Power Management for Computer Systems and Datacenters

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
## Overview of Tutorial

### 1. Introduction and Background
- New focus on power management - why, who, what?
- Understanding the problem
  - Diverse requirements
  - WHAT is the problem?
  - Understanding variability

### 2. Power Management Concepts
- Basic solutions
- Advanced solutions
  - Building blocks - sensors and actuators
  - Feedback-driven and model-assisted solution design

### 3. Industry solutions
- Sample solutions
- In-depth solutions

### 4. Datacenter
- Facilities Management
  - Anatomy of a datacenter
  - Improving efficiency
Scope of this tutorial

- Power Management Solutions for Servers and the Data Center
  - Problems
  - Solution concepts, characteristics and context
  - Tools and approaches for developing solutions
  - Brief overview of some industrial solutions

Outside Scope

- Power-efficient microprocessor design, circuit, process technologies
- Power-efficient middleware, software, compiler technologies
- Embedded systems
Servers and Storage Heat Density Trends

Year of First Product Announcement / Year of First Product Shipment

- 2000 Projection for Servers and Storage
- 2005 Projection for Storage
- 2005 Projections for Servers - 2 RU & Greater
- 2005 Projections for Servers - 1 RU, Blade & Custom

Cost of Power and Cooling

Worldwide Server Market

Source: IDC, The Impact of Power and Cooling on Data Center Infrastructure, May 2006
Government and Organizations in Action

- Regulations and standardization of computer energy efficiency (e.g. Energy Star for computers), because of
  - Spiraling energy and cooling costs at data centers
  - Environmental impact of high energy consumption

- Japan, Australia, EU and US EPA working out joint/global regulation efforts

- Benchmarks for power and performance
  - Metrics: GHz, BW, FLOPS → SpecInt, SpecFP, Transactions/sec → Performance/Power
Benchmarking Power-Performance: SPECPower_ssj_2008

- Based on SPECjbb, a java performance benchmark.
  - Generalization to other workloads, blade/cluster environments underway.

- Range of load levels – **not just peak**
  - Self-calibration phases determine peak throughput on system-under-test
  - Benchmark consists of 11 load levels: 100% of peak throughput to idle, in 10% steps
  - Fixed time interval per load level
  - Random arrival times for transactions to mimic realistic variations within each load level

- Primary benchmark metric

\[
\frac{\sum_{\text{idle}} \text{Throughput per level}}{\sum_{\text{idle}} \text{power per level}} \times 100\%
\]

Source: Heather Hanson, IBM
Outline

1. Introduction and Background
   • New focus on power management - why, who, what?
   • Understanding the problem
     ▪ Diverse requirements
     ▪ WHAT is the problem?
     ▪ Understanding variability

2. Power Management Concepts
   • Basic solutions
   • Advanced solutions
     ▪ Building blocks - sensors and actuators
     ▪ Feedback-driven and model-assisted solution design

3. Industry solutions
   • Sample solutions
   • In-depth solutions

4. Datacenter
   • Facilities Management
     ▪ Anatomy of a datacenter
     ▪ Improving efficiency
User Requirements

- Users desire different goals
  - High performance
  - Safe and reliable operation
  - Low operating costs

- But solution for one can be potentially contradictory to a different goal
  - Increasing frequency for higher performance can lead to unsafe operating temperatures
  - Lowering power consumption by operating at lower active states can help safe operation while increasing execution time and potentially total energy costs
Is a single sub-system the main problem? No

Server power budget breakdown for different classes/types.

- Biggest power consumer varies with server class
- Important to understand composition of target class for delivering targeted solutions
- Important to address power consumption/efficiencies in all main subsystems
Estimated power breakdown for two Petaflop supercomputer HPC node designs, each configuration tailored to application class.

- Depending on configuration and usage, processor or memory sub-system turn out to be dominant.
- Power distribution/conversion losses is a significant fraction.

Note 1: Room A/C energy costs are not captured – significant at the HPC data center level.
Note 2: I/O and storage can also be significant power consumers for commercial computing installations.
Power Variability across Real Systems

- Large variation between different applications
  - 2x between LINPACK and idle, Linpack about 20% higher than memory-bound SPEC CPU2000 swim benchmark.
- Smaller power variation between individual blades (8%)

Source: Charles Lefurgy, IBM

Power measured on 33 IBM HS21 blades

Processor clock modulation

Processor voltage/freq scaling
Impact of Process Variation

- 10% power variation in random sample of five ‘identical’ Intel PentiumM processor chips running Linpack.

Source: Juan Rubio, Karthick Rajamani, IBM
Variations from Design

- Memory power specifications for DDR2 parts with identical performance specifications
  - 2X difference in active power and 1.5X difference in idle power

Source: Karthick Rajamani, IBM
Environmental Conditions: Ambient Temperature

~5 C drop in ambient temperature

~5 C drop in CPU temperature

Take Away

- Power management solutions need to target the computer system as a whole with mechanisms to address all major sub-systems. Further, function-oriented design and workload’s usage of a system can create dramatic differences in the distribution of power by sub-systems.

- User requirements impose a diverse set of constraints which can sometimes be contradictory.

- There is increasing variability and even unpredictability of power consumption due to manufacturing technologies, incorporation of circuit-level power reduction techniques, workload behavior and environmental effects.

- Power management solutions need to be flexible and adaptive to accommodate all of the above.
Outline

1. Introduction and Background
   • New focus on power management - why, who, what?
   • Understanding the problem
     ▪ Diverse requirements
     ▪ WHAT is the problem?
     ▪ Understanding variability

2. Power Management Concepts
   • Basic solutions
   • Advanced solutions
     ▪ Building blocks - sensors and actuators
     ▪ Feedback-driven and model-assisted solution design

3. Industry solutions
   • Sample solutions
   • In-depth solutions

4. Datacenter
   • Facilities Management
     ▪ Anatomy of a datacenter
     ▪ Improving efficiency
Basic Solutions – Save Energy

Function Definition: Reduce power consumption of computer system when not in active use.

Conventional Implementation:

- Exploiting architected idle modes of processor entered by executing special code/instructions with interrupts causing exit.
  - Multiple modes with increasing reduction in power consumption usually with increased latency to entry/exit
  - Exploit circuit-level clock gating of significant portions of the chip and more recently voltage-frequency reductions for idle states

- Exploiting device idle modes through logic in controller or device driver e.g.
  - standby state for disks
  - self-refresh for DRAM in embedded systems
Basic Solutions – Avoid System Failure

Function Definition: Maintain continued computer operation in the face of power or cooling emergencies.

Conventional Implementation

- Using redundant components in power distribution and cooling sub-systems
  - N+M solutions.

- Throttling of components under thermal overload.
  - Employing thermal sensors, components are throttled when temperatures exceed pre-set thresholds.

- Fan speed is adjusted based on heat load detected with thermal sensors – to balance acoustics, power and cooling considerations.
Adaptive Solutions: Key to address variability and diverse conditions

Feedback-driven control provides
1. capability to adapt to environment, workload, varying user requirements
2. regulate to desired constraints even with imperfect information

Models provide ability to
1. estimate unmeasured quantities
2. predict impact for change

Feedback-driven and model-assisted control framework

Providing real-time feedback
- Power, temperature, performance (activity), stability
- Weapon against variability and unpredictability

Regulate
- Component states e.g. voltage/freq, DRAM power-down;
- Activity – instruction/request throughput.
- Manage CPU, memory, fans, disk.
- Tune power-performance levels to environment, workload and constraints.
Thermal Sensors

- Thermal sensor key characteristics
  - Accuracy and precision - lower values require higher tolerance margins for thermal control solutions.
  - Accessibility and speed - Impact placement of control and rate of response.

- Ambient measurement sensors
  - Located on-board, inlet temperature, outlet temperature, at the fan e.g. National Semiconductor LM73 on-board sensor with +/-1 deg C accuracy.
  - Relatively slower response time – observing larger thermal constant effects.
  - Standard interfaces for accessing include PECI, I²C, SMBus, and 1-wire

- On-chip/-component sensors
  - Measure temperatures at specific locations on the processor or in specific units
  - Need more rapid response time, feeding faster actuations e.g. clock throttling.
  - Proprietary interfaces with on-chip control and standard interfaces for off-chip control.
Power Measurement Sensors

- **AC power**
  - External components – Intelligent PDU, SmartWatt
  - Intelligent power supplies – PSMI standard
  - Instrumented power supplies

- **DC power**
  - Most laptops – battery discharge rate
  - IBM Active Energy Manager – system power
  - Intel Foxton technology – processor power

- Sensor must suit the application:
  - Access rate (second, ms, us)
  - Accuracy
  - Precision
  - Accessibility (I²C, ethernet, open source driver)


Activity Monitors

- ‘Performance’ Counters
  - Traditionally part of processor performance monitoring unit
  - Can track microarchitecture and system activity of all kinds
  - A fast feedback for activity, have also been shown to serve as potential proxies for power and even thermals

- Resource utilization metrics in the operating system
  - Serve as useful input to resource state scheduling solutions for power reduction

- Application performance metrics
  - Best feedback for assessing power-performance trade-offs
Actuators for Processor Power Control

- Processor pipeline throttling in IBM Power4 and follow-ons.
- Clock throttling in x86 architectures.
- Dynamic voltage and frequency scaling (DVFS) in modern Intel, AMD, IBM POWER6 processors.

Chart shows DFS/DVFS power-performance trade-offs on an IBM LS-20 blade that uses AMD Opteron processor.

- DVFS significantly more efficient than DFS.
- DVFS trade-offs also near-linear at system-level.

Source: Karthick Rajamani, IBM
Memory Systems Power Management

- DRAM power linearly related to memory bandwidth.
  - Request throttling an effective means to limit memory power with linear power-performance trade-offs
- Incorporation of DRAM idle power-down modes can significantly lower their power.
  - Can be implemented in memory controller
  - Can increase performance in power-constrained systems by reducing throttling on active ranks as idle ranks consume less power

Source: Karthick Rajamani, IBM

Combination of 2-channel grouping and Power-down management attacks both active and idle power consumption yielding the best results.
Low-power Enterprise Storage

- Better components
  - Use disks with fewer higher capacity spindles
  - Flash solid-state drives
  - Variable speed disks – with disk power being proportional to rotational speed (squared), tailoring speed to required performance could improve efficiency

- Massive Array of Idle Disks (MAID) – turn on only 25% of array at one time
  - Targets “persistent data” layer between highly available, low latency disk storage and tape
  - May need to tolerate some latency when a disk turns on, keep some always on
  - Improve performance by tuning applications to target only 25% of array

- ‘Virtualization’
  - “Thin provisioning” or over-subscription of storage can drive utilization toward 100%, delaying purchase of additional storage
  - Physical storage is dedicated only when data is actually written
Outline

1. Introduction and Background
   - New focus on power management - why, who, what?
   - Understanding the problem
     ▪ Diverse requirements
     ▪ WHAT is the problem?
     ▪ Understanding variability

2. Power Management Concepts
   - Basic solutions
   - Advanced solutions
     ▪ Building blocks - sensors and actuators
     ▪ Feedback-driven and model-assisted solution design

3. Industry solutions
   - Sample solutions
   - In-depth solutions

4. Datacenter
   - Facilities Management
     ▪ Anatomy of a datacenter
     ▪ Improving efficiency
Adaptive Power Management Demo

- **Power6 blade prototype power capping demo**
  - Demonstrates ability to adapt at runtime to workload changes and user input (power cap limit) while maintaining blade power below specified value.

- **Power6 blade power savings demo**
  - Demonstrates ability to adapt to load characteristics to provide increased power savings by matching power-performance level to load-demand.
    - Adapting to load level
    - Adapting to load type
Advanced Solutions – Performance Maximization

Usage:

When cooling or power resources are shared by multiple entities, dynamic partitioning of the resource among the sharing entities can increase utilization of the shared resource enabling higher performance. We term this technique as *power shifting*.

Figure shows an example of CPU-memory *power shifting*.

- Points show execution intervals of many workloads with no limit on power budget.

- ‘Static’ dotted rectangle encloses unthrottled intervals for a budget of 40W partitioned statically – 27W CPU, 13W Memory.

- ‘Dynamic’ dashed triangle encloses unthrottled intervals for power shifting with 40W budget
  - Better performance as a much larger set of intervals run unthrottled.
Power Shifting between Processor and Memory

Graph shows the increase in execution time for constrained power budget with SPECCPU2000 workloads

- Proportional-Last-Interval – particular power shifting algorithm
- Static – common budget allocation – using average consumption across workloads.

Models

\[
P_{\text{cpu}} = DPC \cdot C1 + P_{\text{stby}} \cdot \text{cpu}
\]
\[
P_{\text{mem}} = BW \cdot M1 + P_{\text{stby}} \cdot \text{mem}
\]

Re-adjust budgets based on feedback

\[
P_{\text{dynamic}} = P_{\text{budget}} - P_{\text{cpu, stby}} - P_{\text{mem, stby}}
\]
\[
P_{\text{est}} = DPC_0 \cdot C1 + BW_0 \cdot M1
\]

\[
\text{Budget}_{\text{cpu}} = DP_{0} \cdot C1 \cdot P_{\text{est}} \cdot P_{\text{dynamic}} + P_{\text{cpu, stby}}
\]
\[
\text{Budget}_{\text{mem}} = BW_0 \cdot M1 \cdot P_{\text{est}} \cdot P_{\text{dynamic}} + P_{\text{mem, stby}}
\]

New budgets determine new throughput and bandwidth limits.

Take Away

- Advanced functions extend basic power management capabilities
  - Providing adaptive solutions for tackling variability in systems, environment and requirements, and
  - Enabling dynamic power-performance trade-offs.

- They can be implemented by methodologies incorporating
  - Targeted sensors and actuators.
  - Feedback-control systems.
  - Model-assisted frameworks.
Virtualization – Opportunities and Challenges for Power Reduction

- Different studies have shown significant under-utilization of compute resources in large data centers.

- Virtualization enables multiple low-utilization OS images to occupy a single physical server.
  - Multi-core processors with virtualization support and large SMP systems provide a growing infrastructure which facilitates virtualization-based consolidation.

- The common expectation is
  - A net reduction in energy costs.
  - Lower infrastructure costs for power delivery and cooling.

- However:
  - Processor capacity is not the only resource workloads need – memory, I/O..
  - Workloads on partitions sharing a system might have different power-performance needs.

- Isolating and understanding the characteristics and consequent power management needs for each partition is non-trivial, requiring
  - Additional instrumentation and monitoring capabilities.
  - Augmenting VM managers for coordinating, facilitating or managing power management functions
Managing Server Power in a Data Center

## Outline

### 1. Introduction and Background
- New focus on power management - why, who, what?
- Understanding the problem
  - Diverse requirements
  - WHAT is the problem?
  - Understanding variability

### 2. Power Management Concepts
- Basic solutions
- Advanced solutions
  - Building blocks - sensors and actuators
  - Feedback-driven and model-assisted solution design

### 3. Industry solutions
- Sample solutions
- In-depth solutions

### 4. Datacenter
- Facilities Management
  - Anatomy of a datacenter
  - Improving efficiency
Computer Industry Response

Components:
- Thermal sensors
- Idle modes
- DVFS
- Liquid cooling

Server:
- Power measurement
- Efficient power supplies
- Power capping
- Adaptive power savings

Rack:
- Hot-aisle containment
- In-rack cooling
- Cluster power trending
- Virtualization

Data Center:
- Efficiency improvements
- DC-powered data center
- Free cooling
- Dynamic Smart Cooling
- Mobile data center
In-depth Examples

1. Active Energy Manager – for cluster-wide system monitoring and policy management.

2. EnergyScale – for power management of POWER6 systems.
IBM Systems Director Active Energy Manager

- Provides a single view of the actual power usage across multiple platforms.
- Measures, trends, and controls energy usage of all managed systems.
Rack-mount Server Management

Measure and Trend: power, thermals
Control: power cap, energy savings
## Intelligent Power Distribution Unit (iPDU)

### Web interface

**Model:** IBM DPI C19 PDU+ (39M2819)

**IBM DPI**

**Input**
- Voltage (phase A/B/C): 203.8/201.9/201.8V
- Frequency: 60.0Hz

**Output**
- Voltage (phase A/B/C): 203.8/201.9/201.8V
- Frequency: 60.0Hz

#### Current Power

<table>
<thead>
<tr>
<th>J1</th>
<th>J2</th>
<th>J3</th>
<th>J4</th>
<th>J5</th>
<th>J6</th>
<th>1A</th>
<th>3A</th>
<th>5A</th>
</tr>
</thead>
<tbody>
<tr>
<td>203.6</td>
<td>203.6</td>
<td>201.9</td>
<td>201.6</td>
<td>201.6</td>
<td>201.6</td>
<td>Output Voltage(V)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.3</td>
<td>3.8</td>
<td>2.3</td>
<td>Output Current(A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>837</td>
<td>113</td>
<td>111</td>
<td>Min</td>
<td>752</td>
<td>960</td>
<td>Output Power(W)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>837</td>
<td>112</td>
<td>111</td>
<td>622</td>
<td>219</td>
<td>646</td>
<td>PDU Watt Hour Usage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1153</td>
<td>3413</td>
<td>4504</td>
<td>338</td>
<td>3577</td>
<td>2527</td>
<td>Cumulative KW-hrs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Cumulative use**

- Note: A reading of "Min" indicates one of several possible scenarios: nothing is plugged into the load group, the load group device is powered off, a very low power device is plugged in, a circuit breaker has tripped, or a device is malfunctioning. Further investigation is recommended if a low power reading was not expected.

**User defined receptacle names**

- J1: DPM upper 2+3
- J2: DPM upper 2+3
- J3: DPM upper 2+3
- J4: DPM upper 2+3
- J5: DPM upper 2+3
- J6: DPM upper 2+3
- 1A: DPM lower 2+4
- 3A: DPM lower 2+4
- 5A: DPM lower 2+4

**Refresh Rate:** 10 sec
EnergyScale Elements and Functionality

- Thermal / Power measurement
- System health monitoring/maintenance
- Power/thermal capping
- Power saving
- Performance-aware power management
# EnergyScale Sensors

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Feedback On</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>On chip digital temperature sensors (internal use)</td>
<td>POWER6 chip temperature</td>
<td>24 digital temp. sensitive ring oscillators – 8 per core, 8 in nest</td>
</tr>
<tr>
<td>On chip analog temperature sensors</td>
<td>POWER6 chip temperature</td>
<td>3 metal thermistors – 1 per core, 1 in nest – analog</td>
</tr>
<tr>
<td>Critical path monitor (internal use)</td>
<td>POWER6 operational stability</td>
<td>24 sensors providing real-time timing margin feedback</td>
</tr>
<tr>
<td>Voltage Regulation Module (VRM)</td>
<td>Component/voltage-rail power</td>
<td>Voltage and current provided for each voltage-rail in the system</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Temperature of VRM components</td>
</tr>
<tr>
<td>On-board power measurement sensors</td>
<td>System and component power</td>
<td>Calibrated sensors, accurate, real-time feedback</td>
</tr>
<tr>
<td>Discrete temperature sensor</td>
<td>System level ambient temperature</td>
<td>Reports temperature of air seen by components</td>
</tr>
<tr>
<td>Dedicated processor activity counters</td>
<td>Core activity and performance</td>
<td>Per-core activity information</td>
</tr>
<tr>
<td>Dedicated memory controller activity counters</td>
<td>DRAM usage</td>
<td>Per-controller activity and power management information</td>
</tr>
</tbody>
</table>
Temperature Sensors on POWER6

- Digital thermal sensors
  - Quick response time
  - Placed in hot-spots identified during simulation and early part characterization

- Metal thermistors
  - Very accurate
  - Used in previous designs
  - Large area for a 65nm chip

Source: M. Floyd et al., IBM J. R&D, November 2007
## EnergyScale Actuators

<table>
<thead>
<tr>
<th>Actuator/Mode</th>
<th>Controlled By</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Voltage and Frequency</td>
<td>Service processors</td>
<td>Variable frequency oscillator control, voltage control for array and logic</td>
</tr>
<tr>
<td>Scaling (DVFS)</td>
<td></td>
<td>domains, system-wide</td>
</tr>
<tr>
<td>Processor Pipeline Throttling</td>
<td>On-chip thermal protection</td>
<td>6 different modes of pipeline throttling, manageable on a per-core basis</td>
</tr>
<tr>
<td></td>
<td>circuitry, service</td>
<td></td>
</tr>
<tr>
<td></td>
<td>processors</td>
<td></td>
</tr>
<tr>
<td>Processor Standby Modes</td>
<td>OS and Hypervisor</td>
<td>Nap/Active</td>
</tr>
<tr>
<td>Memory Throttling</td>
<td>Service processors</td>
<td>4 different modes of activity/request throttling</td>
</tr>
<tr>
<td>Memory Standby Modes</td>
<td>Memory controller</td>
<td>Per-rank power-down mode enable – FSP/dedicated microcontroller</td>
</tr>
<tr>
<td>Fan Speeds</td>
<td>Service processors</td>
<td>Based on ambient temperature</td>
</tr>
</tbody>
</table>

Service Processors: FSP and EnergyScale Controller on slide 70
## Water cooling

<table>
<thead>
<tr>
<th></th>
<th>Thermal conductivity [W/(m*K)]</th>
<th>Volumetric heat capacity [kJ/(m³*K)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air</td>
<td>0.0245</td>
<td>1.27</td>
</tr>
<tr>
<td>H₂O</td>
<td>0.6</td>
<td>4176</td>
</tr>
</tbody>
</table>

Water-cooled chips

NCAR Bluefire Supercomputer using IBM p575 hydrocluster. Images courtesy of UCAR maintained Bluefire web gallery.
## 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</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

- There is significant effort from industry on power and temperature management of computer systems.

- The current intent is not only to make individual components more energy efficient, but:
  - Enhance functionality in certain components
  - Integrate them into system-level power management solutions
# Outline

1. Introduction and Background
   - New focus on power management - why, who, what?
   - Understanding the problem
     - Diverse requirements
     - WHAT is the problem?
     - Understanding variability

2. Power Management Concepts
   - Basic solutions
   - Advanced solutions
     - Building blocks - sensors and actuators
     - Feedback-driven and model-assisted solution design

3. Industry solutions
   - Sample solutions
   - In-depth solutions

4. Datacenter
   - Facilities Management
     - Anatomy of a datacenter
     - Improving efficiency
The Data Center Raised Floor

No two are the same
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
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
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
Sample Data Center Energy Consumption Breakdown

- HVAC - Air Movement: 25%
- UPS Losses: 10%
- HVAC - Pumps and Chiller: 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

Minimum PUE

Fallacy: Cooling power = IT power.

Reality: Data center efficiency varies.

Outline

1. Introduction and Background
   • New focus on power management - why, who, what?
   • Understanding the problem
     ▪ Diverse requirements
     ▪ WHAT is the problem?
     ▪ Understanding variability

2. Power Management Concepts
   • Basic solutions
   • Advanced solutions
     ▪ Building blocks - sensors and actuators
     ▪ Feedback-driven and model-assisted solution design

3. Industry solutions
   • Sample solutions
   • In-depth solutions

4. Datacenter
   • Facilities Management
     ▪ Anatomy of a datacenter
     ▪ Improving efficiency
Data Center Power Distribution
Maintainability and Availability vs. Energy Efficiency

- Distributing power across a large area requires a decentralized/hierarchical architecture
  - Power transformers (13.2 kV to 480 V or 600 V)
  - Power distribution units (PDU)
  - And a few miles of copper

- 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 (N+1, 2N, 2(N+1), …) to guarantee power and cooling for IT equipment

- Both introduce energy losses
Substation usually receives power from 2 or more points

Substation provides 2 redundant feeds to data center

If utility power is lost:
1. UPS can support critical load for at least 15 minutes
2. EPS generators can be online in 10-20 seconds
3. The transfer switch will switch to EPS generators for power
4. EPS generators are diesel fueled and can run for an extended period of time
Data Center Power Conversion Efficiencies

<table>
<thead>
<tr>
<th>Component</th>
<th>Efficiency Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPS(^{(1)})</td>
<td>88 - 92%</td>
</tr>
<tr>
<td>Power Distribution(^{(2)})</td>
<td>98 - 99%</td>
</tr>
<tr>
<td>Power Supply(^{(3,4)})</td>
<td>55 - 90%</td>
</tr>
<tr>
<td>DC/DC(^{(5)})</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
Data center must wire to nameplate power
However, real workloads do not use that much power
Result: available power is **stranded** and cannot be used
Example: IBM HS20 blade server – nameplate power is 56 W above real workloads.

Source: Lefurgy, IBM
Data Center 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

480 VAC Bulk Power Supply

480 VAC Bulk Power Supply

380 VDC

Loads using Legacy Voltages

Loads using Silicon Voltages

Server

PSU

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM

VRM
## 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
Raised Floor Cooling

**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
  - Air pressure drops in front of the machine
  - Servers fans need to work harder to get cold air across its components.
  - Additionally, hot air can create a high pressure area, and 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, other servers idle.
  - Adjust air flow from perforated tiles

![Graph showing impact of raised floor air flow on server power](image_url)
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

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
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.

<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>

(3) http://www.hp.com/hpinfo/globalcitizenship/gcreport/energy/casestudies.html
Commercial Liquid Cooling Solutions for Racks

- **Purpose:**
  - Localized heat removal – attack it before it reaches the rest of the computer room
  - Allows us to install equipment with high power densities in a room not designed for it

- **Implementation:**
  - Self-contained air cooling solution (water or glycol for taking heat from the air)
  - Air movement

- **Types:**
  - Enclosures – create cool microclimate for selected ‘problem’ equipment
  - Sidecar heat exchanger – to address rack-level hotspots without increasing HVAC load

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.
Next generation solutions – Component to Datacenter

- Integrated fan control for thermal management
- Dynamic and deterministic performance boost
- Power-aware workload management
- Power-aware virtualization for resource and server consolidation
- Combined firmware-OS power management
- Integrated IT and facilities management
- Datacenter modeling and optimization
- Partition-level power management
- Power-aware workload management
- Power-aware virtualization for resource and server consolidation
- Combined firmware-OS power management

- On-chip power measurement
- Process Technologies
- Enhanced processor idle states
- Hardware Acceleration
- Fine-grained, faster clock scaling
- Power-aware micro-architecture
- Power-aware memory power modes
- Enhanced memory power modes
- Power-aware workload management
Conclusion

- There is lot of work going on in the industry and academia to address power and cooling issues – its 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.