Biden Ups Ante to $50 Billion for CHIPS Act
By Alan Patterson  EETimes 04.01.2021

US President Joe Biden’s $2 trillion infrastructure plan announced this week increases the amount of funding that would be allocated to revive the American semiconductor industry to $50 billion.

Announcements

• Assignment 3 due this week
Outline

- Power performance tradeoffs
  - Architecture
  - Circuit-level tradeoffs
  - Lowering supply voltages

5. C Architectural Optimization
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

\[
F_c = (E/E_0)(D/D_0)^\eta \quad 0 \leq \eta < +\infty,
\]

\[
\eta = -\left. \frac{D\delta E}{E_0 D} \right|_y
\]

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

Optimal logic depth in pipelined processors is ~18FO4
Relatively flat in the 16-22FO4 range

Architectural Tradeoffs

• H, Mair, ISSCC’20
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**

\[ E_{SW} = \alpha_0 \to 1 \left(C_{L,i} + C_{int,i}\right)V_{DD}^2 \]

♦ **Leakage**

\[ E_{Lk} = W_{in} I_0 e^{-\frac{(V_{in} - \gamma V_{DD})}{nV_s}} V_{DD} D \]

Sizing, Supply, Threshold Optimization

- Transistor sizing can yield large power savings with small delay penalties
  - Gate sizing
  - Beta-ratio adjustments \( \beta = Wp/Wn \)
  - (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 \( \min D = \sum t_{pi} \)

\[ C_{gn,j} = \sqrt{C_{gn,j-1}C_{gn,j+1}} \quad W_j = \sqrt{W_{j-1}W_{j+1}} \]

Constrained energy: find \( \min D \), under \( E < E_{\text{max}} \)

Where \( E = \sum e_i \)

Constrained Optimization

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

• E.g. Lagrange multipliers

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

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

• Can solve analytically for \( x = W_j, V_{DD}, V_{Th} \)
Inverter Chain: Sizing Optimization

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

$W_j = \frac{W_{j-1} - W_{j+1}}{1 + \lambda W_{j-1}}$

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

$\lambda = -\frac{2KVD^2}{\tau_{nom}S_W}$

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

Stojanovic, ICCAD'02
### Sensitivity to Sizing and Supply

- **Gate sizing** ($W_j$)
  \[
  \frac{\partial E_{sw}}{\partial W_j} = -\frac{e_j}{\tau_{nom} (f_j - f_{j-1})}.
  \]

- **Supply voltage** ($V_{dd}$)
  \[
  \frac{\partial E_{sw}}{\partial V_{DD}} = E_{sw} \frac{1 - x_v}{D} \frac{2}{\alpha - 1 + x_v}.
  \]
  \[
  x_v = \frac{(V_{th} + \Delta V_{th})}{V_{dd}}
  \]

\[\infty \text{ for equal } f_{eff} \text{ (D}_{min}\text{)}\]

### 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**

\[\Rightarrow \text{speedup comes for “free”}\]
### 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></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><strong>Leakage</strong></td>
<td>Stack effects</td>
<td>Sleep T's</td>
</tr>
<tr>
<td></td>
<td>Trans sizing</td>
<td>Multi-$V_{DD}$ Variable $V_{Th}$</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>

### Energy-Performance Tradeoffs

<table>
<thead>
<tr>
<th>Enable Time/ Perf. Impact</th>
<th>Design Time</th>
<th>Run Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Near-zero perf. penalty</strong></td>
<td>Clock gating</td>
<td>Dynamic $V_{DD}$</td>
</tr>
<tr>
<td></td>
<td>Architectural switching</td>
<td>Dynamic $V_{Th}$</td>
</tr>
<tr>
<td></td>
<td>reduction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multi-$V_{Th}$</td>
<td></td>
</tr>
<tr>
<td><strong>True tradeoffs</strong></td>
<td>Fine-granularity clock</td>
<td>Power gating</td>
</tr>
<tr>
<td></td>
<td>gating</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$V_{DD}$, $V_{Th}$ adjustments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multi-$V_{DD}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sizing, logic styles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stack forcing</td>
<td></td>
</tr>
</tbody>
</table>
### 5.5 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><strong>Active</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logic design</td>
<td></td>
<td>Clock gating</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></td>
<td></td>
</tr>
<tr>
<td>Stack effects</td>
<td></td>
<td>Sleep T's</td>
</tr>
<tr>
<td>Trans sizing</td>
<td></td>
<td>Multi-$V_{DD}$ Variable $V_{Th}$</td>
</tr>
<tr>
<td>Scaling $V_{DD}$</td>
<td></td>
<td>+ Input control</td>
</tr>
<tr>
<td>+ Multi-$V_{Th}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- DFS: Dynamic Frequency Scaling
- DVS: Dynamic Voltage Scaling
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 $V_{dd}$

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

32nm process

![Graph showing the relationship between $V_{DD}$ and delay, switching power, and leakage power.]

Lower $V_{DD}$ Increases Delay

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

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

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

- Relatively independent of logic function and style.
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 \( T_{\text{adder}} + T_{\text{comparator}} = 25 \text{ns} \)
  \[ \Rightarrow f_{\text{ref}} = 40 \text{MHz} \]
- 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 [Chandrakasan92] (IEEE JSSC)

**Parallel Datapath**

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

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

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-frequency scaling