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
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%

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

- Capacity upgrade
- Demand increase

Probe provisioned power capacity
- Provisioned power capacity
- Peak power estimate
- Actual power consumption

- 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

Time for which current should be sustained before tripping the circuit breaker

Current normalized to circuit-breaker’s capacity

- Hand drawn figure
Profiling Application Power Consumption

- Application
- Virtual Machine
- Xen VMM

Accuracy: 1 µA
Granularity: 1 ms

Signametrics Multimeter (SM2040)

PDF

Power (W)

Probabilty

Idle power ~ 160 W
Max power ~ 300 W
Power Profiles - 2 ms Granularity

- **TPC-W**
  - Emulates a two-tiered implementation of an e-commerce bookstore with front-end JBoss web server and back-end MySQL database.

![Graph showing Power Profiles](image-url)
Statistical Multiplexing Based Sustained Power Prediction

Raritan PDU

Measurement
Accuracy: 0.1 A
Granularity: 1 s

Compare

Predicted aggregate power distribution

Individual application power profiles

Less than 10% error - Upper bound

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

PDU (B Watts)

Servers

\[ \sum_{i=1}^{n} u_{i}^{100} \leq B \]

Sum of peaks

Might still be conservative - peaks are rare for bursty applications
Under-provisioning Based on Power Profile Tails

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

PDU (B Watts)

Servers

Sum high percentile power needs

Not all peaks happen at the same time
Statistical-multiplexing Based Provisioning

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

\[ U^{100-P} \leq B \]
Provisioning Techniques - Evaluation

No. Servers connected to 1200 W PDU

Application agnostic provisioning

Faceplate rating calculators (450W) (385W)

Vendor

Peak-based provisioning TPC-W

Under-provision 90th percentile TPC-W

Stat-multiplex 100th percentile TPC-W

Stat-multiplex 90th percentile TPC-W

Application aware provisioning

90th percentile TPC-W

100th percentile TPC-W
Threshold-based Soft-fuse Enforcement

PDU (1200 W, 5 s) → Periodic power measurement (1 s) → Threshold-based Enforcer

Soft fuse (1200 W, 3 s)

Runtime power consumption of the PDU

- No throttling

Power (W) vs. Time (s)

1200 W
Threshold-based Soft-fuse Enforcement

PDU (1200 W, 5 s)

Periodic power measurement (1s)

Threshold-based Enforcer

Soft fuse (1200 W, 3 s)

Guarantee ??

Power (W)

1200

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 Time (s)

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