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

Bhuvan Urgaonkar
Sriram Govindan, Shiva Chaitanya, Jeonghwan Choi, Anand Sivasubramanian

Computer Systems Lab, Penn State
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
  - Doubled since 2000, further expected to double by 2011

Contributing Factors

- Massive growth of installed hardware resources
  - Server sprawl: By 2010, servers expected to triple from 2000
  - Similar trends for storage, networking equipment, etc.
  - Techniques for consolidation and dynamic power modulation

- Conservative provisioning of power supply infrastructure cause data center sprawl
  - Provisioned power capacity: Maximum power available to the data center as negotiated with the electricity provider
  - Relatively less attention

Data Center Power Provisioning

- Hand drawn figure
Data Center Power Provisioning

- Current practices take an excessively pessimistic approach to dealing with
  - Reliability and performance concerns
  - Based on accommodating worst-case demands
Over-provisioned Power Infrastructure

Over-provisioning hurts profitability of data centers due to

- Unnecessary proliferation of data centers
- Efficiency is worse at lower loads
- Utility cost decoupling
  - Penalizes consumers that spend less energy than they have asked for via higher recurring cost

Goal: Improve utilization of power infrastructure while adhering to reliability and performance constraints

Adopted in 4 states so far in U.S including CA and MD
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 could be sustained before tripping the circuit breaker

Sustained Power Budget
(X Watts, T seconds)

Current normalized to circuit-breaker’s capacity

- Hand drawn figure
Average and Sustained Power

Average power = Area under the curve / Total time

Sustained power is S1 and S2 for intervals L1 and L2

Power

0 W  150 W  170 W  190 W  210 W  230 W

Time

L1  L2

min = S1

min = S2
Profiling Application Power Consumption

Application

Virtual Machine

Xen VMM

Accuracy: 1 µA
Granularity: 1 ms

Signametrics Multimeter (SM2040)

PDF

Idle power ~ 160 W
Max power ~ 300 W

Power (W)

Probability

0

160

300
Power Profile

PDF

CDF

Power Consumption (W)

Power (W)

Probability

Cumulative Probability

PDF

CDF

Power Profile

0

1

0

1

0.99

A

B

0

A

B
Power Profiles @ 2 ms granularity

TPC-W

Power when workload is blocked

Non-CPU-Saturating Application

CPU-Saturating Application

Bzip2
Existing Power Provisioning Techniques

- **Face-plate rating/Name-plate rating**
  - Assumes all components are populated in the server
    - E.g.,: 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
  - However, workload-agnostic
Intuitions

• **Observation 1:** Applications with long power tails
  – Implication: Under-provisioning would yield benefits *(Safe?)*
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.
Intuitions

• Observation 1: Applications with long power tails
  – Implication: Under-provisioning would yield benefits (Safe?)

• Observation 2: Self-similarity in power usage
  – E.g., Hurst parameter for TPC-W is 0.86
  – Implication: Burstiness (and hence under-provisioning) likely to be seen even at higher power elements
A Workload with Bursty Power Needs

TPC-W

Self-similarity:
Hurst parameter
0.86

7 TPC-W
Intuitions

• **Observation 1:** Applications with long power tails
  – Implication: Under-provisioning would yield benefits (Safe?)

• **Observation 2:** Self-similarity in power usage
  – E.g., Hurst parameter for TPC-W is 0.86
  – Implication: Burstiness (and hence under-provisioning) likely to be seen even at higher power elements

• **Observation 3:** Inter-application variance in power usage
  – Implication: Statistical multiplexing gains likely
Statistical Multiplexing Based Sustained Power Prediction

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

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

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

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

No. Servers connected to 1200 W PDU

- Application agnostic provisioning
- Application aware provisioning

- Faceplate rating calculators (450W)
- Vendor (385W)
- Peak-based provisioning TPC-W
- Under-provision 90th percentile TPC-W
- Stat-multiplex 100th percentile TPC-W
- Stat-multiplex 90th percentile TPC-W
Threshold-based Soft-fuse Enforcement

PDU (1200 W, 5 s)

Periodic power measurement (1s)

Threshold-based Enforcer

Soft fuse (1200 W, 3 s)

- Hand drawn figure
Threshold-based Soft-fuse Enforcement

PDU
(1200 W, 5 s)

Periodic power measurement (1s)

Threshold-based Enforcer

Soft fuse
(1200 W, 3 s)

Guarantee ??

Throttling initiated

- Hand drawn figure
### Soft-fuse Enforcement

<table>
<thead>
<tr>
<th>Power State</th>
<th>6 Servers</th>
<th>7 Servers</th>
<th>8 Servers</th>
<th>9 Servers</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4 Ghz</td>
<td>1191.0 W</td>
<td>1300.0 W</td>
<td>1481.0 W</td>
<td>1672.0 W</td>
</tr>
<tr>
<td>2.8 Ghz</td>
<td>976.6 W</td>
<td>1138.6 W</td>
<td>1308.2 W</td>
<td>1478.2 W</td>
</tr>
<tr>
<td>1.4 Ghz</td>
<td>861.7 W</td>
<td>1011.7 W</td>
<td>1162.7 W</td>
<td>1313.6 W</td>
</tr>
</tbody>
</table>

- **Threshold-based**: Choose appropriate throttling state that satisfies reliability constraint (1200W, 5s)
- **More sophisticated**: MDP formulation that yields efficient control rules dependent only on #jobs

---

CSL

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
  - 95\textsuperscript{th} percentile response time of TPC-W increased from 1.59 s to 1.78 s (12% degradation)
Concluding Remarks

• Desirable and feasible to provision close to power demands
  – Characterization of hardware reliability constraints
  – Statistical profiling and prediction of aggregate power needs
  – Provisioning techniques with tunable performance impact

• For further information: http://csl.cse.psu.edu/hotmap
  – Papers
    • Sigmetrics 2008: Control rules for enforcing power budgets
    • MASCOTS 2008: Profiling and prediction techniques
    • Eurosyst 2009: Statistical multiplexing gains
  – Software:
    • Power profiling and prediction scripts
    • Xen-based soft-fuse enforcer
Related work

- **Power/Thermal provisioning**
  - **Server-Level Power Control.** *ICAC ’07.* Lefurgy et al., IBM Austin.
  - **Power-shifting:** A Performance-Conserving Approach for Reducing Peak Power Consumption in Server Systems. *ICS ’05.* Felter et al., IBM Austin.
  - **Cluster-level feedback power control** for performance optimization. *HPCA ’08.* Wang et al., Univ of Tennessee, Knoxville.
  - **Ensemble-level Power Management** for Dense Blade Servers. *ISCA 2006.* Ranganathan et al., HP.
Comparison of measured and predicted sustained power consumption (T=1 sec) for a server consolidation TPC-W, Bzip2 and TPC-W with reservation 60%, 20% and 20%
Under-provisioning power capacity

- It is profitable to provision for less than worst case for bursty applications

Provision for the 100\textsuperscript{th} Percentile of TPC-W. Provision up to 5 Servers.

PDU (1200 W)

Provision for the 90\textsuperscript{th} Percentile of TPC-W. Provision up to 6 Servers.
Statistical-multiplexing based Provisioning

- Not all peaks happen at the same time

Statistical-multiplexing based TPC-W peak (6 Servers)

Statistical-multiplexing based 90\textsuperscript{th} percentile of TPC-W power (7 Servers)

PDU (1200 W)
Sustained power estimation
- Space Sharing Results

Comparison of measured and predicted sustained power consumption
(T=5 sec) for a PDU consolidation 7 servers each running
an instance of TPC-W
Profiling

- **Workload size estimator – IBM link and others**
Applications

- **TPCW**
  - E-commerce book-store. 2 tier implementation – front-end Jboss Web server and back-end Database

- **SPECJBB 2005**
  - Emulates a 3-tier java client/server system

- **Streaming Media Server**
  - Transmits UDP packets to specified number of clients at specified rate

- **SPECCPU 2000**
  - Art, Mesa, Bzip2, Mcf
Power Profiles - 2 ms Granularity

TPC-W
(60 sessions)

Bzip2

Non-CPU-Saturating Application

CPU-Saturating Application
Commercial Circuit Breaker Details
Vendors power calculator - Links

- **IBM** -

- **SUN** -
  - [http://malaysia.sun.com/servicesolutions/ecoinnovation/powercalculators.jsp](http://malaysia.sun.com/servicesolutions/ecoinnovation/powercalculators.jsp)

- **HP** -

- **DELL**
Our Experimental Servers

• DELL PowerEdge SC1425
  – 2 - Xeon Processor
  – 4 - 512MB DIMM slot
  – 2 - Hard drive

• Face plate of our servers: 450W

• Vendor Power Calculator:
  – **Processor intensive applications:**
    System Heat/Power: 465 W
    Total Current: 2.24 Amps
  – **I/O intensive applications:**
    System Heat/Power: 446.4 W
    Total Current: 2.15 Amps
  – **Average load:**
    System Heat/Power: 385.9 W
    Total Current: 1.86 Amps
  – **Idle:**
    System Heat/Power: 315W
    Total Current: 1.51 Amps
Sustained Power Prediction - Time Sharing

Individual power profiles (from off-line profiling)

Prediction

Measurement

Predicted server power distribution

\[ R_{A1}, R_{A2}, \ldots, R_{An} \]

\[ A_1, A_2, \ldots, A_n \]
Profiling Infrastructure

Virtual Machine

Xen VMM

Application

Signametrics Multimeter
SM2040

Accuracy: 1 µA
Granularity: 1 ms

Raritan PDU
DPCR 20-20

Accuracy: 0.1 A
Granularity: 1 s

Idle power ~ 160 W
Max power ~ 300 W
Growing Energy Demands

- In 2006, U.S data centers
  - Spent $4.5 billion just for powering their servers
  - 1.5% of the total electricity consumed in U.S
  - Expected to double in the next five years
- Solution: Consolidation through virtualization
  - Companies like IBM, Sun, Cassatt, etc., offer virtualization services for reducing energy consumption
  - But consolidation also increases the power density/demands of the underlying servers
Predicting power demands of consolidated applications

- For systems working within these power budgets, we try to answer,
  - What will be the energy (average power) consumption of a data center hosting a set of applications?
  - What is the peak power requirement of a set applications consolidated within a power supply?
Peak Power – characterization

• We define peak power budget using the notion of *Sustained power consumption* which is a tuple, (X Watts, T seconds)

  – ‘X’ Watts of power or more that has to be
Average and Sustained Power

Average power = Area under the curve / Total time

Sustained power is S1 and S2 for intervals L1 and L2
Profiling Infrastructure

- Intel Xeon processors
- Network Interface Card
- Disk

Major contributor of power

Application

Virtual Machine

Xen VMM

Multimeter - 1 ms granularity

Idle power ~ 160 W

Idle power ~ 160 W
Sustained power prediction

**Input:**
- *From Profiling* - Individual application resource usage and power distributions, Idle Power consumption of the servers
- CPU reservations for the applications

**Output:**
- Predict the tail of sustained power distribution
  - **Time-sharing**: For a server consolidating 'n' applications.
  - **Space-sharing**: For a PDU supplying power to 'm' servers
Sustained power prediction – Time Sharing

$R_{A1} \rightarrow R_{A2} \rightarrow R_{An}$

Individual power profiles (from off-line profiling)

Measurement

Compare

Predicted server power distribution
Time sharing prediction

- Usage of each app exceeds 0.5.
- Therefore add probabilities to (0.5,0.5).
- Sum of utilization of $A_1$ and $A_2$ is less than 1.
- Therefore cannot contribute to sustained power violation.
Power Budgets

- Power demands of a system take different forms (budgets)

  - **Energy Budget:** X joules (days or weeks)

    - Revenue
    - Electricity cost trade-off: limit on the cost expended towards electricity bills

  - **Sustained Power Budget:** X watts, Y seconds

    - Fuses/circuit-breakers which are used to protect the circuit from over current situations

  - **Thermal Budget:** X watts, Y seconds

- A system is typically designed for a thermal capacity that depends on spatial granularity (chip to rack)