Lecture 20 – Low Power Design

Borivoje Nikolić, Vladimir Stojanović, Sophia Shao

IEEE MICRO, Nov/Dec 2021
Microprocessor at 50: Looking
Back and Looking Forward
Special issue on 50 years of
a microprocessor

Advertisement in the Electronics News Weekly in November 1971
announcing the Intel 4004.
Recap

• Power is a primary design constraint
  • In both cloud and edge systems

• Excess performance traded off for power savings
Architectural Optimizations
Optimal Processors

• Processors used to be optimized for performance
  • Optimal logic depth was found to be 8-11 FO4 delays in superscalar processors
  • 1.8-3 FO4 in sequentials, rest in combinatorial
    • Kunkel, Smith, ISCA’86
    • Hriskesh, Jouppi, Farkas, Burger, Keckler, Shivakumar, ISCA’02
    • Harstein, Puzak, ISCA’02
    • Sprangle, Carmean, ISCA’02

• But those designs have very high power dissipation
  • Need to optimize for both performance and power/energy
From System View: What is the Optimum?

• How do sensitivities relate to more traditional metrics:
  • Power per operation (MIPS/W, GOPS/W, TOPS/W)
  • Energy per operation (Joules per op)
  • Energy-delay product

• Can be reformatted as a goal of optimizing power x delay^n
  • n = 0 – minimize power per operation
  • n = 1 – minimize energy per operation
  • n = 2 – minimize energy-delay product
  • n = 3 – minimize energy-(delay)^2 product
Optimization Problem

• Set up optimization problem:
  • Maximize performance under energy constraints
  • Minimize energy under performance constraints

• Or minimize a composite function of $E^n D^m$
  • What are the right $n$ and $m$?

• $n = 1, m = 1$ is EDP – improves at lower $V_{DD}$

• $n = 1, m = 2$ is invariant to $V_{DD}$
  • $E \sim CV_{DD}^2$
  • $D \sim 1/V_{DD}$
Hardware Intensity

• Introduced by Zyuban and Strenski in 2002.
• Measures where is the design on the Energy-Delay curve
• Parameter in cost function optimization

\[ F_c = \left(\frac{E}{E_0}\right)\left(\frac{D}{D_0}\right)^\eta \quad 0 \leq \eta < +\infty, \]

\[ \eta = -\frac{D\partial E}{E\partial D} \bigg|_{v}, \]

Slope of the optimal E-D curve at the chosen design point
Optimum Across Hierarchy Layers

Optimal logic depth in pipelined processors is $\sim 18\text{FO4}$
Relatively flat in the 16-22FO4 range

Zyuban et al, TComp’04
Architectural Tradeoffs

• H, Mair, ISSCC’20
Architectural Tradeoffs: Tri-Gear

- **HP**: High performance (ARM Cortex A78, optimized for speed, 3.0GHz)
- **BP**: Balanced performance (ARM Cortex A78, optimized for power, 2.6GHz)
- **HE**: High efficiency (ARM A55, 2.0GHz)
Announcements

• Quiz 2 today
• Homework 3 due next week
Circuit-Level Tradeoffs
Alpha-Power Based Delay Model

\[ t_{pi} = \frac{K_d V_{DD}}{(V_{DD} - V_{Th})^\alpha \left(1 + \frac{C_{L,i}}{C_{in,i}}\right)} \]

\[ D = \sum t_{pi} = \sum \frac{K_d V_{DD}}{(V_{DD} - V_{Th})^\alpha \left(1 + \frac{W_{L,i}}{W_{in,i}}\right)} \]
Energy Models

♦ **Switching**

\[ E_{SW} = \alpha_0 \rightarrow_1 \left( C_{L,i} + C_{int,i} \right) V_{DD}^2 \]

♦ **Leakage**

\[ E_{Lk} = W_{Ln} I_0 e^{-\frac{(V_{Th} - \gamma V_{DD})}{nV_t}} V_{DD}\]
Sizing, Supply, Threshold Optimization

• Transistor sizing can yield large power savings with small delay penalties
  • Gate sizing
  • Beta-ratio adjustments \( \beta = W_p/W_n \)
  • (Stack resizing)

• Supply voltage affects both active and leakage energy

• Threshold voltage affects primarily the leakage
Apply to Sizing of an Inverter Chain

Unconstrained energy: find min $D = \Sigma t_{pi}$

$$C_{gin,j} = \sqrt{C_{gin,j-1}C_{gin,j+1}} \quad W_j = \sqrt{W_{j-1}W_{j+1}}$$

Constrained energy: find min $D$, under $E < E_{max}$

Where $E = \Sigma e_i$
Constrained Optimization

• Find min($D$) subject to $E = E_{\text{max}}$
  • Constrained function minimization

• E.g. Lagrange multipliers

\[ \Lambda(x) = D(x) + \lambda(E(x) - E_{\text{max}}) \]

\[ \frac{\partial \Lambda}{\partial x} = 0 \]

• Can solve analytically for $x = W_{ij}, V_{DD}, V_{Th}$

Or dual:

\[ K(x) = E(x) + \lambda(D - D_{\text{max}}) \]
**Inverter Chain: Sizing Optimization**

- **Variable taper achieves minimum energy**
- **Reduce number of stages at large $d_{\text{inc}}$**

\[ W_j = \sqrt{W_{j-1}W_{j+1}} / (1 + \lambda W_{j-1}) \]

[Ma, Franzon, *IEEE JSSC*, 9/94]

\[ \lambda = -\frac{2KV_{DD}^2}{\tau_{\text{nom}}S_W} \]

\[ S_W \propto \frac{e_j}{f_j - f_{j-1}} \]

Stojanovic, ICCAD’02

**effective fanout, $f$**

<table>
<thead>
<tr>
<th>stage</th>
<th>0%</th>
<th>1%</th>
<th>10%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$e_i$ – energy per stage
$f_i$ – fanout per stage

---

EECS251B L20 LOW-POWER DESIGN
Sensitivity to Sizing and Supply

- **Gate sizing** \((W_i)\)

\[
\frac{\partial E_{sw}}{\partial W_j} \cdot \frac{\partial D}{\partial W_j} = \frac{e_j}{\tau_{nom}(f_j - f_{j-1})}
\]

- **Supply voltage** \((V_{dd})\)

\[
\frac{\partial E_{sw}}{\partial V_{dd}} \cdot \frac{\partial D}{\partial V_{dd}} = \frac{E_{sw}}{D} \cdot \frac{2}{\alpha - 1 + x_v}
\]

\[
x_v = \frac{(V_{Th} + \Delta V_{Th})}{V_{dd}}
\]

\(\infty\) for equal \(f_{eff}\)

\((D_{min})\)
Sensitivity to $V_{th}$

- Threshold voltage ($V_{th}$)

\[
\frac{\partial E}{\partial \Delta V_{Th}} = P_{Lk} \left( \frac{V_{DD} - V_{Th} - \Delta V_{Th}}{\alpha n V_{t}} - 1 \right)
\]

Low initial leakage

$\Rightarrow$ speedup comes for “free”
Scaling Supplies
Reducing \( V_{dd} \)

\[
P \times t_d = E_t = C_L \times V_{dd}^2
\]

\[
\frac{E_{(V_{dd}=2)}}{E_{(V_{dd}=5)}} = \frac{(C_L) \times (2)^2}{(C_L) \times (5)^2}
\]

\[
E_{(V_{dd}=2)} \approx 0.16 \times E_{(V_{dd}=5)}
\]

- Strong function of voltage \((V^2\) dependence).
- Relatively independent of logic function and style.
- Power Delay Product Improves with lowering \( V_{DD} \).

Chandrakasan, JSSC'92
Lower $V_{DD}$ Increases Delay

\[ T_d = \frac{C_L \cdot V_{dd}}{I} \]

\[ I \sim (V_{dd} - V_t)^2 \]

\[
\frac{T_d(V_{dd}=2)}{T_d(V_{dd}=5)} = \frac{(2) \cdot (5 - 0.7)^2}{(5) \cdot (2 - 0.7)^2} \\
\approx 4
\]

- Relatively independent of logic function and style.
Trade-off Between Power and Delay

\[ \text{Power} = a \cdot f \cdot C \cdot V_{DD}^2 + I_0 \cdot 10^{-10} \cdot \frac{V_{TH}}{s} \cdot V_{DD} \]

\[ \text{Delay} \propto \frac{C \cdot V_{DD}}{(V_{DD} - V_{TH})^{1.3}} \]

50nm node, FO3 INV
Architecture Trade-off for Fixed-rate Processing
Reference Datapath

- Critical path delay $\Rightarrow T_{\text{adder}} + T_{\text{comparator}} (= 25\text{ns})$
  $\Rightarrow f_{\text{ref}} = 40\text{Mhz}$
- Total capacitance being switched $= C_{\text{ref}}$
- $V_{\text{dd}} = V_{\text{ref}} = 5\text{V}$
- Power for reference datapath $= P_{\text{ref}} = C_{\text{ref}} V_{\text{ref}}^2 f_{\text{ref}}$
  from [Chandakasan92] (IEEE JSSC)
Parallel Datapath

- The clock rate can be reduced by half with the same throughput \( f_{\text{par}} = f_{\text{ref}} / 2 \)
- \( V_{\text{par}} = V_{\text{ref}} / 1.7 \), \( C_{\text{par}} = 2.15C_{\text{ref}} \)
- \( P_{\text{par}} = (2.15C_{\text{ref}}) \left( V_{\text{ref}} / 1.7 \right)^2 (f_{\text{ref}}/2) \approx 0.36 \ P_{\text{ref}} \)

Area = 1476 x 1219 \( \mu \)²
Pipelined Datapath

- Critical path delay is less $\Rightarrow \max [T_{\text{adder}}, T_{\text{comparator}}]$
- Keeping clock rate constant: $f_{\text{pipe}} = f_{\text{ref}}$
  Voltage can be dropped $\Rightarrow V_{\text{pipe}} = V_{\text{ref}} / 1.7$
- Capacitance slightly higher: $C_{\text{pipe}} = 1.15 C_{\text{ref}}$
- $P_{\text{pipe}} = (1.15 C_{\text{ref}}) (V_{\text{ref}}/1.7)^2 f_{\text{ref}} \approx 0.39 P_{\text{ref}}$

Area = 640 x 1081 $\mu^2$
## A Simple Datapath: Summary

<table>
<thead>
<tr>
<th>Architecture type</th>
<th>Voltage</th>
<th>Area</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple datapath (no pipelining or parallelism)</td>
<td>5V</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pipelined datapath</td>
<td>2.9V</td>
<td>1.3</td>
<td>0.39</td>
</tr>
<tr>
<td>Parallel datapath</td>
<td>2.9V</td>
<td>3.4</td>
<td>0.36</td>
</tr>
<tr>
<td>Pipeline-Parallel</td>
<td>2.0V</td>
<td>3.7</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Multiple Supplies
Multiple Supply Voltages

• Block-level supply assignment (“power domains” or “voltage islands”)
  • Higher throughput/lower latency functions are implemented in higher $V_{DD}$
  • Slower functions are implemented with lower $V_{DD}$
  • Often called “Voltage islands”
  • Separate supply grids, level conversion performed at block boundaries

• Multiple supplies inside a block
  • Non-critical paths moved to lower supply voltage
  • Level conversion within the block
  • Physical design challenging
  • (Not used in practice)
Power Domains
Practical Examples

• Intel Skylake (ISSCC’16)
  • Four power planes indicated by colors
Practical Examples

- Intel 28-core Skylake-SP (ISSCC’18)

- 9 primary VCC domains are partitioned into 35 VCC planes
Leakage Issue

- Driving from $V_{DDL}$ to $V_{DDH}$

Level converter

\[ |V_{GS}| > V_{Th} \quad \rightarrow \quad I_{Leak} \]

\[ V_{OH} < V_{DDH} \]

\[ V_{DDH} \]

\[ V_{in} \quad \rightarrow \quad V_{DDL} \]

\[ V_{out} \]
Multiple Supplies Within A Block

- Downsizing, lowering the supply on the critical path will lower the operating frequency
- Downsize (lowering supply) non-critical paths
  - Narrows down the path delay distribution
  - Increases impact of variations
Multiple Supplies in a Block

Conventional Design

CVS Structure

Lower $V_{DD}$ portion is shaded

“Clustered voltage scaling”
Multiple Supplies in a Block

CVS

Layout:

Usami’98
Level-Converting Flip-Flop
Summary

• Power-performance tradeoffs
  • Sizing
  • Supplies
  • Thresholds

• Lowering supplies

• Multiple supply voltages
Next Lecture

• Low-power design
  • Dynamic voltage-frequency scaling
  • Clock gating