ESE534: Computer Organization

Day 11: February 20, 2012
Instruction Space Modeling

Last Time

• Instruction Requirements
• Instruction Space

Architecture Taxonomy

<table>
<thead>
<tr>
<th>PCs</th>
<th>Pints/PC</th>
<th>depth</th>
<th>width</th>
<th>Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>N</td>
<td>1</td>
<td>1</td>
<td>FPGA</td>
</tr>
<tr>
<td>1</td>
<td>N (48,640)</td>
<td>8</td>
<td>1</td>
<td>Tabula ABAX (A1EC04)</td>
</tr>
<tr>
<td>1</td>
<td>1024</td>
<td>32</td>
<td></td>
<td>Scalar Processor (RISC)</td>
</tr>
<tr>
<td>1</td>
<td>N</td>
<td>D</td>
<td>W</td>
<td>VLIW (superscalar)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Small</td>
<td>W*N</td>
<td>SIMD, GPU, Vector</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>(47)</td>
<td>2048</td>
<td>16-core</td>
</tr>
<tr>
<td>N</td>
<td>D</td>
<td>W</td>
<td></td>
<td>MIMD</td>
</tr>
</tbody>
</table>

Today

• Model Architecture from Instruction Parameters
  – implied costs
  – gross application characteristics

Quotes

• If it can’t be expressed in figures, it is not science; it is opinion.  -- Lazarus Long

Modeling

• Why do we model?
Motivation

• Need to understand
  – How costly is a solution
    • Big, slow, hot, energy hungry….
  – How compare to alternatives
  – Cost and benefit of flexibility

What we really want:

• Complete implementation of our application
• For each architectural alternatives
  – In same implementation technology
  – w/ multiple area-time points

Reality

• Seldom get it packaged that nicely
  – much work to do so
  – technology keeps moving
• We must deal with
  – estimation from components
  – technology differences
  – few area-time points

Modeling Instruction Effects

• Restrictions from "ideal"
  + save area and energy
  – limit usability (yield) of PE
    • May cost more energy, area in the end…
• Want to understand effects
  – area model [today] (energy model on HW5)
  – utilization/yield model

Preclass

• Energies?
  – 8-bit, 16-bit, 32-bit
• 16-bit on 32-bit?
  – Sources of inefficiency?
• 8-bit operations per 16-bit operation?
• 16-bit on 8-bit?
  – Sources of inefficiency?

Efficiency/Yield Intuition

• What happens when
  – Datapath is too wide?
  – Datapath is too narrow?
  – Instruction memory is too deep?
Efficiency/Yield Intuition

- What happens when
  - Datapath is too wide?
  - Datapath is too narrow?
  - Instruction memory is too deep?
  - Instruction memory is too shallow?

Computing Device

- Composition
  - Bit Processing elements
  - Interconnect: space
  - Interconnect: time
  - Instruction Memory

Relative Sizes

- Bit Operator: 3-5KF
- Bit Operator Interconnect: 200K-250KF
- Instruction (w/ interconnect): 20KF
- Memory bit (SRAM): 250-500F

Model Area

\[ A_{bit.elm} = A_{fixed} + N_{SW} (N_p, w_p) \cdot A_{SW} \]

\[ + \left( \frac{c}{d} \cdot n_{bits} \cdot A_{mem.cell} \right) \]

\[ + d \cdot A_{mem.cell} \]
Architectures Fall in Space

Calibrate Model

Calibration model for architectures:
- FPGA: \( w = 1, d = c = 1, k = 4 \) 880K\( \mu \)m\(^2\)
- Xilinx 4K: 630K\( \mu \)m\(^2\)
- Altera 8K: 930K\( \mu \)m\(^2\)
- SIMD: \( w = 1000, c = 0, d = 64, k = 3 \) 170K\( \mu \)m\(^2\)
- Abacus: 190K\( \mu \)m\(^2\)
- Processor model: \( w = 32, d = 32, c = 1024, k = 2 \) 2.6M\( \mu \)m\(^2\)
- MIPS-X: 2.1M\( \mu \)m\(^2\)

Peak Densities from Model

Peak Densities from Model

- Only 2 of 4 parameters
  - small slice of space
  - 100x density across

- Large difference in peak densities
  - large design space!

Architectural parameters \( \rightarrow \) Peak Densities

Efficiency

- What do we really want to maximize?
  - Not peak, “guaranteed not to exceed” performance, but…
  - Useful work per unit silicon [per Joule]

- Yield Fraction / Area
- (or minimize (Area/Yielded performance))
Efficiency

- For comparison, look at relative efficiency to ideal.
- Ideal = architecture exactly matched to application requirements
- Efficiency = \( \frac{A_{\text{ideal}}}{A_{\text{arch}}} \)
- \( A_{\text{arch}} = \text{Area Op/Yield} \)

Width Mismatch Efficiency Calculation

\[
E = \frac{\text{Area(Task – on – matched – Architecture)}}{\text{Area(Task – on – this – Architecture)}}
\]

\[
E = \frac{W_{\text{task}} \times A_{\text{ideal}}}{W_{\text{arch}} \times \left[ \frac{W_{\text{task}}}{W_{\text{arch}}} \right] \times A_{\text{ideal}}}
\]

Efficiency: Width Mismatch

\( c=1, \quad 16\text{K PEs} \)

Efficiency for Preclass

\[
E = \frac{\text{Energy(Task – on – matched – Architecture)}}{\text{Energy(Task – on – this – Architecture)}}
\]

- Preclass 6 table

Application vs. Architecture

- \( W_{\text{task}} \) vs. \( W_{\text{arch}} \)
- Path Length vs. Context Depth

Path Length

- How many primitive-operator delays before can perform next operation?
  - Reuse the resource
Reuse

How many times can I reuse each primitive operator?

**Path Length:** How much sequentialization is allowed (required)?

*E.g.* Want meet 30ns real time rate with 1.5ns cycle time, can afford to issue 15 sequential ops.

---

Context (Instruction) Depth

---

Efficiency with fixed Width

- $w=1$, 16K PEs

---

Ideal Efficiency (different model)

- Two resources here:
  - active processing elements
  - operation description/state

Applications need in different proportions.

**Robust Point**

- What is Energy Robust Point for preclass model?

---

Robust Point depend on Width

- $w=1$
- $w=8$
- $w=64$
Processors and FPGAs
(architecture vs. two application axes)

FPGA $c=d=1, w=1, k=4$

"Processor" $c=d=1024, w=64, k=2$

Application Needs

- What are common application datawidths?
- What are common application path lengths?

Examples

<table>
<thead>
<tr>
<th>Application</th>
<th>Wapp</th>
<th>Lcritpath</th>
<th>Lpath</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conway LIFE</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Run as fast as possible</td>
</tr>
<tr>
<td>Entropy Code</td>
<td>1</td>
<td>1-10</td>
<td>100</td>
<td>100ns memory interface</td>
</tr>
<tr>
<td>Video</td>
<td>8</td>
<td>1-6</td>
<td>24</td>
<td>1GHz for 1024x1024 x30 frames/s</td>
</tr>
<tr>
<td>Audio</td>
<td>16</td>
<td>1-10</td>
<td>20,000</td>
<td>44KHz for 1GHz</td>
</tr>
<tr>
<td>FDTD</td>
<td>35</td>
<td>1-5</td>
<td>1-5</td>
<td></td>
</tr>
</tbody>
</table>

Intermediate Architecture

$w=8$
$c=64$
16K PEs

Hard to be robust across entire space...

Caveats

- Model abstracts away many details that are important
  - interconnect (day 15--18)
  - control (day 22)
  - specialized functional units (day 14)
- Applications are a heterogeneous mix of characteristics

Modeling Message

- Architecture space is huge
- Easy to be very inefficient
- Hard to pick one point robust across entire space
- Why we have so many architectures?
General Message

- Parameterize architectures
- Look at continuum
  - costs
  - benefits
- Often have competing effects
  - leads to maxima/minima

Admin

- Should now have all background for HW5
  - Problem 2 similar (looking for robust point)
  - Different: Interconnect parameter, Energy
- No class Wednesday
- No office hours Tuesday
- Next class Monday
  - Reading online
- HW 6 out
  - 1 and 2 due Friday 2/3
  - Should be able to do 1 now

Big Ideas

[MSB Ideas]

- Applications typically have structure
- Exploit this structure to reduce resource requirements
- Architecture is about understanding and exploiting structure and costs to reduce requirements

Big Ideas

[MSB Ideas]

- Instruction organization induces a design space (taxonomy) for programmable architectures
- Arch. structure and application requirements mismatch ⇒ inefficiencies
- Model ⇒ visualize efficiency trends
- Architecture space is huge
  - can be very inefficient
  - need to learn to navigate