CS184a: Computer Architecture
(Structures and Organization)

Day6: October 11, 2000
Instruction Taxonomy
VLSI Scaling

Last Time

- Computing requirements
- Instruction requirements
- Structure
Today

- Instruction Taxonomy
- VLSI Scaling

Instruction Distribution

- Beyond 64 PE, instruction bandwidth dictates PE size

\[
\frac{\sqrt{\text{PE}_{\text{area}}} \times 4 \times \sqrt{N}}{(64 \times 8 \lambda)} = N
\]

\[
\text{PE}_{\text{area}} = 16K\lambda^2 \times N
\]

- Build larger arrays
  \[\Rightarrow\] processing elements become less dense
Instruction Memory Requirements

• **Idea:** put instruction memory in array

• **Problem:** Instruction memory can quickly dominate area, too
  
  – Memory Area = 64 × 1.2Kλ²/instruction

  – PE_{area} = 1Mλ² + (Instructions) × 80Kλ²

Instruction Pragmatics

• Instruction requirements *could* dominate array size.

• Standard architecture trick:
  
  – Look for structure to exploit in “typical computations”
Two Extremes

- **SIMD Array** (microprocessors)
  - Instruction/cycle
  - share instruction across array of PEs
  - uniform operation in space
  - operation variance in time

- **FPGA**
  - Instruction/PE
  - assume temporal locality of instructions (same)
  - operation variance in space
  - uniform operations in time

Hybrids

- **VLIW (SuperScalar)**
  - Few *pins* / cycle
  - Share instruction across *w* bits

- **DPGA**
  - Small instruction store / PE
## Architecture Instruction Taxonomy

<table>
<thead>
<tr>
<th>Control Threads (PCs)</th>
<th>$p$/insts per Control Thread</th>
<th>$n$/insts per Control Thread</th>
<th>$c$/$w$ Granularity</th>
<th>$m$/$n$ Granularity</th>
<th>Architecture/Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>n/a</td>
<td>Hardwired Functional Unit (e.g., ECC/EDC Unit, FP MPY)</td>
<td>1</td>
<td>FPGA</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>$n_e$/$w$</td>
<td>Reconfigurable ALUs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>1</td>
<td>DPGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>n</td>
<td>c $n_x$/$w$</td>
<td>HSRA/SCORE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>$c/w$ Vector Processors</td>
<td>MSIMD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>VEGA</td>
<td>PADDI-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>w</td>
<td>MIMD (traditional)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Instruction Message

- Architectures fall out of:
  - general model too expensive
  - look for structure in common problems
  - exploit structure to reduce resource requirements
- Architectures can be viewed in a unified design space
VLSI Scaling

Why Care?

• In this game, we must be able to predict the future
• Rapid technology advance
• Reason about changes and trends
• re-evaluate prior solutions given technology at time X.
Why Care

• Cannot compare against what competitor does today
  – but what they can do at time you can ship

• Careful not to fall off curve
  – lose out to someone who can stay on curve

Scaling

• **Premise:** features scale “uniformly”
  – everything gets better in a predictable manner

• **Parameters:**
  \( \lambda \) (lambda) -- Mead and Conway (class)
  – S -- Bohr
  – \( 1/\kappa \) -- Dennard
Feature Size

λ is half the minimum feature size in a VLSI process

[minimum feature usually channel width]

Scaling

- Channel Length (L)
- Channel Width (W)
- Oxide Thickness (T_{ox})
- Doping (N_a)
- Voltage (V)
Scaling

- Channel Length (L) \( \lambda \)
- Channel Width (W) \( \lambda \)
- Oxide Thickness (\( T_{\text{ox}} \)) \( \lambda \)
- Doping (\( N_a \)) \( 1/\lambda \)
- Voltage (V) \( \lambda \)

Effects?

- Area
- Capacitance
- Resistance
- Threshold (\( V_{\text{th}} \))
- Current (\( I_d \))
- Gate Delay (\( \tau_{\text{gd}} \))
- Wire Delay (\( \tau_{\text{wire}} \))
- Power
Area

- $\lambda \rightarrow \lambda/\kappa$
- $A = L \times W$
- $A \rightarrow A/\kappa^2$

- $0.35\mu m \rightarrow 0.25\mu m$
- 50% area
- 2x capacity same area

Area Perspective

[2000 tech.]
18mm x 18mm
0.18$\mu m$
60G $\lambda^2$
Capacitance

- Capacitance per unit area
  - \( C_{ox} = \varepsilon_{SiO_2} / T_{ox} \)
  - \( T_{ox} \rightarrow T_{ox} / \kappa \)
  - \( C_{ox} \rightarrow \kappa C_{ox} \)

Capacitance

- Gate Capacitance
  - \( C_{gate} = A * C_{ox} \)
  - \( A \rightarrow A / \kappa^2 \)
  - \( C_{ox} \rightarrow \kappa C_{ox} \)
  - \( C_{gate} \rightarrow C_{gate} / \kappa \)
Threshold Voltage

Before:

\[ V_{th} = \frac{1}{C_{OX}} \left( -Q_{eff} + \left( 2\epsilon_{ox} \cdot q \cdot N_A (\phi_s + V_{sub}) \right)^{1/2} \right) + (W_f + \phi_b) \]

\( (W_f + \phi_b) \approx 0 \)

Adjust \( V_{sub} \) so \( (\phi_s + V_{sub}) \to \left( \phi_s + \frac{V_{sub}}{\kappa} \right) \)

After:

\[ V'_{th} = \frac{1}{\kappa C_{OX}} \left( -Q_{eff} + \left( 2\epsilon_{ox} \cdot q \cdot N_A \left( \phi_s + \frac{V_{sub}}{\kappa} \right) \right)^{1/2} \right) \]

\[ V'_{th} \approx \frac{V_{th}}{\kappa} \]

Threshold Voltage

- \( V_{TH} \to V_{TH} / \kappa \)
Current

- Saturation Current
  \[ I_d = \left( \mu C_{\text{ox}}/2 \right)(W/L)(V_{gs} - V_{TH})^2 \]
  - \( V_{gs} \rightarrow V/\kappa \)
  - \( V_{TH} \rightarrow V_{TH}/\kappa \)
  - \( W \rightarrow W/\kappa \)
  - \( C_{\text{ox}} \rightarrow \kappa C_{\text{ox}} \)
  - \( I_d \rightarrow I_d/\kappa \)

Gate Delay

\[ \tau_{gd} = Q/I = (CV)/I \]

- \( V \rightarrow V/\kappa \)
- \( I_d \rightarrow I_d/\kappa \)
- \( C \rightarrow C/\kappa \)

\[ \tau_{gd} \rightarrow \tau_{gd}/\kappa \]
Resistance

- \( R = \frac{\rho L}{W \cdot t} \)
- \( W \rightarrow \frac{W}{\kappa} \)
- \( L, t \) similar
- \( R \rightarrow \kappa R \)

Wire Delay

\[ \tau_{\text{wire}} = R \cdot L \cdot C \]

- \( R \rightarrow \kappa R \)
- \( C \rightarrow \frac{C}{\kappa} \)
- \( \tau_{\text{wire}} \rightarrow \tau_{\text{wire}} \)

- ...assuming (logical) wire lengths remain constant...
Power Dissipation (Static)

- Resistive Power
  - \( P = V \times I \)
  - \( V \rightarrow V/\kappa \)
  - \( I_d \rightarrow I_d/\kappa \)
  - \( P \rightarrow P/\kappa^2 \)

Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - \( P = (1/2)CV^2f \)
  - \( V \rightarrow V/\kappa \)
  - \( C \rightarrow C/\kappa \)
  - \( P \rightarrow P/\kappa^3 \)

- Increase Frequency?
  - \( f \rightarrow \kappa f \)
  - \( P \rightarrow P/\kappa^2 \)
Effects?

- Area \( \frac{1}{\kappa^2} \)
- Capacitance \( \frac{1}{\kappa} \)
- Resistance \( \kappa \)
- Threshold \( V_{th} \) \( \frac{1}{\kappa} \)
- Current \( I_d \) \( \frac{1}{\kappa} \)
- Gate Delay \( \tau_{gd} \) \( \frac{1}{\kappa} \)
- Wire Delay \( \tau_{wire} \) \( 1 \)
- Power \( \frac{1}{\kappa^2} \rightarrow \frac{1}{\kappa^3} \)

Delays?

- If delays in gates/switching?
- If delays in interconnect?
- Logical interconnect lengths?
Delays?

• If delays in gates/switching?
  – Delay reduce with $1/\kappa [\lambda]$

Delays

• Logical capacities growing
• Wirelengths?
  – No locality $\rightarrow \kappa$
  – Rent’s Rule
    • $L \rightarrow n^{(p-0.5)}$
    • [$p>0.5$]
Capacity

- Rent: \( IO = C \cdot N^p \)
- \( p > 0.5 \)
- \( A = C \cdot N^{2p} \)
- Logical Area \( \rightarrow \kappa^2 \)
  - \( \kappa^2 A = C \cdot N_2^{2p} \)
  - \( \kappa^2 N^{2p} = N_2^{2p} \)
  - \( N_2 = \kappa^{(1/p)} N \)
- Sanity Check
  - \( p = 1 \)
  - \( N_2 = \kappa N \)
  - \( p \sim 0.5 \)
  - \( N_2 \sim \kappa^2 N \)

Compute Density

- Density = compute / (Area * Time)
- \( \kappa^3 \) > compute density scaling > \( \kappa \)
  - \( \kappa^3 \): gates dominate, \( p < 0.5 \)
  - \( \kappa^2 \): moderate \( p \), good fraction of gate delay
  - \( \kappa \): large \( p \) (wires dominate area and delay)
Power Density

- $P \rightarrow P/\kappa^2$ (static, or increase frequency)
- $P \rightarrow P/\kappa^3$ (dynamic, same freq.)
  \[ A \rightarrow A/\kappa^2 \]
- $P/A \rightarrow P/A \ldots$ or $\ldots P/\kappa A$

Physical Limits

- Doping?
- Features?
Physical Limits

• Depended on
  – bulk effects
    • doping
    • current (many electrons)
    • mean free path in conductor
  – localized to conductors

• Eventually
  – single electrons, atoms
  – distances close enough to allow tunneling

Finishing Up...
Big Ideas
[MSB Ideas]

• Instruction organization induces a design space (taxonomy) for programmable architectures

• Moderately predictable VLSI Scaling
  – unprecedented capacities/capability growth for engineered systems
  – change
  – be prepared to exploit
  – account for in comparing across time

Big Ideas
[MSB-1 Ideas]

• Uniform scaling reasonably accurate for past couple of decades
• Area increase $\kappa^2$
  – Real capacity maybe a little less?
• Gate delay decreases ($1/\kappa$)
• Wire delay not decrease, maybe increase
• Overall delay decrease less than ($1/\kappa$)