Lecture 17: Design for Low Power

Slides courtesy of Deming Chen

Slides based on the initial set from David Harris
Outline

- Power and Energy
- Dynamic Power
- Static Power

- Readings: 5.1-5.3
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: $P_{avg} = \frac{E}{T} = \frac{1}{T} \int_0^T P(t)dt$
Power in Circuit Elements

\[ P_{VDD}(t) = I_{DD}(t)V_{DD} \]

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

\[ E_C = \int_{0}^{\infty} I(t)V(t)\,dt = \int_{0}^{\infty} C\frac{dV}{dt}V(t)\,dt \]

\[ = C \int_{0}^{V_c} V(t)\,dV = \frac{1}{2}CV_C^2 \]
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
Example: $V_{DD} = 1.0 \text{ V}, C_L = 150 \text{ fF}, f = 1 \text{ GHz}$
Switching Power

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

\[ = \frac{V_{DD}}{T} \int_0^T i_{DD}(t) dt \]

\[ = \frac{V_{DD}}{T} \left[ T f_{sw} C V_{DD} \right] \]

\[ = C V_{DD}^2 f_{sw} \]
Activity Factor

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where $\alpha$ = activity factor
  - If the signal is a clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = \frac{1}{2}$

- Dynamic power:
  $$P_{\text{switching}} = \alpha CV_{DD}^2 f$$
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

- $P_{\text{total}} = P_{\text{dynamic}} + P_{\text{static}}$
- **Dynamic power:** $P_{\text{dynamic}} = P_{\text{switching}} + P_{\text{shortcircuit}}$
  - Switching load capacitances
  - Short-circuit current
- **Static power:** $P_{\text{static}} = (I_{\text{sub}} + I_{\text{gate}} + I_{\text{junct}} + I_{\text{contention}})V_{\text{DD}}$
  - Subthreshold leakage
  - Gate leakage
  - Junction leakage
  - Contention current
Dynamic Power Example

- 1 billion transistor chip
  - 50M logic transistors
    - Average width: 12 \( \lambda \)
    - Activity factor = 0.1
  - 950M memory transistors
    - Average width: 4 \( \lambda \)
    - Activity factor = 0.02
- 1.0 V 65 nm process (50nm channel length)
- \( C = 1 \text{ fF/\( \mu\text{m} \)} \) (gate) + 0.8 fF/\( \mu\text{m} \) (diffusion)

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

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

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

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

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

- Let $P_i = \text{Prob(node } i = 1)$
  - $\bar{P}_i = 1 - P_i$
- $\alpha_i = \bar{P}_i \cdot P_i$
- Completely random data has $P = 0.5$ and $\alpha = 0.25$
- 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

<table>
<thead>
<tr>
<th>Gate</th>
<th>$P_Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND2</td>
<td>$P_A P_B$</td>
</tr>
<tr>
<td>AND3</td>
<td>$P_A P_B P_C$</td>
</tr>
<tr>
<td>OR2</td>
<td>$1 - \overline{P_A} \overline{P_B}$</td>
</tr>
<tr>
<td>NAND2</td>
<td>$1 - P_A P_B$</td>
</tr>
<tr>
<td>NOR2</td>
<td>$\overline{P_A} \overline{P_B}$</td>
</tr>
<tr>
<td>XOR2</td>
<td>$P_A \overline{P_B} + \overline{P_A} P_B$</td>
</tr>
</tbody>
</table>
Example

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

\[ P = \frac{3}{4} \quad \alpha = \frac{3}{16} \quad P = \frac{1}{16} \quad \alpha = \frac{15}{256} \]

\[ P = \frac{3}{4} \quad \alpha = \frac{3}{16} \]

\[ P = \frac{3}{4} \quad \alpha = \frac{3}{16} \]
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
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

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

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

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

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

\[ P_{\text{static}} = (584 \text{ mA} + 275 \text{ mA})(1.0 \text{ V}) = 859 \text{ mW} \]
Subthreshold Leakage

- For $V_{ds} > 50$ mV

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

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

Typical values in 65 nm:

- $I_{off} =$ 100 nA/\(\mu\)m @ $V_t = 0.3$ V
- $I_{off} =$ 10 nA/\(\mu\)m @ $V_t = 0.4$ V
- $I_{off} =$ 1 nA/\(\mu\)m @ $V_t = 0.5$ V

- $\eta = 0.1$
- $k_{\gamma} = 0.1$
- $S = 100$ mV/decade
Stack Effect

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

\[
I_{sub} = I_{off} 10^{\frac{\eta(V_x - V_{DD})}{4}} \frac{S}{N1} = I_{off} 10^{\frac{-V_x + \eta((V_{DD} - V_x) - V_{DD}) - k_y V_x}{4}} \frac{S}{N2}
\]

\[
V_x = \frac{\eta V_{DD}}{1 + 2\eta + k_y} \left(1 + \frac{1 + \eta + k_y}{1 + 2\eta + k_y}\right)
\]

\[
I_{sub} = I_{off} 10^\frac{-\eta V_{DD}}{S} \approx I_{off} 10^\frac{-\eta V_{DD}}{S}
\]

- 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 $V_b$
    • *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
- $I_{gn} = 6.3 \text{ nA}$
- $I_{gp} = 0$
- $I_{offn} = 5.63 \text{ nA}$
- $I_{offp} = 9.3 \text{ nA}$

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

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
Summary

- Power is the limiting factor for modern circuits design
- Power estimation
  - Dynamic power: main challenge -- switching activities and capacitances
  - Static power: various physics-related factors
- Power minimization

- Next lecture: VLSI Design Styles