CHAOS: Composable Highly Accurate OS-based Power Models

John D. Davis, Suzanne Rivoire, Moises Goldszmidt, Ehsan Ardestani

(john.d@microsoft.com)

IISWC 2012

Software power modeling?

- Prediction and forecasting
- Reduce Capex and Opex
  - Data Center
  - System

Cluster power modeling challenges

- Systems

systems

3X Servers

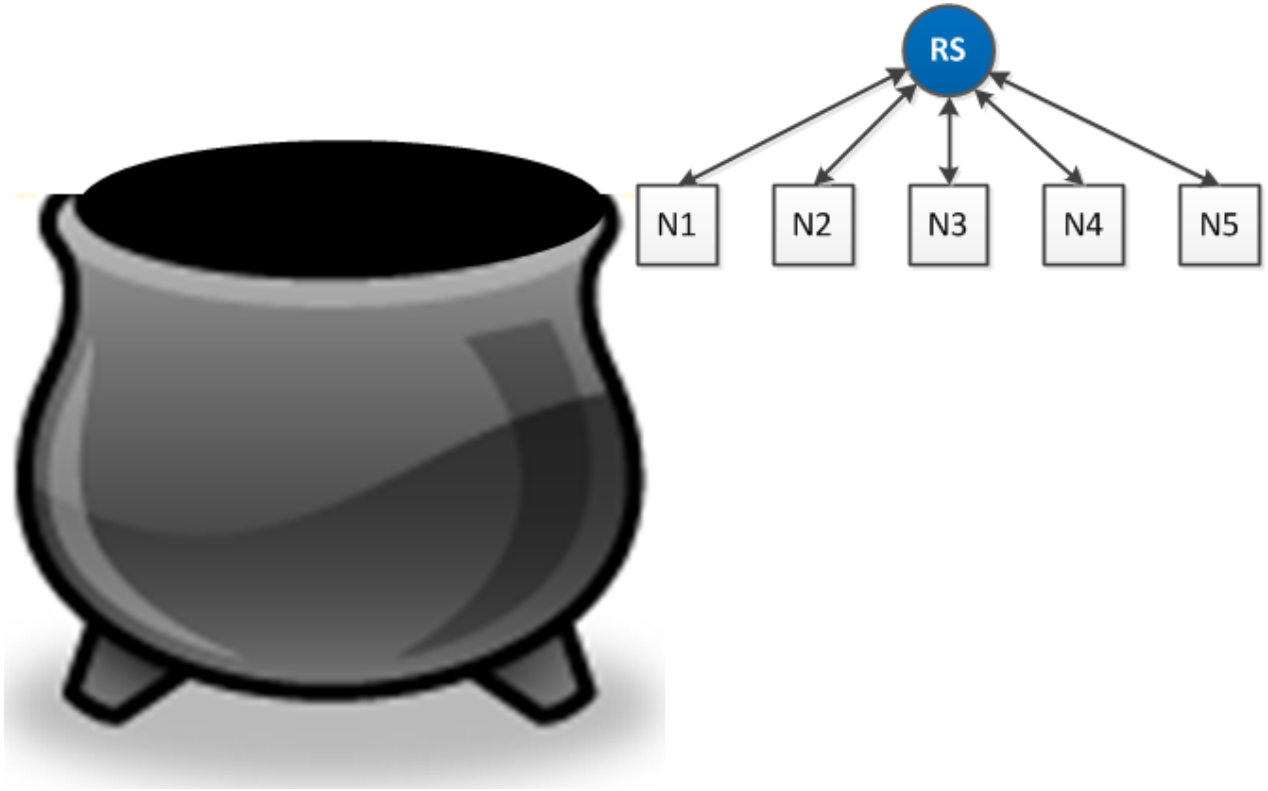
Laptop

Embedded

Desktop

HDDs

SSDs



Cluster power modeling challenges

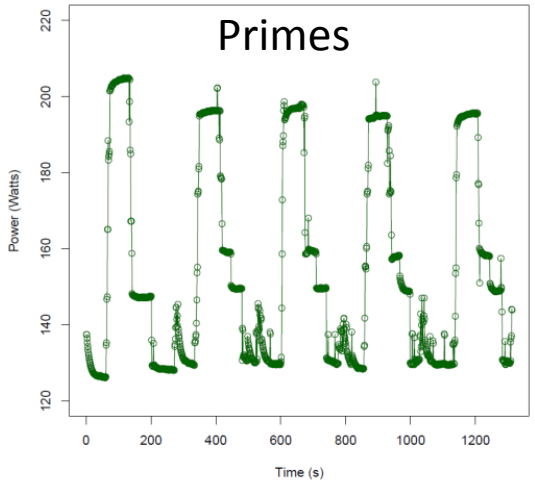
- Applications

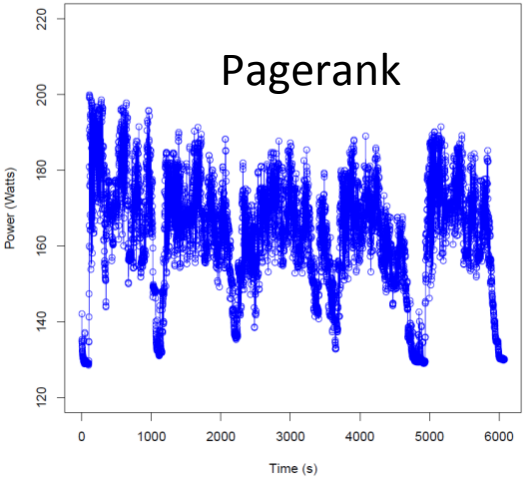
Primes (CPU)

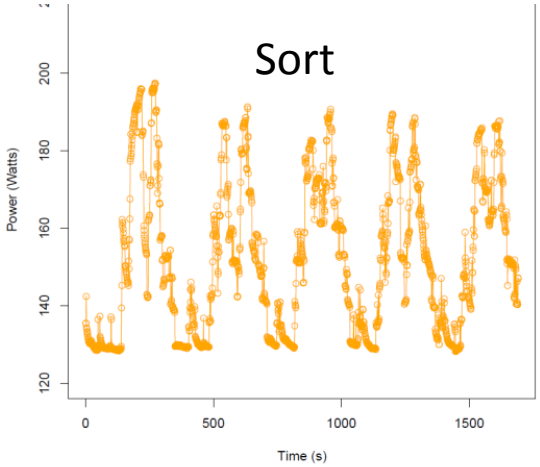
Pagerank (Net)

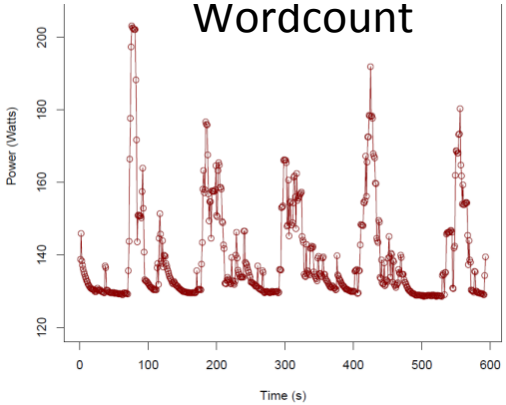
Wordcount (Disk)

Sort (Disk+Net)



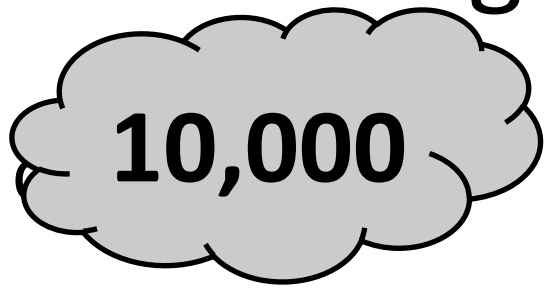






Cluster power modeling challenges

- OS performance



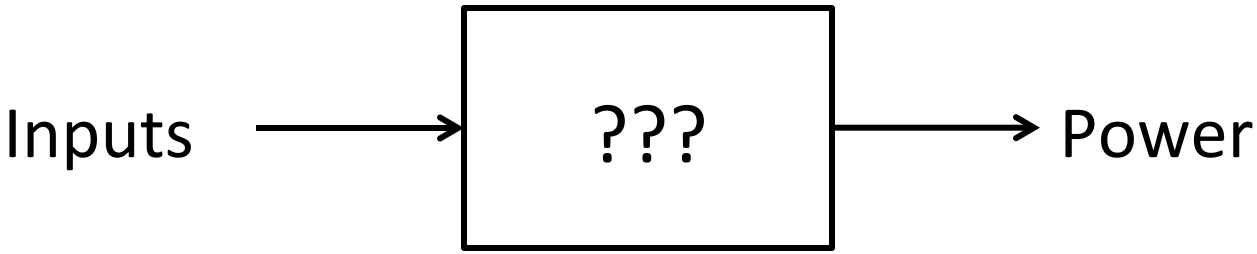
Cluster power modeling challenges

- Apply machine learning??

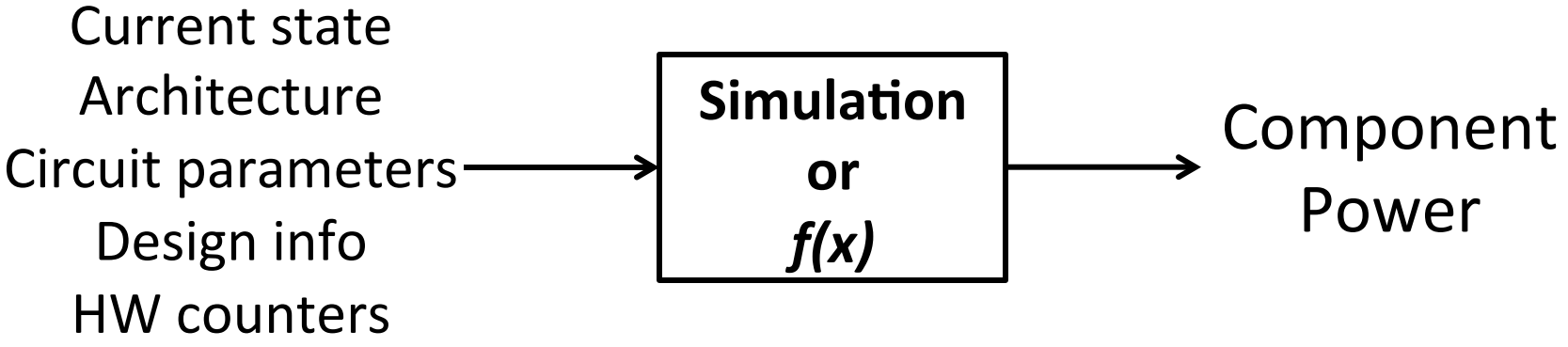
CHAOS power models

- Portability
  - OS performance counters
- Accuracy
  - Dynamic Rang
- Scalability
  - Online
  - Machine mod
- Inexpensive

What is a Power Model?

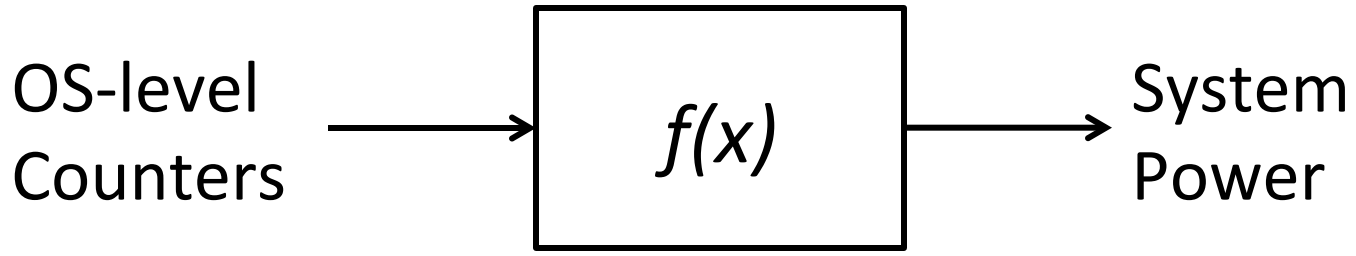


Previous power models



- Fidelity spectrum (low ↔ high)
- Not full-system power
- Slow (not real-time) and/or complex, require specialized knowledge
- Not portable

CHAOS power models

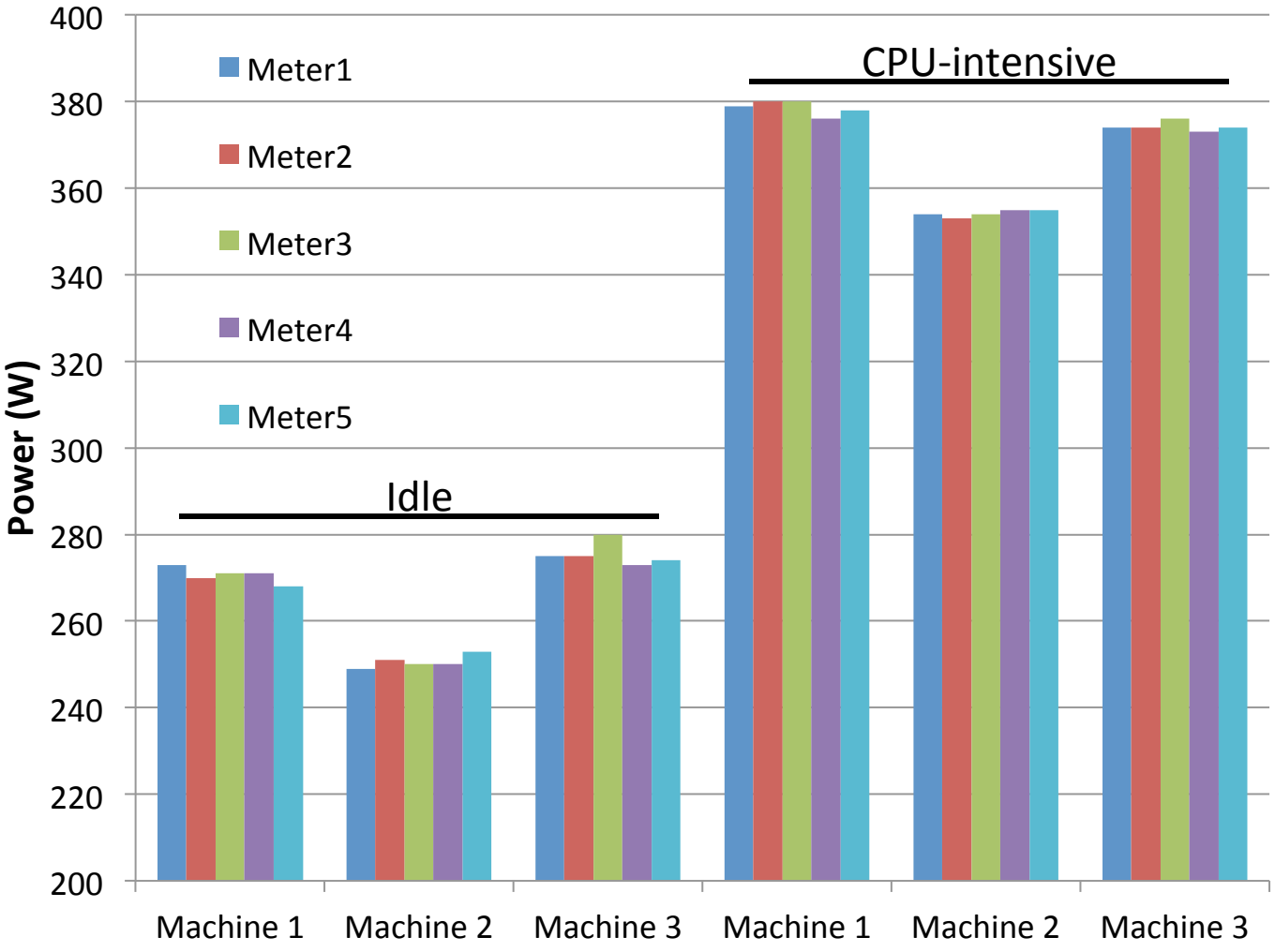


- Portable
  - Framework, model features, etc.
- How accurate?
  - Machine-to-machine power variability
  - Tradeoff between model parameters/complexity and accuracy
- How scalable?
  - Low-overhead, sampling theory and cluster model

Cluster power modeling methods

- (A) Cluster Power = $f_{machine}(x_{machine})$ *X N*
  - *Single model, inputs from a single machine*
  - Implicitly or explicitly assumed by previous work
  - 1 machine model
- (B) Cluster Power = $\Sigma_i f_{machine}(x_{machine_i})$
  - *Single model, inputs from each machine*
  - *How many machines to train the model?*
- (C) Cluster Power = $\Sigma_i f_{machine_i}(x_{machine_i})$
  - *Model per machine, inputs from each machine*
  - *Too many models*

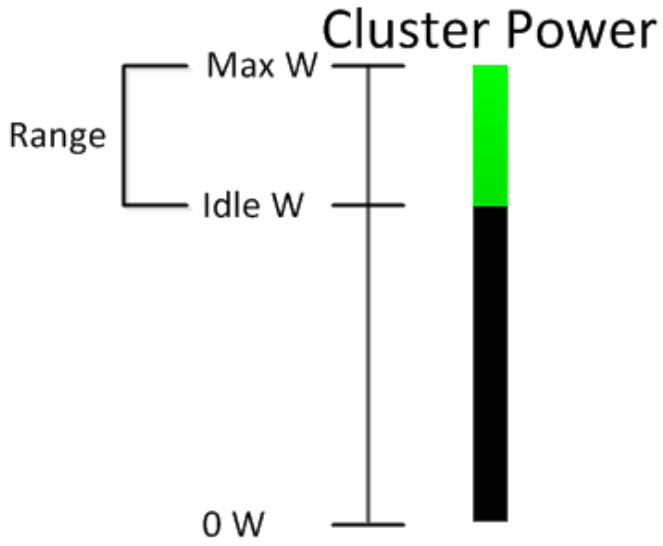
Server power variability



Dynamic Range Error

- New power model error metric
  - Replaces MSE or average/median error

$$Error\,(DRE) = \sqrt{Mean\ Square\ error} \,/\, Max\ Power{\downarrow}Cluster - Min\ Power{\downarrow}Cluster$$



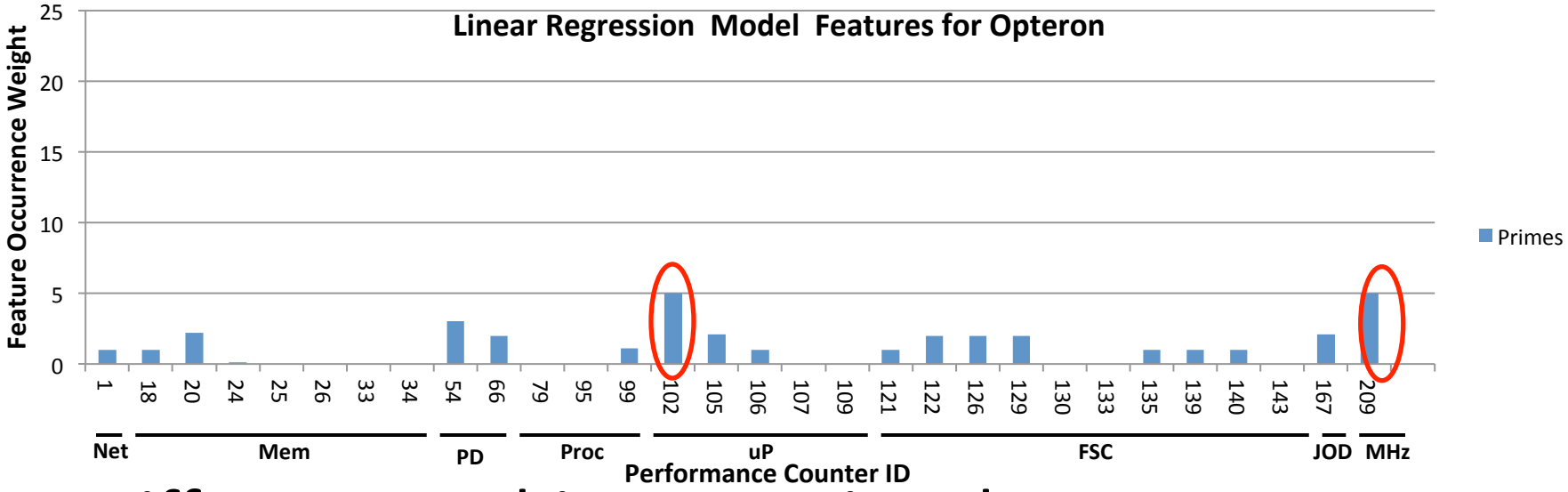
Feature selection

- ETW (Event Tracing for Windows)
  - Performance counters: **~10,000**
- Architecture selected: **~250**
  - Processor and frequency, physical and logical disk, network, memory, application
- Remove dependent performance counters: **~45**
  - Correlation Matrix (> |0.95|)
  - Performance counter definitions
- Linear regression to select model features
- Stepwise regression to remove insignificant features: **~10**

Selected features across all models

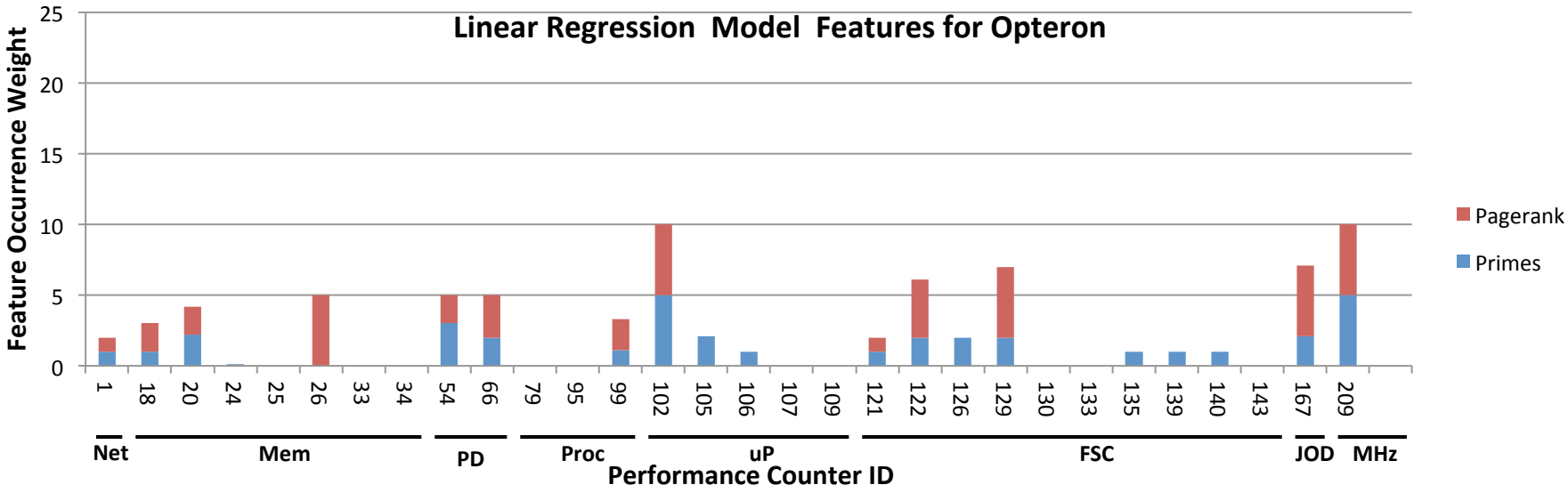
Category	Performance counter	counter ID
Network (Net)	Datagram/sec	1
Memory (Mem)	Page Faults/sec	18
	Committed Bytes	20
	Cache Faults/sec	24
	Pages/sec	26
	Page Reads/sec	28
	Pool Nonpaged Allocs	34
Physical Disk (PD)	Disk Total Disk Time %	54
	Disk Total Disk Bytes/sec	66
Process (Proc)	Total Page Faults/sec	79
	Total IO Data Bytes/sec	99
Processor (uP)	Total Processor Time % (Utilization)	102
	Total Processor Interrupts/sec	105
	Total Processor % DPC Time	106
File System Cache (FSC)	Data Map Pins/sec	121
	Pin Reads/sec	122
	Pin Read Hits %	125
	Copy Reads/sec	126
	Fast Reads not Possible/sec	139
	Lazy Write Flushes/sec	140
Job Object Details (JOD)	Total Page File Bytes Peak	167
Processor Performance (MHz)	Processor 0 Processor Frequency	209

Opteron (server) features

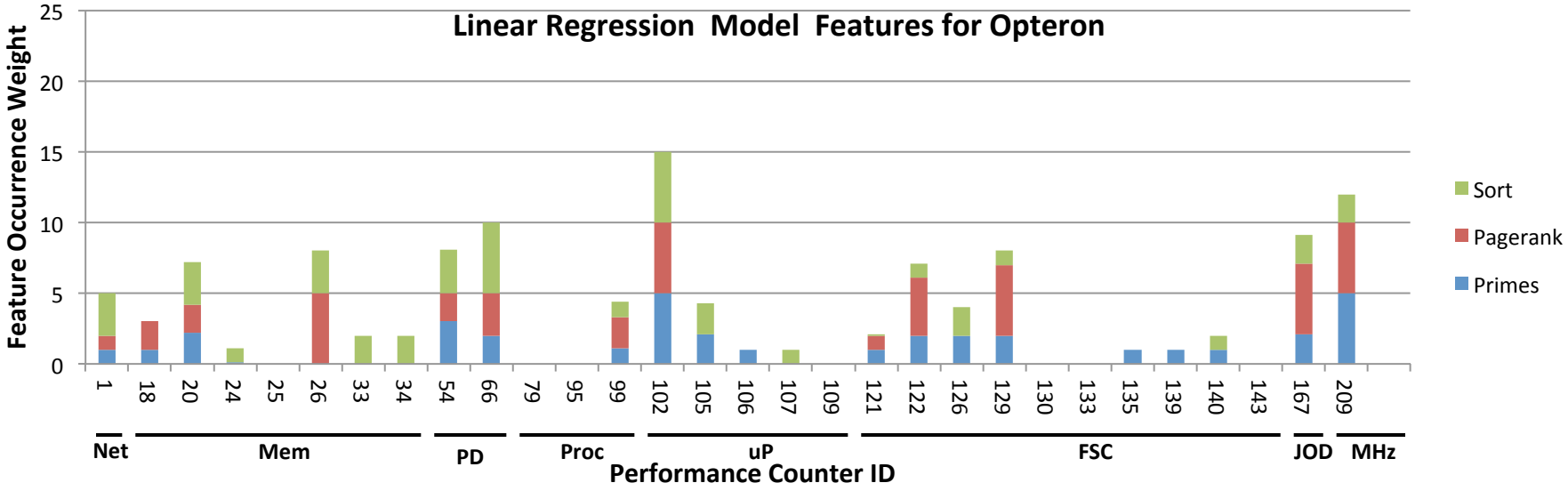


- Different machines running the same workload identify different significant features (height of the bars)

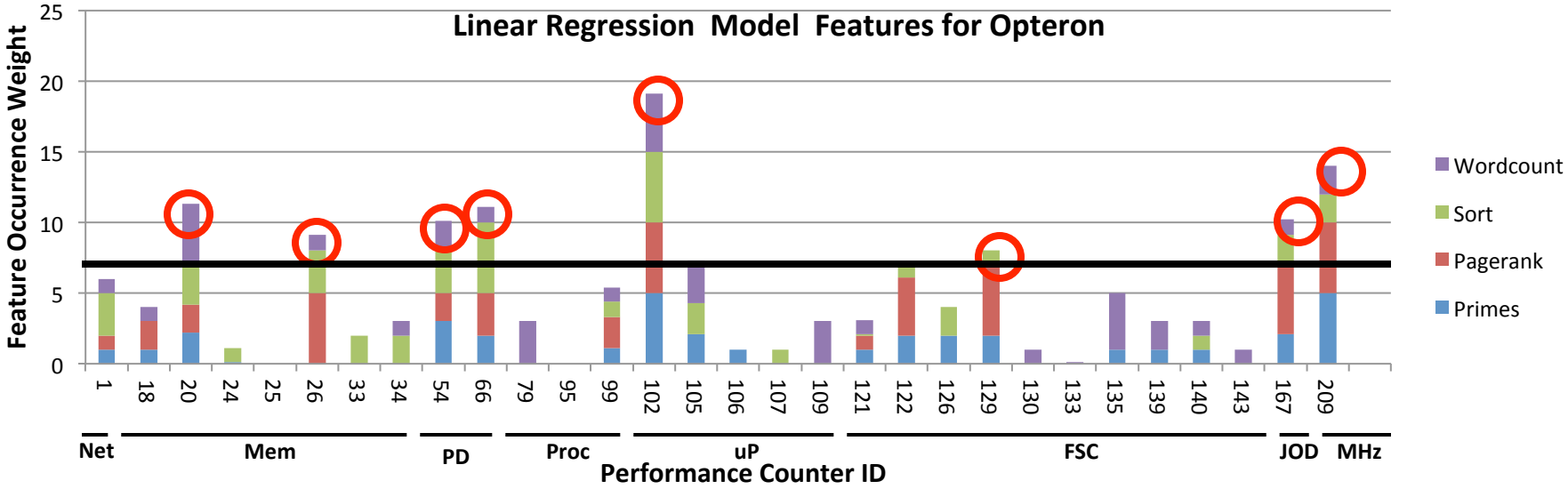
Opteron (server) features



Opteron (server) features

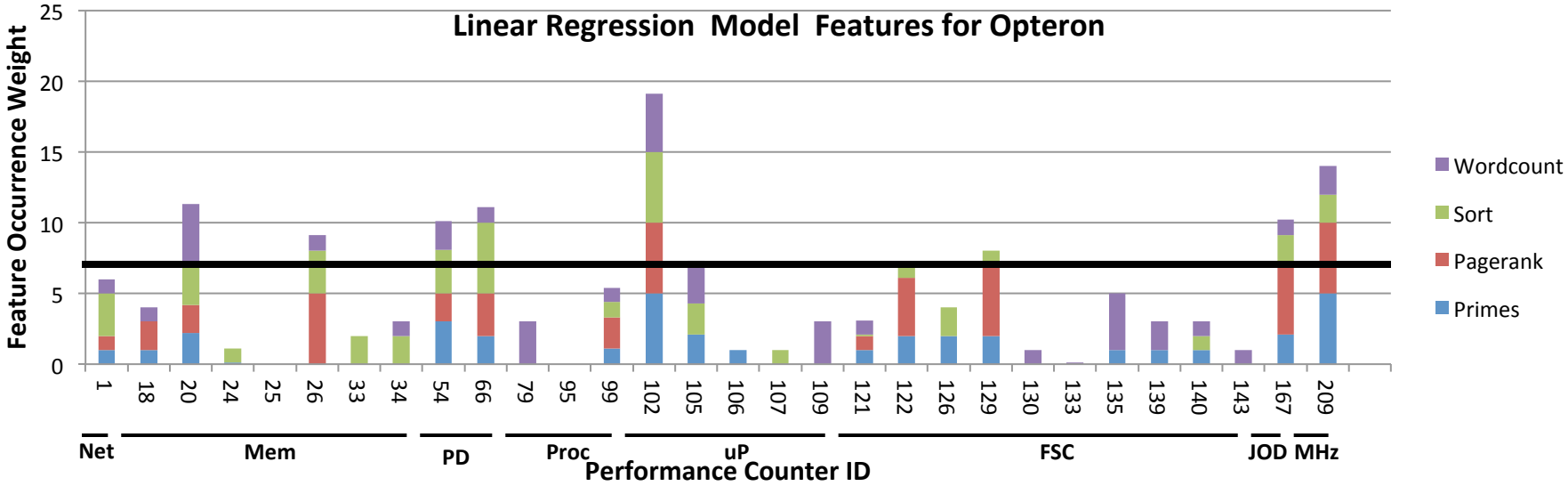


Opteron (server) features



Model

Opteron (server) features



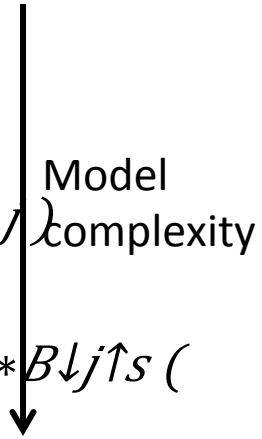
- Stepwise regression removes features that are insignificant across the cluster (defines the threshold value)

Platform model features

Category	Performance counter	Atom	Core2	Athlon	Opteron	Xeon SATA	Xeon SAS	General
Network	Datagram/sec					X	X	
Memory	Page Faults/sec			X		X	X	
	Committed Bytes					X	X	
	Cache Faults/sec	X	X		X			X
	Pages/sec			X		X	X	X
	Page Reads/sec					X	X	
	Pool Nonpaged Allocs	X	X					X
Physical Disk	Disk Total Disk Time %		X		X	X	X	
	Disk Total Disk Bytes/sec	X			X		X	X
Process	Total Page Faults/sec					X	X	
	Total IO Data Bytes/sec			X				
Processor	Total Processor Time % (Utilization)	X	X	X	X	X	X	X
	Total Processor Interrupts/sec					X		
	Total Processor % DPC Time					X		
File System Cache	Data Map Pins/sec	X			X	X	X	
	Pin Reads/sec		X		X	X	X	X
	Pin Read Hits %						X	
	Copy Reads/sec	X						
	Fast Reads not Possible/sec	X					X	
	Lazy Write Flushes/sec	X				X	X	
Job Object Details	Total Page File Bytes Peak	X	X	X	X	X	X	X
Processor Performance	Processor_0 Processor Frequency		X	X	X	X	X	X

Server power models

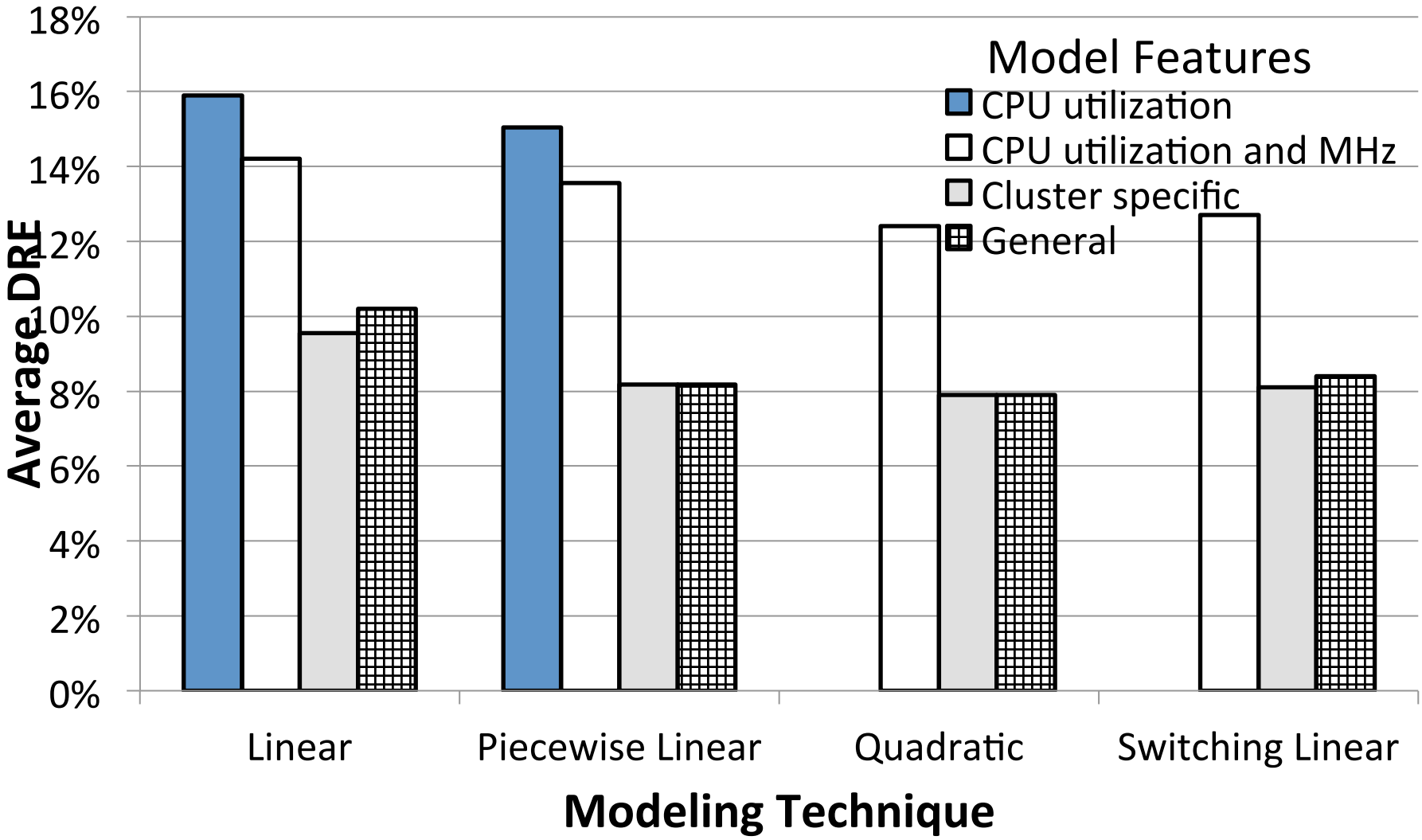
- *Baseline linear power model*:
  - $f(x_1,\ldots,x_n) = a_0 + \sum_i a_i * x_i$
- *Piecewise linear (PWL) power model*:
  - $f(x_1,\ldots,x_n) = a_0 + \sum_i \sum_j a_{i,j} * B_{i,j}^s(x_i, t_{i,j})$
- *Quadratic power model*:
  - $f(x_1,\ldots,x_n) = a_0 + \sum_i \sum_j a_{i,j} * B_i^s(x_i, t_i) * B_j^s(x_j, t_j)$
- *Switching power model*:
  - $f(x_1,\ldots,x_n) = I(f)(a_0 + \sum_i a_i * x_i) + (1 - I(f))(a'_0 + \sum_i a'_i * x'_i)$
- where *I(f) = 1 iff the frequency < threshold; otherwise I(f) = 0.*



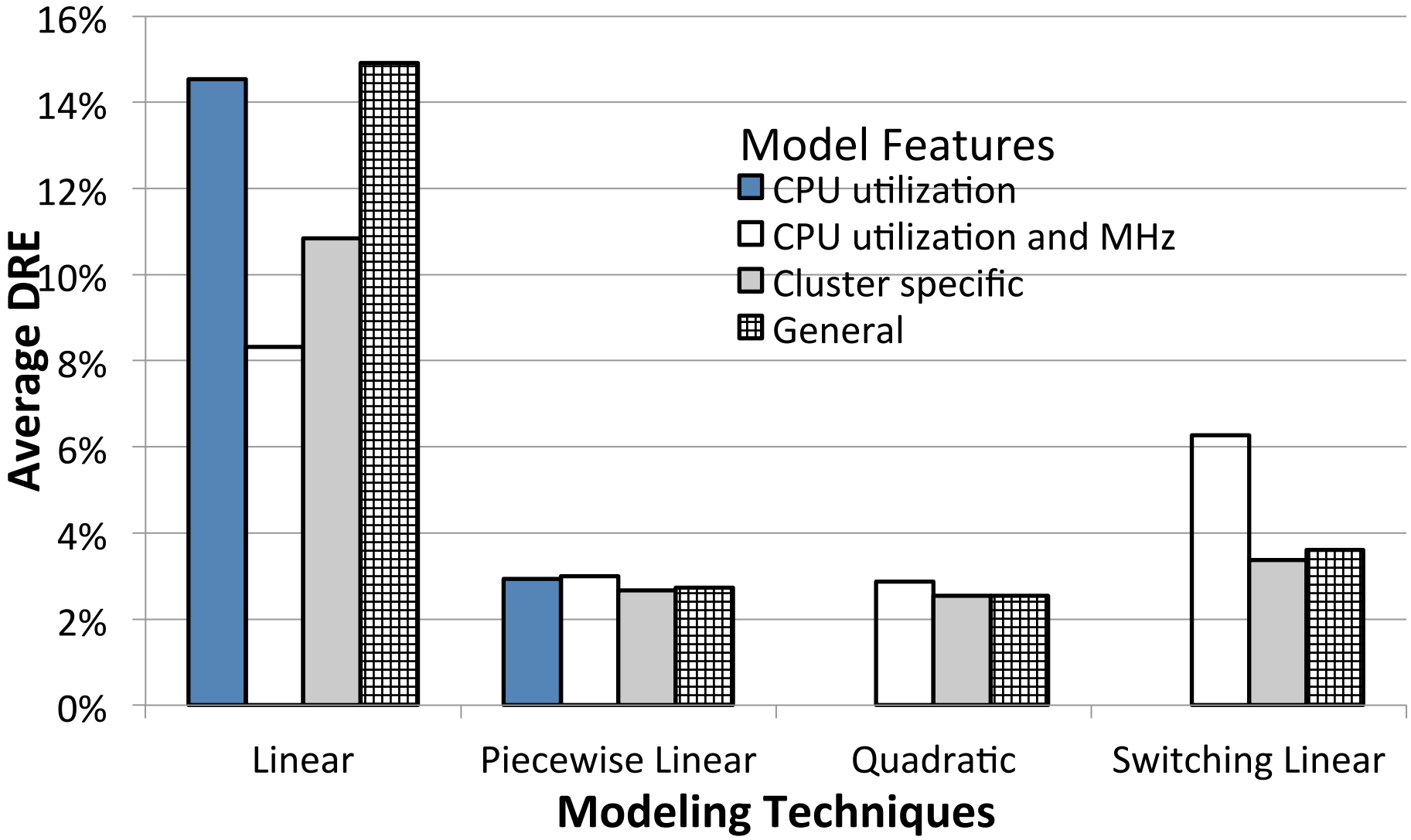
Cluster power model evaluation

- 16 models per cluster
  - Tradeoff feature and modeling complexity
  - CPU utilization to cluster specific features, including frequency history
  - Empirically developed a general set of features.
    - ~1% increase in DRE

Feature selection matters



Modeling technique matters



Cluster model error

Workload	Atom	Core 2	Athlon	Opteron	Xeon SATA	Xeon SAS
PageRank	9.2%, PU	7.4%, QC	8.9%, QC	7.7%, QCP	9.6%, QCP	8.1%, QCP
Prime	10.7%, QC	4.9%, QC	3.6%, QC	2.5%, QC	8.6%, QC	9.9%, QC
Sort	10.2%, QC	7.4%, QC	6.1%, QC	7.9%, QC	11.0%, QG	10.5%, QC
WordCount	11.4%, LC	9.8%, SC	6.0%, QG	7.6%, QC	9.8%, QC	9.2%, QC

- Model|Feature set
  - QC[P]: Quadratic Cluster [Past MHz (t-1)]
  - QG: Quadratic General
  - LC: Linear Cluster
  - PU: Piecewise Linear CPU Utilization
  - SC: Switching Cluster

CHAOS conclusions

- Automatic modeling building framework
- Portable framework → ETW performance counters
- High fidelity models → multiple machines required to build the machine model
- Account for variability in building large-scale power models:
  - Feature selection & measured machine power
  - # of machine required to build the machine model
- Sampling theory → technique scales

- Framework to build high fidelity cluster power models
  - < 12% DRE (or <2.5% median error)

Questions?

Back up

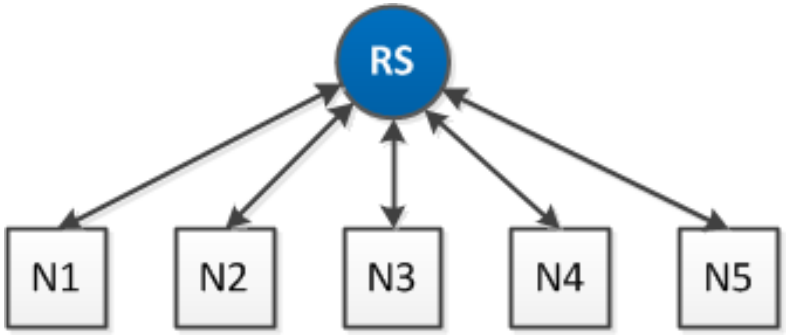
Summary

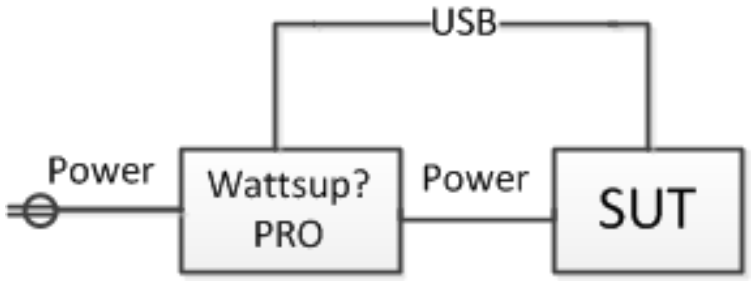
- Composable Highly Accurate OS-based (CHAOS) power models
  - Full-system cluster power models
  - MapReduce-style applications
  - 6 different clusters, each with 5 machines less than 10% DRE (new error metric)
  - Generalized model feature extraction heuristic
  - Automatic power model building framework

Hardware Infrastructure

- 6 clusters or platforms, each 5 machines or nodes

Cluster	Atom (embedded)	Intel Core2 Duo (laptop)	Athlon (desktop)	Opteron (server)	Xeon (server)	Xeon (server)
CPU	Intel Atom X2 1.6 GHz	Intel Core 2Duo X2 2.26 GHz	AMD Athlon X2 2.8 GHz	AMD Opteron 2X4 2.0 GHz	Intel Xeon 2X4 2.33 GHz	Intel Xeon 2X4 2.66 GHz
Storage	SSD	SSD	SSD	HDD	HDD	HDD:SAS
Idle Power (W)	22	25	54	135	260	275
Dyn Power range (W)	4	20	50	55	100	110
OS	Windows Server 2008 R2					





Machine Power Variability

Machine Power Variability

- Is one machine enough to build a cluster model?
- Machine-to-machine power range:

Clusters	3% Error in Power (+/- 1.5%)			Average benchmark power range				
	Average	Minimum	Maximum	@ Idle	primes	pagerank	sort	wordcount
Opteron	4.5	4	5.7	3.0	3.1	0.2	0.8	2.6
Athlon	2.3	1	3.3	2.9	7.7	6.5	3.8	2.2
Intel Core2 Duo	1	1	1	3.1	3.8	0.8	0.9	0.5
Atom	1	1	1	2.0	0.1	0.2	0.2	0.2

Machine Power Variability

- Is one machine enough to build a cluster model?
- Machine-to-machine power range:

Clusters	3% Error in Power (+/- 1.5%)			Average benchmark power range				
	Average	Minimum	Maximum	@ Idle	primes	pagerank	sort	wordcount
Opteron	4.5	4	5.7	3.0	3.1	0.2	0.8	2.6
Athlon	2.3	1	3.3	2.9	7.7	6.5	3.8	2.2
Intel Core2 Duo	1	1	1	3.1	3.8	0.8	0.9	0.5
Atom	1	1	1	2.0	0.1	0.2	0.2	0.2

Machine Power Variability

- Is one machine enough to build a cluster model?
- Machine-to-machine power range:

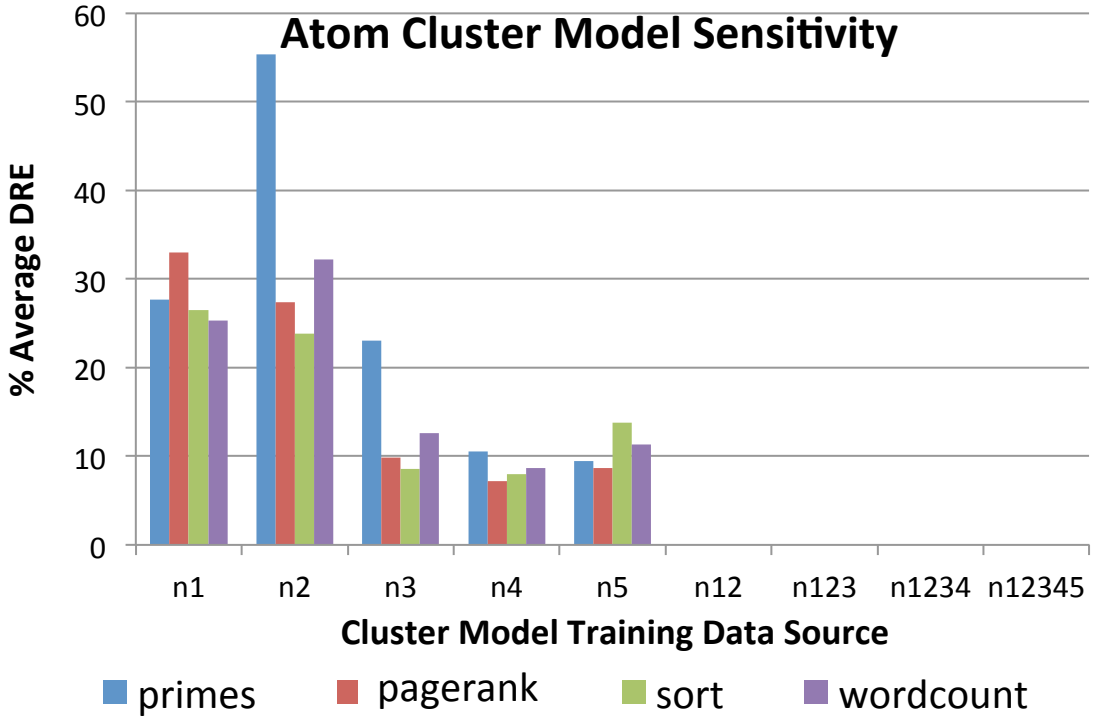
Clusters	3% Error in Power (+/- 1.5%)			Average benchmark power range				
	Average	Minimum	Maximum	@ Idle	primes	pagerank	sort	wordcount
Opteron	4.5	4	5.7	3.0	3.1	0.2	0.8	2.6
Athlon	2.3	1	3.3	2.9	7.7	6.5	3.8	2.2
Intel Core2 Duo	1	1	1	3.1	3.8	0.8	0.9	0.5
Atom	1	1	1	2.0	0.1	0.2	0.2	0.2

- Open questions:
  - How many machines to sample?

Machine Variability

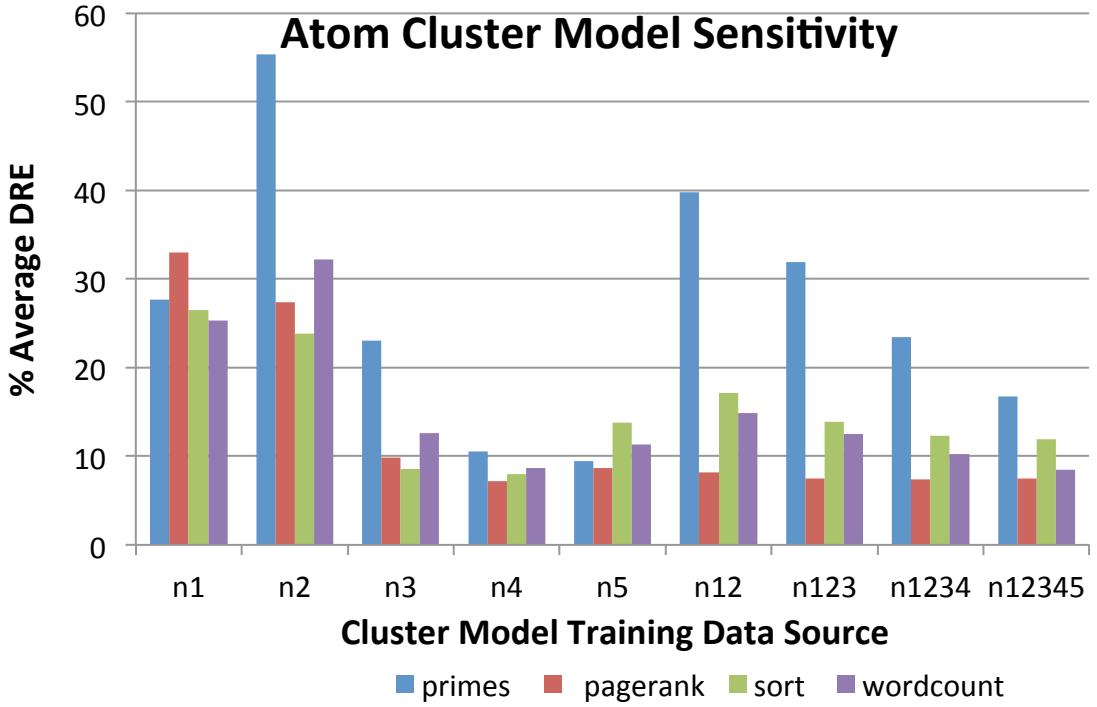
- (A) Worst-case DRE ~150%
- (B) →

  Error is node dependent



Machine Variability

- (A) Worst-case DRE ~150%
- (B) Error is node dependent
- (C) → Removes "luck"



How do you achieve the ease of cluster model (B) and the accuracy of (C) for large-scale systems?

Scaling the Cluster Model

- Chernoff-Hoeffding bound: $\Pr[|S-\mathcal{S}|\geq N\delta]\leq 2e^{-2\delta^2/I^2}$ q

Clusters	Idle workload power range (W)				# of sampled machines (@85%)			
	primes	pagerank	sort	wordcount	Machines	Machines	Machines	Machines
Opteron	3.0	2.6	2.8	3.0	16	13	15	16
Athlon	1.0	0.9	1.2	1.1	5	5	6	6
Intel Core2 Duo	2.5	3.1	2.5	3.1	13	16	13	16
Atom	1.2	2.0	1.3	1.8	6	10	7	9

- Small number of machines required to build a model & population independent
- Use technique (B), 1 machine model with machine specific inputs