Efficiency Trends and Limits from Comprehensive Microarchitectural Adaptivity

Benjamin C. Lee, David M. Brooks

www.seas.harvard.edu/∼bclee
School of Engineering and Applied Sciences
Harvard University

Architectural Support for Programming Languages and Operating Systems
3 March 2008

Adaptive Microarchitectures

- **Technology Trends**
  - Increasing transistor budgets
  - Abundant microarchitectural resources
  - Power as constraining design metric
- **Design Paradigm**
  - Allocate hardware resources at run-time
  - Match hardware to application dynamics
  - Enhance performance, localize power costs
- **Cost-Benefit Analysis**
  - Costs :: implementation complexity, control overheads
  - Benefits :: performance, power efficiency

Dimensionality Challenges

- **Temporal Adaptivity**
  - Frequency of hardware reconfiguration
  - Increases responsiveness to application
- **Spatial Adaptivity**
  - Scope of hardware reconfiguration
  - Exposes parameter synergies, interactions
- **Optimization Framework**
  - Mechanisms to sample workloads, designs
  - Models to estimate design metrics
  - Heuristics to traverse design topology

Outline

Motivation & Background
- Adaptive Microarchitectures
- Dimensionality Challenges

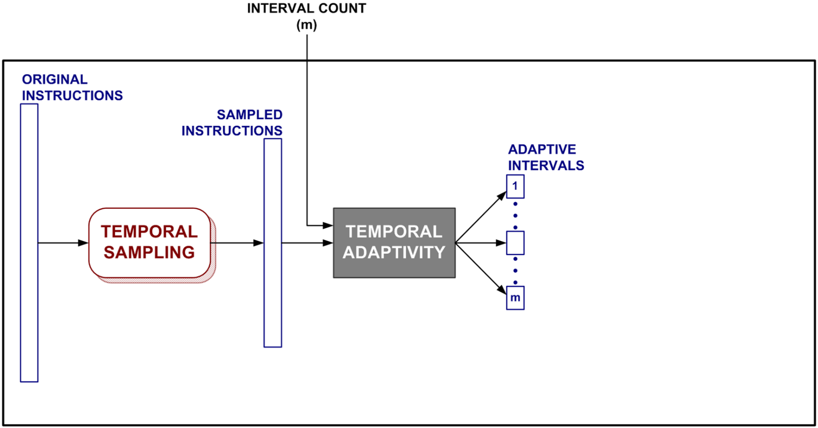
Analysis Framework
- Sampling
- Modeling
- Optimization

Comprehensive Adaptivity
- Temporal Adaptivity
- Spatial Adaptivity

Outline

- Motivation & Background
  - Adaptive Microarchitectures
  - Dimensionality Challenges
- Analysis Framework
  - Sampling
  - Modeling
  - Optimization
- Comprehensive Adaptivity
  - Temporal Adaptivity
  - Spatial Adaptivity

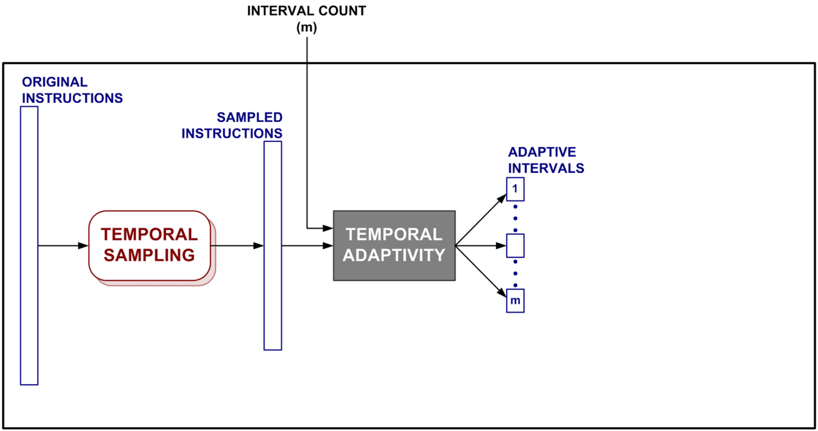
Temporal Adaptivity & Sampling



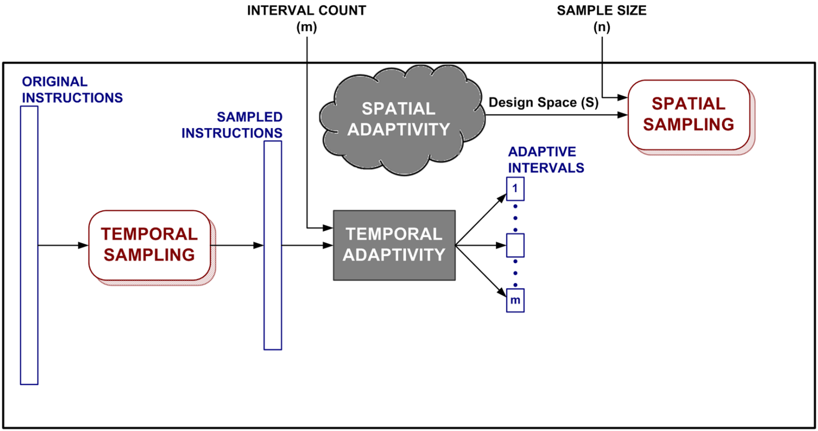
Temporal Adaptivity & Sampling

- **Benchmarks**
  - SPECcpu :: compute intensive
  - SPECjbb :: Java server
  - SPLASH :: numerical methods, scientific computing
  - BIOPERF :: bioinformatics
- **Temporal Adaptivity**
  - 100M representative instructions
  - 1,125 adaptive intervals
  - 80K instructions per interval

Spatial Adaptivity & Sampling



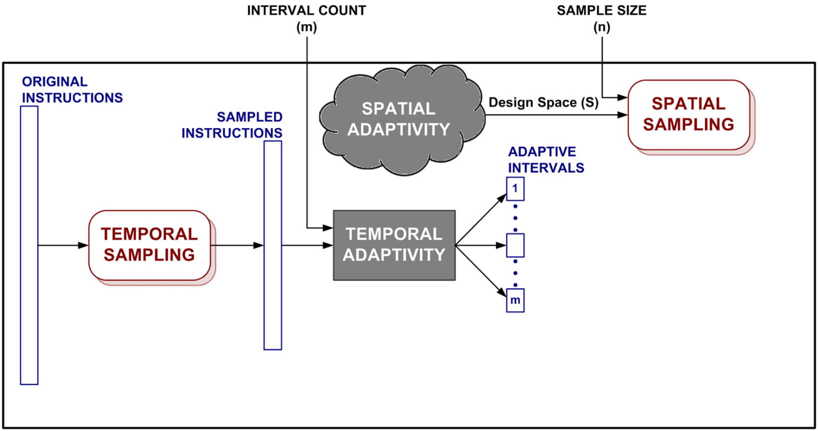
Spatial Adaptivity & Sampling



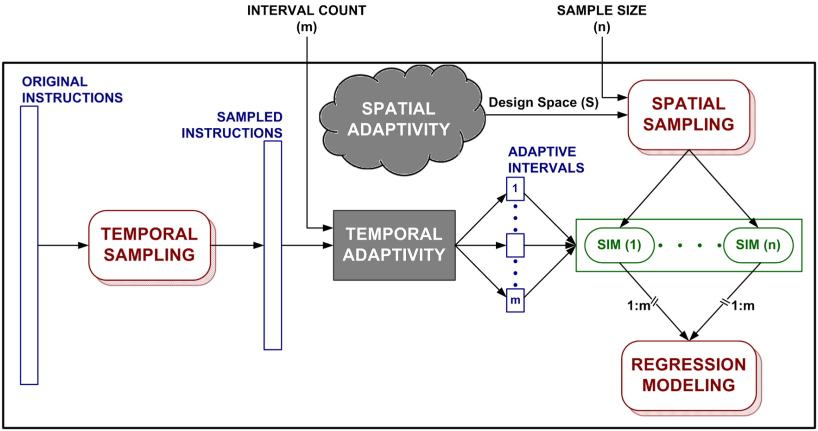
Spatial Adaptivity & Sampling

	Set	Parameters	Measure	Range	|S|
S_1	Depth	depth	FO4	9::3::36	10
S_2	Width	width	inst decode b/w	2,4,8	3
		functional units	count	1,2,4	
S_3	Branch Predictor	BTB associativity	sets	1,2,4,8	4
		BTB size	log_2(entries)	12::1::15	
S_4	Load/Store Queue	load/store queue	entries	9::5::54	10
S_5	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_6	Reservation Stations	branch	entries	6::1::15	10
		fixed-point/memory	entries	10::2::28	
		floating-point	entries	5::1::14	
S_7	I-L1 Cache	i-L1 cache size	KB	16::2x::256	5
S_8		i-L1 cache assoc.	sets	1,2,4,8	4
S_9	D-L1 Cache	d-L1 cache size	KB	8::2x::128	5
S_{10}		d-L1 cache assoc.	sets	1,2,4,8	4
S_{11}		load/store latency	cycles	1::1::5	5
S_{12}	L2 Cache	L2 cache size	MB	0.25::2x::4	5
S_{13}		L2 cache assoc.	sets	1,2,4,8	4
S_{14}		L2 cache latency	cycles	8::2::16	5
S_{15}	Main Memory	main memory latency	cycles	70::5::115	10

Simulation & Regression



Simulation & Regression



Simulation Framework

- **Simulation Paradigm**[1]
  - Comprehensively understand design space
  - Selectively simulate modest number of designs
  - Efficiently leverage simulation data with inference
- **Simulator**
  - Turandot :: a cycle-accurate trace driven simulator
  - PowerTimer :: power models derived from circuit analyses
  - Baseline simulator models POWER4/POWER5 architecture
- **Statistical Framework**
  - R :: software environment for statistical computing

[1] Lee+[ASPLOS'06]

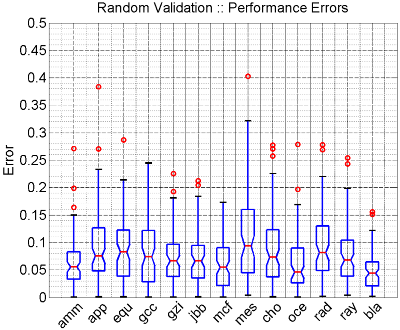
Statistical Inference

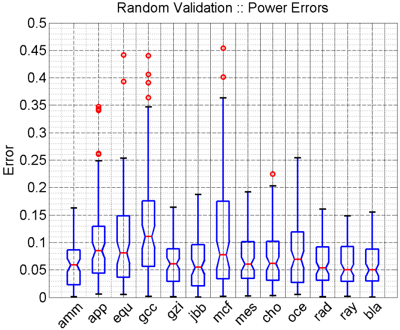
- **Regression Models**
  - y :: design metrics (performance, power)
  - X :: design parameters (depth, width, ...)

$$F(y) = G(X)\beta + \varepsilon$$

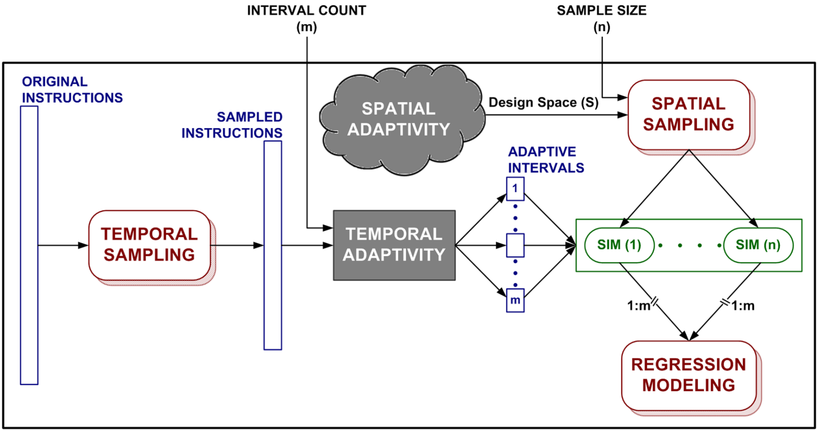
  - Low formulation costs (500 samples from 240B designs)
  - Accurate inference (6% median error)
  - Efficient computation (100's of predictions per second)
- **Validation**
  - Obtain 100 additional random samples for validation
  - Simulator-reported versus regression-predicted

Model Accuracy

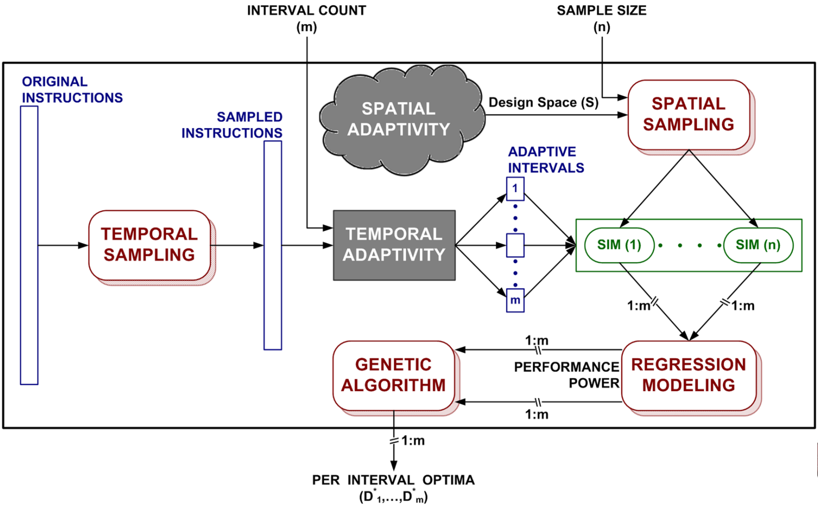




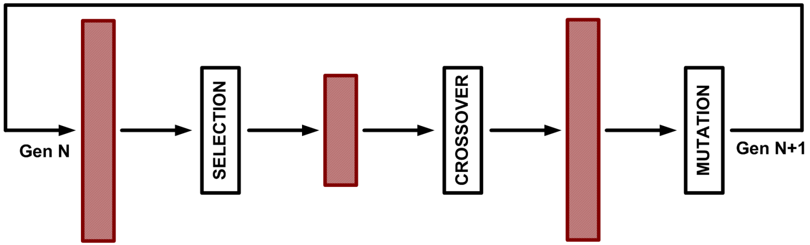
Genetic Optimization



Genetic Optimization

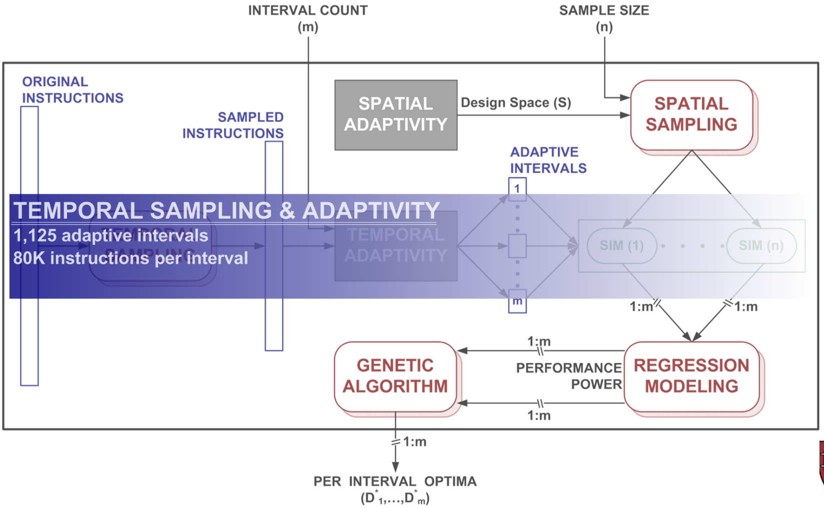


Genetic Optimization

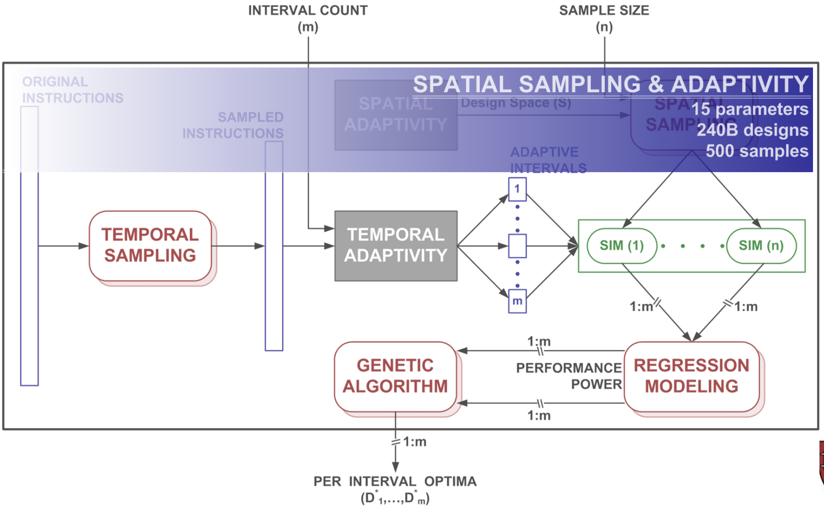


- Improves population of designs across multiple generations
- Population size = 10^2, Generations = 10^2
- Cost = 10^4 predictions per interval, Intervals = 10^3

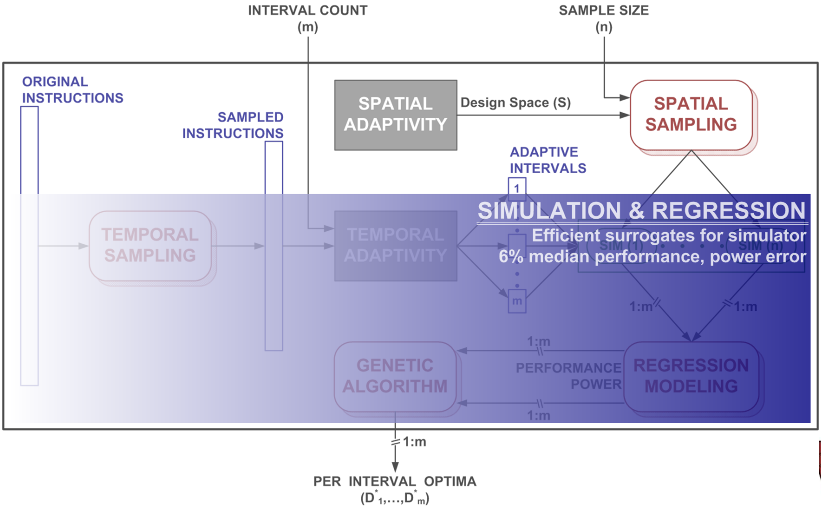
Analysis Framework



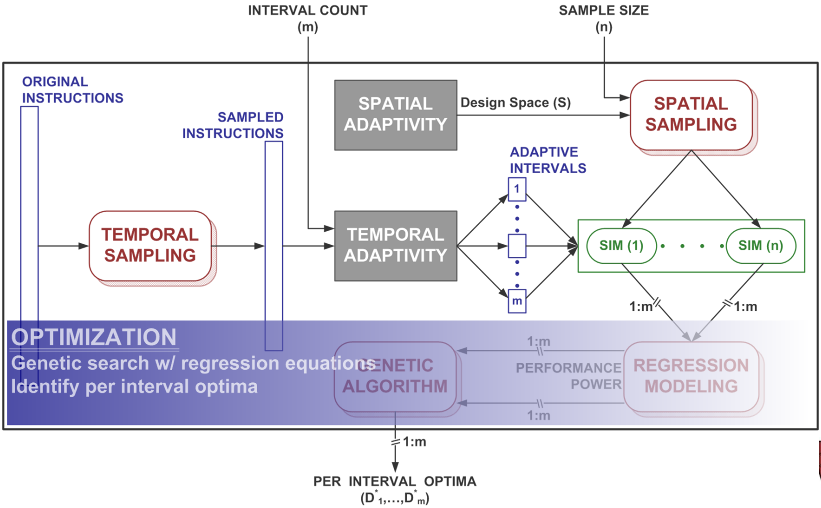
Analysis Framework



Analysis Framework



Analysis Framework



OPTIMIZATION

Genetic search w/ regression equations

Identify per interval optima

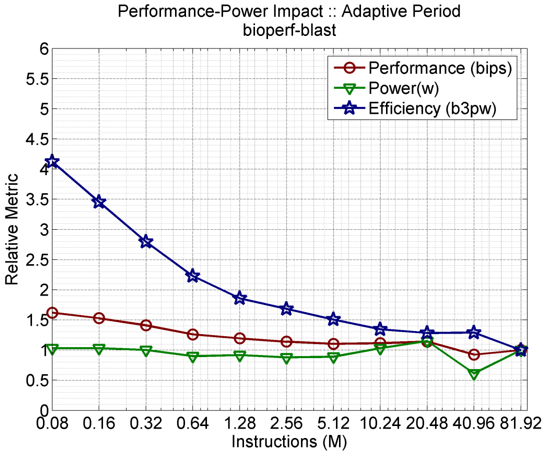
Outline

- Motivation & Background
  - Adaptive Microarchitectures
  - Dimensionality Challenges
- Analysis Framework
  - Sampling
  - Modeling
  - Optimization
- Comprehensive Adaptivity
  - Temporal Adaptivity
  - Spatial Adaptivity

Temporal Adaptivity

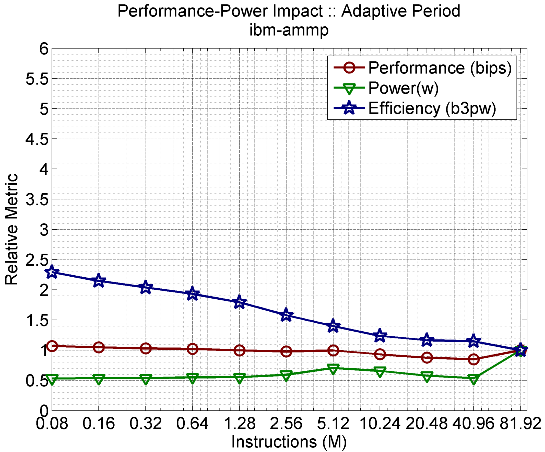
- **Definition**
  - Frequency of hardware reconfiguration
  - High adaptivity :: short intervals
- **Analysis**
  - Assume high spatial adaptivity (15 parameters)
  - Vary interval size from 80K to 80M instructions
- **Evaluation**
  - Efficiency increases with temporal adaptivity
  - 5.3x maximum, 2.4x median $bips^3/w$ gain

Varying Temporal Adaptivity I



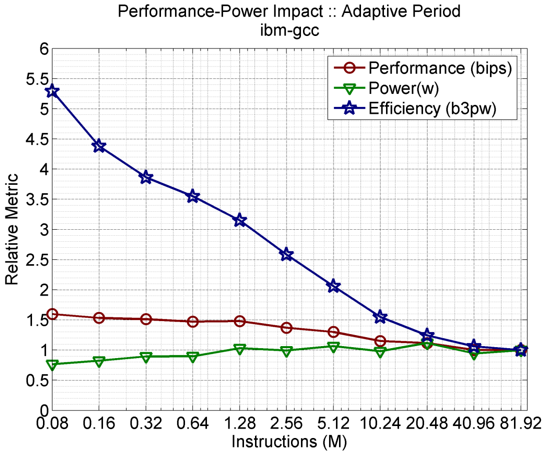
- ($\Uparrow$) Performance, ($\Longleftrightarrow$) Power

Varying Temporal Adaptivity II



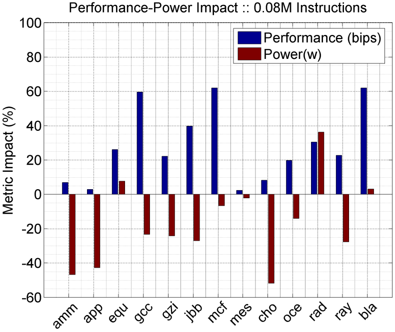
- ($\Longleftrightarrow$) Performance, ($\Downarrow$) Power

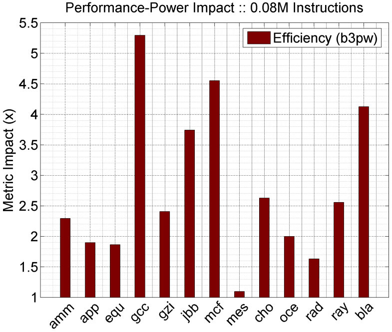
Varying Temporal Adaptivity III



- ($\Uparrow$) Performance, ($\Downarrow$) Power

Performance & Power Impact

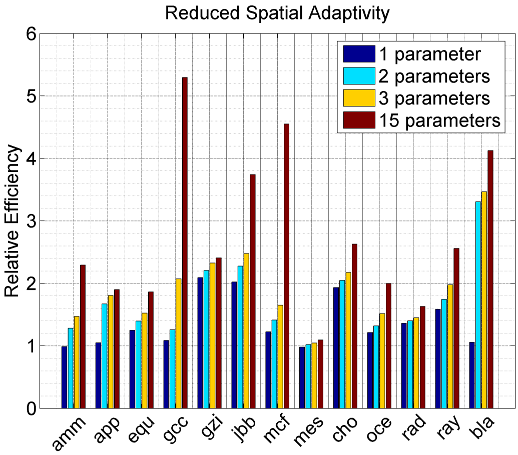




Spatial Adaptivity

- **Definition**
  - Scope of hardware reconfiguration
  - High adaptivity :: many parameters
- **Analysis**
  - Assume high temporal adaptivity (80K-instruction intervals)
  - Identify $k = 1, 2, 3$ most significant adaptive parameters
  - Evaluate each of $\binom{15}{k}$ possible combinations
- **Evaluation**
  - Achieve 77% of potential with 3 of 15 parameters
  - Significant parameters differ across workloads

Reduced Spatial Adaptivity



- Average 60%, 71%, 77% of potential with 1, 2, 3 parameters

Parameter Significance

	amm	app	equ	gcc	gzi	jbb	mcf	mes	cho	oce	rad	ray	bla
depth	1	2	1	1	1	1	1	1	1	2	1	1	2
width									2	3	2		
bp													
lsq											3		
reg													
resv											2*		
i1Size													
i1Assoc													
d1Size					2			2				2	
d1Assoc						3			3				
d1Lat						2							3
l2Size		3			3							3	
l2Assoc	3		2	2*			3		2*				
l2Lat			3	2			2						
memLat	2	1		3				3		1		2*	1

Parameter Significance

	amm	app	equ	gcc	gzi	jbb	mcf	mes	cho	oce	rad	ray	bla
depth	1	2	1	1	1	1	1	1	1	2	1	1	2
width									2	3	2		
bp													
lsq											3		
reg													
resv											2*		
i1Size													
i1Assoc													
d1Size					2			2				2	
d1Assoc						3			3				
d1Lat						2							3
l2Size		3			3							3	
l2Assoc	3		2	2*			3		2*				
l2Lat			3	2			2						
memLat	2	1		3				3		1		2*	1

Parameter Significance

	amm	app	equ	gcc	gzi	jbb	mcf	mes	cho	oce	rad	ray	bla
depth	1	2	1	1	1	1	1	1	1	2	1	1	2
width									2	3	2		
bp													
lsq											3		
reg													
resv											2*		
i1Size													
i1Assoc													
d1Size					2			2				2	
d1Assoc						3			3				
d1Lat						2							3
l2Size		3			3							3	
l2Assoc	3		2	2*			3		2*				
l2Lat			3	2			2						
memLat	2	1		3				3		1		2*	1

Outline

Conclusion
- Paper Details
- Conclusion

Also in the paper...

- **Analysis Framework**
  - Genetic algorithms
  - Framework synergies
- **Temporal Adaptivity**
  - Number, range of adapted parameters
  - Source of performance gains, power savings
- **Spatial Adaptivity**
  - Discussion for $3 < k < 15$ parameters
  - Dynamic voltage/frequency scaling

Conclusion

- **Analysis Framework**
  - Sampling :: sparsely sampled instructions, designs
  - Modeling :: accurate performance, power regression
  - Optimization :: efficient genetic algorithms
- **Potential Efficiency**
  - High temporal, spatial adaptivity
  - 5.3x maximum, 2.4x median $bips^3/w$ gain
  - Motivates rigorous cost analysis
- **Hardware Implications**
  - Achieve 77% of potential with 3 of 15 parameters
  - Significant parameters differ across workloads
  - Motivates comprehensive adaptive hardware substrate

Efficiency Trends and Limits from Comprehensive Microarchitectural Adaptivity

Benjamin C. Lee, David M. Brooks

www.seas.harvard.edu/∼bclee
School of Engineering and Applied Sciences
Harvard University

Architectural Support for Programming Languages and Operating Systems
3 March 2008

Further Reading

www.seas.harvard.edu/∼bclee

- B.C. Lee and D.M. Brooks.
Efficiency trends and limits from comprehensive microarchitectural adaptivity
ASPLOS-XIII: International Conference on Architectural Support for Programming Languages and Operating Systems, March 2008.

- B.C. Lee and D.M. Brooks.
Roughness of microarchitectural topologies and its implications for optimization
HPCA-14: International Symposium on High Performance Computer Architecture, Feb 2008.

- 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.