April 2, AnandTech: Intel Details 10th Gen Comet Lake-H for 45 W Notebooks: Up to 5.3 GHz*

*This CPU can hit this frequency on two cores, when the system is both within its secondary power limits but also Intel’s Thermal Velocity Boost is enabled, which means there has to be additional thermal headroom in the system (and it has to be enabled by the OEM). This allows the CPU to go from 5.1 GHz to 5.3 GHz. Every Intel Thermal Velocity Boost enabled CPU requires OEM support in order to get those extra two bins on the single core frequency.
Announcements

• Assignment 3 due today, April 2.
• Quiz next Tuesday, end of class
Outline

• Module 5
  • Circuit-level power-performance tradeoffs
  • Reducing supply voltage
Architectural Tradeoffs

- H, Mair, ISSCC'20

![Graph showing architectural tradeoffs between performance and power for Cortex-A77 and Cortex-A55.]
5.D 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**

\[ E_{Lk} = W \ln I_0 e^{-\frac{(V_{Th} - \gamma V_{DD})}{nV_t}} V_{DD} D \]
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 $D = \sum t_{pi}$

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

Constrained energy: find min $D$, under $E < E_{max}$
Where $E = \sum e_i$
Constrained Optimization

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

- E.g. Lagrange multipliers

Or dual:

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

$$\frac{\partial \Lambda}{\partial x} = 0$$

$$K(x) = E(x) + \lambda (D - D_{\text{max}})$$

- Can solve analytically for $x = W_j, V_{DD}, V_{Th}$
Inverter Chain: Sizing Optimization
Inverter Chain: Sizing Optimization

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

\[ W_j = \sqrt{\frac{W_{j-1} W_{j+1}}{1 + \lambda W_{j-1}}} \]

\[ \lambda = -\frac{2K V_{DD}^2}{\tau_{nom} S_W} \]

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

Stojanovic, ICCAD’02

[Ma, Franzon, IEEE JSSC, 9/94]
Sensitivity to Sizing and Supply

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

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

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

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

\[x_v = (V_{th} + \Delta V_{th})/V_{dd}\]
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”
## Power/Energy Optimization Space

<table>
<thead>
<tr>
<th>Energy</th>
<th>Constant Throughput/Latency</th>
<th>Variable Throughput/Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Active</strong></td>
<td>Design Time</td>
<td>Sleep Mode</td>
</tr>
<tr>
<td>Logic design</td>
<td>Clock gating</td>
<td>DFS, DVS</td>
</tr>
<tr>
<td>Scaled $V_{DD}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans. sizing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi-$V_{DD}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Leakage</strong></td>
<td>Stack effects</td>
<td>Sleep T's</td>
</tr>
<tr>
<td>Trans sizing</td>
<td>Multi-$V_{DD}$ Variable $V_{Th}$</td>
<td>Variable $V_{Th}$</td>
</tr>
<tr>
<td>Scaling $V_{DD}$</td>
<td>Variable $V_{Th}$</td>
<td></td>
</tr>
<tr>
<td>+ Multi-$V_{Th}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Energy-Performance Tradeoffs

<table>
<thead>
<tr>
<th>Enable Time/Perf. Impact</th>
<th>Design Time</th>
<th>Run Time</th>
</tr>
</thead>
</table>
| Near-zero perf. penalty  | Clock gating  
Architectural switching reduction  
Multi-$V_{Th}$ | Dynamic $V_{DD}$  
Dynamic $V_{Th}$ |
| True tradeoffs           | Fine-granularity clock gating  
$V_{DD}$, $V_{TH}$ adjustments  
Multi-$V_{DD}$  
Sizing, logic styles  
Stack forcing | Power gating |
5.E Scaling Supplies
## Power/Energy Optimization Space

<table>
<thead>
<tr>
<th>Energy</th>
<th>Constant Throughput/Latency</th>
<th>Variable Throughput/Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Design Time</td>
<td>Sleep Mode</td>
</tr>
<tr>
<td>Active</td>
<td>Logic design</td>
<td>Clock gating</td>
</tr>
<tr>
<td></td>
<td>Scaled $V_{DD}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trans. sizing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multi-$V_{DD}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stack effects</td>
<td>Sleep T’s</td>
</tr>
<tr>
<td></td>
<td>Trans sizing</td>
<td>Multi-$V_{DD}$</td>
</tr>
<tr>
<td></td>
<td>Scaling $V_{DD}$</td>
<td>+ Input control</td>
</tr>
<tr>
<td></td>
<td>+ Multi-$V_{Th}$</td>
<td></td>
</tr>
</tbody>
</table>
Supply Voltage Adjustment

• How to maintain throughput under reduced supply?

• Introducing more parallelism/pipelining
  • Area increase
  • Cost/power tradeoff

• Multiple voltage domains
  • Separate supply voltages for different blocks
  • Lower VDD for slower blocks
  • Cost of DC-DC converters

• Dynamic voltage scaling – with variable throughput

• Reducing $V_{TH}$ to improve speed
  • Leakage issues
Reducing $V_{dd}$

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

- 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
Reducing $V_{DD}$

32nm process

![Graph showing the relationship between $V_{DD}$ and various parameters such as delay, switching power, and leakage power.](image-url)
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

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

Power [W/gate] vs. Delay [ps] for 50nm node, FO3 INV

Equi-delay

\[ P = k \cdot C \cdot V_{DD} \cdot V_{TH} \]

\[ f \cdot C \cdot V_{DD} \cdot 10^{\frac{V_{DD}}{V_{TH}}} \]
Two Types of Processing

• Fixed-rate processing (e.g. signal processing for multimedia or communications)
  • Stream-based computation
  • No advantage in obtaining throughput in excess of the real-time constraint

• Variable-rate or burst-mode computation (e.g. general purpose computation)
  • Mostly idle (or low-load) with bursts of computation
  • Faster is better
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 = $C_{\text{ref}}$
- $V_{dd} = V_{\text{ref}} = 5\text{V}$
- Power for reference datapath = $P_{\text{ref}} = C_{\text{ref}} V_{\text{ref}}^2 f_{\text{ref}}$

from [Chandarakasan92] (IEEE JSSC)
Parallel Datapath

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

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

• Low-power design
  • Multiple supplies
  • Dynamic voltage scaling