Statistical Inference for Efficient Microarchitectural and Application Analysis

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

CHALLENGES

Costly parameter space exploration, optimization

- ❑ Exponentially increasing parameter space size
- ❑ Parameters :: Hardware design, software tuning
- ❑ Hardware :: Cycle-accurate simulation
- ❑ Software :: Execution-based profiling

OBJECTIVES

Comprehensively understand parameter space

- ❑ Specify large, high-resolution parameter space

Selectively measure modest number of points

- ❑ Sample points randomly from space for measurement

Efficiently leverage measured data with inference

- ❑ Enable prediction for metrics of interest from sparse sampling

REGRESSION & SPLINE MODELS

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

- ❑ Response (y) modeled as weighted sum of predictors (x)
- ❑ Interaction specified by products ($x_3 = x_1 x_2$)
- ❑ Non-linearity captured by restricted cubic splines ($g = rcs(x,k)$)

Derivation Overview

- ❑ Hierarchical Clustering :: eliminate redundant predictors
- ❑ Correlation Analysis :: assess predictor strength
- ❑ Model :: specify predictor interaction, non-linearity
- ❑ Residual Analysis :: assess model bias
- ❑ Significance Testing :: assess predictive ability of model terms

Optimizations

- ❑ Regional Sampling :: train model with most relevant samples

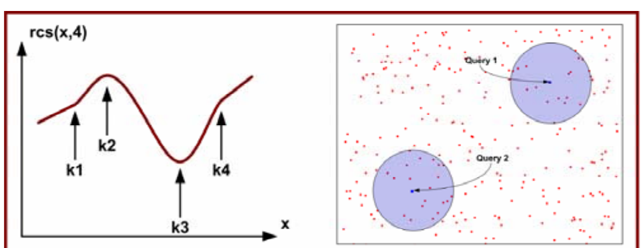

Figure 1 :: Splines & Regional Sampling

MICROARCHITECTURAL DESIGN

- ❑ Depth, width, register file, reservation stations, L1/L2 cache
- ❑ Simulated with Turandot/PowerTimer based on POWER4/5

Validation :: Regression vs Simulation

- ❑ Performance :: 7.4% median error
- ❑ Power :: 4.3% median error

Case Study :: Heterogeneous Multiprocessors

1. Simulate 1K samples from 375K designs to formulate model
2. Identify modeled bips^3/w optimal design for each benchmark
3. K-means cluster optima to identify compromise designs

[mcf] [jbb mesa] [gcc gzip] [ammp applu equake twolf]

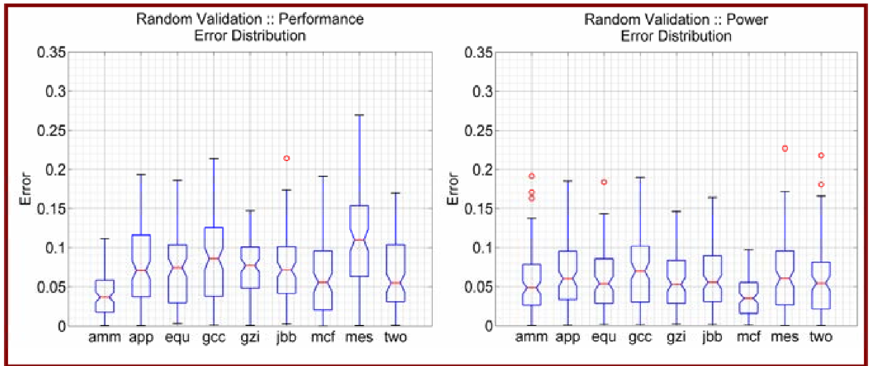

Figure 2 :: Microarchitectural Validation

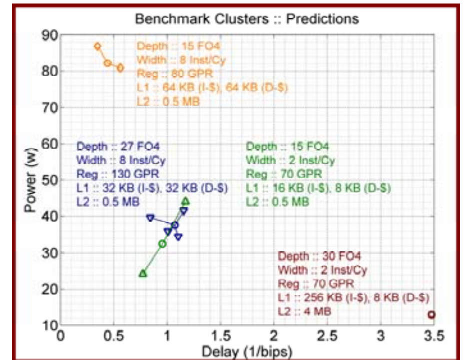

Figure 3 :: Heterogeneous Clusters

APPLICATION TUNING

- ❑ Semicoarsening Multigrid (SMG) :: node topology, workload size
- ❑ High-Perf. Linpack (HPL) :: block size, node topology, factoring
- ❑ Executed on IBM Blue Gene/L, Intel Xeon clusters (ALC/MCR)

Validation :: Regression vs Execution

- ❑ SMG Performance :: 8.5% median error
- ❑ HPL Performance :: 3.1% median error

Case Study :: Application Performance Gradients

1. Run 600 samples from 3K parameter sets to formulate model
2. Compute modeled execution time for every point in space
3. Compute numerical performance gradients with local differences

Future Work

- ❑ Combine microarchitecture, application models in joint prediction

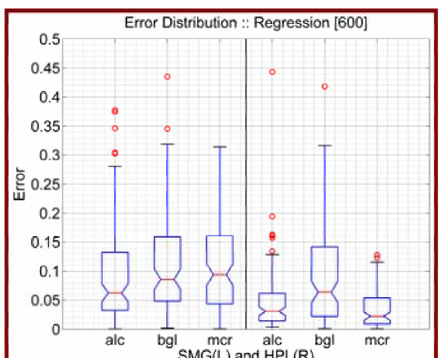

Figure 4 :: Application Validation

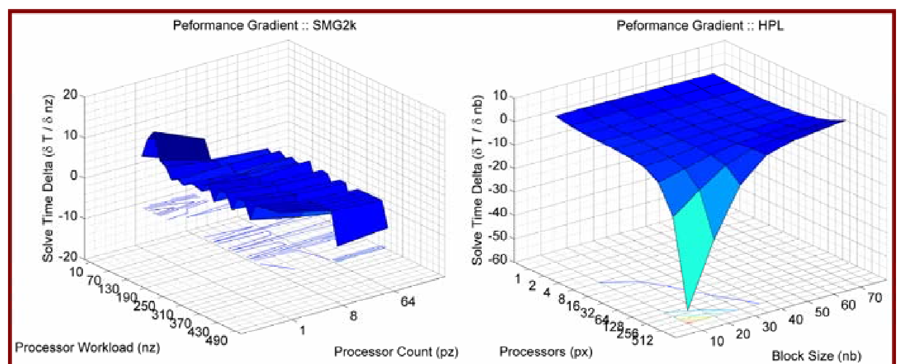

Figure 5 :: Performance Gradients