CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture

Yanqi Zhou\textsuperscript{1} Henry Hoffmann\textsuperscript{2} David Wentzlaff\textsuperscript{1}

\textsuperscript{1} PRINCETON UNIVERSITY
\textsuperscript{2} THE UNIVERSITY OF CHICAGO

Department of Computer Science

NSF
Web Services Have Latency Requirements

How One Second Could Cost Amazon $1.6 Billion In Sales

It’s Official – ‘Web Stress’ is Bad for Business
A Limited Pallet of Choices in EC2

<table>
<thead>
<tr>
<th>Region: US West (Oregon)</th>
</tr>
</thead>
</table>

### Standard On-Demand Instances

<table>
<thead>
<tr>
<th>Size</th>
<th>Linux/UNIX Usage</th>
<th>Windows Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small (Default)</td>
<td>$0.105 per Hour</td>
<td>$0.105 per Hour</td>
</tr>
<tr>
<td>Medium</td>
<td>$0.132 per Hour</td>
<td>$0.132 per Hour</td>
</tr>
<tr>
<td>Large</td>
<td>$0.165 per Hour</td>
<td>$0.165 per Hour</td>
</tr>
</tbody>
</table>

### Micro On-Demand Instances

<table>
<thead>
<tr>
<th>Size</th>
<th>Linux/UNIX Usage</th>
<th>Windows Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro</td>
<td>$0.020 per Hour</td>
<td>$0.020 per Hour</td>
</tr>
</tbody>
</table>

### High-Memory On-Demand Instances

<table>
<thead>
<tr>
<th>Size</th>
<th>Linux/UNIX Usage</th>
<th>Windows Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra Large</td>
<td>$0.570 per Hour</td>
<td>$0.570 per Hour</td>
</tr>
<tr>
<td>Double Extra Large</td>
<td>$1.140 per Hour</td>
<td>$1.140 per Hour</td>
</tr>
<tr>
<td>Quadruple Extra Large</td>
<td>$2.280 per Hour</td>
<td>$2.280 per Hour</td>
</tr>
</tbody>
</table>

### High-CPU On-Demand Instances

<table>
<thead>
<tr>
<th>Size</th>
<th>Linux/UNIX Usage</th>
<th>Windows Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>$0.285 per Hour</td>
<td>$0.285 per Hour</td>
</tr>
<tr>
<td>Extra Large</td>
<td>$1.140 per Hour</td>
<td>$1.140 per Hour</td>
</tr>
</tbody>
</table>

### Cluster Compute Instance

<table>
<thead>
<tr>
<th>Size</th>
<th>Linux/UNIX Usage</th>
<th>Windows Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eight Extra Large</td>
<td>$2.400 per Hour</td>
<td>$2.970 per Hour</td>
</tr>
</tbody>
</table>

**I need more performance**
**I need to reduce costs**

Image Credit: Amazon
<table>
<thead>
<tr>
<th>Features</th>
<th>Distributed ILP</th>
<th>TRIPS, CLP</th>
<th>Core Fusion</th>
<th>WiDGET</th>
<th>Conjoined Cores</th>
<th>Clustered big. LITTLE</th>
<th>Sharing Arch.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale up &amp; down</td>
<td>✔</td>
<td>✔</td>
<td>✗</td>
<td>✔</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Distributed/switched</td>
<td>✔</td>
<td>✔</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Symmetric</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Partition L2</td>
<td>✔</td>
<td>✔</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Dynamic OoO</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
<td>✗</td>
<td>✔</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>ISA Compatible</td>
<td>✔</td>
<td>✗</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Multiple Metrics</td>
<td>✗</td>
<td>✔</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
</tr>
</tbody>
</table>
Managing Fine-grain Configurability is Hard

True Maxima
Managing Fine-grain Configurability is Hard
Managing Fine-grain Configurability is Hard
Managing Fine-grain Configurability is Hard
Observations:

- Fine-grain configurability is extremely helpful given current constraints
- Such configurability should produce non-convex optimization spaces
- Much research needs to be done on managing these non-convex spaces
CASH: Architecture and Runtime for Managing Fine-grain Configurability

Solution:
- Statically homogeneous, dynamically heterogeneous architecture
- Lightweight runtime system manages configurability online
Architecture: Sharing Architecture

[Zhou & Wentzlaff ASPLOS 2014]
Aggregating Performance Data Across Slices

CASH Runtime Interface

Register Flush

L2 Flush

L1 Icache

Fetch

Decode

Global Rename

Local Rename

Instruction Window

ALU

LS Hash

LD/ST

L1 Dcache

Br_pred & btb

ROB

Scoreboard

Scalar Operand Network

Local RF

LS Sorting Network

Fetch&BTB Sync

Global Rename &Scoreboard Sync

Operand Network

LS bank sorting

L1/L2 Switched Crossbar

Switched

Switched

Switched

Switched

Fetch&BTB Sync

Global Rename &Scoreboard Sync

Operand Network

LS bank sorting

L1/L2 Switched Crossbar

Switched

Switched

Switched

Switched

Performance Counter

CASH Runtime Interface

Register Flush

L2 Flush

L1 Icache

Fetch

Decode

Global Rename

Local Rename

Instruction Window

ALU

LS Hash

LD/ST

L1 Dcache

Br_pred & btb

ROB

Scoreboard

Scalar Operand Network

Local RF

LS Sorting Network

Fetch&BTB Sync

Global Rename &Scoreboard Sync

Operand Network

LS bank sorting

L1/L2 Switched Crossbar

Switched

Switched

Switched

Switched

Performance Counter
Controlling Application Performance

All problems in computer science can be solved by another level of indirection…

-- David Wheeler

Corollary: All optimization problems can be solved by another level of feedback…
Controlling Application Performance
Adapting to Phases

- PID Controller
- Kalman Filter
- Phase-specific Coefficients
- Performance Feedback
- Performance Requirement
- Sharing Architecture
- Slices
- L2$
Controlling Application Performance
Learn Online

- PID Controller
- Kalman Filter
- Optimizer
- Sharing Architecture

Performance Requirement

Phase-specific Coefficients

Performance Feedback

Slices
L2$
Benefit of Managing Fine-grain Configurability

Cost

- race
- convex
- CASH

Performance

- race
- convex
- CASH

$/$s

Normalized Performance

Time (M Cycles)

Time (M Cycles)
Comparison of Fine-grain Management Techniques

**Cost**

- Optimal
- Convex
- Race
- CASH

**QoS Violations**

- Optimal
- Convex
- Race
- CASH

---

**Benchmarks**

- apache
- astar
- bzip
- ferret
- gcc
- h264ref
- hmmmer
- lib
- mailserver
- mcf
- omnetpp
- sjeng
- x264
- GEOMEAN

---

**Benchmark**

- apache
- astar
- bzip
- ferret
- gcc
- h264ref
- hmmmer
- lib
- mailserver
- mcf
- omnetpp
- sjeng
- x264
- GEOMEAN
Comparisons to Coarse-grain Reconfigurable

**Cost**
- Coarse Race
- Coarse Adapt
- Fine Race
- CASH Adapt

**QoS Violations**
- Coarse Race
- Coarse Adapt
- Fine Race
- CASH Adapt

Benchmark: apache, astar, bzip, ferret, gcc, h264ref, hmmer, lib, mailserver, mcf, omnetpp, sjeng, x264, GEOMEAN

Benchmark: apache, astar, bzip, ferret, gcc, h264ref, hmmer, lib, mailserver, mcf, omnetpp, sjeng, x264, geomean
CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture

Yanqi Zhou  Henry Hoffmann David Wentzlaff

Department of Computer Science