Statistical Inference for Efficient Microarchitectural Analysis

Benjamin C. Lee, David M. Brooks

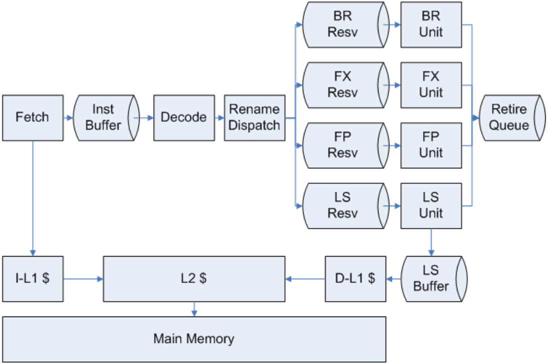
www.deas.harvard.edu/∼bclee
Division of Engineering and Applied Sciences
Harvard University

Boston Area Architecture Workshop
26 January 2007

Outline

Outline

Microarchitectural Design Space



- Increasing diversity of interesting, viable designs
- Examples :: Power 4, Pentium 4, UltraSPARC T1
- Tractably quantify trends across comprehensive design space

Microarchitectural 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 scale exponentially (m^p)
    - p :: parameter count
    - m :: parameter resolution

Microarchitectural Sampling

- **Temporal Sampling**
  - Sample from instruction traces in time domain
  - Reduce simulation costs via size of inputs
  - Synthetic traces from profiled workloads [1]
  - Sampled traces from phase analysis [2]
- **Spatial Sampling**
  - Sample from design space
  - Reduce simulation costs via number of simulations

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

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**
  - Low formulation costs (1K samples from 1B designs)
  - Accurate inference ($5-7\%$ median error)
  - Efficient computation (100's of predictions per second)

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[3]

[3] Harrell [Springer,'01]

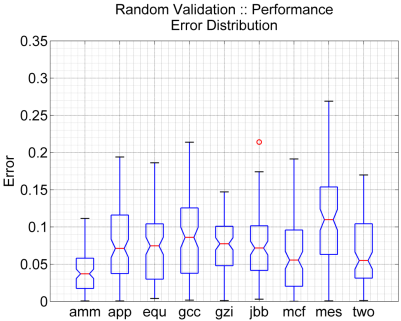
Predictors :: Microarchitecture

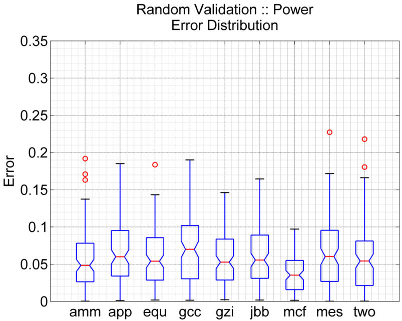
	Set	Parameters	Measure	Range	|S|
S_1	Depth	depth	FO4	9::3::36	10
S_2	Width	width	insn b/w	4,8,16	3
		L/S reorder queue	entries	15::15::45	
		store queue	entries	14::14::42	
		functional units	count	1,2,4	
S_3	Physical Registers	general purpose (GP)	count	40::10::130	10
		floating-point (FP)	count	40::8::112	
		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(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





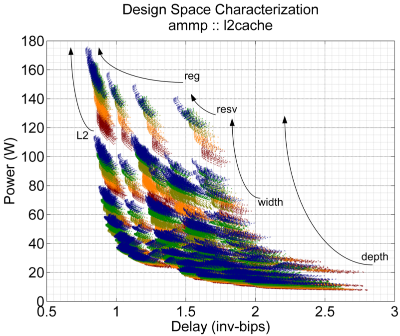
Outline

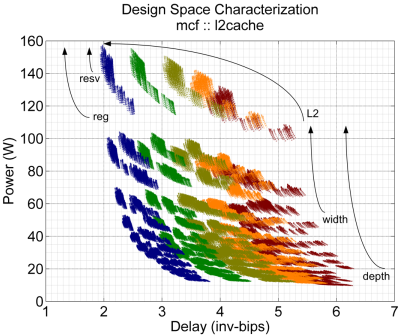
- Motivation & Background
  - Simulation Challenges
  - Simulation Paradigm
  - Regression Theory
- Sampling and Modeling
  - Experimental Methodology
  - Model Evaluation
- Design Optimization
  - Pareto Frontier
  - Multiprocessor Heterogeneity
- Conclusion

Pareto Frontier Analysis

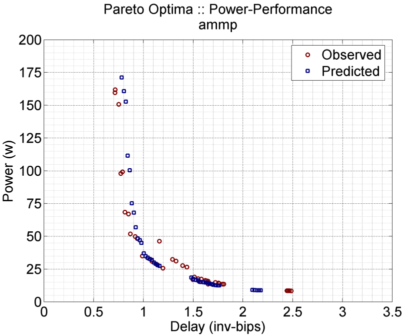
- **Background**
  - Pareto optimization improves at least one metric without negatively impacting any other metric
  - Pareto frontier is set of pareto optima
- **Objective**
  - Construct pareto frontiers for the power-delay space
- **Approach**
  - Simulate 1K samples from design space
  - Formulate regression models for performance, power
  - Exhaustively evaluate models to identify frontier

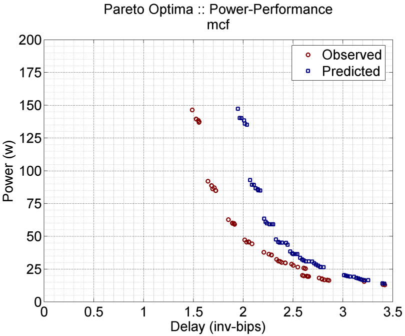
Design Space Characterization





Pareto Frontier



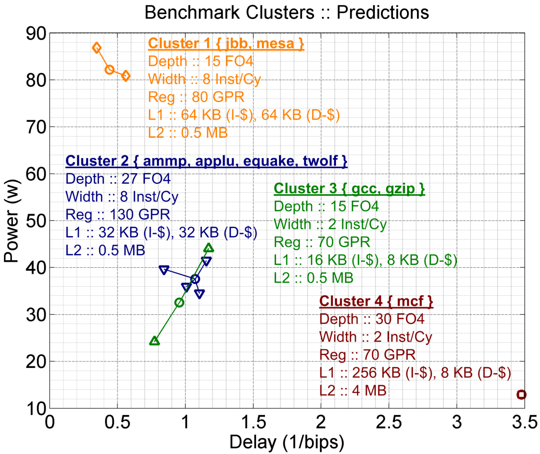


Multiprocessor Heterogeneity

- **Background**
  - Prior heterogeneity studies constrained design options[4]
- **Objective**
  - Identify efficient heterogeneous design compromises
  - Mitigate penalties from single design 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

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

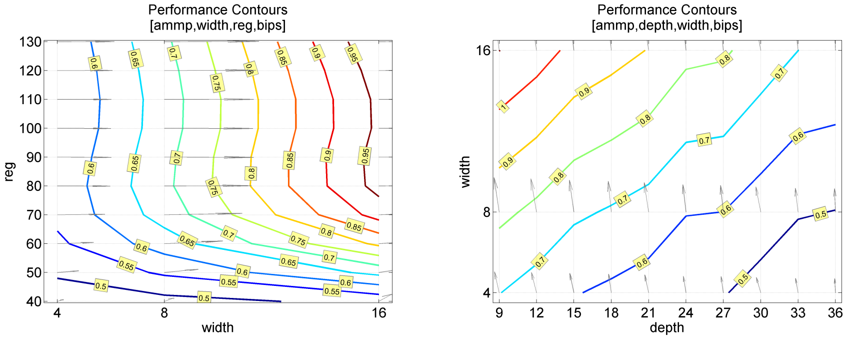
Multiprocessor Heterogeneity II



Outline

Continuing Work I

- **Topology Visualization**



Continuing Work II

- **Topology Roughness Metrics**[5]

$$
\begin{aligned}
R_1 &= \int_x f''(x)^2 dx \\
R_2 &= \int_{x_2}\int_{x_1}\left\{\left(\frac{\delta^2 f}{\delta x_1^2}\right)^2+\left(\frac{\delta^2 f}{\delta x_1 \delta x_2}\right)^2+\left(\frac{\delta^2 f}{\delta x_2^2}\right)^2\right\}dx_1 dx_2
\end{aligned}
$$

- **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)

[5] Green+[Monographs Stat & App Prob]

Conclusion

- **Exploration Paradigm**
  - Comprehensively understand design space
  - Selectively simulate modest number of designs
  - Efficiently leverage simulation data with inference
- **Regression Modeling**
  - Statistical analysis for robust, efficient models
  - Low formulation costs with accurate inference
  - Computationally efficient
- **Model Evaluation**
  - 7.2%, 5.4% median errors for performance, power
  - Applicable to practical design optimization

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.