# Accounting for Variability in Large-Scale Cluster Power Models

**John D. Davis,**
**Microsoft Research, Silicon Valley Lab**,
Suzanne Rivoire, Moises Goldszmidt, and Ehsan K. Ardestani

March 6, 2011

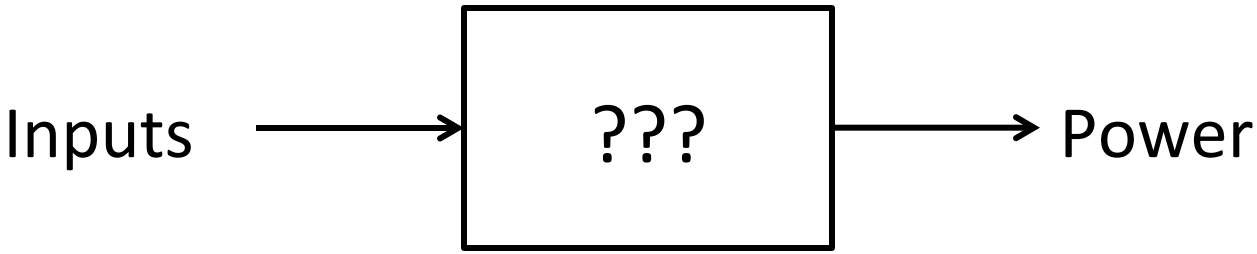
# Why do we need power models?

- Data Center
  - What workloads can I run together?
  - What should the capacity of my new facility be?
  - How much power will new systems consume?
  - System utilization vs. power?
  - What racks can I move where?
  - How much do you charge internal or external customers? (right-sizing power)
  - Measurement infrastructure too expensive, non-existent, error prone, …

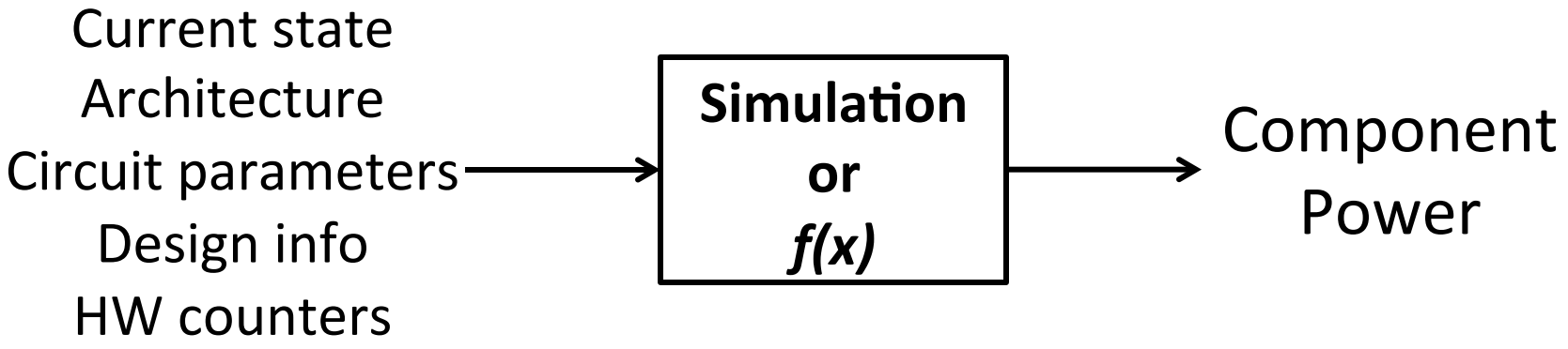
# Power modeling goals

- Goal: Online, full-system power models
- Model requirements
  - Non-intrusive and low-overhead
  - Easy to develop and use
  - Fast enough for online use
  - Reasonably accurate (within 10%)
  - Inexpensive
  - Scalable
  - Generic and portable

# What is a Power Model?

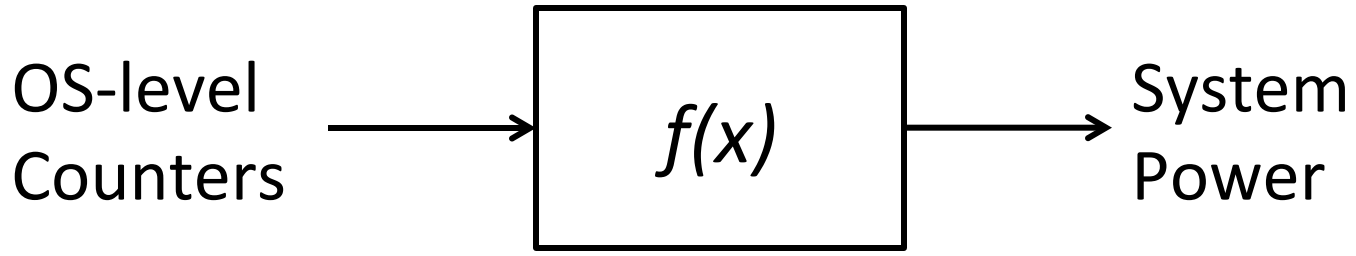


# What is a Power Model?




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

# Building a Full-System Power Model




- How portable?
  - Framework, model features, etc. ?
- How accurate?
  - Machine-to-machine power variability?
- How scalable?
  - Sampling theory and cluster model?
- Future work:
  - Tradeoff between model parameters/complexity and accuracy?

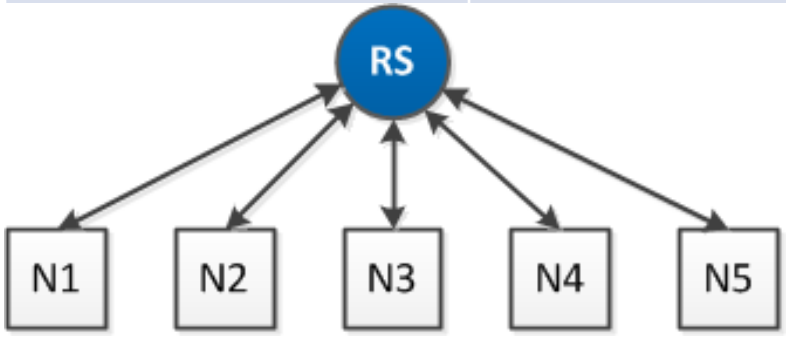
# Talk Overview

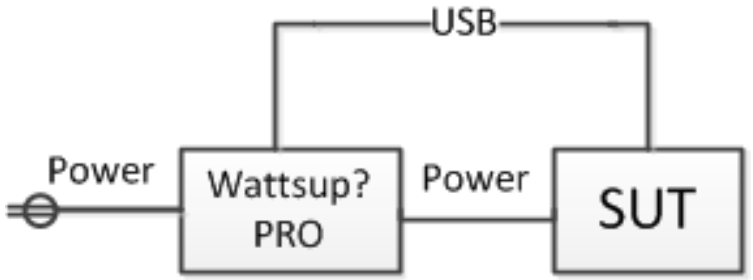
- Motivation and Background
- Hardware and Software Infrastructure
- Model Feature Selection and Variability
- Machine Power Variability
- Cluster Models
- Model Scalability
- Discussion & Future Work

# Hardware Infrastructure

- 4 clusters or platforms, each 5 machines or nodes

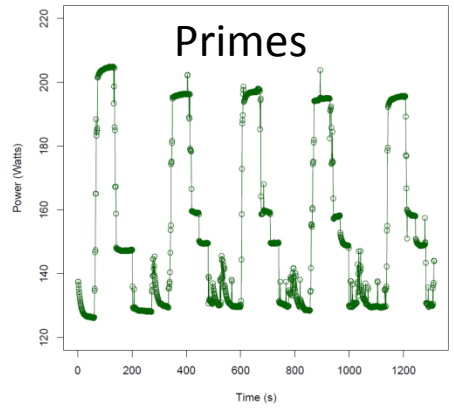
| Cluster | Atom [24] (embedded) | Intel Core2 Duo (laptop) [13] | Athlon (desktop) [9] | Opteron (server) |
|---|---|---|---|---|
| CPU | Intel Atom X2 1.6 GHz | Intel Core 2Duo X2 2.26 GHz | AMD Athlon X2 2.8 GHz | AMD Opteron 2X4 2.0 GHz |
| Storage | SSD | SSD | SSD | HDD |
| Idle Power (W) | 22 | 25 | 54 | 135 |
| Dyn Power range (W) | 4 | 20 | 50 | 55 |
| OS | Windows Server 2008 R2 | | | |

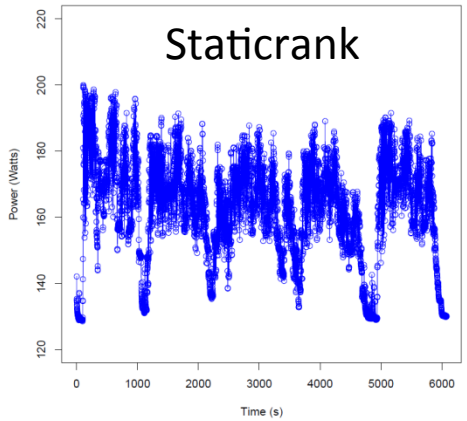


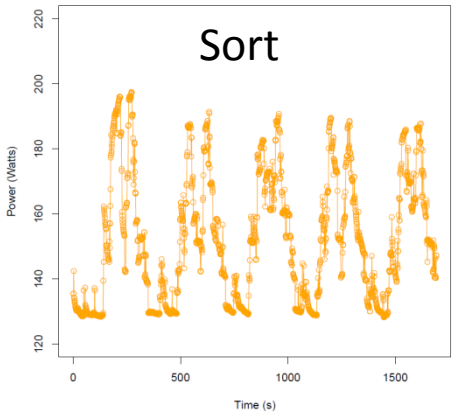



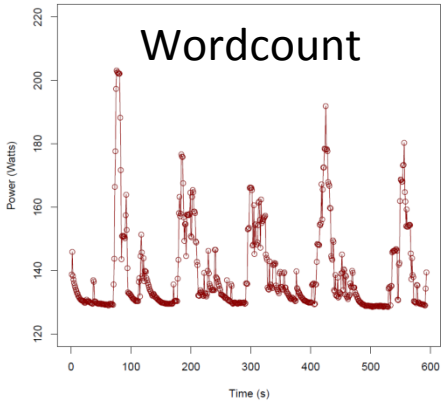
# Software Infrastructure

- Workloads
  - Primes (CPU)
  - Staticrank (Net)
  - Sort (Disk + Net)
  - Wordcount (Disk)




- Tools
  - Dryad, DryadLinq, Artemis, ETW, Joulemeter, R

# Feature Selection

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

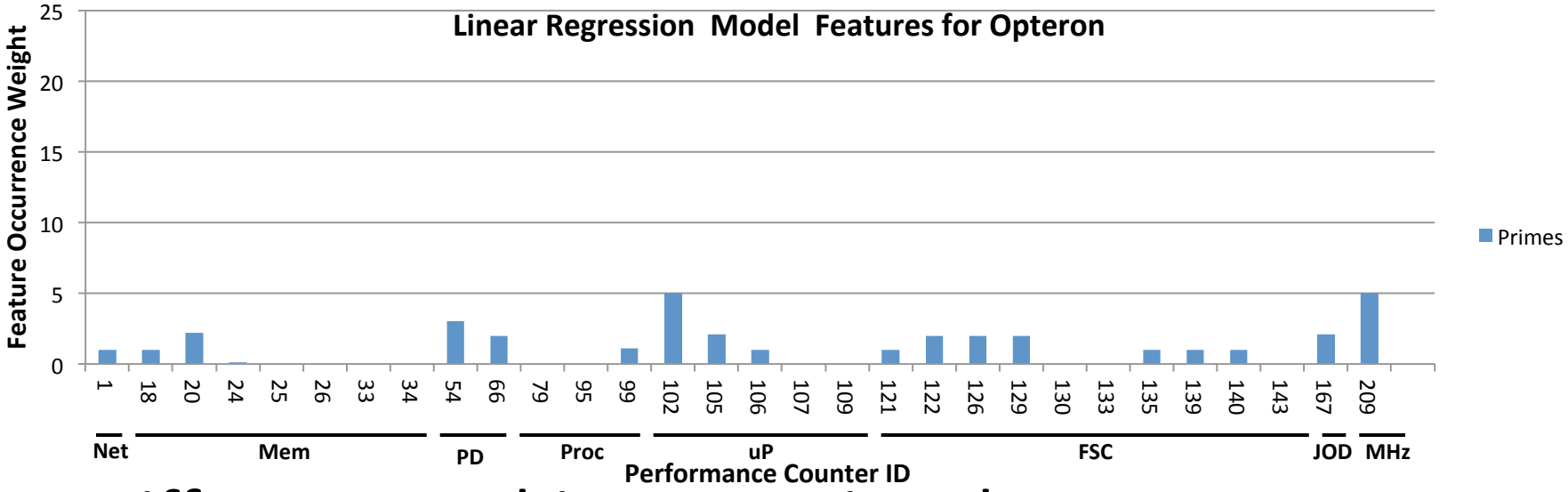
# Selected Features

- 5 machines and 4 workloads

| Category | Performance counter | Ctr. ID |
|---|---|---|
| Memory (Mem) | Page Faults/sec | 18 |
| | Cache Faults/sec | 24 |
| | Pages/sec | 26 |
| | Pool Nonpaged Allocs | 34 |
| Physical Disk (PD) | Disk Total Disk Time % | 54 |
| | Disk Total Disk Bytes/sec | 66 |
| Process (Proc) | Total IO Data Bytes/sec | 99 |
| Processor (uP) | Total Processor Time % | 102 |
| File System Cache (FSC) | Data Map Pins/sec | 121 |
| | Pin Reads/sec | 122 |
| | 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 |

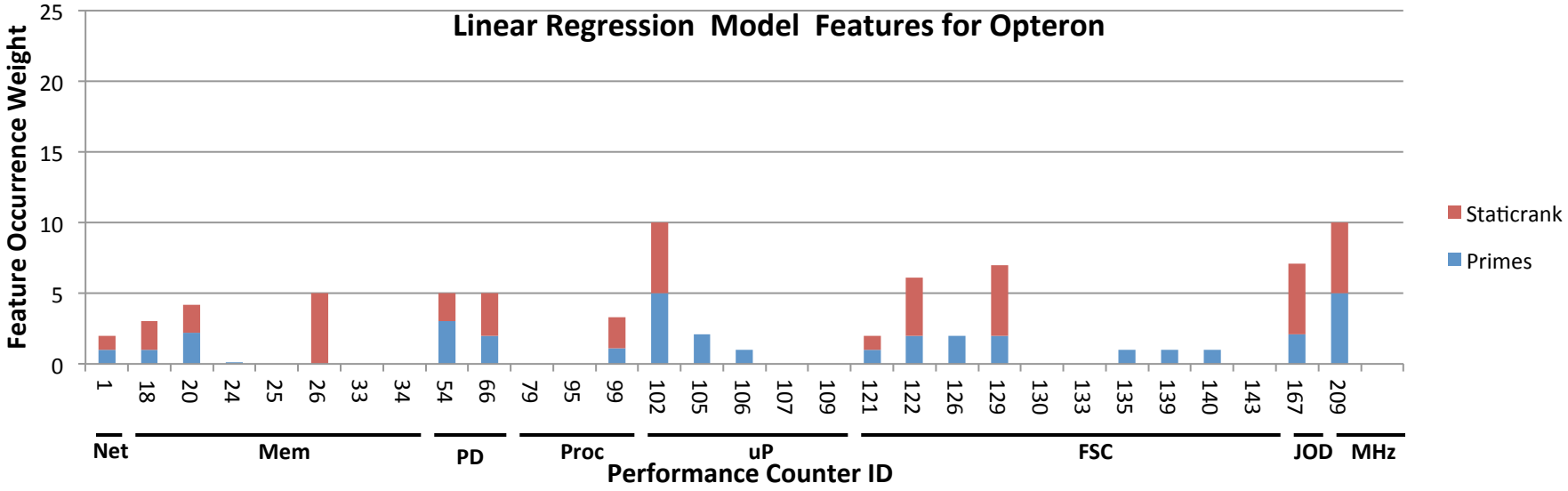
# Model Feature Variability

# Opteron Features

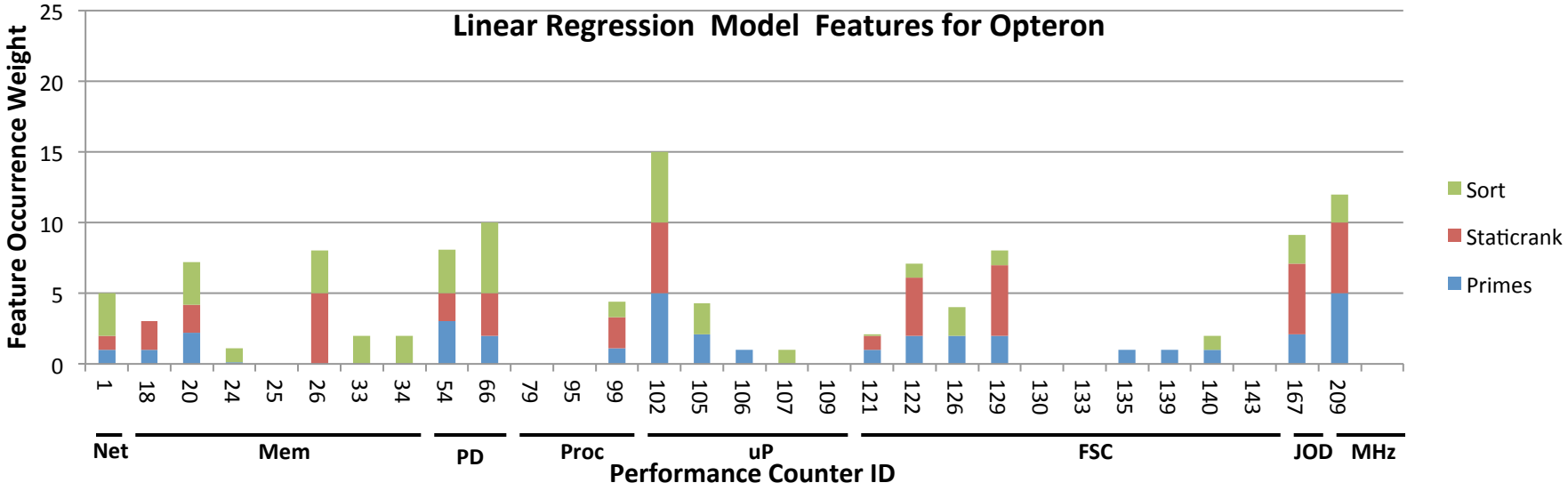



- Different machines running the same workload identify different significant features (height of the bars)

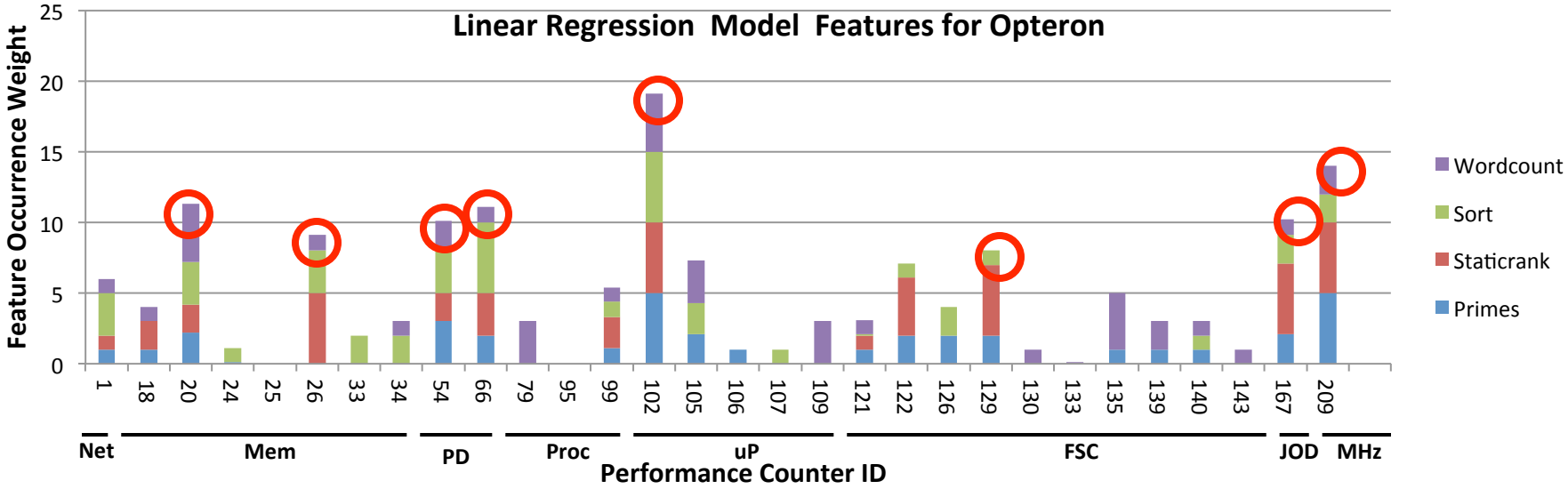
# Opteron Features



# Opteron Features

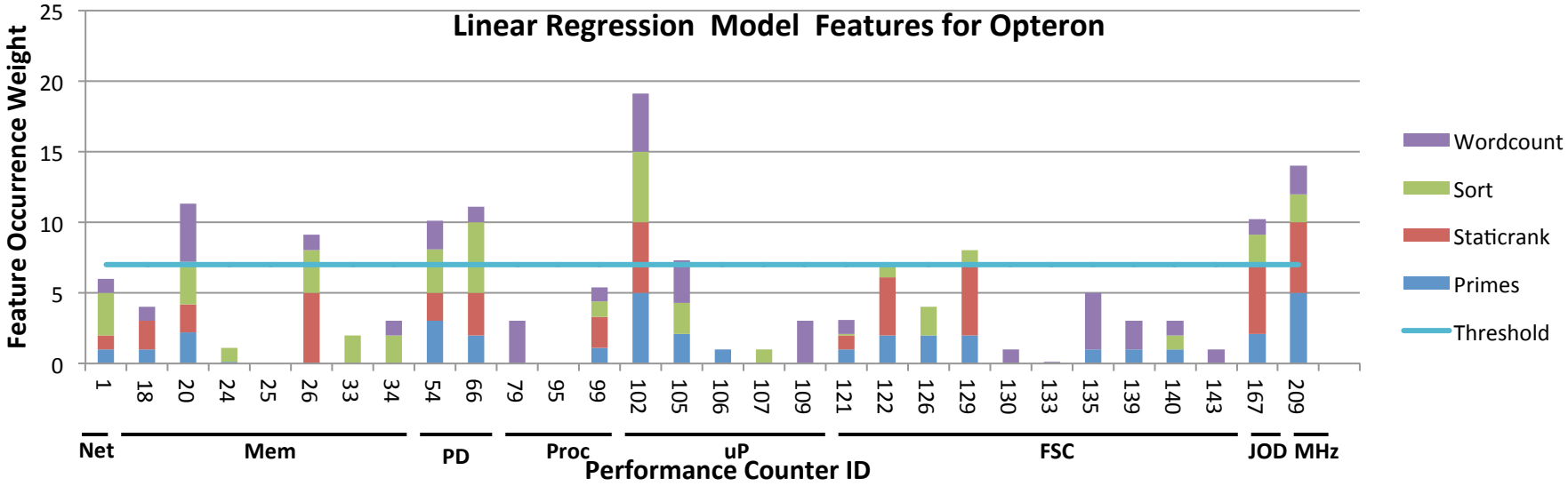


# Opteron Features




Model

# Opteron Features




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

# 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 | staticrank | 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 | staticrank | 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 | staticrank | 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 |

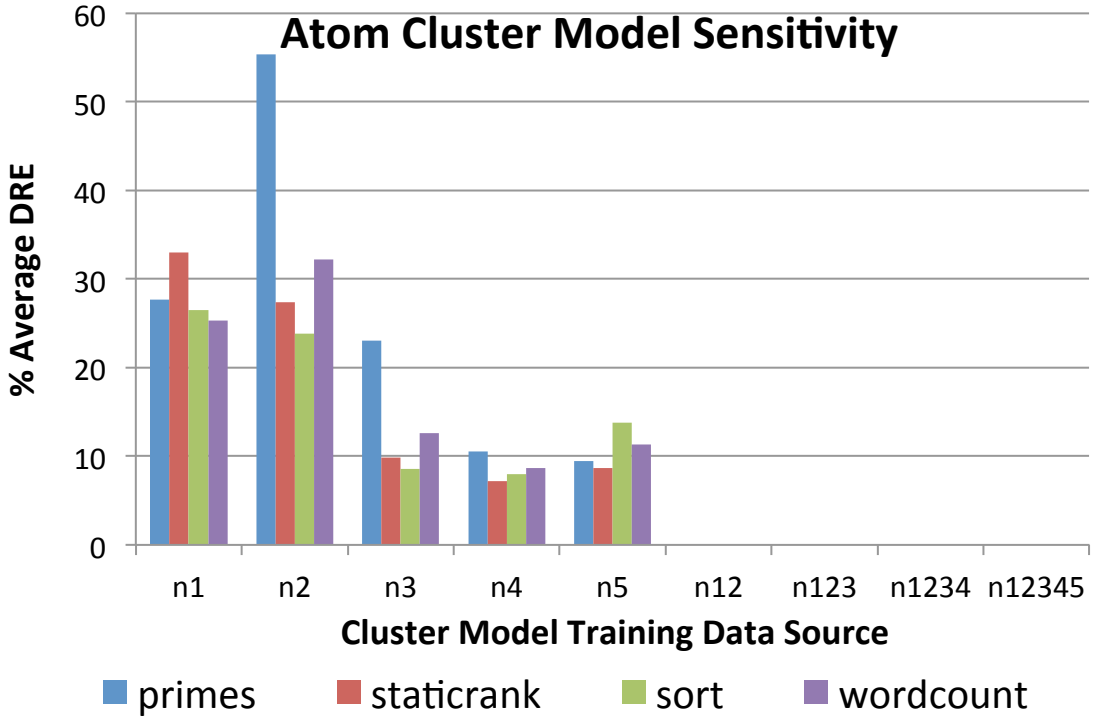
- Open questions:
  - How many machines to sample?
  - Sources of variability? (Future Work)

# Cluster Power Modeling Methods

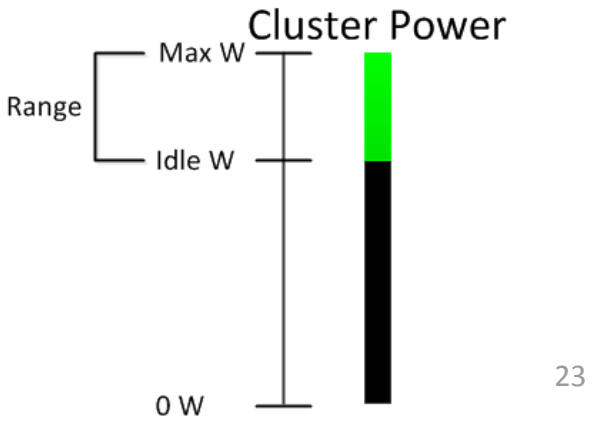
- (A) Cluster Power = $f_{machine}(x_{machine})$ *X N*
  - *Single model, inputs from a single machine*
  - Implicitly or explicitly assumed by previous work
  - 1 machine model
- (B) Cluster Power = $\Sigma_i f_{machine}(x_{machine_i})$
  - *Single model, inputs from each machine*
  - *How many machines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

# Machine Variability

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

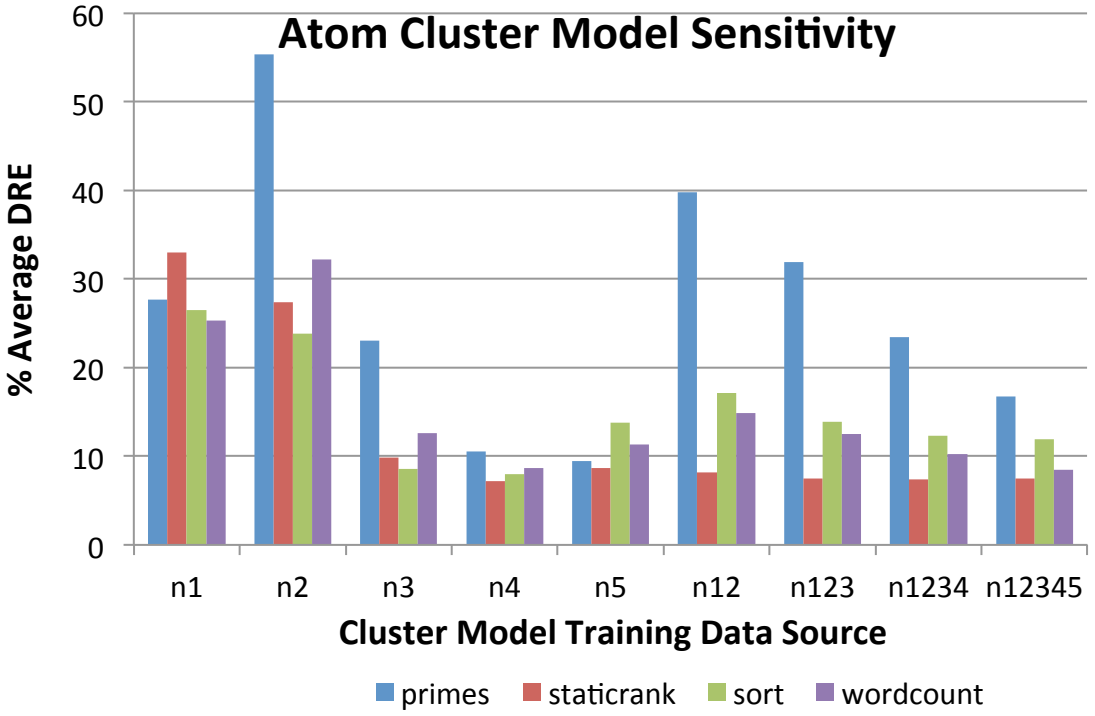



$$Error\,(DRE) = \sqrt{Mean\ Square\ error} \,/ Max\ Power{\downarrow}Cluster - Min\ Power{\downarrow}Cluster$$



# 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: $\Pr[|S-S'|\geq N\delta]\leq 2e^{-2\delta^2/I^2}$
  - Guards against worst-case (more machines)

| Clusters | Idle workload power range | | | | # of sampled machines (@85%) | | | |
|---|---|---|---|---|---|---|---|---|
| | primes | staticrank | 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

# Discussion

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

- Framework to build high fidelity cluster power models

# Future work

- More machines …
- More models (beyond linear)…
- More workloads …
- Sources of power variation?
- Tradeoff between number of model parameters, model complexity and accuracy

# Questions?

# Back-up

# Cluster Machine Models

- Server (Opteron):
  - *Features: 2 memory , 2 disk , 2 file system, 1 CPU utilization, and 1 CPU frequency*
- Desktop (Athlon):
  - Features*: 3 memory , 1 disk , 1 CPU utilization, and 1 CPU frequency*
- Mobile (Intel Core2Duo):
  - Features*: 3 memory , 1 disk , 1 file system, 1 CPU utilization, and 1 CPU frequency*
- Embedded (Atom):
  - Features*: 2 memory , 1 disk , 5 file system, and 1 CPU utilization*

# Cluster Machine Models

- Server (Opteron):
  - $f = \beta\downarrow 0 + \beta\downarrow 26\ a\downarrow 26 + \beta\downarrow 54\ a\downarrow 54 + \beta\downarrow 66\ a\downarrow 66 + \beta\downarrow 102\ a\downarrow 102 + \beta\downarrow 121\ a\downarrow 121 + \beta\downarrow 122\ a\downarrow 122 + \beta\downarrow 167\ a\downarrow 167 + \beta\downarrow 209\ a\downarrow 209$
- Desktop (Athlon):
  - $f = \beta\downarrow 0 + \beta\downarrow 18\ a\downarrow 18 + \beta\downarrow 26\ a\downarrow 26 + \beta\downarrow 99\ a\downarrow 99 + \beta\downarrow 102\ a\downarrow 102 + \beta\downarrow 167\ a\downarrow 167 + \beta\downarrow 209\ a\downarrow 209$
- Mobile (Intel Core2Duo):
  - $f = \beta\downarrow 0 + \beta\downarrow 24\ a\downarrow 24 + \beta\downarrow 34\ a\downarrow 34 + \beta\downarrow 54\ a\downarrow 54 + \beta\downarrow 102\ a\downarrow 102 + \beta\downarrow 122\ a\downarrow 122 + \beta\downarrow 167\ a\downarrow 167\ + \beta\downarrow 209\ a\downarrow 209$
- Embedded (Atom):
  - $f = \beta\downarrow 0 + \beta\downarrow 24\ a\downarrow 24 + \beta\downarrow 34\ a\downarrow 34 + \beta\downarrow 66\ a\downarrow 66 + \beta\downarrow 102\ a\downarrow 102 + \beta\downarrow 121\ a\downarrow 121 + \beta\downarrow 126\ a\downarrow 126 + \beta\downarrow 139\ a\downarrow 139 + \beta\downarrow 140\ a\downarrow 140 + \beta\downarrow 167\ a\downarrow 167$