Illustrative Design Space Studies with Microarchitectural Regression Models

Benjamin C. Lee, David M. Brooks

www.deas.harvard.edu/~bclee
Engineering and Applied Sciences
Harvard University

International Symposium on High-Performance Computer Architecture
14 February 2007
Outline

Motivation & Background
  Exploration Challenges
  Simulation Paradigm
  Regression Theory

Microarchitectural Modeling
  Experimental Methodology
  Model Evaluation

Design Optimization
  Pareto Frontier
  Multiprocessor Heterogeneity

Conclusion
Outline

Motivation & Background
- Exploration Challenges
- Simulation Paradigm
- Regression Theory

Microarchitectural Modeling
- Experimental Methodology
- Model Evaluation

Design Optimization
- Pareto Frontier
- Multiprocessor Heterogeneity

Conclusion
Exploration Challenges

- **Metric Diversity**
  - Differentiated market segments and metric emphases
  - Examples :: latency, throughput, power, temperature

- **Design Diversity**
  - Diversity of interesting, viable designs
  - Examples :: Power, Pentium, UltraSPARC

- **Comprehensive Design Exploration**
  - Location of optima depend on workload, metrics
  - Multiprocessor design increases diversity
Simulation Challenges

- **Cycle-Accurate Simulation**
  - Accurately identifies trends in design space
  - Tracks instructions’ progress through microprocessor
  - Estimates performance, power, temperature, . . .

- **Simulation Costs**
  - Long simulation times (minutes, hours per design)
  - Number of potential simulations scales exponentially ($m^p$)
    - $p$ :: parameter count
    - $m$ :: parameter resolution
Temporal Sampling

- **Instruction Sampling from Time Domain**
  - Reduce simulation costs via size of inputs
  - Synthetic traces from profiled workloads
  - Sampled traces from phase analysis

---

1. Eeckhout+[ISPASS’00]
2. Sherwood+[ASPLOS’02], Wunderlich+[ISCA’03]
Spatial Sampling

- Design Sampling from Comprehensive Space
  - Reduce simulation costs via number of simulations
Spatial Sampling

- **Design Sampling from Comprehensive Space**
- Reduce simulation costs via number of simulations
Simulation Paradigm

- **Comprehensively understand design space**
  - Specify large, high-resolution design space
  - Consider all design parameters simultaneously

- **Selectively simulate modest number of designs**
  - Sample points randomly from design space for simulation
  - Decouple resolution of design space and simulation

- **Efficiently leverage simulation data with inference**
  - Reveal trends, trade-offs from sparse sampling
  - Enable predictions for metrics of interest
Regression Theory

- **Statistical Inference**
  - Models approximate solutions to intractable problems
  - Requires initial data to train, formulate model
  - Leverages correlations from initial data for prediction

- **Regression Models\(^3\)**
  - Low formulation costs (1K samples from 1B designs)
  - Accurate inference (5 – 7% median error)
  - Efficient computation (100’s of predictions per second)

\(^3\)Lee+ [ASPLOS’06]
Model Formulation

**Notation**

- \( n \) observations \( \{\text{simulated design samples}\} \)
- Response :: \( \bar{y} = y_1, \ldots, y_n \) \( \{\text{e.g., performance, power}\} \)
- Predictor :: \( \bar{x}_i = x_{i,1}, \ldots, x_{i,p} \) \( \{\text{e.g., depth, cache}\} \)
- Regression Coefficients :: \( \beta = \beta_0, \ldots, \beta_p \)
- Random Error :: \( \bar{e} = e_1, \ldots, e_n \) where \( e_i \sim N(0, \sigma^2) \)
- Transformations :: \( f, \bar{g} = g_1, \ldots, g_p \)

**Model**

\[
f(y) = \beta_0 + \sum_{j=1}^{p} \beta_j g_j(x_j) + e
\]
Predictor Interaction

**Modeling Interaction**
- Suppose effects of predictors $x_1$, $x_2$ cannot be separated
- Construct predictor $x_3 = x_1x_2$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1x_2 + e_i$$

**Example**
- Let $x_1$ be pipeline depth, $x_2$ be L2 cache size
- Performance impact of pipelining affected by cache size

$$\text{Speedup} = \frac{\text{Depth}}{1 + \text{Stalls/Inst}}$$
Predictor Non-Linearity

**Restricted Cubic Splines**
- Divide predictor domain into intervals separated by knots
- Piecewise cubic polynomials joined at knots
- Higher order polynomials provide better fits

\[ rcs(x,5) \]

\[ k1, k2, k3, k4, k5 \]

---

\[ ^4 \text{Stone [SS'86]} \]
Prediction

**Expected Response**
- $\beta$ are known from least squares
- $x_{i,1}, \ldots, x_{i,p}$ are known for a given query $i$
- Expected response is weighted sum of predictor values

\[
E[y] = E[\beta_0 + \sum_{j=1}^{p} \beta_j x_j] + E[e] \\
= \beta_0 + \sum_{j=1}^{p} \beta_j x_j
\]
Outline

Motivation & Background
  Exploration Challenges
  Simulation Paradigm
  Regression Theory

Microarchitectural Modeling
  Experimental Methodology
  Model Evaluation

Design Optimization
  Pareto Frontier
  Multiprocessor Heterogeneity

Conclusion
Tools and Benchmarks

- **Simulation Framework**
  - Turandot :: a cycle-accurate trace driven simulator
  - PowerTimer :: power models derived from circuit analyses
  - Baseline simulator models POWER4/POWER5 architecture

- **Benchmarks**
  - SPEC2kCPU :: compute-intensive benchmarks
  - SPECjbb :: Java server benchmark

- **Statistical Framework**
  - R :: software environment for statistical computing
  - Hmisc and Design packages\(^5\)

\(^5\)Harrell [Springer,’01]
### Predictors :: Microarchitecture

| Set   | Parameters                  | Measure                 | Range            | |S| |
|-------|-----------------------------|-------------------------|------------------|------------------|
| $S_1$ | Depth                       | depth                   | FO4              | 9::3::36         | 10 |
| $S_2$ | Width                       | width                   | insn b/w         | 4,8,16           | 3  |
|       |                             | L/S reorder queue       | entries          | 15::15::45       |    |
|       |                             | store queue             | entries          | 14::14::42       |    |
|       |                             | functional units        | count            | 1,2,4            |    |
| $S_3$ | Physical Registers          | general purpose (GP)    | count            | 40::10::130      | 10 |
|       |                             | floating-point (FP)     | count            | 40::8::112       |    |
|       |                             | special purpose (SP)    | count            | 42::6::96        |    |
| $S_4$ | Reservation Stations        | branch                  | entries          | 6::1::15         | 10 |
|       |                             | fixed-point/memory      | entries          | 10::2::28        |    |
|       |                             | floating-point          | entries          | 5::1::14         |    |
| $S_5$ | I-L1 Cache                  | i-L1 cache size         | $\log_2(\text{entries})$ | 7::1::11 | 5  |
| $S_6$ | D-L1 Cache                  | d-L1 cache size         | $\log_2(\text{entries})$ | 6::1::10 | 5  |
| $S_7$ | L2 Cache                    | L2 cache size           | $\log_2(\text{entries})$ | 11::1::15 | 5  |
Model Evaluation I

- **Framework**
  - Formulate models with $n = 1,000$ samples
  - Obtain 100 additional random samples for validation
  - Quantify percentage error, $100 \times |\hat{y}_i - y_i|/y_i$

- **Comparison**
  - Simulator-reported performance, power
  - Regression-predicted performance, power
Model Evaluation II

Random Validation :: Performance
Error Distribution

Random Validation :: Power
Error Distribution
Outline

Motivation & Background
- Exploration Challenges
- Simulation Paradigm
- Regression Theory

Microarchitectural Modeling
- Experimental Methodology
- Model Evaluation

Design Optimization
- Pareto Frontier
- Multiprocessor Heterogeneity

Conclusion
Design Optimization

- **Pareto Frontier**
  - Characterize comprehensive design space
  - Identify pareto frontier

- **Pipeline Depth**
  - Vary all parameters simultaneously with depth
  - Identify most efficient designs at each depth

- **Multiprocessor Heterogeneity**
  - Identify most efficient designs for each benchmark
  - Identify multiple design compromises
Design Optimization

- **Pareto Frontier**
  - Characterize comprehensive design space
  - Identify pareto frontier

- **Pipeline Depth**
  - Vary all parameters simultaneously with depth
  - Identify most efficient designs at each depth

- **Multiprocessor Heterogeneity**
  - Identify most efficient designs for each benchmark
  - Identify multiple design compromises
Pareto Frontier

**Background**
- Optimization improves at least one metric without negatively impacting any other metric

**Objective**
- Construct pareto frontier in power-delay space

**Approach**
- Simulate 1K samples from design space
- Formulate regression models for performance, power
- Characterize design space via regression
- Identify frontier from characterization
Design Space Characterization

Design Space Characterization
ammp :: depth

Power (W)

Delay (inv-bips)

depth
Design Space Characterization

Design Space Characterization
ammp :: width

- Power (W)
- Delay (inv-bips)
Design Space Characterization
Design Space Characterization

Design Space Characterization
ammp :: resv

Power (W)

Delay (inv-bips)

0.5 1 1.5 2 2.5 3

Benjamin C. Lee, David M. Brooks

27 :: HPCA :: 14 Feb 07
Design Space Characterization

![Design Space Characterization](image)

Power (W) vs. Delay (inv-bips) for different configurations of L2 cache in the design space.
Workload Characterization

Design Space Characterization
ammp :: l2cache

Design Space Characterization
mcf :: l2cache

Benjamin C. Lee, David M. Brooks
Multiprocessor Heterogeneity

- **Background**
  - Prior heterogeneity studies constrained design options

- **Objective**
  - Identify efficient heterogeneous compromises
  - Mitigate penalties from homogenous compromise

- **Approach**
  - Simulate 1K samples from design space
  - Formulate regression models for performance, power
  - Identify per benchmark optima \( (bips^3/w) \) via regression
  - Identify compromises via K-means clustering

---

\(^6\text{Kumar+}[\text{ISCA'04}, \text{Kumar+}[\text{PACT'06}]]\)
Heterogeneous Efficiency
Heterogeneous Clusters

Benchmark Clusters :: Predictions

Cluster 1 \{ jbb, mesa \}
Depth :: 15 FO4
Width :: 8 Inst/Cy
Reg :: 80 GPR
L1 :: 64 KB (I-$\cdot$), 64 KB (D-$\cdot$)
L2 :: 0.5 MB

Cluster 2 \{ ammp, applu, equake, twolf \}
Depth :: 27 FO4
Width :: 8 Inst/Cy
Reg :: 130 GPR
L1 :: 32 KB (I-$\cdot$), 32 KB (D-$\cdot$)
L2 :: 0.5 MB

Cluster 3 \{ gcc, gzip \}
Depth :: 15 FO4
Width :: 2 Inst/Cy
Reg :: 70 GPR
L1 :: 16 KB (I-$\cdot$), 8 KB (D-$\cdot$)
L2 :: 0.5 MB

Cluster 4 \{ mcf \}
Depth :: 30 FO4
Width :: 2 Inst/Cy
Reg :: 70 GPR
L1 :: 256 KB (I-$\cdot$), 8 KB (D-$\cdot$)
L2 :: 4 MB
Future Directions

- **Topological Analysis**
  - Visualization with contour maps
  - Roughness metrics quantify observed trends

- **Optimization**
  - Heuristic search (e.g., gradient descent)
  - Symbolic optimization

- **Chip Multiprocessor Design**
  - Decoupled models (e.g., core and interconnect)
  - Larger parameter space (e.g., in-order execution)
Conclusion

- **Simulation Paradigm**
  - Comprehensively understand design space
  - Selectively simulate modest number of designs
  - Efficiently leverage simulation data with inference

- **Design Optimization**
  - New capabilities in practical design optimization
  - Characterize comprehensive design spaces
  - Identify diverse optima and compromises

- **ISCA 2007 Tutorial**
  - Inference and Learning for Large Scale Microarchitectural Analysis
Further Reading

www.deas.harvard.edu/~bcee

Methods of inference and learning for performance modeling of parallel applications

B.C. Lee and D.M. Brooks.
Illustrative design space studies with microarchitectural regression models

B.C. Lee and D.M. Brooks.
Accurate, efficient regression modeling for microarchitectural performance, power prediction.

B.C. Lee and D.M. Brooks.
Statistically rigorous regression modeling for the microprocessor design space.

B.C. Lee and D.M. Brooks.
Regression modeling strategies for microarchitectural performance and power prediction.