

PENNSTATE



Profiling, Prediction, and Capping of Power in Consolidated Environments

Bhuvan Uргаonkar
Computer Systems Laboratory
The Penn State University



CGRS, March 5, 2008

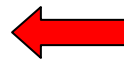
Data Center Growth

- ❑ Explosive growth in both size and numbers
- ❑ Serious implications on
 - Robustness of operation
 - Heterogeneity of hardware/software, workloads, ...
 - Potential lack of scalability of existing resource management solutions
 - Flash crowds
 - Cost of operation
 - Administrative costs
 - Power consumption



Data Center Growth

- ❑ Explosive growth in both size and numbers
- ❑ Serious implications on
 - Robustness of operation
 - Heterogeneity of hardware/software, workloads, ...
 - Potential lack of scalability of existing resource management solutions
 - Flash crowds
 - Cost of operation
 - Administrative costs
 - **Power consumption**





Growing Power Consumption in Data Centers

- ❑ Significant energy consumption and growing!
 - Up to 1.2% of overall power consumption within the US
 - Growing @ 8-18% every year

- ❑ Growing number of servers main culprit
 - Increase-per-unit less significant contributor

- ❑ Lot of interest in dampening this growth

- ❑ **Key technique: Consolidation**

Consolidation in Data Centers

- ❑ **Goal:** Operating/provisioning fewest possible hardware resources while meeting service-level agreements
 - Our focus: Servers

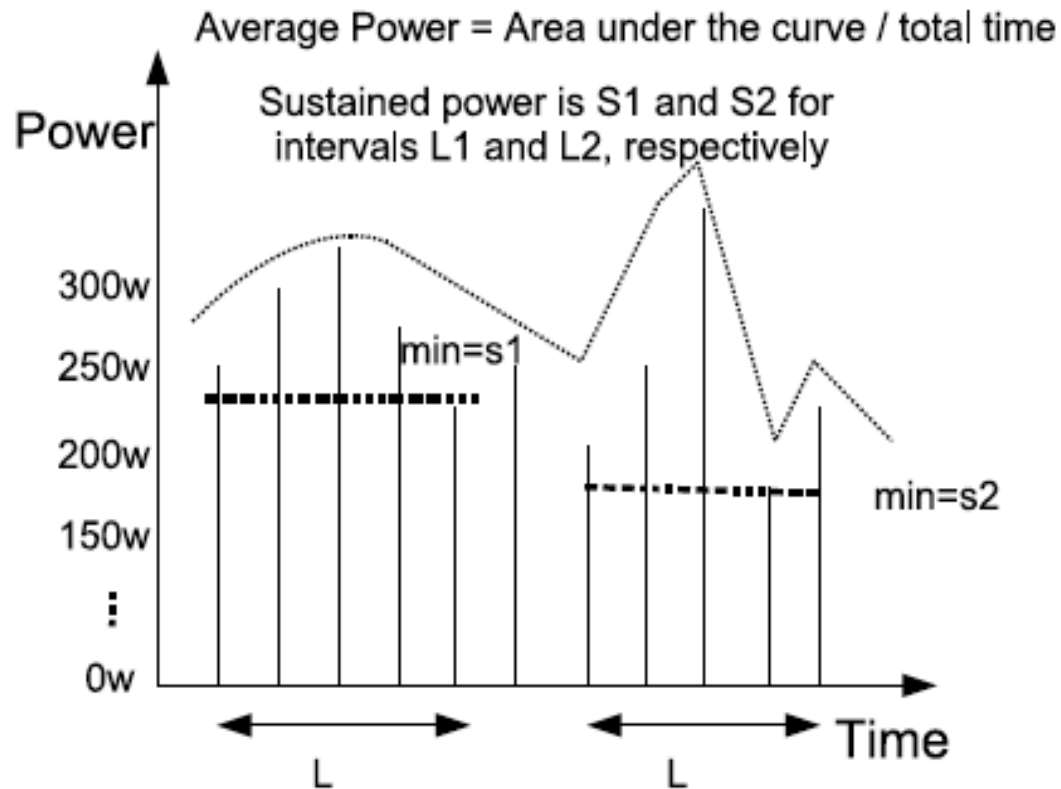
- ❑ Multiple spatial scales
 - Packing multiple applications on a server
 - Reducing the number of data centers operated by a company

- ❑ Key ingredients of existing consolidation solutions
 - Workload characterization and prediction
 - Resource requirement inference
 - Dynamic resource provisioning
 - Efficient statistical multiplexing

Power Consumption in Consolidated Data Centers

- ❑ How does consolidation affect power consumption?
 - How does the power consumed by consolidated aggregates relate to those of individuals?
 - Spatial
 - Servers, racks, room, ...
 - Temporal
 - Long-term averages: energy consumed, thermal profiles
 - Short-term surges: fuses, circuit breakers
 - Can we effectively predict these phenomena?
 - How should we characterize power consumption of individuals?
 - Such that we can meaningfully infer behavior upon consolidation
- ❑ Benefits/utility of such characterization and prediction
 - Enable consolidation that adheres to “power budgets”
 - Adapt placement to changes in workloads to obtain desired performance/power behavior
 - Determine optimal “power states” (if any) exposed by hardware
 - E.g., CPU DVFS states

Average and Sustained Power Consumption



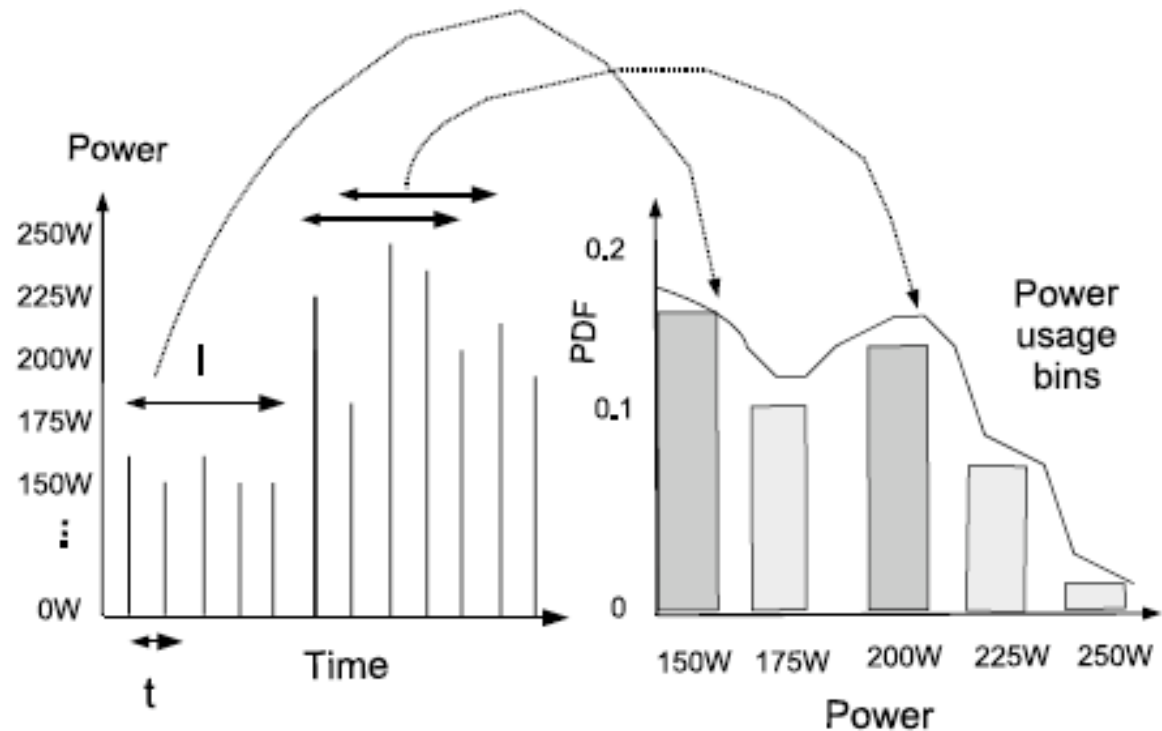
- Two quantities of interest
 - Average power consumption
 - Sustained power consumption
- Average and sustained power "budgets" or "caps" of interest at various spatial levels
- Our focus: single server consolidating multiple applications

- ✓ Motivation
- **Power Profiles**
- Power Prediction
- Preliminary Evaluation
 - Power Capping
- Concluding Remarks

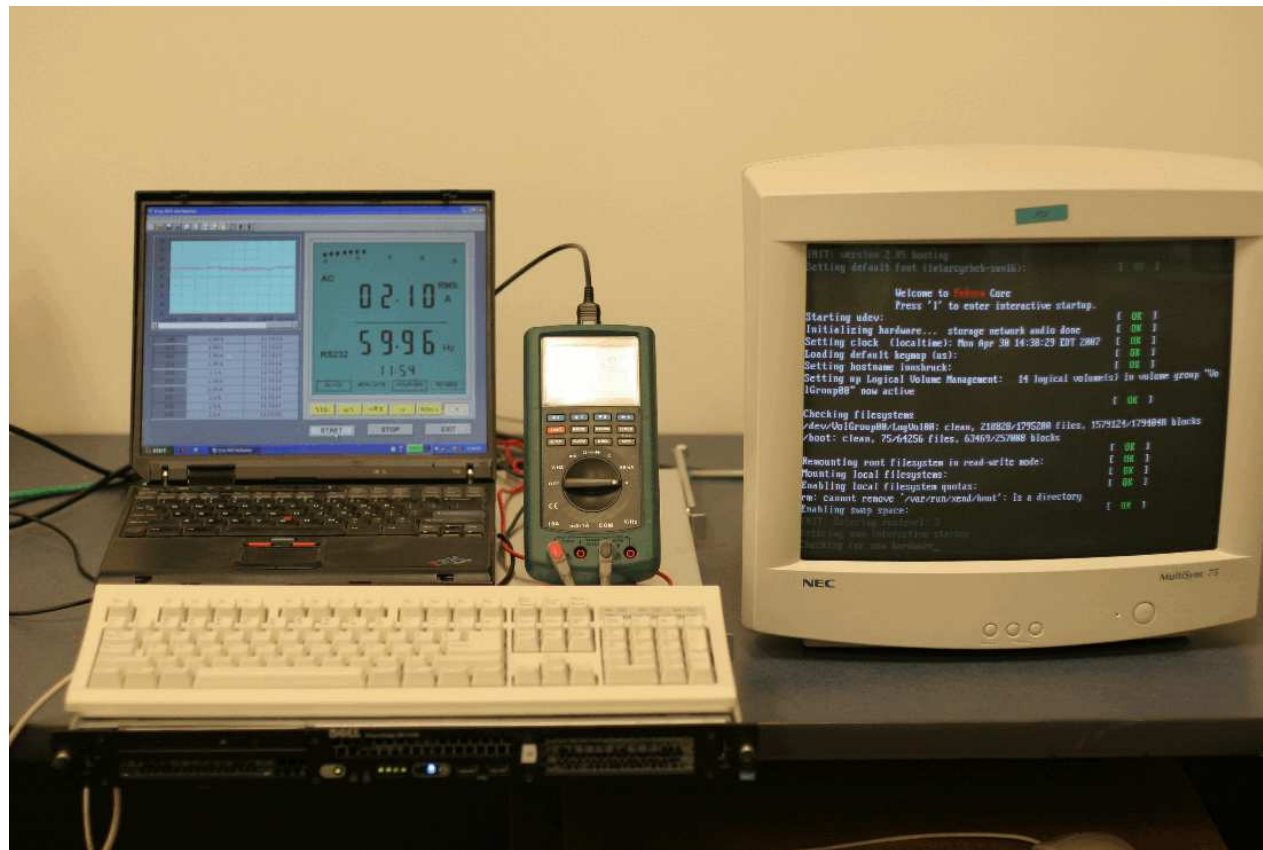
Characterizing Power Consumption

- Desirable features
 - Easy/efficient to realize
 - Amenable to meaningful statistical aggregation

- **Offline profiling**
 - Run application in isolation and subject it to realistic workload
 - PDF of power consumed during intervals of chosen length
 - **Power profile**



Offline Profiling Setup



- ❑ Signametrics SM2040
 - Measurement rates: 0.2/sec - 1000/sec
 - Measurement range (AC): 2.5A
 - Interface: PCI

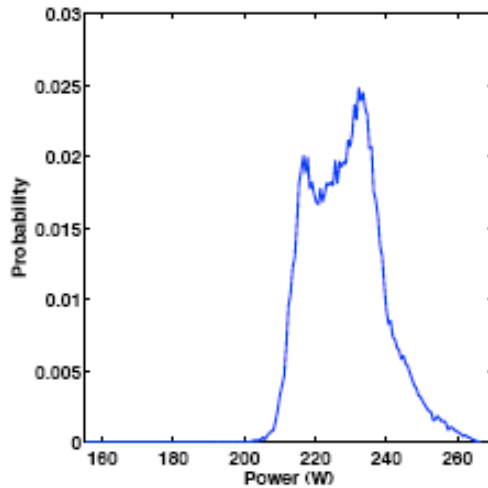


Characterizing Power Consumption

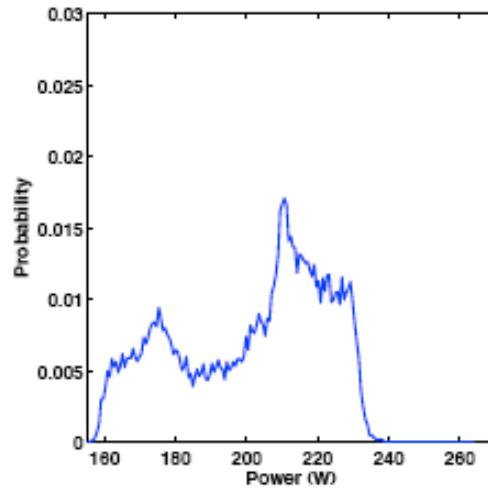
□ Other noteworthy points

- Xen-based virtualized hosting
 - Each application hosted within a Xen domain
- Also measure resource usage and provision enough resources when consolidating
 - Appropriate resource managers within the Xen VMM
- Dell PowerEdge server (DVFS-capable CPU)

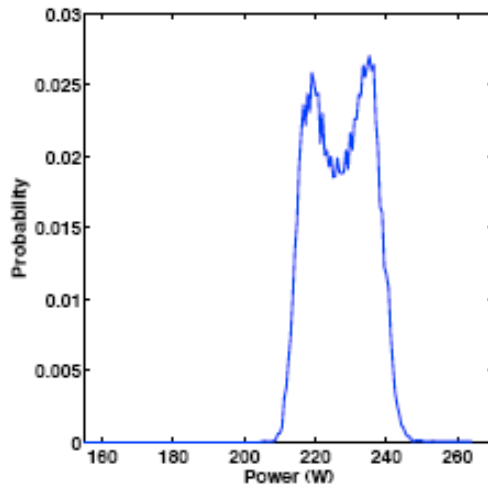
Power Profiles of Real Applications



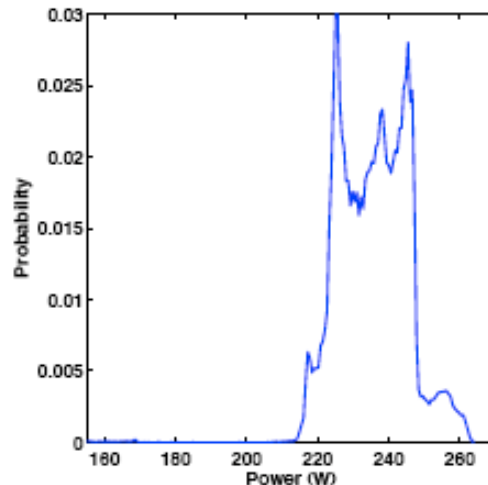
(a) SPECjbb



(b) Streaming, 60 clients



(c) Bzip2



(d) Mcf

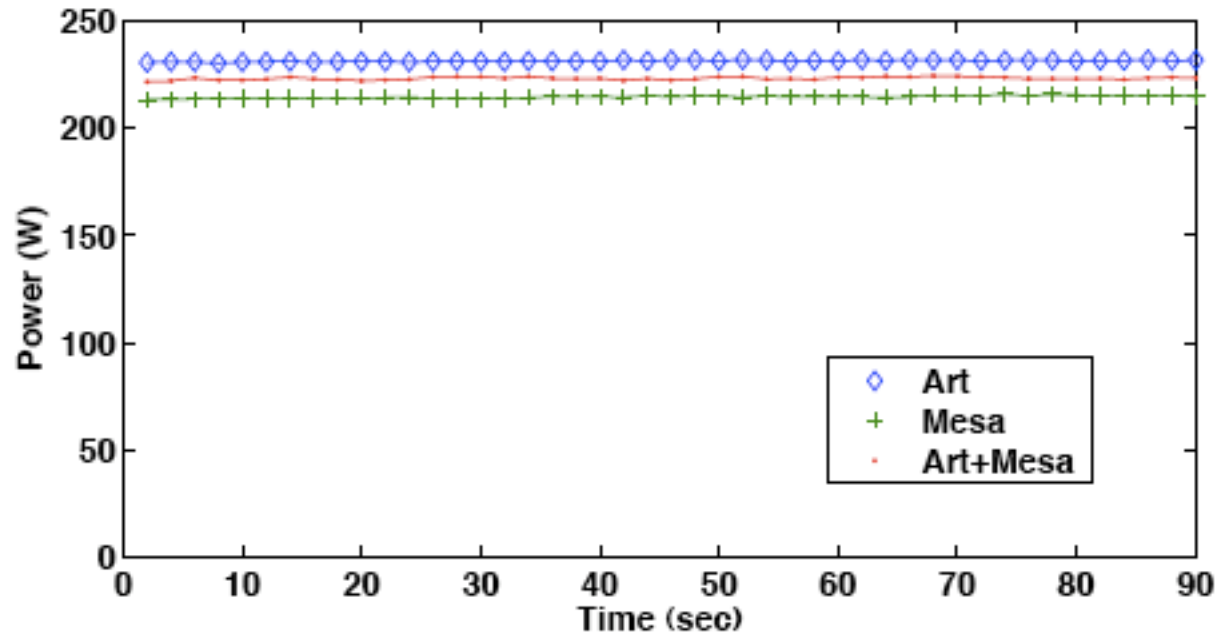
Applications

- SPECjbb
- Streaming
- SPECInt
 - Bzip, MCF
- TPC-W

□ $t = I = 2$ msec

- ✓ Motivation
- ✓ Power Profiles
- **Power Prediction**
- Preliminary Evaluation
 - Power Capping
- Concluding Remarks

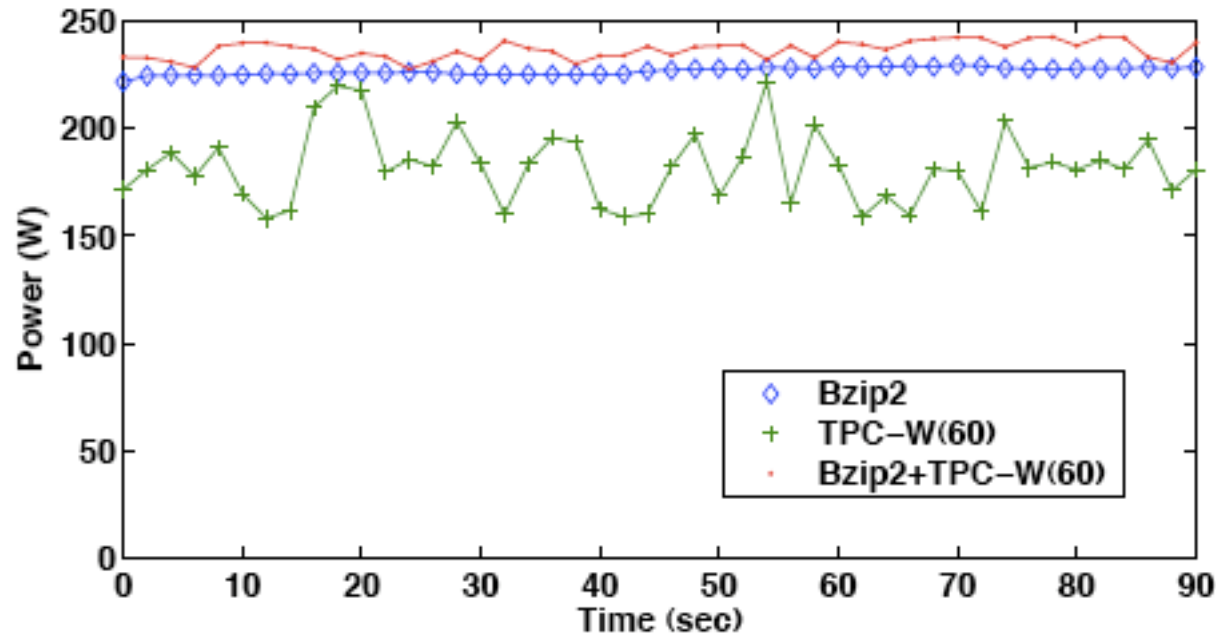
Average Power Upon Consolidation



(a) Two CPU-saturating applications: Art and Mesa

- Consolidation of CPU saturating applications
 - Average of individual power consumptions

Average Power Upon Consolidation



(b) A CPU-saturating (Bzip2) and a non CPU-saturating (TPC-W) application

- Consolidation of CPU-saturating and non CPU-saturating
 - More complex: Some kind of additive effect

Average Power: What's Going On?

- ❑ CPU+CPU:
 - Sole significant consumer of power is being time-shared
- ❑ CPU+non-CPU:
 - CPU being time-shared
 - Though not equally, since non-CPU apps block
 - CPU and I/O devices being used simultaneously

- ❑ **Insight #1:** Separate out power due to resources (such as CPU and I/O)
- ❑ **Insight #2:** Also consider the utilization of relevant resources

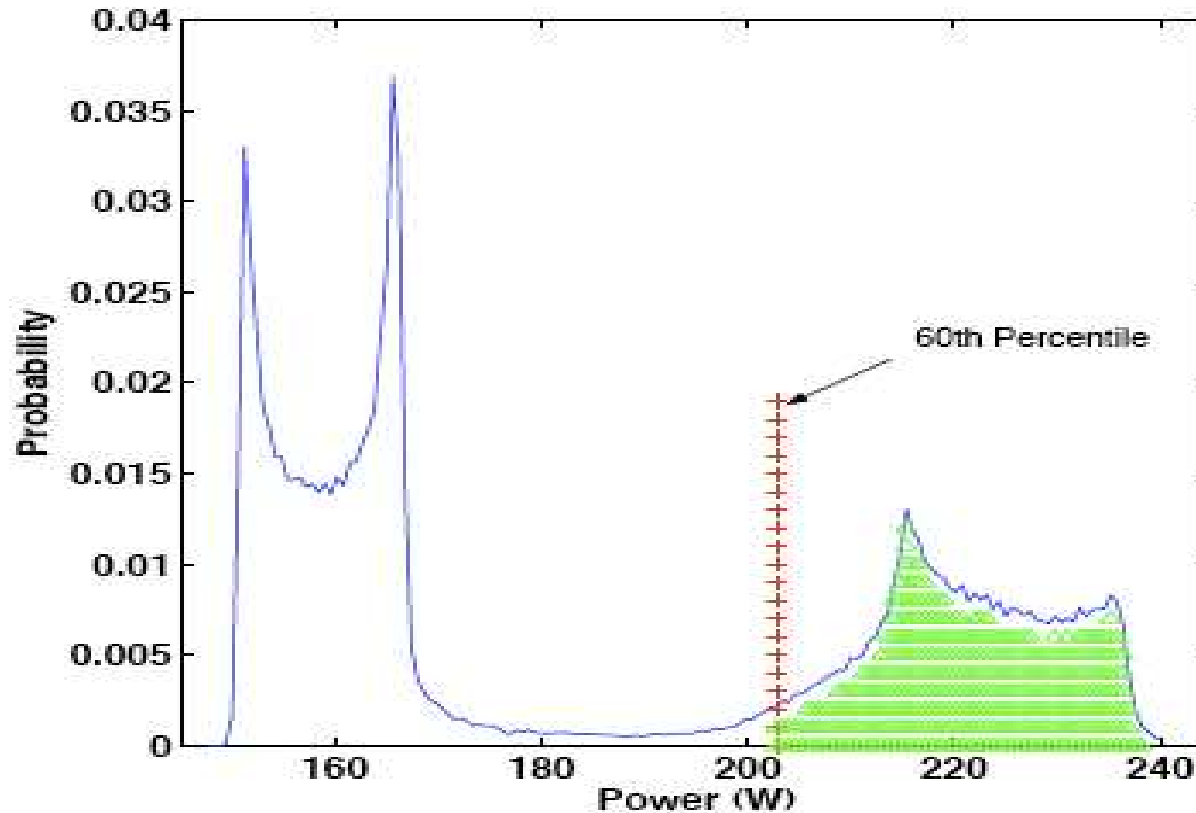
- ❑ Bottom-up approach: Start simple and add complexity incrementally

A Simple Predictor for Average Power

$$\bar{P}_{A_1, \dots, A_n} = \begin{cases} \sum_{i=1}^n (\bar{P}_{A_i}^{busy/cpu} \cdot R_{A_i}^{cpu}) & // \text{ power when CPU busy} \\ + \\ \bar{P}^{idle} \cdot (1 - \sum_{i=1}^n R_{A_i}^{cpu}) & // \text{ power when CPU idle} \\ + \\ \sum_{i=1}^n (\bar{P}_{A_i}^{i/o} \cdot R_{A_i}^{i/o}) & // \text{ I/O power} \end{cases}$$

- \bar{P}^{idle} : Power when no application/VM running
 - Note difference from leakage power
- $\bar{P}^{busy/cpu}$ and $\bar{P}^{i/o}$ for an application
 - CPU Power when application running
 - I/O power due to the application

Improved Estimate of Active Power



- Capturing non-idle power portion for TPC-W, 60 clients
 - CPU utilization was 40%

Average Power Prediction: Some Results

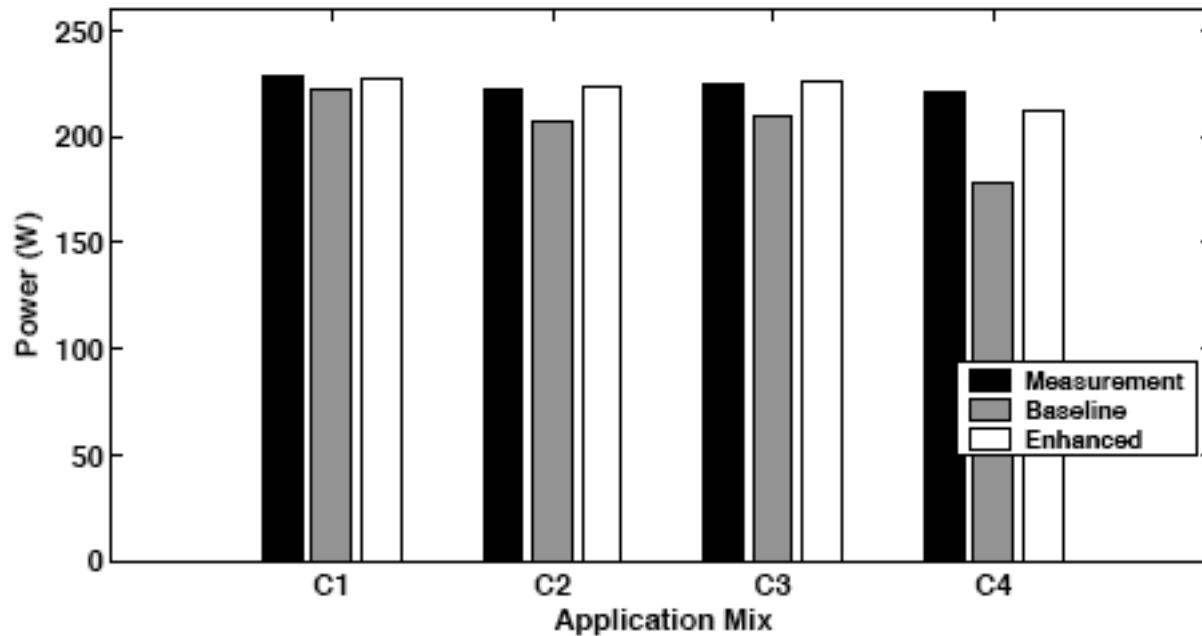


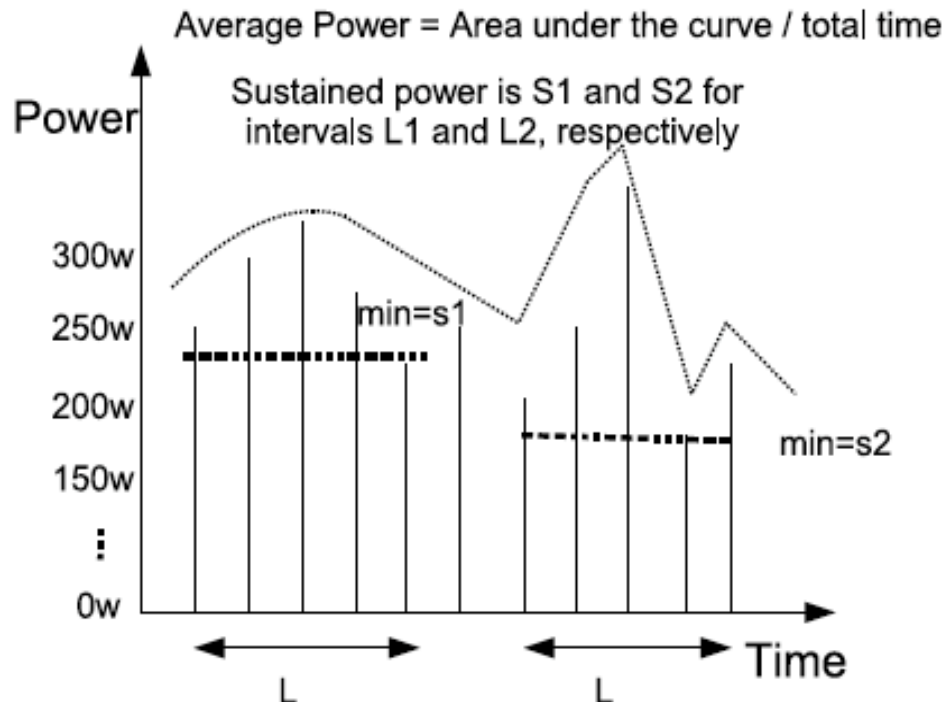
Figure 8: Comparison of our predictors for a variety of consolidation scenarios. C1:TPC-W(10)+Art, C2:TPC-W(30)+Mesa, C3:TPC-W(60)+Bzip2 C4:TPC-W(60)+TPC-W(60)

□ Prediction accuracy of 2% !

- Disclaimer: Pretty small degrees of consolidation

Sustained Power Prediction

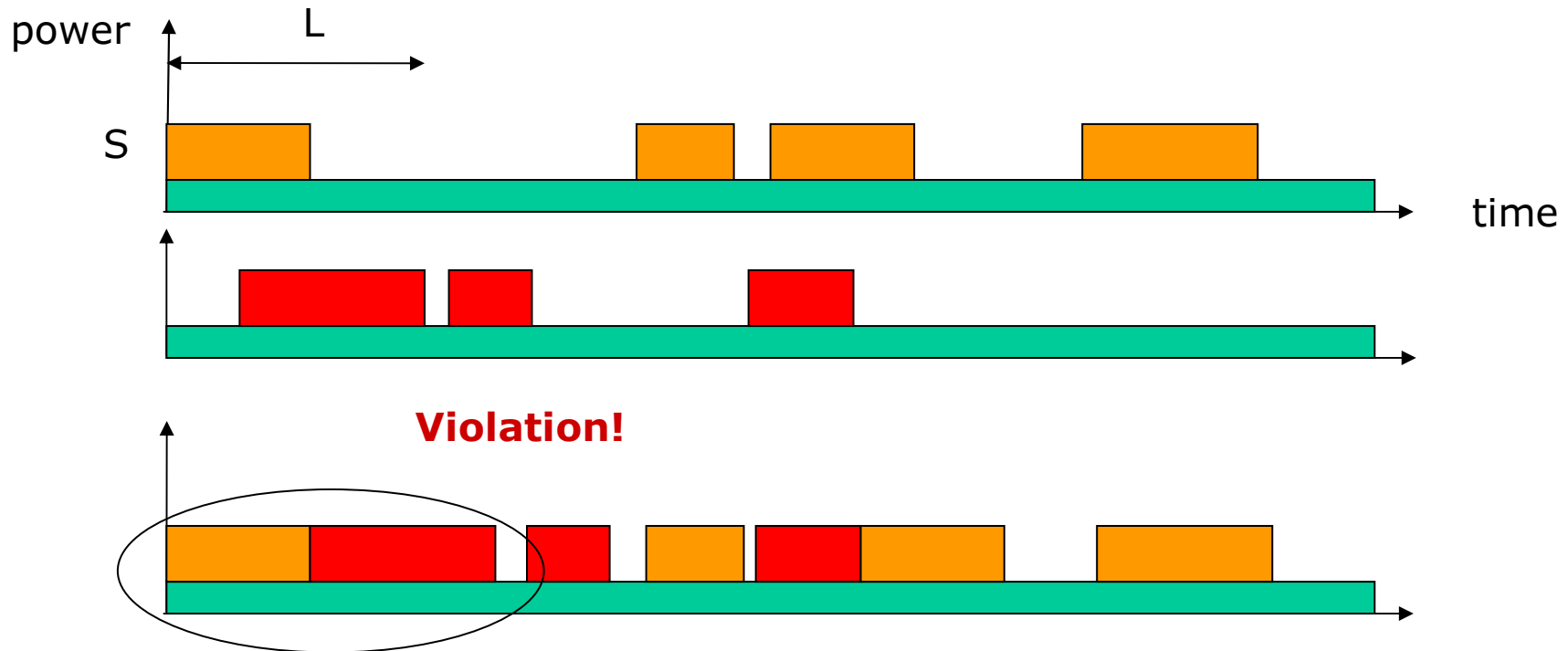
- ❑ **Goal:** Predict the probability that at least S units of power would be consumed for L **consecutive** seconds
- ❑ What's difficult?
 - Applications that individually do not violate a sustained budget can do so upon consolidation

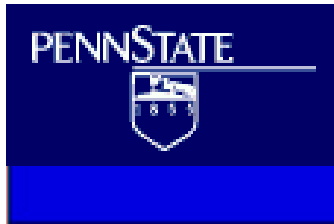


Sustained Power Prediction

□ What's difficult?

- Applications that individually do not violate a sustained budget can do so upon consolidation





Sustained Power Prediction: Some Results

- Harder to predict than average
 - More sophisticated statistical techniques
 - Omitting details, please find them in our technical report
 - Overview:
 - Count various ways in which multiplexing can occur
 - Estimate tail of aggregate using techniques such as Chernoff bounds

- **Encouraging news:** Our profile-based techniques appear to do a good job

Example: Power-aware Application Packing

Applications consolidated		DVFS0		DVFS3	
		avg. (W)	sust. (W)	avg. (W)	sust. (W)
S1	TCP-W(20)+TPC-W(20)	178.0	190.1	176.2	174.5
S2	TPC-W(20)+TPC-W(20)+Streaming	193.5	193.2	177.0	172.0

Table 8: Predicted values for sustained and average power consumption for two subset of applications at two processor power states

- ❑ Average budget 180 W
- ❑ Sustained budget 185 W, 1 sec
- ❑ **Questions:** Which apps can be consolidated and at what DVFS state?

Example: Power-aware Application Packing

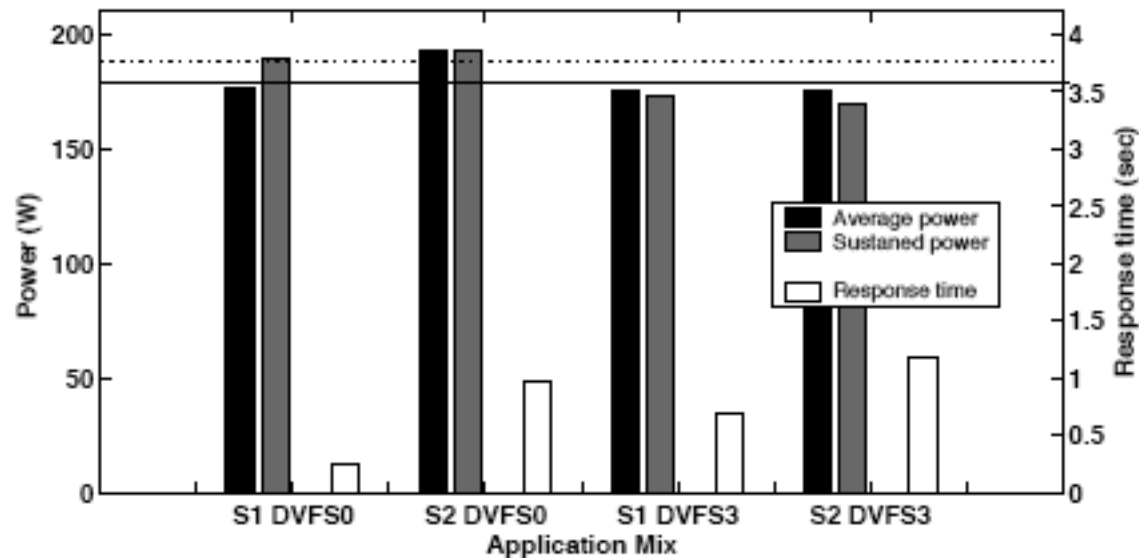


Figure 11: Illustration of packing decisions made by our predicting technique involving 3 applications.

- Prediction techniques allow systematic answers to previous questions

- ✓ Motivation
- ✓ Power Profiles
- ✓ Power Prediction
- ✓ Preliminary Evaluation
- **Concluding Remarks**

Key Take-aways

- ❑ Statistical **power profiles** allow simple yet effective prediction of average and sustained power consumption
- ❑ Likely to be useful in the design and operation of power-aware data centers

- ❑ Reactive capping mechanisms at the server-level to complement predictive placement
 - E.g., dynamic DVFS control [Sigmetrics 2008, to appear]
 - MDP-based control
 - Salient feature: DVFS state solely decided by queue length

- ❑ Online profiling and prediction
 - Low-overhead VMM mechanisms to separate out per-VM power consumption

- ❑ Extending prediction and capping to the rack-level and beyond
 - Employing Raritan's PDU model DPCR 20-20
 - Use PDU measurements for online profiling
 - Prediction of power behavior based on these profiles

- ❑ Reactive capping mechanisms at the server-level to complement predictive placement
 - E.g., dynamic DVFS control [Sigmetrics 2008, to appear]
 - MDP-based control
 - Salient feature: DVFS state solely decided by queue length

- ❑ Online profiling and prediction
 - Low-overhead VMM mechanisms to separate out per-VM power consumption

- ❑ Extending prediction and capping to the rack-level and beyond
 - Employing Raritan's PDU model DPCR 20-20
 - Use PDU measurements for online profiling
 - Prediction of power behavior based on these profiles

- ❑ Reactive capping mechanisms at the server-level to complement predictive placement
 - E.g., dynamic DVFS control [Sigmetrics 2008, to appear]
 - MDP-based control
 - Salient feature: DVFS state solely decided by queue length

- ❑ Online profiling and prediction
 - Low-overhead VMM mechanisms to separate out per-VM power consumption

- ❑ Extending prediction and capping to the rack-level and beyond
 - Employing Raritan's PDU model DPCR 20-20
 - Use PDU measurements for online profiling
 - Prediction of power behavior based on these profiles

- ❑ Reactive capping mechanisms at the server-level to complement predictive placement
 - E.g., dynamic DVFS control [Sigmetrics 2008, to appear]
 - MDP-based control
 - Salient feature: DVFS state solely decided by queue length

- ❑ Online profiling and prediction
 - Low-overhead VMM mechanisms to separate out per-VM power consumption

- ❑ Extending prediction and capping to the rack-level and beyond
 - Employing Raritan's PDU model DPCR 20-20
 - Use PDU measurements for online profiling
 - Prediction of power behavior based on these profiles

- ❑ **Acknowledgements**

- Jeonghwan Choi
- Sriram Govindan
- Anand Sivasubramaniam

- ❑ **More information:**

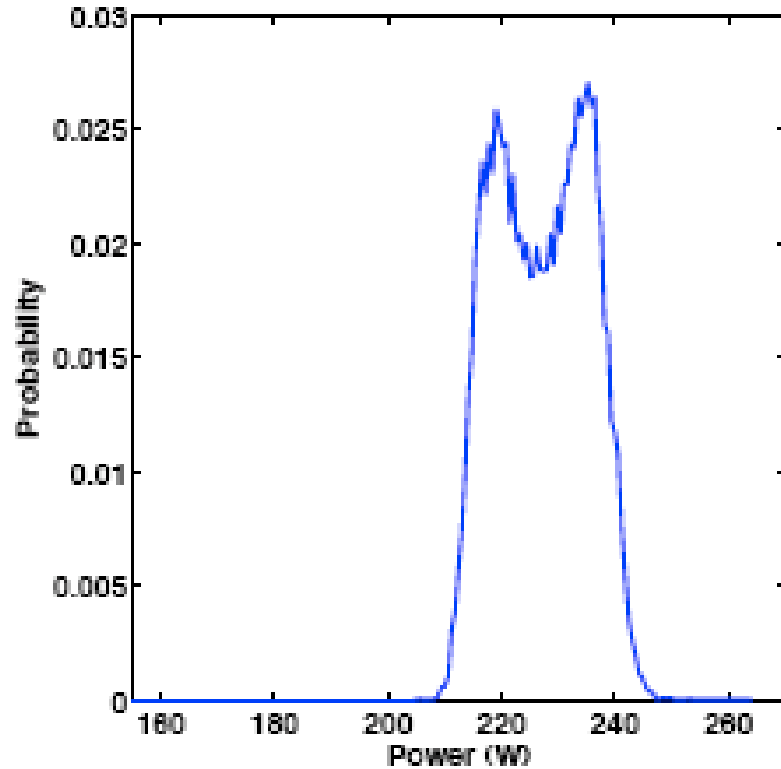
- <http://csl.cse.psu>

- ❑ **Questions or comments?**

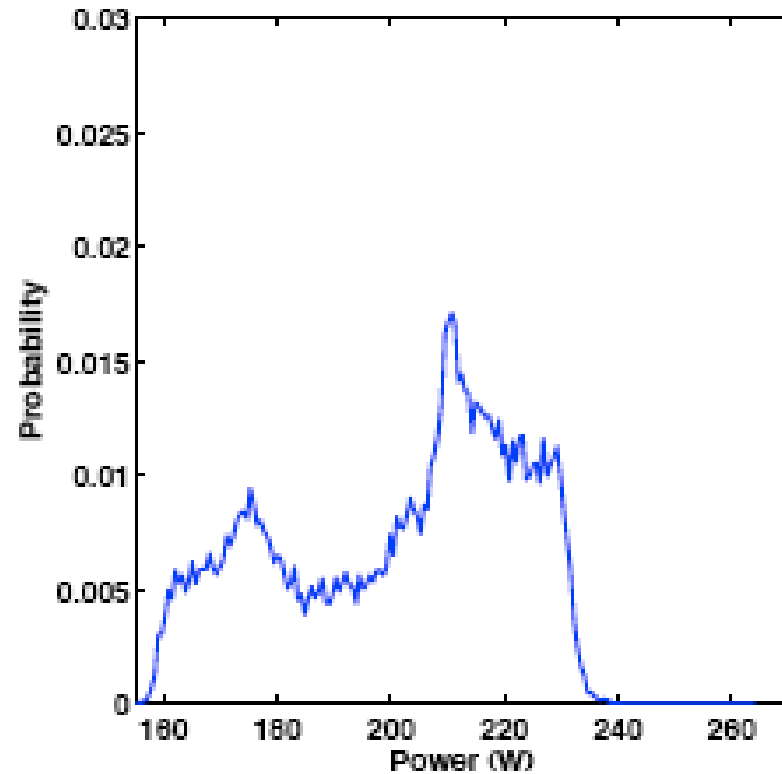


Additional Slides Follow ...

Variance of Power Profiles



Bzip2



Streaming

- Higher variance (longer tails) for non CPU-saturating applications

Impact of DVFS State

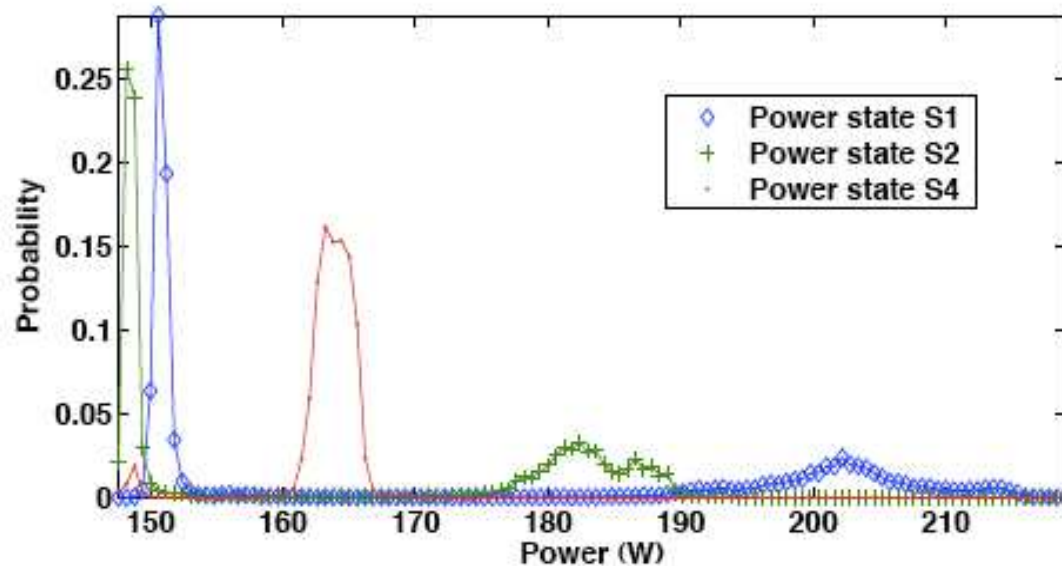
Power state	Bzip2			TPC-W(20)		
	Power (W)	CPU (frac.)	Norm. degrad.	Power (W)	CPU (frac.)	Norm. degrad.
S1	224.9	0.98	1	164.7	0.14	1
S2	200.1	0.99	1.10	161.9	0.15	1.07
S3	189.2	0.99	1.19	160.5	0.16	1.12
S4	172.1	0.99	2.19	161.3	0.33	2.02

- ❑ Power/performance trade-offs depend significantly on how CPU-saturating the application is

Power state	CPU Utilization				Normalized Performance degradation	Power (Watt)
	Average	95th	99th	Peak		
S1	0.41	0.92	0.93	0.95	1	185.6
S2	0.44	0.93	0.95	0.98	1.18	175.3
S4	0.92	0.97	0.98	0.99	15.69	173.2

TPC-W, 60 clients

Impact of DVFS State



- Non CPU-saturating apps at lower power states
 - CPU utilization increases
 - Power profile less bursty

Power state	CPU Utilization				Normalized Performance degradation	Power (Watt)
	Average	95th	99th	Peak		
S1	0.41	0.92	0.93	0.95	1	185.6
S2	0.44	0.93	0.95	0.98	1.18	175.3
S4	0.92	0.97	0.98	0.99	15.69	173.2

TPC-W, 60 clients