Informed Microarchitecture Design Space Exploration Using Workload Dynamics

Chang-Burm Cho, Wangyuan Zhang, Tao Li
(MICRO-40)

December 4th, 2007

IDEAL Research
(Intelligent Design of Efficient Architectures Laboratory)
Electrical and Computer Engineering
University of Florida
Motivation

- Workload dynamics prediction

- Workload dynamics prediction for worst-case scenario

Introduction

• Workload dynamic variations

• Slow, detailed simulation
Previous research

• Sherwood (ASPLOS ’02)
  – Phase analysis and Simpoint
• Lee (ASPLOS ’06)
  – Linear regression model
• Ipek (ASPLOS ’06)
  – Neural network model
• Joseph (HPCA ’06, MICRO ’06)
  – Linear and non-linear predictive models
Contributions

• Build accurate and informative predictive models
  – Combine the wavelet and neural network

• Evaluate efficiency
  – Performance, power and reliability domains

• Present case studies
  – Scenario-aware prediction
  – DVM (Dynamic Vulnerability Management)
Background: wavelet transform (1/3)

Original Data
3, 4, 20, 25, 15, 5, 20, 3

Scaling Filter ($G_0$)
3.5, 22.5, 10, 11.5

Wavelet Filter ($H_0$)
-0.5, -2.5, 5, 8.5

Scaling Filter ($G_1$)
13, 10.75

Wavelet Filter ($H_1$)
-9.5, -0.75

Scaling Filter ($G_2$)
11.875

Wavelet Filter ($H_2$)
1.125

Haar wavelet transform ->
Scaling filter = (even+odd)/2
Wavelet filter = (even-odd)/2
Background: wavelet transform (2/3)

- Workload behavior can be represented at different scales using wavelet transform

![Time domain program behavior (64 samples)](image)

(a) 1 wavelet coefficient

(b) 2 wavelet coefficients

(c) 4 wavelet coefficients

(d) 8 wavelet coefficients

(e) 16 wavelet coefficients

(f) 64 wavelet coefficients
Background: neural network (3/3)

- **Output layer:** f(x)
- **Hidden layer:** H_1(x), H_2(x), ..., H_n(x)
- **Input layer:** X_1, X_2, ..., X_n

Radial Basis Function (RBF)

(distance)
Combining wavelet and NN

- Wavelet decomposition of workload behavior
- Wavelet coefficients are predicted by NNs
- Reconstruct workload behavior using predicted wavelet coefficients
Experimental setup

• Benchmarks (SPEC CPU 2000)
  – bzip2, crafty, eon, gap, gcc, mcf, parser, perlbmk, swim, twolf, vortex and vpr

• Metrics
  – Performance (CPI)
  – Power (Wattch-based model)
  – Reliability (Architectural Vulnerability Factor: AVF)
Explore the design space

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Ranges</th>
<th># of Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fetch_width</td>
<td>2, 4, 8, 16</td>
<td>4</td>
</tr>
<tr>
<td>ROB_size</td>
<td>96, 128, 160</td>
<td>3</td>
</tr>
<tr>
<td>IQ_size</td>
<td>32, 64, 96, 128</td>
<td>4</td>
</tr>
<tr>
<td>LSQ_size</td>
<td>16, 24, 32, 64</td>
<td>4</td>
</tr>
<tr>
<td>L2_size</td>
<td>256, 1024, 2048, 4096 KB</td>
<td>4</td>
</tr>
<tr>
<td>L2_lat</td>
<td>8, 12, 14, 16, 20</td>
<td>5</td>
</tr>
<tr>
<td>il1_size</td>
<td>8, 16, 32, 64 KB</td>
<td>4</td>
</tr>
<tr>
<td>dl1_size</td>
<td>8, 16, 32, 64 KB</td>
<td>4</td>
</tr>
<tr>
<td>dl1_lat</td>
<td>1, 2, 3, 4</td>
<td>4</td>
</tr>
</tbody>
</table>

- 9 microarchitectural parameters
- LHS (Latin Hypercube Sampling) + L2-star discrepancy
Selecting wavelet coefficients

• Order-based wavelet coefficients
  – Select first $k$ coefficients

• Magnitude-based wavelet coefficients
  – Select the largest $k$ coefficients
Evaluation : MSE

- Mean Square Error

\[ MSE = \frac{1}{N} \sum_{k=1}^{N} ((x(k) - \hat{x}(k)))^2 \]
Evaluation: sensitivity analysis

- Number of wavelet coefficients
- Number of samples

![Graph showing sensitivity analysis results for wavelet coefficients and samples.]
Evaluation: significance

• Which parameters are dominant?
• Which observations show similar behavior?

(a) Performance (CPI)

(b) Reliability (AVF)
Case study – 1 (1/2)

- Scenario-based prediction
  - Allows architect to quickly examine the application worst-case scenarios

- Directional Symmetry (DS) metric

\[
DS = \frac{1}{N} \sum_{k=1}^{N} \varphi(x(k) \cdot \hat{x}(k))
\]
Case study – 1 (2/2)

• Directional Asymmetry (1-DS)
• DVM(Dynamic Vulnerability Management)
  – A set of strategies to control hardware runtime soft-error susceptibility under a tolerable threshold
Case study – 2 (2/3)

(a) DVM_Disabled

(b) DVM_Enabled

Configuration-A
Case study – 2 (3/3)

(a) DVM_Disabled

(b) DVM_Enabled

Configuration-B
Conclusions

- Propose accurate and informative prediction models
  - Wavelet based multiresolution decomposition
  - Neural network based non-linear regression modeling
- Predictive models achieve high accuracy in revealing workload dynamic behavior across a large microarchitecture design space
- The proposed techniques can be used to efficiently explore workload scenario-driven architecture optimizations
Questions

• Thank you for listening