Statistical Profiling-based Techniques for Effective Power Provisioning in Data Centers

Sriram Govindan,

Jeonghwan Choi, Bhuvan Urgaonkar,

Anand Sivasubramaniam, Andrea Baldini

Penn State, KAIST, Tata Consultancy Services, Cisco Systems

Eurosys 2009 , March 31st – April 3rd 2009

Growing Energy Demands

- In 2006, U.S data centers
  - Spent $4.5 billion just for powering their infrastructure
  - 1.5% of the total electricity consumed in the U.S
  - Has more than doubled since 2000 - further expected to double by 2011
- Massive growth of installed hardware resources
  - By 2010, servers expected to triple from 2000
  - Average utilization of servers between 5% and 15%

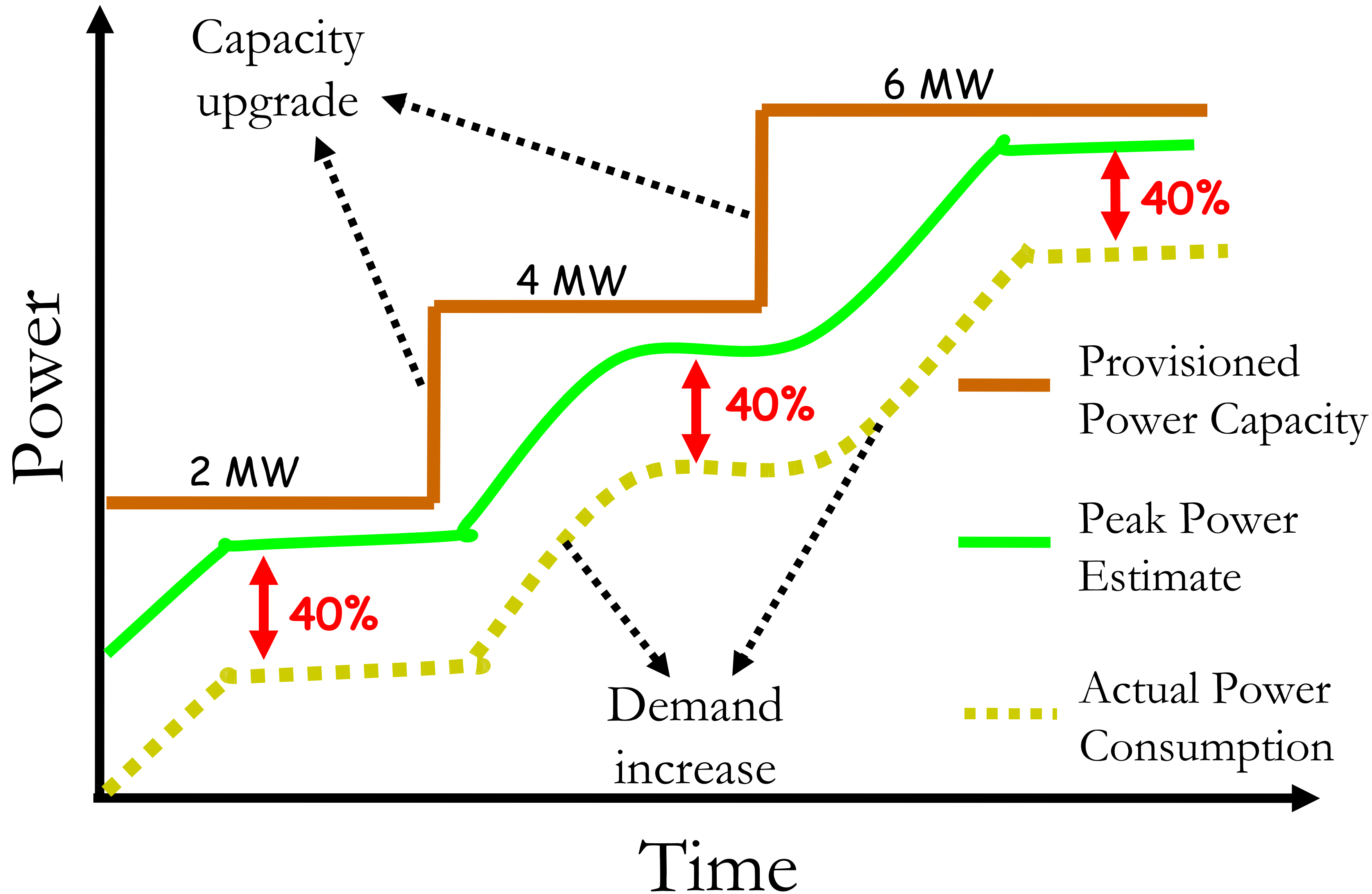
COMPUTER SYSTEMS LAB

Reference: *EPA Data center report, 2007*

Data Center Energy Management

- Tackle server sprawl
  - *Server virtualization:* Consolidates workload on to fewer number of servers and switch off remaining idle servers
  - Growth in number of data centers – provisioning power infrastructure of a data center
    - *Provisioned power capacity:* Maximum power available to the data center as negotiated with the electricity provider
    - *Provisioning:* How many IT equipments (servers, disk arrays, etc.) can be hosted within a data center ?

Data Center Power Provisioning



- Hand drawn figure

Over-provisioned Data Centers

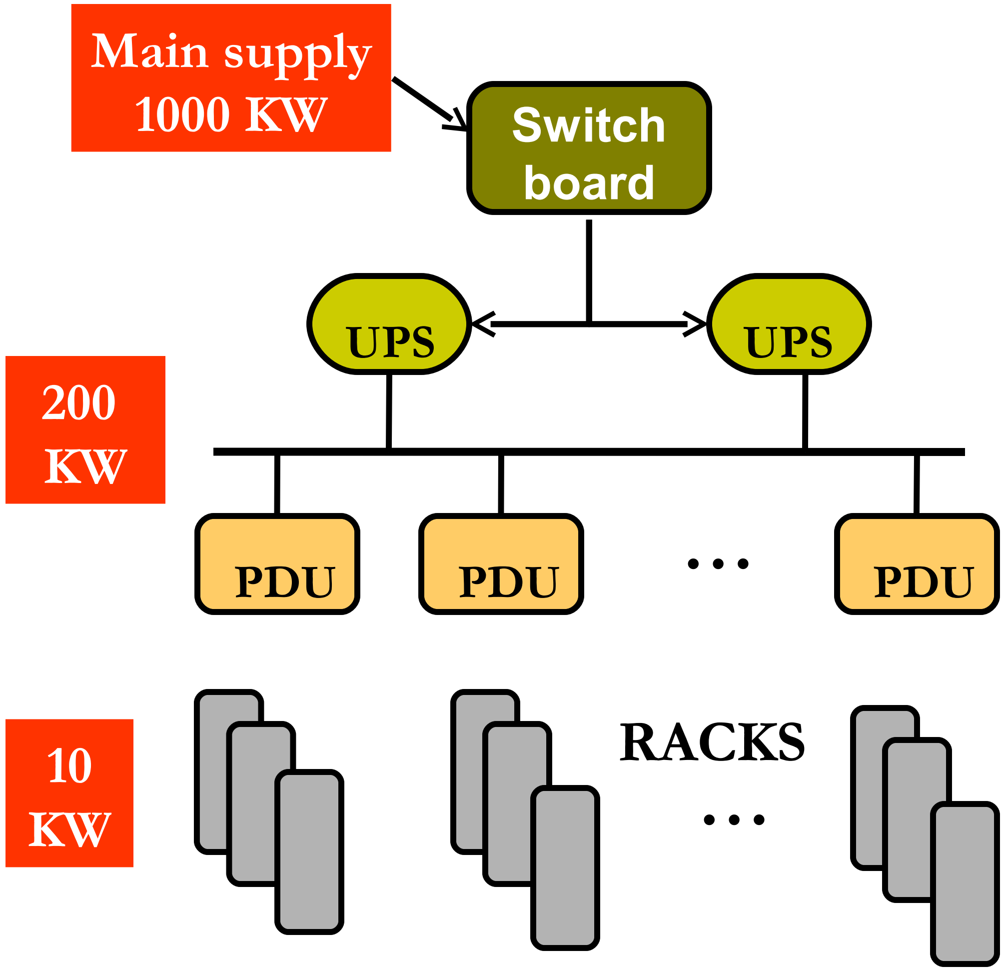
- Current provisioning practices render data centers' power infrastructure highly under-utilized
    - **Reliability concerns**
- Over-provisioning hurts profitability of data centers due to
    - Unnecessary proliferation of data centers
        - Increase in management and installation costs
    - Electrical and cooling inefficiency
        - Efficiency is worse at lower loads

- **Goal: Improve utilization of the power infrastructure in data centers while adhering to reliability constraints**

Talk Outline

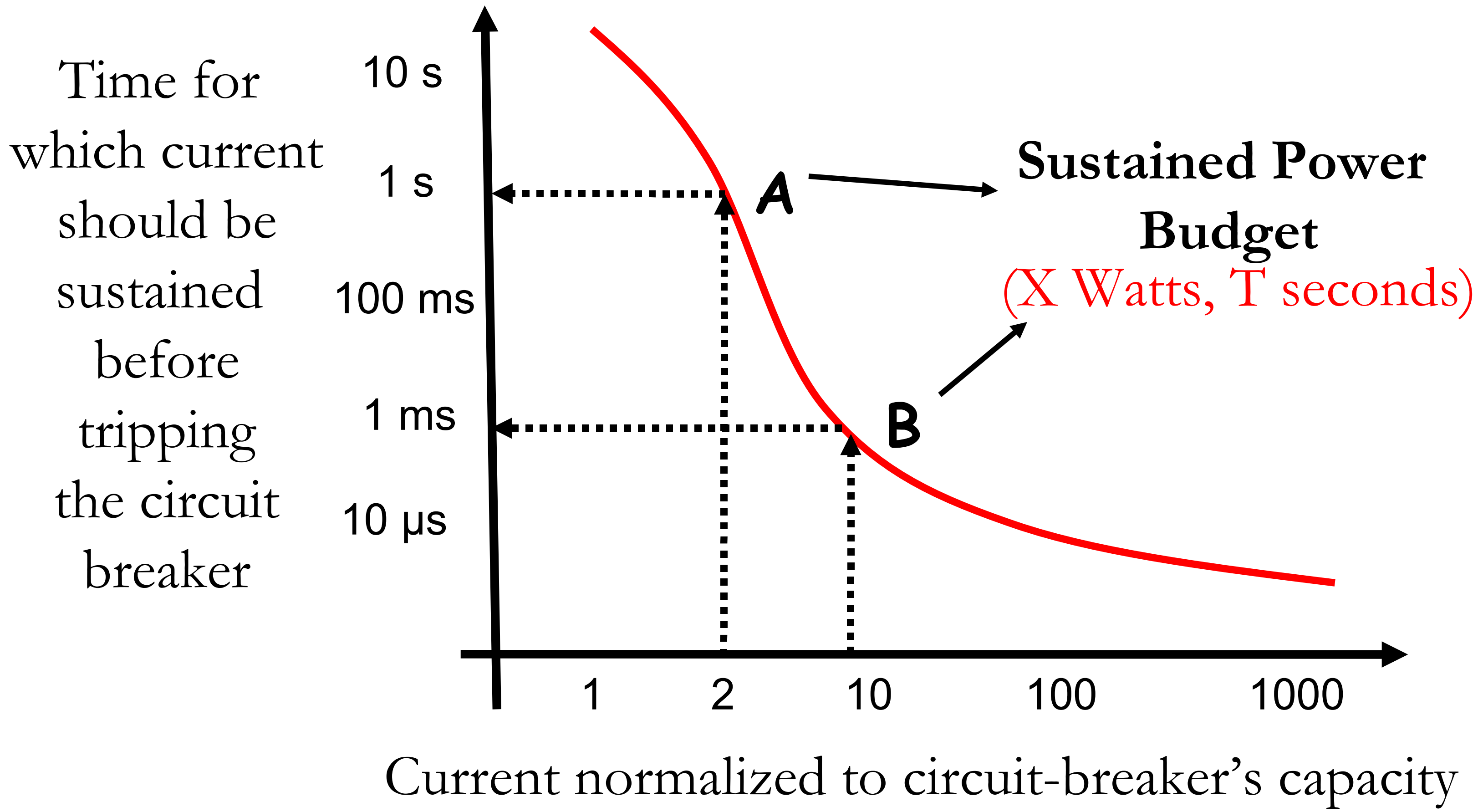
- Data Center Power Hierarchy
  - Hardware reliability constraints
- Application Power Profiles
- Improved Power Provisioning
  - Threshold-based power budget enforcer
- Evaluation

Data center Power Supply Hierarchy



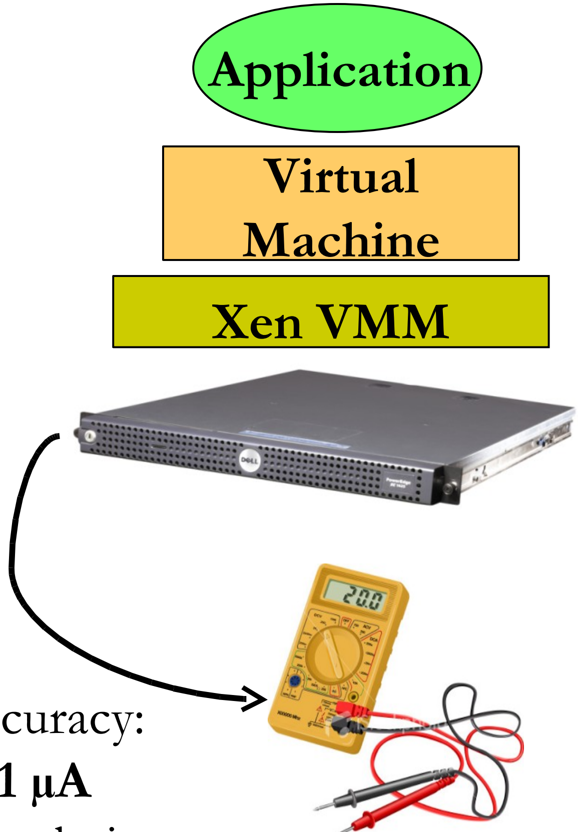
- *Circuit breakers* placed at each element of a data center power hierarchy to protect the underlying circuit from current overdraw or short-circuit situations

Time-current characteristics Curve of a typical Circuit-breaker



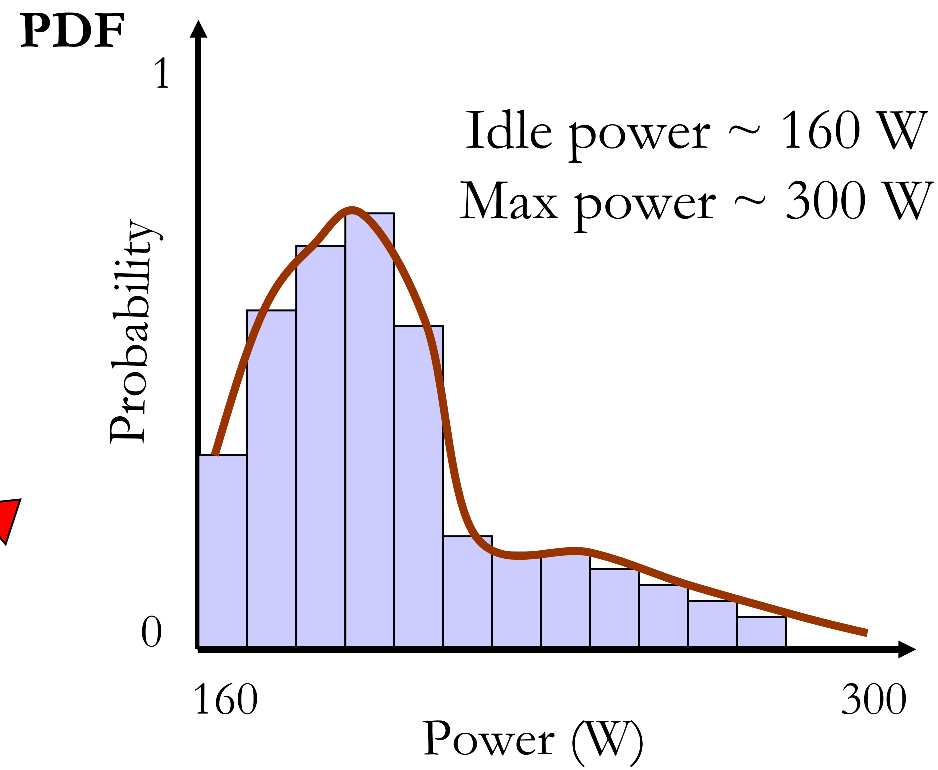
- Hand drawn figure

Profiling Application Power Consumption

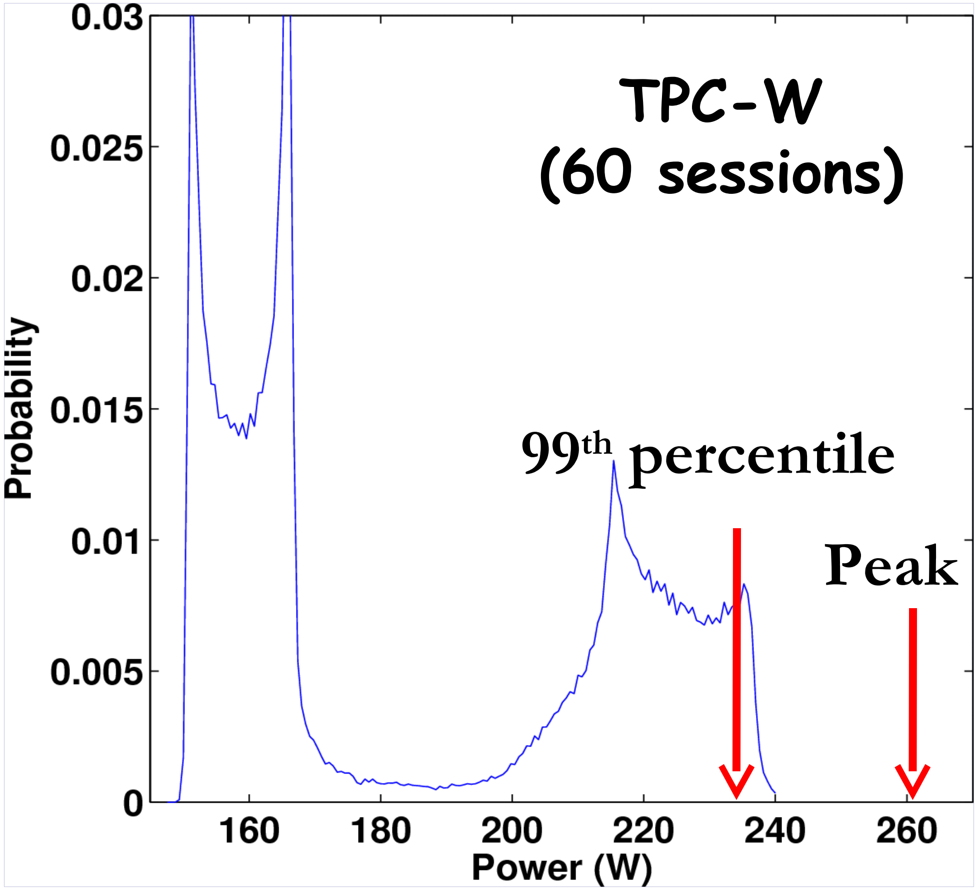


Accuracy:
1 µA
Granularity:
1 ms

Signametrics Multimeter (SM2040)

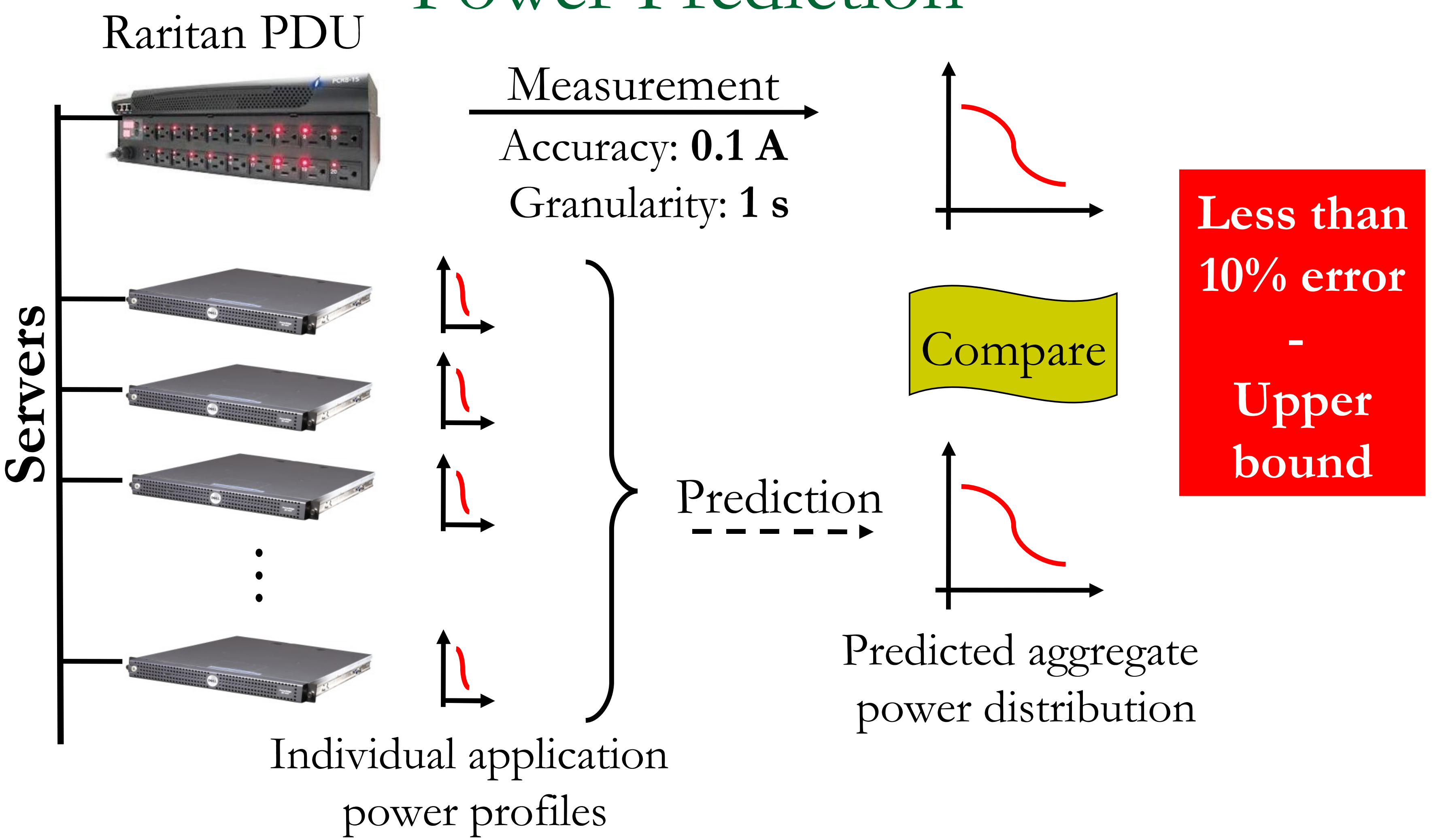


Power Profiles - 2 ms Granularity



- TPC-W
  - Emulates a two-tiered implementation of an e-commerce book-store with front-end jboss web server and back-end mysql database.

Statistical Multiplexing Based Sustained Power Prediction



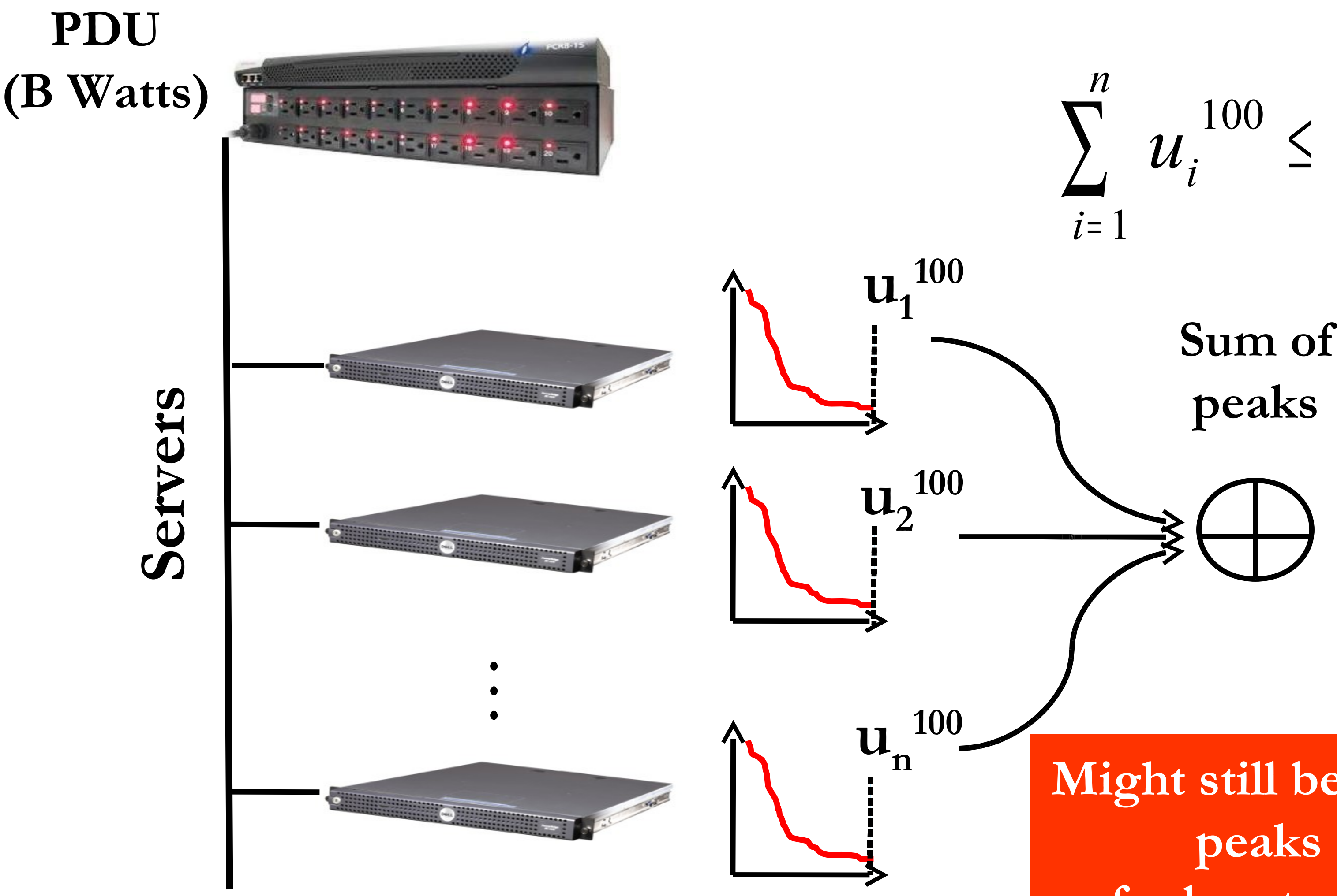
Reference: Profiling, prediction and capping of power-consumption for Consolidated Data-center environment, Choi et al., MASCOTS 2008

Existing Power Provisioning Techniques

- Face-plate rating/Name-plate rating
  - Assumes all components are populated in the server
    - Eg: All processor sockets, DIMM slots, HDDs etc.,
  - Assumes all components consume peak power at the same time
- Vendor power calculators
  - Dell, IBM, HP etc.
  - Tuned for current server's configuration and coarse-level application load information.
  - Less conservative than Face-plate Rating

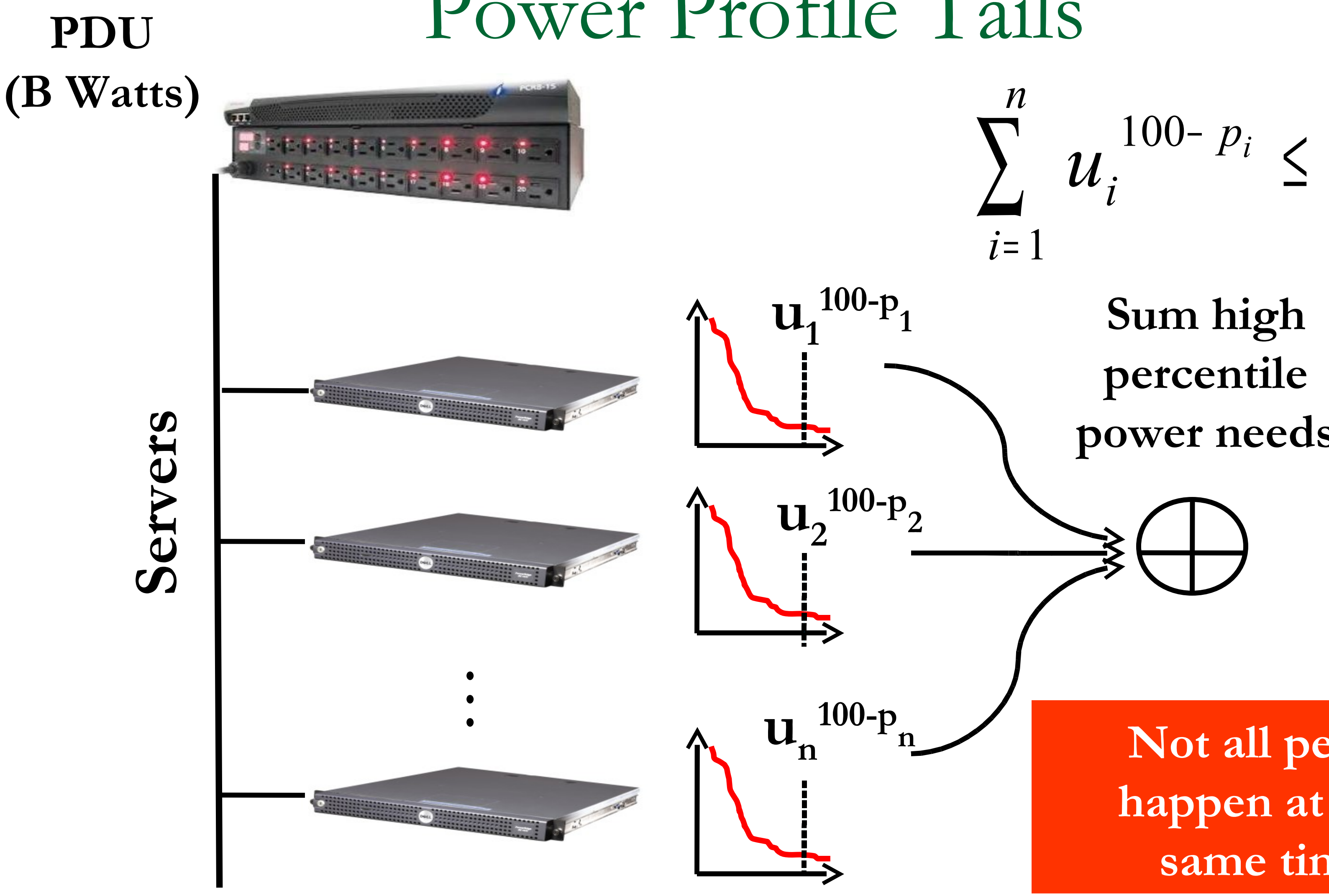
Provisioning for Peak Power Needs

$$\sum_{i=1}^{n} u_i^{100} \leq B$$



Might still be conservative - peaks are rare for bursty applications

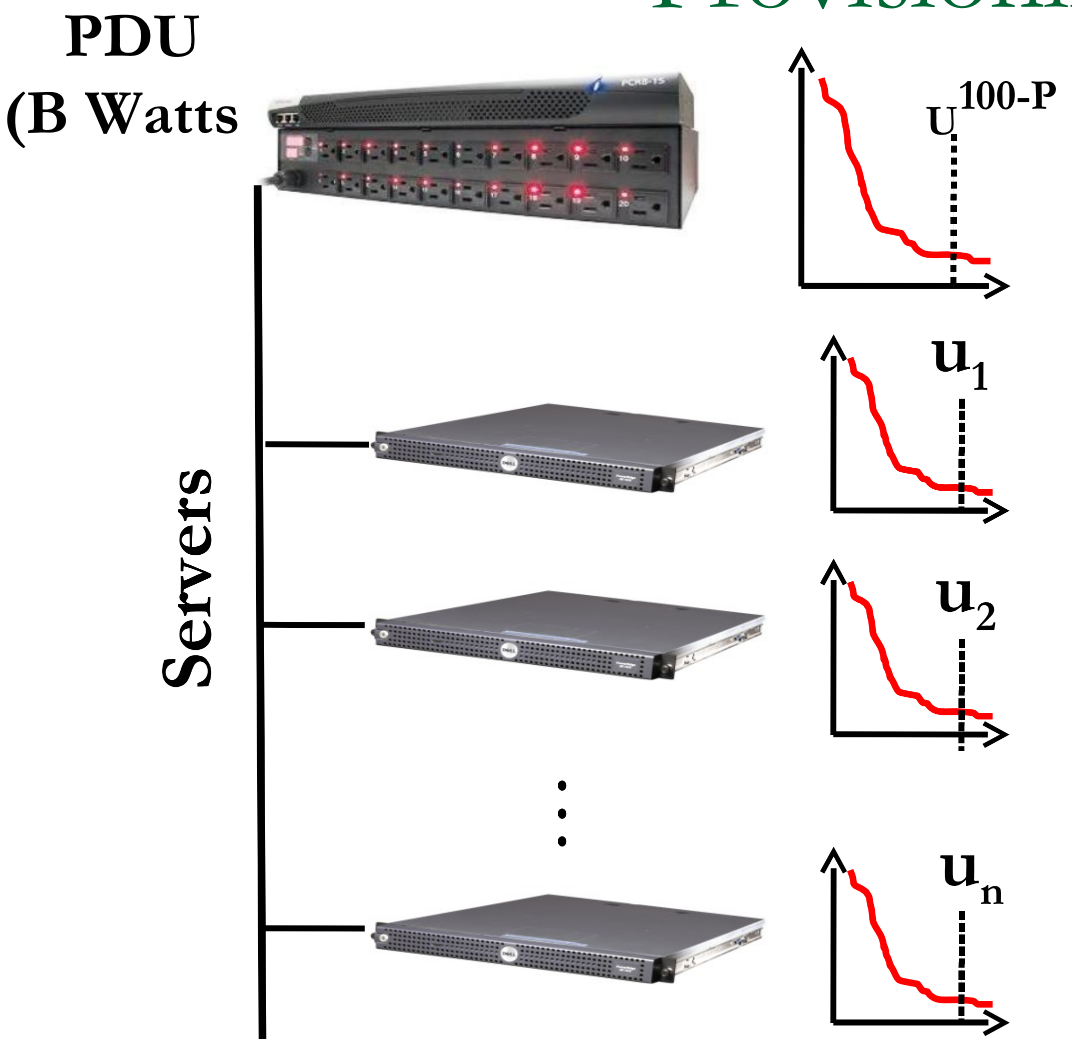
Under-provisioning Based on Power Profile Tails



$$\sum_{i=1}^{n} u_i^{100-p_i} \leq B$$

Not all peaks happen at the same time

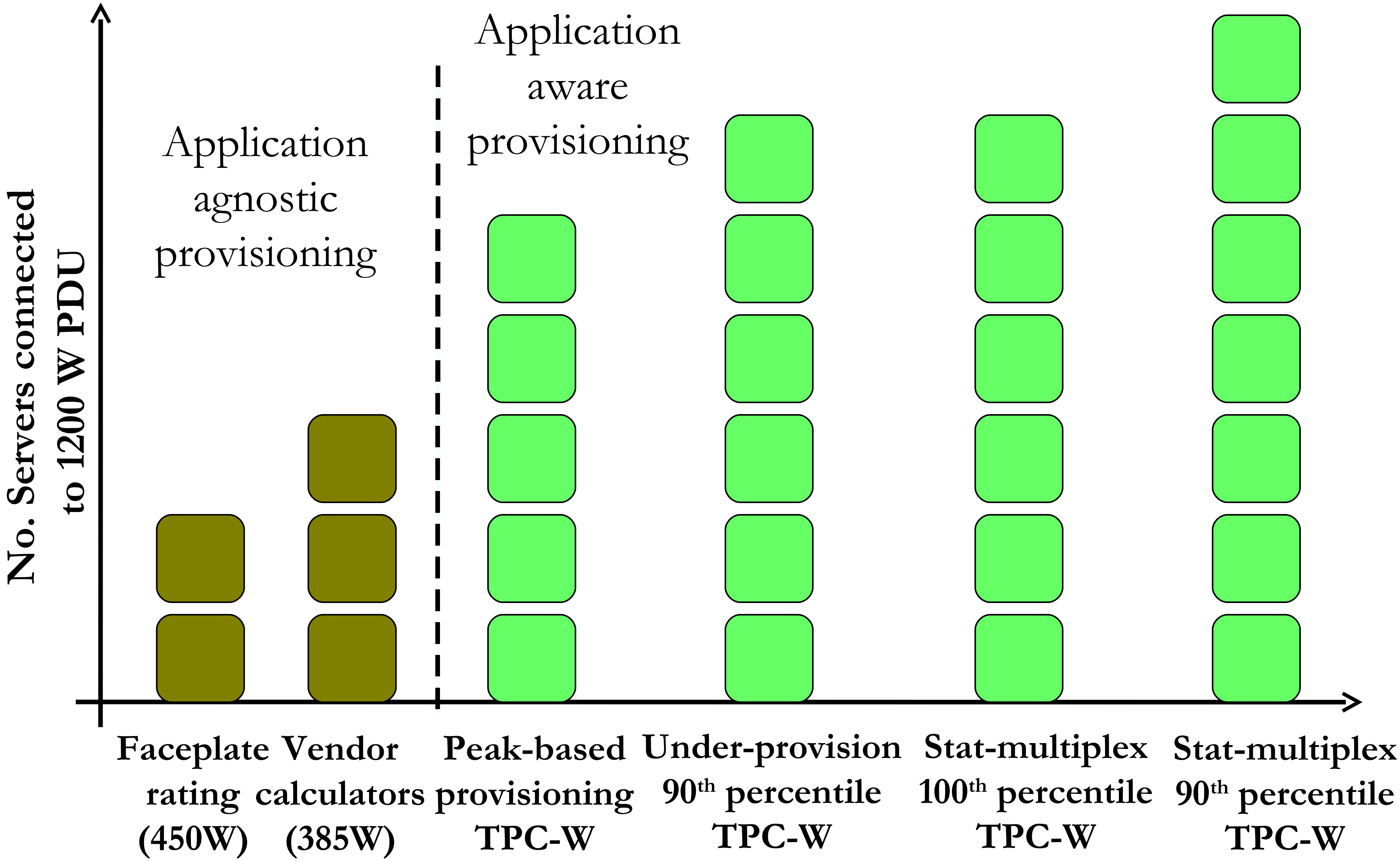
Statistical-multiplexing Based Provisioning



$$U^{100-P} \leq B$$

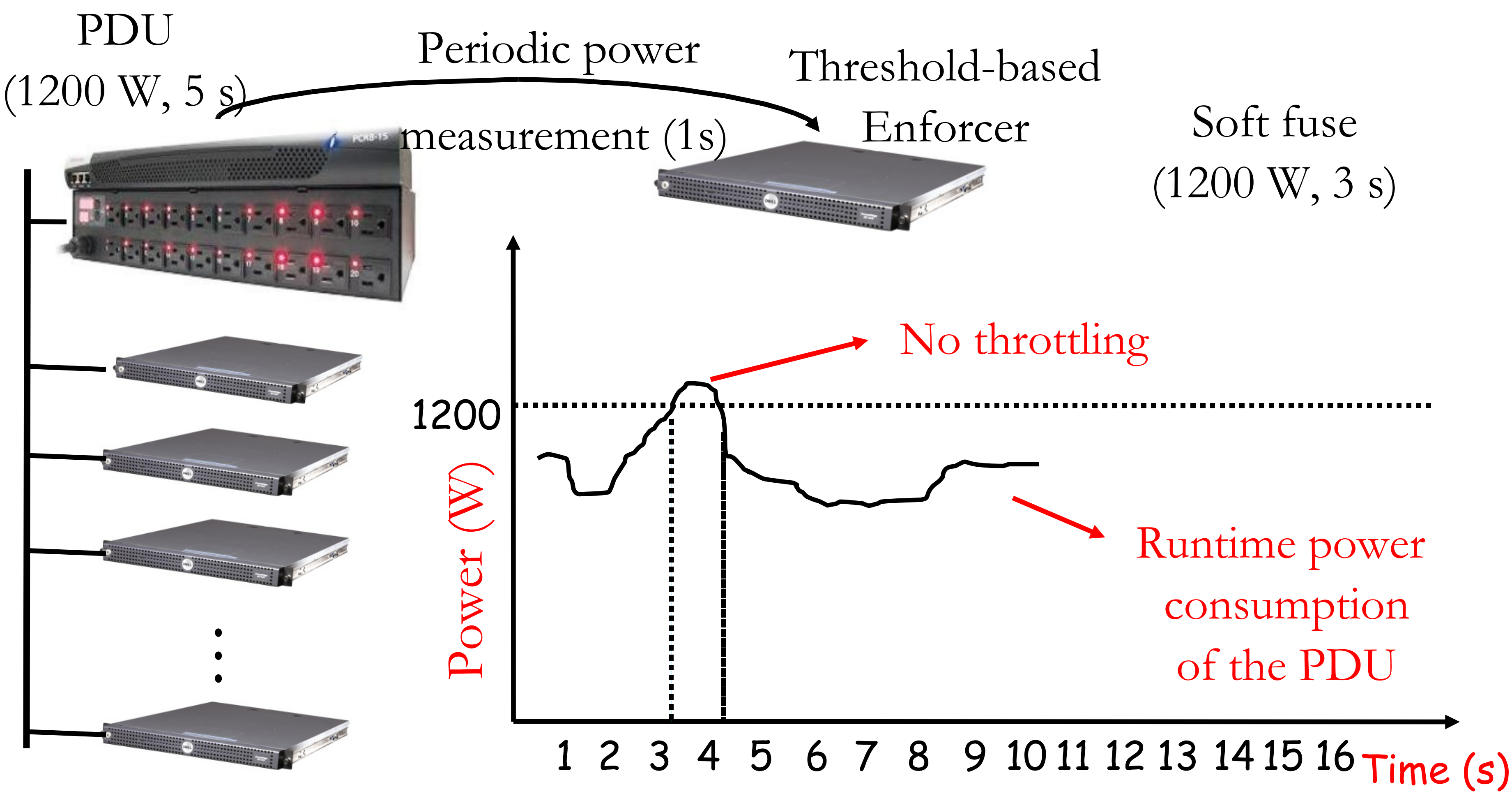
Provision for the aggregated power profile of the PDU, 'U' as predicted by our sustained power prediction technique

Provisioning Techniques -Evaluation



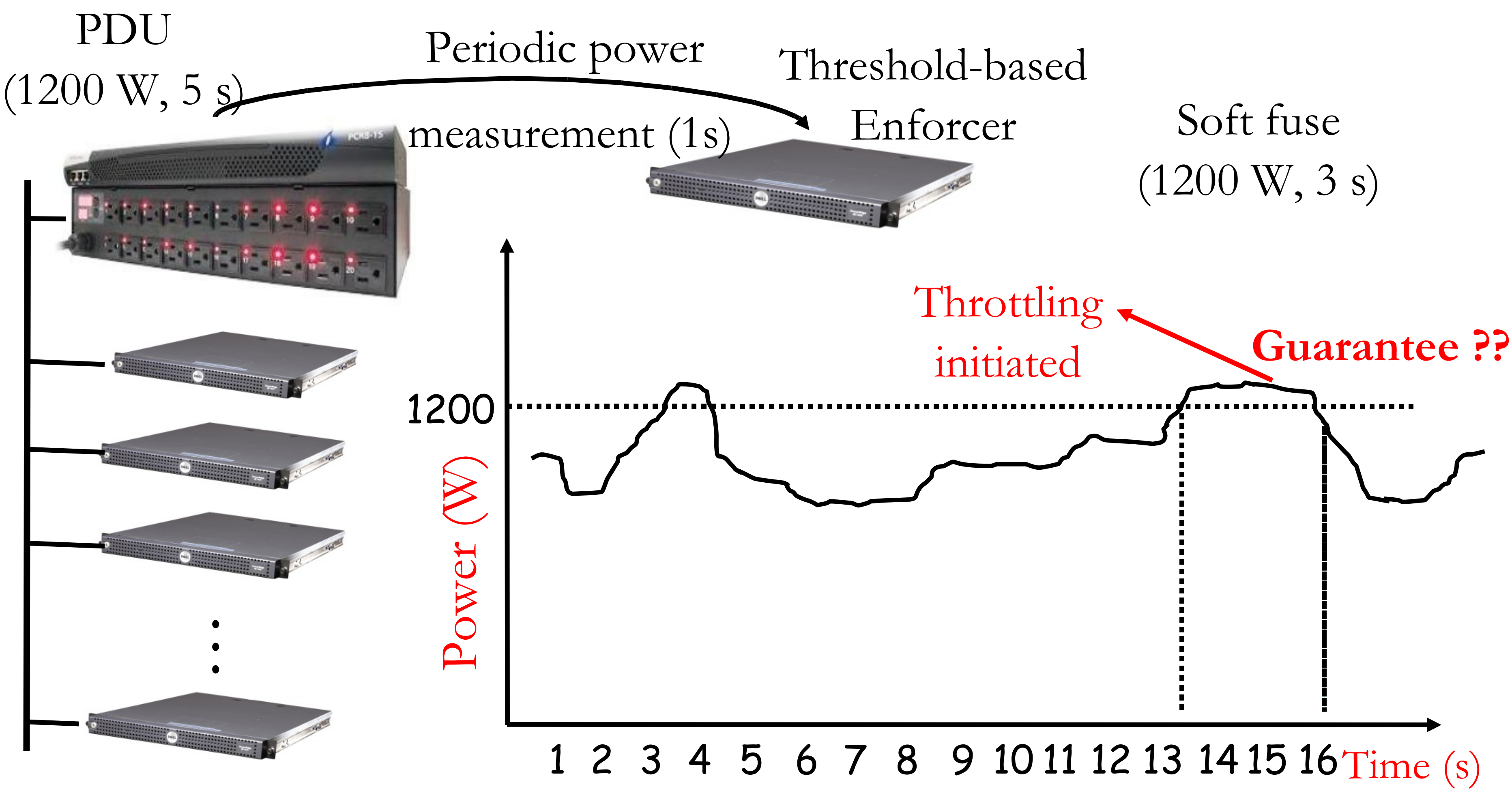
COMPUTER SYSTEMS LAB

Threshold-based Soft-fuse Enforcement



- Hand drawn figure

Threshold-based Soft-fuse Enforcement



- Hand drawn figure

Threshold-based Soft-fuse Enforcement

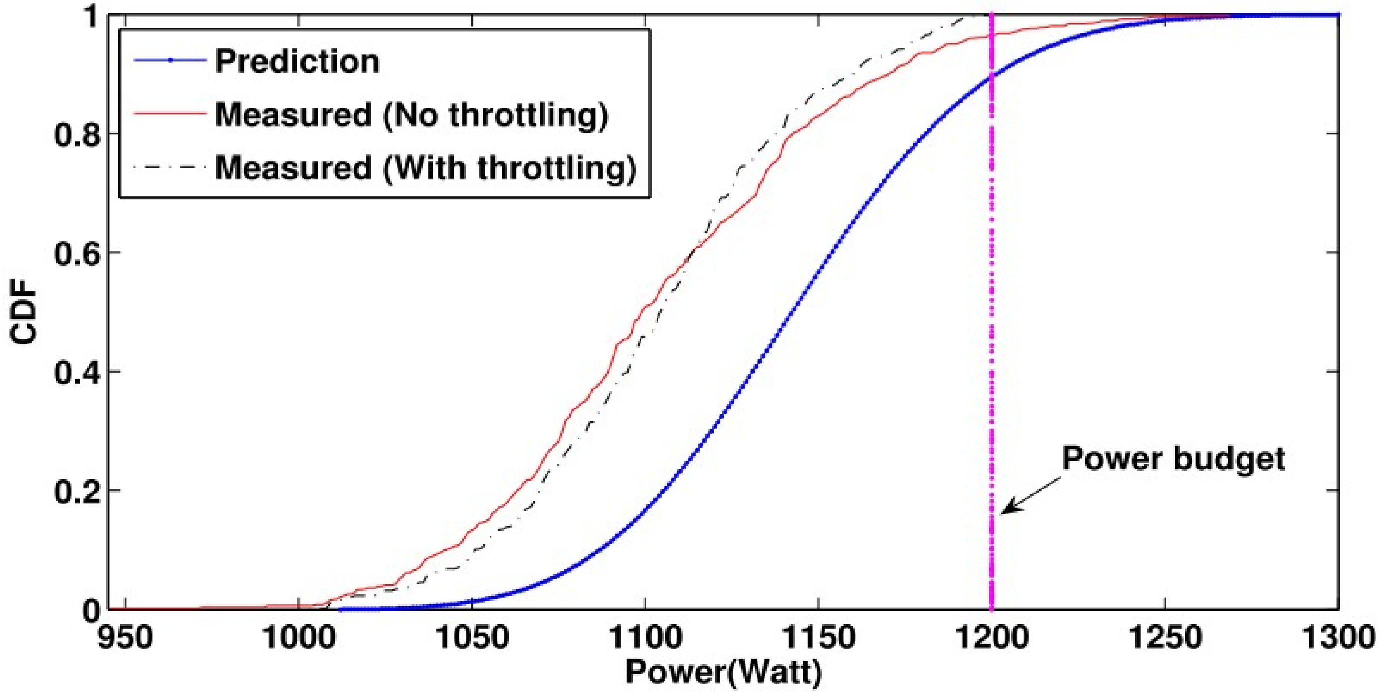
Sustained power consumption (100th percentile) of a PDU connected to servers hosting TPC-W

Power State	6 Servers	7 Servers	8 Servers	9 Servers
3.4 Ghz	**1191.0 W**	1300.0 W	1481.0 W	1672.0 W
2.8 Ghz	976.6 W	**1138.6 W**	1308.2 W	1478.2 W
1.4 Ghz	861.7 W	1011.7 W	**1162.7 W**	1313.6 W

- Choose appropriate throttling state that satisfies reliability constraint (1200W, 5s) as highlighted in the table

COMPUTER SYSTEMS LAB

Threshold-based Soft-fuse Enforcement



- **Provisioning for the 90th percentile power needs:** Threshold based enforcer is successfully able to enforce soft fuse of the PDU connected to 7 TPC-W servers

Gains vs Performance Degradation

- Experiment: 7 TPC-W servers connected to 1200 W PDU
- Gains: Computation per Provisioned Watt
  - Increase in number of servers (computation cycles) hosted in the data center
  - Decrease in number of computation cycles due to throttling
  - CPW increased by 120% from vendor-based provisioning
- Performance Degradation:
  - Average response time of TPC-W not affected
  - 95th percentile response time of TPC-W increased from 1.59 s to 1.78 s (12% degradation)

Concluding Remarks

- Power provisioning in data centers
  - Characterize hardware reliability constraints
  - Profile application power consumption
  - Improve provisioning of data center power infrastructure
- Future work
  - Correlated power peaks across servers
  - Handle dynamically varying workload phases
- Software URL: http://csl.cse.psu.edu/hotmap
  - Sustained power prediction scripts
  - Threshold-based soft-fuse enforcer
  - Xen kernel patch for enabling MSR writes