially sidewall between drain & channel when *halo doping* is used to increase V_t
- Increases junction leakage to significant levels

$$I_{BTBT} = WX_j A \frac{E_j}{E_g^{0.5}} V_{dd} \, e^{-B\frac{E_g^{1.5}}{E_j}}$$

$$E_j = \sqrt{\frac{2qN_{halo}N_{sd}}{\varepsilon\left(N_{halo}+N_{sd}\right)}}\left(V_{DD} + v_T \ln\frac{N_{halo}N_{sd}}{n_i^2}\right)$$

  - X_j: sidewall junction depth
  - E_g: bandgap voltage
  - A, B: tech constants

Junction Leakage

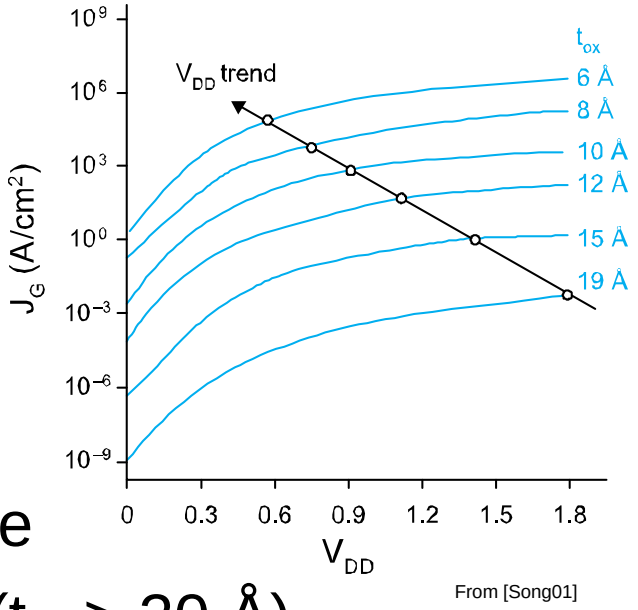
- From reverse-biased p-n junctions
  - Between diffusion and substrate or well
- Ordinary diode leakage is negligible
- Band-to-band tunneling (BTBT) can be significant
  - Especially in high-V_t transistors where other leakage is small
  - Worst at $V_{db} = V_{DD}$
- Gate-induced drain leakage (GIDL)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

Gate Leakage

- Carriers tunnel thorough very thin gate oxides
- Exponentially sensitive to t_{ox} and V_{DD}

$$I_{gate} = WA\left(\frac{V_{DD}}{t_{ox}}\right)^2 e^{-B\frac{t_{ox}}{V_{DD}}}$$

  - A and B are tech constants
  - Greater for electrons
    - So nMOS gates leak more



From [Song01]

- Negligible for older processes (t_{ox} > 20 Å)
- Critically important at 65 nm and below (t_{ox} ≈ 10.5 Å)

Gate Leakage

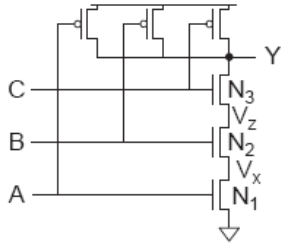
- Extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches subthreshold leakage at 65 nm and below in some processes
- An order of magnitude less for pMOS than nMOS
- Control leakage in the process using t_{ox} > 10.5 Å
  - High-k gate dielectrics help
  - Some 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	0.7	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]

Contention Current

- Static CMOS Circuits have no contention current
- However, pseudo nMOS gates experience contention between the nMOS pulldown and the always-on pMOS pull-ups, when the output is 0.
- Current mode logic and many analog circuits also draw static current
- Such circuits should be turned off in sleep mode by disabling the pull-ups or current source.

Static Power Estimation

- Static power estimation is done by estimating the total width of transistors that are leaking , multiplying by the leakage current per width and multiplying by the fraction of transistors that are in their leaky state (usually one-half)
- Add the contention current if applicable
- The static power is the supply voltage times the static 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

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

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

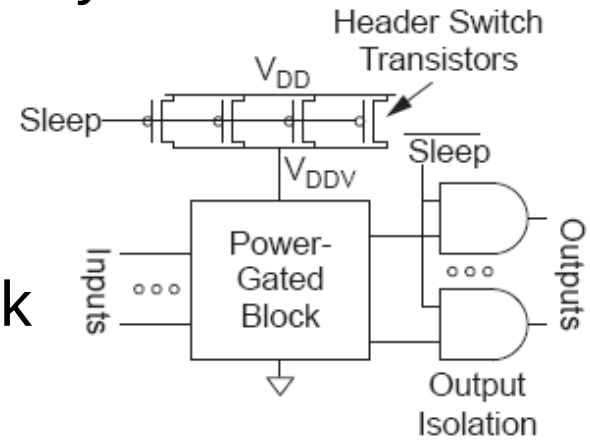
$$I_{sub} = \left[W_{\text{normal-V}_t}\times100\ \text{nA}/\mu\text{m} + W_{\text{high-V}_t}\times10\ \text{nA}/\mu\text{m}\right]/2 = 584\ \text{mA}$$

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

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

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