Power Management for Data Centers

Karthick Rajamani, Charles Lefurgy, Soraya Ghiasi, Juan C Rubio, Heather Hanson, Tom Keller

{karthick, lefurgy, sghiasi, rubioj, hlhanson, tkeller}@us.ibm.com

IBM Austin Research Labs
Scope of this talk

- Anatomy of a Data Center
  - Power distribution
  - Cooling distribution
- Solutions
  - Opportunities for improving efficiency
  - Typical solutions being employed
  - Control-theoretic approaches
My introduction to power management
Power is a significant and growing problem

Reported to US Congress in August 2007

- 45 Billion KWH in the US for direct power consumption of servers, cooling and auxiliary equipment
- 1.2% of US retail electricity sales, costing $2.7B. World consumption is 2.5 times US.
- World average annual growth rate for server electricity use is about 16% (2000 to 2005)


Dilbert (February 12, 2008)

http://www.unitedmedia.com/comics/dilbert/archive/images/dilbert20183362080212.gif
A Typical Data Center Raised Floor

- Networking equipment (switches)
- Racks (computers, storage, tape)
- Secured Vault
- Network Operating Center
- Fiber Connectivity Terminating on Frame Relay Switch
The Data Center Raised Floor

No two are the same
Data Center Power Distribution
Power Delivery Infrastructure for a Typical Large Data Center (30K sq ft of raised-floor and above)

Several pounds of copper

Power Distribution Unit (PDU)

Uninterruptible Power Supply (UPS) modules

UPS batteries

Transfer panel switch

Diesel generators

Diesel tanks

Power feed substation

© 2008 IBM Corporation
Sample Data Center Energy Consumption Breakdown

- **HVAC - Pumps and Chiller**: 25%
- **HVAC - Air Movement**: 10%
- **UPS Losses**: 5%
- **Lighting**: 1%
- **Computer Loads**: 59%

Fans in the servers already consume 5-20% of the computer load.

Data Center Efficiency Metrics

- Need metrics to indicate energy efficiency of entire facility
  - Metrics do not include quality of IT equipment

- Most commonly used metrics

\[
\text{Power Usage Effectiveness (PUE)} = \frac{\text{Total facility power}}{\text{IT equipment power}}
\]

\[
\text{Data Center Efficiency (DCE)} = \frac{\text{IT equipment power}}{\text{Total facility power}}
\]

Study of 22 sample data centers

Efficiency Loss: Reliability

- Maintaining the uptime of a data center requires the use of redundant components
  - Uninterrupted Power Supplies (UPS)
  - Emergency Power Supply (EPS) – e.g. diesel power generators
  - Redundant configurations to guarantee power and cooling for IT equipment
    - N+1
    - 2(N+1)
Real workload maximum

Nameplate power: 308 W

Stranded power: 56 W

Example: IBM HS20 blade server – nameplate power is 56 W above real workloads.

Result: Available power is stranded and cannot be used

However, real workloads do not use that much power

Data center must wire to nameplate power

Efficiency Loss: Stranded Power
Efficiency Loss: Power Conversion

<table>
<thead>
<tr>
<th>UPS(^{(1)})</th>
<th>Power Distribution(^{(2)})</th>
<th>Power Supply(^{(3,4)})</th>
<th>DC/DC(^{(5)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>88 - 92%</td>
<td>98 - 99%</td>
<td>55 - 90%</td>
<td>78% - 93%</td>
</tr>
</tbody>
</table>

The heat generated from the losses at each step of power conversion requires additional cooling power

(1) [http://hightech.lbl.gov/DCTraining/graphics/ups-efficiency.html](http://hightech.lbl.gov/DCTraining/graphics/ups-efficiency.html)
(3) [http://hightech.lbl.gov/documents/PS/Sample_Server_PSTest.pdf](http://hightech.lbl.gov/documents/PS/Sample_Server_PSTest.pdf)
(5) IBM internal sources
Direct Current Power Distribution

- **Goal:**
  - Reduce unnecessary conversion losses

- **Approach:**
  - Distribute power from the substation to the rack as DC
  - Distribute at a higher voltage than with AC to address voltage drops in transmission lines

- **Challenges:**
  - Requires conductors with very low resistance to reduce losses
  - Potential changes to server equipment

- **Prototype:**
  - Sun, Berkeley Labs and other partners.

(1) [http://hightech.lbl.gov/dc-powering/](http://hightech.lbl.gov/dc-powering/)
AC System Losses Compared to DC

9% measured improvement

2-5% measured improvement

© 2008 IBM Corporation
# Typical Industry Solutions in 2008: Power Consumption

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configurator</td>
<td>Estimate power/thermal load of system before purchase</td>
<td>Sun: Sim Datacenter</td>
</tr>
<tr>
<td>Measurement</td>
<td>Servers with built-in sensors measure power, inlet temperature, outlet temperature.</td>
<td>HP: server power supplies that monitor power</td>
</tr>
<tr>
<td>Power capping</td>
<td>Set power consumption limit for individual servers to meet rack/enclosure constraints.</td>
<td>IBM: Active Energy Manager</td>
</tr>
<tr>
<td>Energy savings</td>
<td>Performance-aware modeling to enable energy-savings modes with minimal impact on application performance.</td>
<td>IBM: POWER6 EnergyScale</td>
</tr>
<tr>
<td>Power off</td>
<td>Turn off servers when idle. Based on user-defined policies (load, time of day, server interrelationships)</td>
<td>Cassatt: Active Response</td>
</tr>
<tr>
<td>Virtualization</td>
<td>Consolidate computing resources for increased efficiency and freeing up idle resources to be shutdown or kept in low-power modes.</td>
<td>VMware: ESX Server</td>
</tr>
<tr>
<td>DC-powered data center</td>
<td>Use DC power for equipment and eliminate AC-DC conversion.</td>
<td>Validus DC Systems</td>
</tr>
<tr>
<td>Component-level control</td>
<td>Enable control of power-performance trade-offs for individual components in the system.</td>
<td>AMD: PowerNow, Intel: Enhanced Speedstep</td>
</tr>
</tbody>
</table>

Solutions shown in example column are representative ones incorporating the specific function/technique. Many of these solutions also provide other functions.

No claim is being made regarding superiority of any example shown over any alternatives.
Data Center Cooling
Cooling Infrastructure for a Typical Large Data Center (30K sq ft of raised-floor and above)

Computer Room Air Conditioning (CRAC) units

Water pumps

Water chillers

Cooling towers
Raised Floor Cooling

- Racks
  - Arranged in a hot-aisle cold-aisle configuration

- Computer room air conditioning (CRAC) units
  - Blower moves air across the raised floor and across cooling element
  - Most common type in large data centers uses chilled water (CW) from facilities plant
  - Adjusts water flow to maintain a constant return temperature
  - Often raised floors have a subset of CRACs that also control humidity in floor
  - Usually on floor, but occasionally on ceiling
  - Located in raised-floor room or right outside of raised-floor room
**How to Save Energy by Best Practices**

**Thermodynamic part of cooling:**
Hot spots (high inlet temperatures) impact CRAC efficiency (~ 1.7% per °F)

**Transport part of cooling:**
Low CRAC utilization impacts CRAC blower efficiency (~3 kW/CRAC)

Source: Hendrik Hamann, IBM
Impact of Raised Floor Air Flow on Server Power

- When there is not enough cold air coming from the perforated tiles
  - Servers fans need to work harder to get cold air across its components.
  - Additionally, increase in rack airflow may cause hot air to overflow into the cold aisle.

- Basic experiment
  - Create enclosed micro-system – rack, 2 perforated tiles and path to CRAC.
  - Linpack running on single server in bottom half of the rack.
  - Adjust air flow from perforated tiles.
  - System power increases as fans ramp up to maintain processor temperature.
Air Flow Management

- Equipment
  - Laid out to create hot and cold aisles

- Tiles
  - Standard tiles are 2’ x 2’
  - Perforated tiles are placed according to amount of air needed for servers
  - Cold aisles usually 2-3 tiles wide
  - Hot aisles usually 2 tiles wide

- Under-floor
  - Floor cavity height sets total cooling capability
  - 3’ height in new data centers

Modeling the Data Center

- Computational Fluid Dynamics (CFD)
  - Useful for initial planning phase and what-if scenarios
  - Input parameters such as rack flows, etc. are very difficult/expensive to come by (garbage in – garbage out problem).
  - Coupled partial differential equations require long-winding CFD calculations

- Measurement-based
  - Find problems in existing data centers
  - Measure the temperature and air flow throughout the data center
  - Highlights differences between actual data center and the ideal data center modeled by CFD
Analysis for Improving Efficiencies, MMT

(a) CFD model results @ 5.5 feet
(b) Experimental data @ 5.5 feet
(c) Difference between model and data

Temperature legend for model and data contours

Source: Hendrik Hamann, IBM
Typical Data Center Raised Floor

Problems!

1. Flow control
2. Rack layout
3. Leak
4. Flow control (overprovisioned)
5. Hot air
6. Hot/cold aisle problem
7. Layout
8. Intermixing
9. Recirculation
10. CRAC layout

Source: Hendrik Hamann, IBM
Chilled Water System

- Two separate water loops
- Chilled water (CW) loop
  - Chiller(s) cool water which is used by CRAC(s) to cool down the air
  - Chilled water usually arrives to the CRACs near 45°F-55°F
- Condensation water loop
  - Usually ends in a cooling tower
  - Needed to remove heat out of the facilities

Sample chilled water circuit
Air-Side and Water-Side Economizers (a.k.a. Free Cooling)

- Air-side Economizer (1)
  - A control algorithm that brings in outside air when it is cooler than the raised floor return air
  - Needs to consider air humidity and particles count
  - One data center showed reduction of ~30% in cooling power

- Water-side Economizer (2)
  - Circulate chilled water (CW) thru an external cooling tower (bypassing the chiller) when outside air is significantly cold
  - Usually suited for climates that have wetbulb temperatures lower than 55°F for 3,000 or more hours per year, and chilled water loops designed for 50°F and above chilled water

- Thermal energy storage (TES) (3)
  - Create chilled water (or even ice) at night.
  - Use to assist in generation of CW during day, reducing overall electricity cost for cooling
  - Reservoir can behave as another chiller, or be part of CW loop

**Typical Industry Solutions in 2008: Cooling**

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hot aisle containment</td>
<td>Close hot aisles to prevent mixing of warm and cool air. Add doors to ends of aisle and ceiling tiles spanning over aisle.</td>
<td>American Power Conversion Corp.</td>
</tr>
<tr>
<td>Sidecar heat exchange</td>
<td>Sidecar heat exchange uses water/refrigerant to optimize hot/cold aisle air flow. Closed systems re-circulate cooled air in the cabinet, preventing mixing with room air.</td>
<td>Emerson Network Power: Liebert XD products</td>
</tr>
<tr>
<td>Air flow regulation</td>
<td>Control inlet/outlet temperature of racks by regulating CRAC airflow. Model relationship between individual CRAC airflow and rack temperature.</td>
<td>HP: Dynamic Smart Cooling</td>
</tr>
<tr>
<td>Cooling economizers</td>
<td>Use cooling tower to produce chilled water when outside air temperature is favorable. Turn off chiller’s compressors.</td>
<td>Wells Fargo &amp; Co. data center in Minneapolis</td>
</tr>
<tr>
<td>Cooling storage</td>
<td>Generate ice or cool fluid with help of external environment, or while energy rates are reduced</td>
<td>IBM ice storage</td>
</tr>
<tr>
<td>Modular data center</td>
<td>Design data center for high-density physical requirements. Data center in a shipping container. Airflow goes rack-to-rack, with heat exchangers in between.</td>
<td>Sun: Project Blackbox</td>
</tr>
</tbody>
</table>

Solutions shown in example column are representative ones incorporating the specific function/technique. Many of these solutions also provide other functions.

No claim is being made regarding superiority of any example shown over any alternatives.
## Typical Industry Solutions in 2008: Other Related

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>On demand</td>
<td>Purchase cycles on demand (avoid owning idle resources)</td>
<td>Amazon: Elastic Compute Cloud (EC2)</td>
</tr>
<tr>
<td>Data center assessment</td>
<td>Measure power/thermal/airflow trends. Use computational fluid dynamics to model data center. Recommend changes to air flow, equipment placement, etc.</td>
<td>IBM, HP, Sun, and many others</td>
</tr>
<tr>
<td>Certification for carbon offsets</td>
<td>3rd party verifies energy reduction of facilities. Trade certificates for money on certificate trading market.</td>
<td>Neuwing Energy Ventures</td>
</tr>
<tr>
<td>Utility rebates</td>
<td>Encourage data centers to use less power (e.g. by using virtualization)</td>
<td>PG&amp;E: offer $0.08/kWh of server power removed</td>
</tr>
</tbody>
</table>

Solutions shown in example column are representative ones incorporating the specific function/technique. Many of these solutions also provide other functions.

No claim is being made regarding superiority of any example shown over any alternatives.
Take Away

- Power distribution and cooling can account for a significant portion of the data center energy consumption.

- Improvements on the power distribution need to be assessed carefully with respect to the data center uptime requirements.

- Data center cooling has two main components: air-side and water-side.
  - Air-side cooling can be improved with static air flow management, automated tools, and air-side economizers.
  - Water-side cooling can be improved with plant modeling and free-cooling (ice generation).
Closed-loop control solutions
Control-theoretic approaches

- Monitor power, thermal, performance metrics continuously
- Adapt to changing workloads and input rates
- Meet power budgets, thermal limits, and SLAs
- Adopt closed-loop control techniques

Source: IBM GTO 2008
Metrics for closed-loop control

- Accuracy
  - Does the system converge or is there steady-state error?

- Stability
  - If the input is bounded, then the output should be bounded.
  - Behavior under non-ideal ideal conditions?

- Settling time
  - Short settling time helps reject disturbances

- Overshoot
**HP Dynamic Smart Cooling**

- Deploy air temperature sensor(s) on each rack
- Collect temperature readings at centralized location
- Apply model to determine setting of CRAC fans to maximize cooling of IT equipment

**Challenges:**
- Difficult to determine impact of particular CRAC(s) on temperature of a given rack – using offline principal component analysis (PCA) and online neural networks to assist logic engine
- Requires CRACs with variable frequency drives (VFD) – not standard in most data centers, but becoming available with time.

![CRAC Regions of Influence](image)

<table>
<thead>
<tr>
<th>Category</th>
<th>Small (air cooling)</th>
<th>Medium (air and chilled water cooling)</th>
<th>Large (air and chilled water cooling)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical size</td>
<td>10K sq ft</td>
<td>30K sq ft</td>
<td>&gt;35K sq ft</td>
</tr>
<tr>
<td>Energy savings (% of cooling costs)</td>
<td>40%</td>
<td>30%</td>
<td>15%</td>
</tr>
<tr>
<td>Estimated MWh saved</td>
<td>5,300</td>
<td>9,100</td>
<td>10,500</td>
</tr>
</tbody>
</table>


In-depth Example – Power Capping

Function Definition: Operate computer system/sub-system/component within specific power consumption threshold irrespective of load or environmental conditions. It is an extension to avoiding systems failure (basic function) by power/current oversubscription, by allowing the power cap to be a programmable quantity that can be altered at run time.

Usage:
- Temporarily reduce budgets of individual systems for brown-out tolerance or power distribution maintenance/re-organization.
- Dynamically manage individual system power consumption to fit in allocated power budgets/costs.
- Ensure consumption of server is restricted to known value and free up unused power for alternate purposes.
- Provide increased component and sub-system protection against oversubscription with increased flexibility for replacement while avoiding huge margins for safety.
Server-level Power Control

Charles Lefurgy, Xiaorui Wang, Malcolm Ware
The problem

- **Server power consumption is not well controlled.**
  - System variance (workload, configuration, process, etc.)
  - Design for worst-case power

- **Results:**
  - Power supplies are significantly over-provisioned
  - Therefore, datacenters provision for power that cannot be used
    - Stranded power
  - High cost, with no benefit in most environments
Our approach

- **Use “better-than-worst-case” design**
  - Example: Intel’s Thermal Design Power (TDP)
  - Power, like temperature, can be controlled

- **Reduce design-time power requirements**
  - Run real workloads at full performance
  - Use smaller, cost-effective power supplies

- **Enforce run-time power constraint with feedback control**
  - Slow system when running power virus
Our contributions

- Control of peak server-level power (to 0.5 W in 1 second)
- Derivation and analysis [see paper]
  - Guaranteed accuracy and stability
- Verified on real hardware
- Better application performance than previous methods
Power measurement

HS20 8843 (Intel Xeon blade)

Measure 12V bulk power
0.1 W precision, 2% error

controller firmware on service processor (Renesas H8 2168)

Measurement/calibration circuit

Sense resistors
Options for power control

- **Open-loop**
  - No measurement of power
  - Chooses fixed speed for a given power budget
  - Based on most power hungry workload

- **Ad-hoc**
  - Measures power and compares to power budget
  - +1/-1 adjustments to processor clock throttle register

- **Proportional Controller ("P control")**
  - Designed using control theory
  - Guaranteed controller performance
Open loop design

- P4MAX workload used as basis for open-loop controller
- Graph shows maximum 1 second power for workload
Proportional controller design

- **Settle to within 0.5 W of desired power in 1 second**
  - Based on BladeCenter power supply requirements

- **Every 64 ms**
  - Compare power to target power
  - Use proportional controller to select desired processor speed
    - 12.5% - 100% in units of 0.1%

- **Clock throttling**
  - Intel processor: 8 settings in units of 12.5% (12.5% - 100%)
  - Use delta-sigma modulation to achieve finer resolution
System diagram and code

// Controller code
error = setpoint - power_measurement;
ideal_throttle = throttle + (1/A) * error;

// Actuator code (First-order delta-sigma modulation)
throttle = truncate(ideal_throttle);
frac = ideal_throttle - throttle;
total_fraction = total_fraction + frac;
if (total_fraction > 1) {
    throttle = throttle + 1;
    total_fraction = total_fraction - 1;
}

// Actuator saturation handling
if (throttle > 7) throttle = 7;
if (throttle < 0) throttle = 0;
Why not use ad-hoc control?

Ad-hoc

Set point = 211.0 W

P Controller

Settles to 216.0 W  5 W Violation
CPU speed: 68.8%

Settles to 211.0 W  No violation
CPU speed: 65.8%
Steady-state error

- **P controller** has no steady-state error \((x=y)\)
- **Ad-hoc controller** has steady-state error
  - Add safety margin of 6.1 W to ad-hoc
Comparison of 3 controllers

- Run each controller with 5 power budgets
- Compare throughput of workloads

Table shows settings used for each controller

<table>
<thead>
<tr>
<th>Power budget</th>
<th>Open-loop processor performance setting</th>
<th>Ad-hoc (with safety margin) set point</th>
<th>P control set point</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 W</td>
<td>75%</td>
<td>238.9 W</td>
<td>245.0 W</td>
</tr>
<tr>
<td>240 W</td>
<td>62.5%</td>
<td>229.1 W</td>
<td>235.2 W</td>
</tr>
<tr>
<td>230 W</td>
<td>62.5%</td>
<td>219.3 W</td>
<td>225.4 W</td>
</tr>
<tr>
<td>220 W</td>
<td>50%</td>
<td>209.5 W</td>
<td>215.6 W</td>
</tr>
<tr>
<td>210 W</td>
<td>37.5%</td>
<td>199.7 W</td>
<td>205.8 W</td>
</tr>
</tbody>
</table>
Application performance summary

- **P controller**
  - 31-82% higher performance than open-loop
  - 1-17% higher performance than ad-hoc
    - Quicker settling time
    - Zero steady state error

1 W = 1.1% performance!
Conclusions

- **Power is a 1st class resource that can be managed.**
  - Power is no longer the accidental result component configuration, manufacturing variation, and workload.

- **Reduce power supply capacity, safely.**
  - Relax design-time constraints, enforce run-time constraints.
  - Install more servers per rack.

- **Power control is a fundamental mechanism for power management in a power-constrained datacenter.**
  - Move power to critical workloads.
Power Capping Demo

- Blade DVFS Capping for tutorial.wmv
Summary

- There is lot of work going on in the industry and academia to address power and cooling issues – it's a very hot topic!

- Much of it has been done in the last few years and we're nowhere near solving all the issues.

- Scope of the problem is vast from thermal failures of individual components to efficiency of data centers and beyond.

- There is no silver bullet – the problem has to be attacked right from better manufacturing technologies to coordinated facilities and IT management.

- Key lies in adaptive solutions that are real-time information driven, and incorporate adequate understanding of the interplay between diverse requirements, workloads and system characteristics.