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

Exploration Challenges

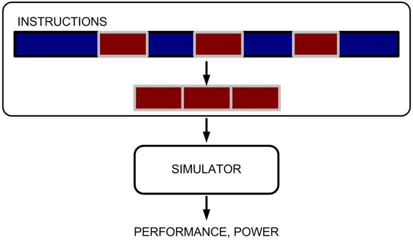
- **Metric Diversity**
  - Differentiated market segments and metric emphases
  - Examples :: latency, throughput, power, temperature
- **Design Diversity**
  - Diversity of interesting, viable designs
  - Examples :: Power, Pentium, UltraSPARC
- **Comprehensive Design Exploration**
  - Location of optima 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

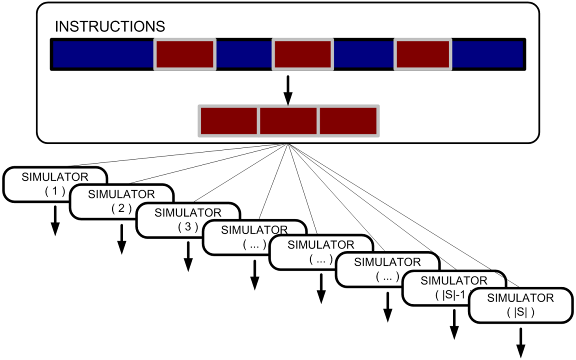
- **Instruction Sampling from Time Domain**
  - Reduce simulation costs via size of inputs
  - Synthetic traces from profiled workloads [1]
  - Sampled traces from phase analysis [2]



[1] Eeckhout+[ISPASS'00]
[2] Sherwood+[ASPLOS'02], Wunderlich+[ISCA'03]

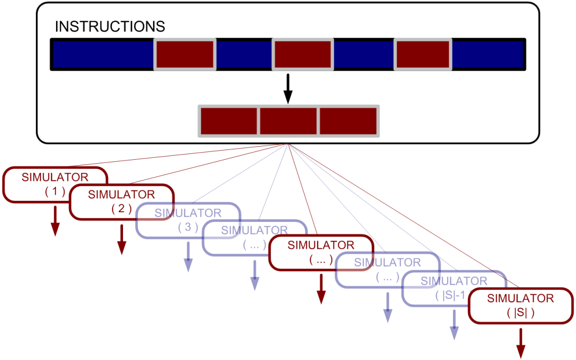
Spatial Sampling

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

$$f(y) \quad = \quad \beta_0 + \sum_{j=1}^{p} \beta_j g_j(x_j) + e$$

Predictor Interaction

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

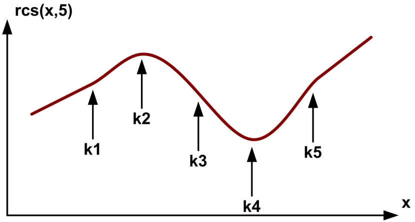
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e_i$$

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

$$Speedup = \frac{Depth}{1 + Stalls/Inst}$$

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 [4]



[4] Stone [SS'86]

Prediction

- **Expected Response**
  - β 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

$$
\begin{aligned}
E\left[y\right] &= E\left[\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right] + E\left[e\right] \\
&= \beta_0 + \sum_{j=1}^{p} \beta_j x_j
\end{aligned}
$$

Outline

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]

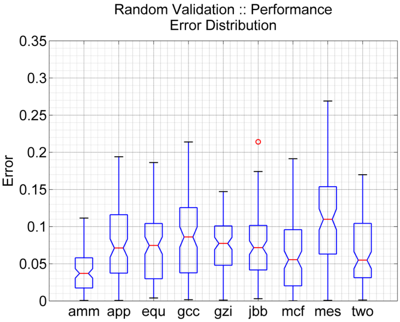
Predictors :: Microarchitecture

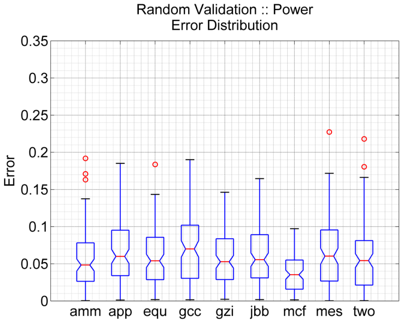
	Set	Parameters	Measure	Range	|S|
S_1	Depth	depth	FO4	9::3::36	10
S_2	Width	width	insn b/w	4,8,16	3
		L/S reorder queue	entries	15::15::45	
		store queue	entries	14::14::42	
		functional units	count	1,2,4	
S_3	Physical	general purpose (GP)	count	40::10::130	10
	Registers	floating-point (FP)	count	40::8::112	
		special purpose (SP)	count	42::6::96	
S_4	Reservation	branch	entries	6::1::15	10
	Stations	fixed-point/memory	entries	10::2::28	
		floating-point	entries	5::1::14	
S_5	I-L1 Cache	i-L1 cache size	log_2(entries)	7::1::11	5
S_6	D-L1 Cache	d-L1 cache size	log_2(entries)	6::1::10	5
S_7	L2 Cache	L2 cache size	log_2(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, $100 * |\hat{y}_i - y_i|/y_i$
- **Comparison**
  - Simulator-reported performance, power
  - Regression-predicted performance, power

Model Evaluation II





Outline

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

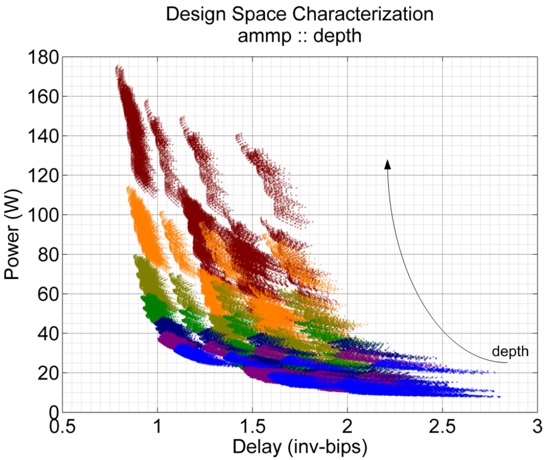
Design Optimization

- **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

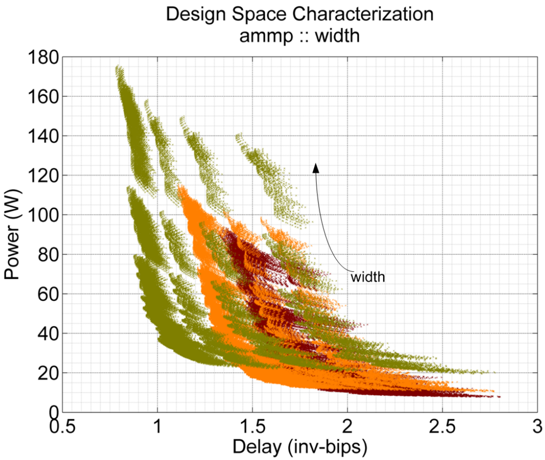
Pareto Frontier

- **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

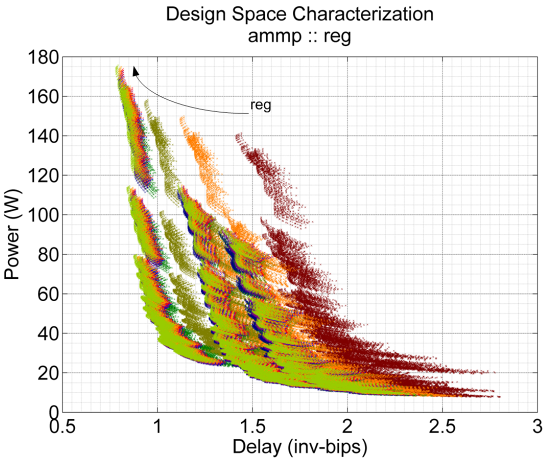
Design Space Characterization



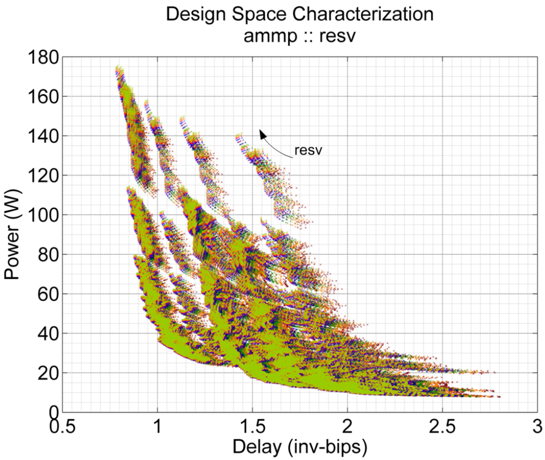
Design Space Characterization



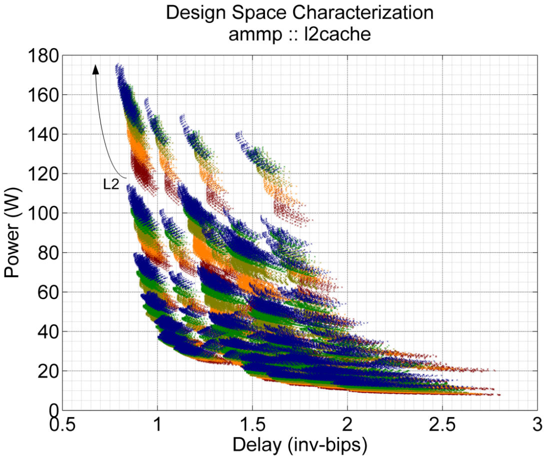
Design Space Characterization



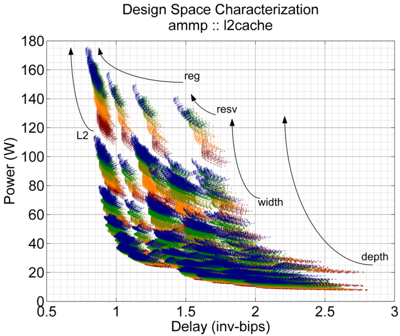
Design Space Characterization

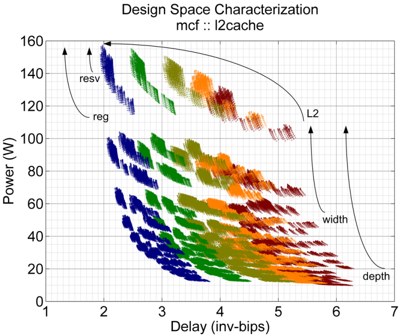


Design Space Characterization

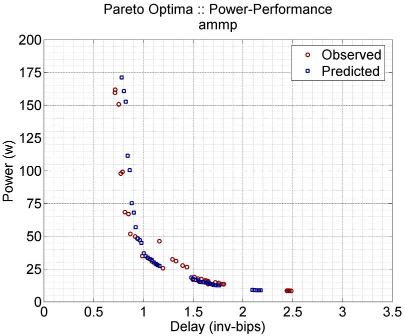


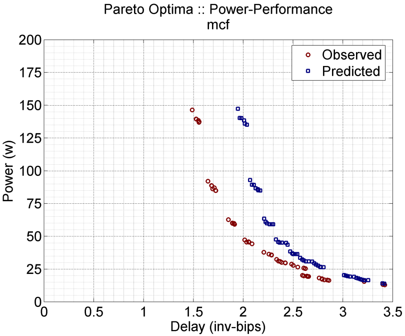
Workload Characterization





Pareto Frontier



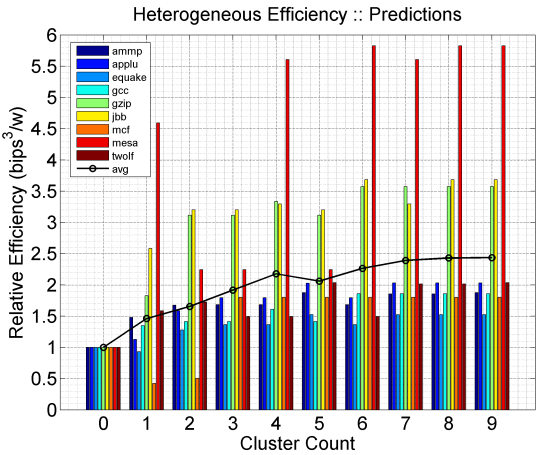


Multiprocessor Heterogeneity

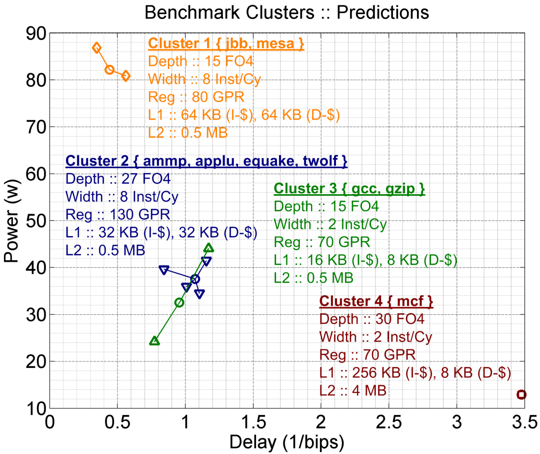
- **Background**
  - Prior heterogeneity studies constrained design options[6]
- **Objective**
  - Identify efficient heterogeneous compromises
  - Mitigate penalties from homogenous compromise
- **Approach**
  - Simulate 1K samples from design space
  - Formulate regression models for performance, power
  - Identify per benchmark optima ($bips^3/w$) via regression
  - Identify compromises via K-means clustering

[6] Kumar+[ISCA'04], Kumar+[PACT'06]

Heterogeneous Efficiency



Heterogeneous Clusters



Outline

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/∼bclee

- B.C. Lee and D.M. Brooks and B.R. de Supinski and M. Schulz and K. Singh and S.A. McKee.
  Methods of inference and learning for performance modeling of parallel applications
  PPoPP'07: Symposium on Principles and Practice of Parallel Programming, March 2007.

- B.C. Lee and D.M. Brooks.
  Illustrative design space studies with microarchitectural regression models
  HPCA-13: International Symposium on High Performance Computer Architecture, Feb 2007.

- B.C. Lee and D.M. Brooks.
  Accurate, efficient regression modeling for microarchitectural performance, power prediction.
  ASPLOS-XII: International Conference on Architectural Support for Programming Languages and Operating Systems, Oct 2006.

- B.C. Lee and D.M. Brooks.
  Statistically rigorous regression modeling for the microprocessor design space.
  MoBS-2: Workshop on Modeling, Benchmarking, and Simulation, June 2006.

- B.C. Lee and D.M. Brooks.
  Regression modeling strategies for microarchitectural performance and power prediction.
  Harvard University Technical Report TR-08-06, March 2006.