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

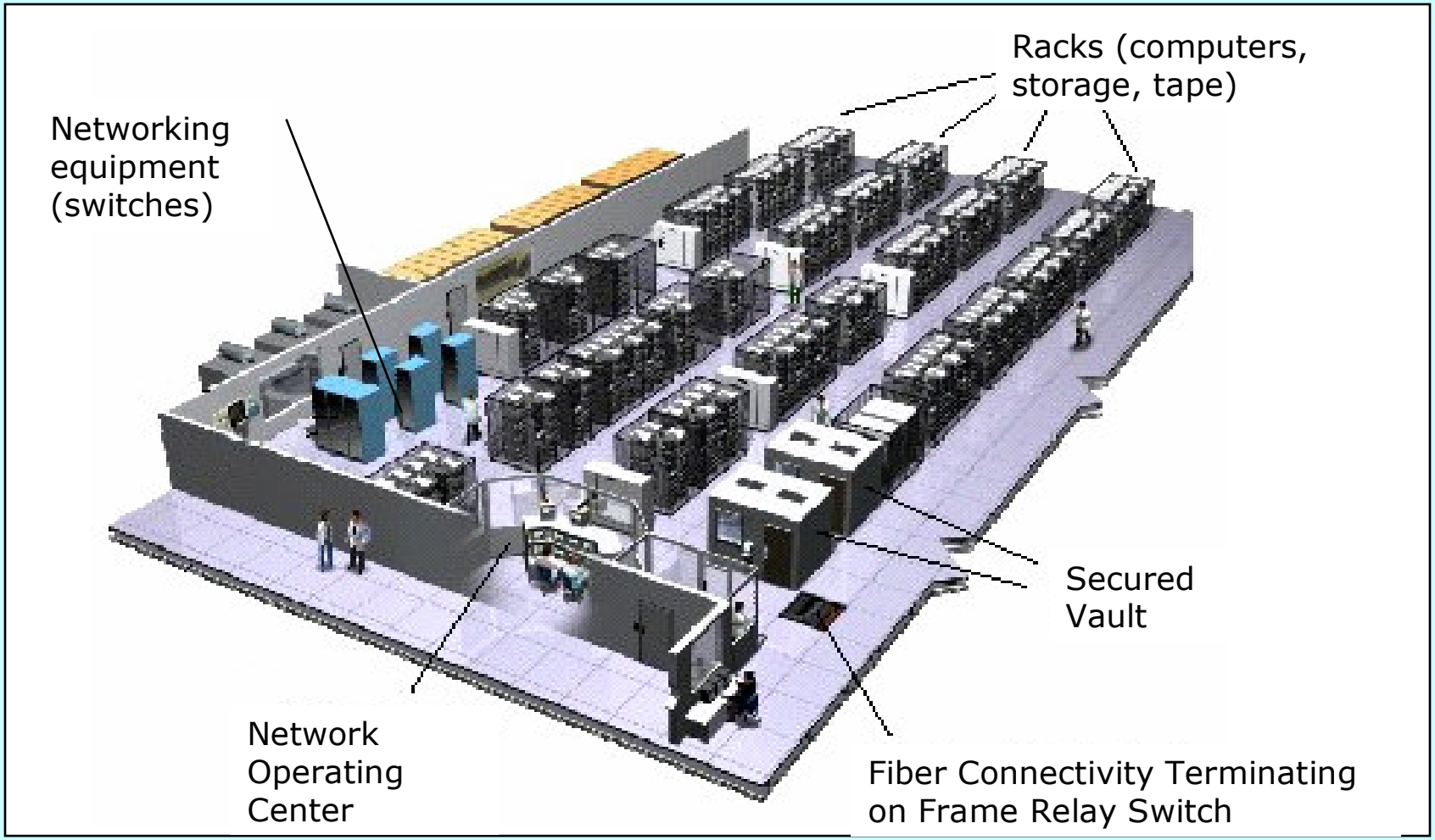
- 45 Billion KWH in the US for direct power consumption of servers, cooling and auxiliary equipment
- 1.2% of US retail electricity sales, costing $2.7B. World consumption is 2.5 times US.
- World average annual growth rate for server electricity use is about 16% (2000 to 2005)

Source: J Koomey, LBNL, *Estimating Total Power Consumption of Servers in the US and the World*, 2007

Dilbert (February 12, 2008)

http://www.unitedmedia.com/comics/dilbert/archive/images/dilbert20183362080212.gif

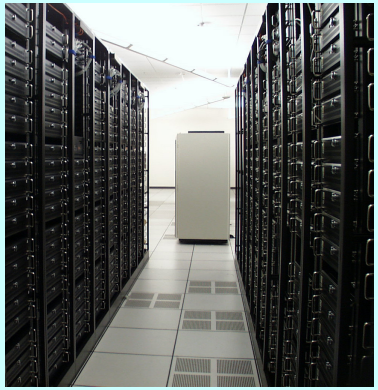
A Typical Data Center Raised Floor

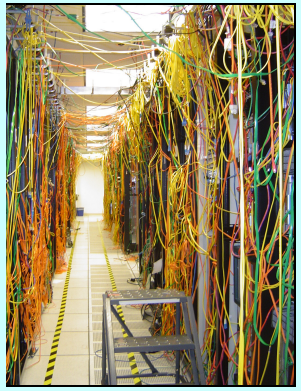


The Data Center Raised Floor







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)

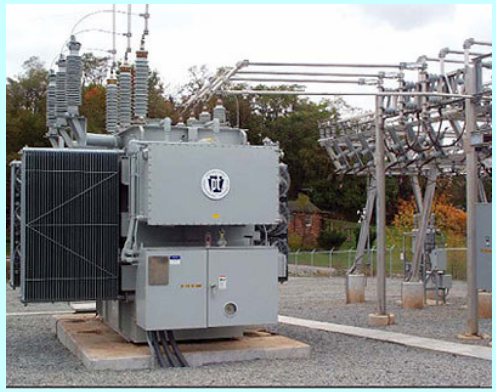
Uninterruptible Power Supply (UPS) modules

UPS batteries

Transfer panel switch

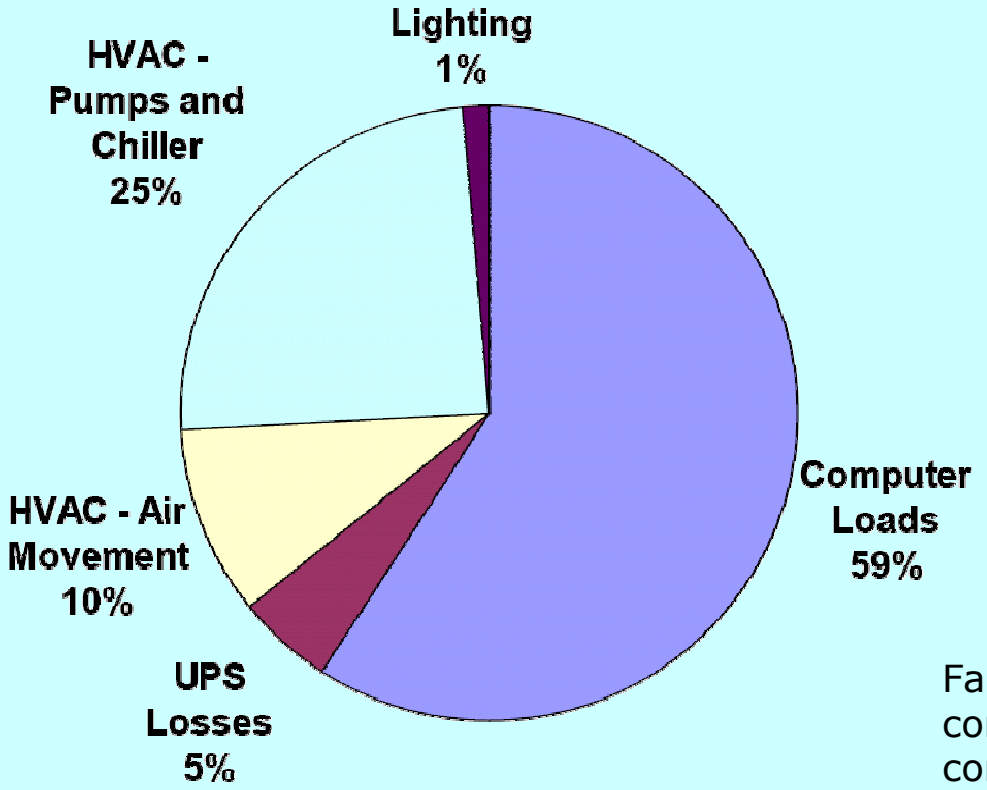
Diesel generators

Diesel tanks



Power feed substation

Sample Data Center Energy Consumption Breakdown



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

Reference: Tschudi, et al., "Data Centers and Energy Use – Let's Look at the Data", ACEEE 2003

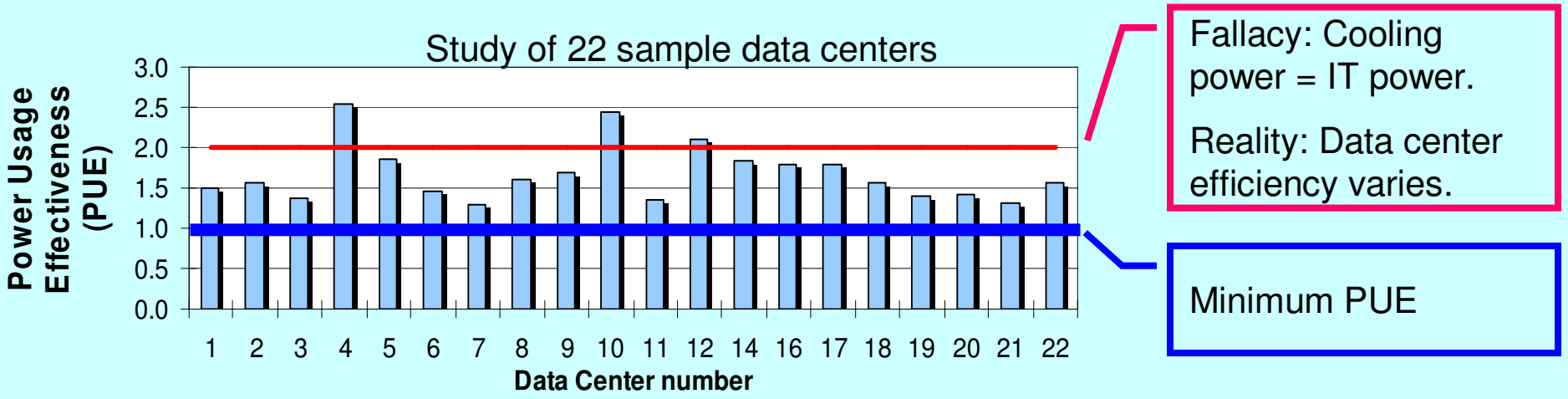
Data Center Efficiency Metrics

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

- Most commonly used metrics

$$Power\ Usage\ Effectiveness\,(PUE) = \frac{Total\ facility\ power}{IT\ equipment\ power}$$

$$Data\ Center\ Efficiency\,(DCE) = \frac{IT\ equipment\ power}{Total\ facility\ power}$$



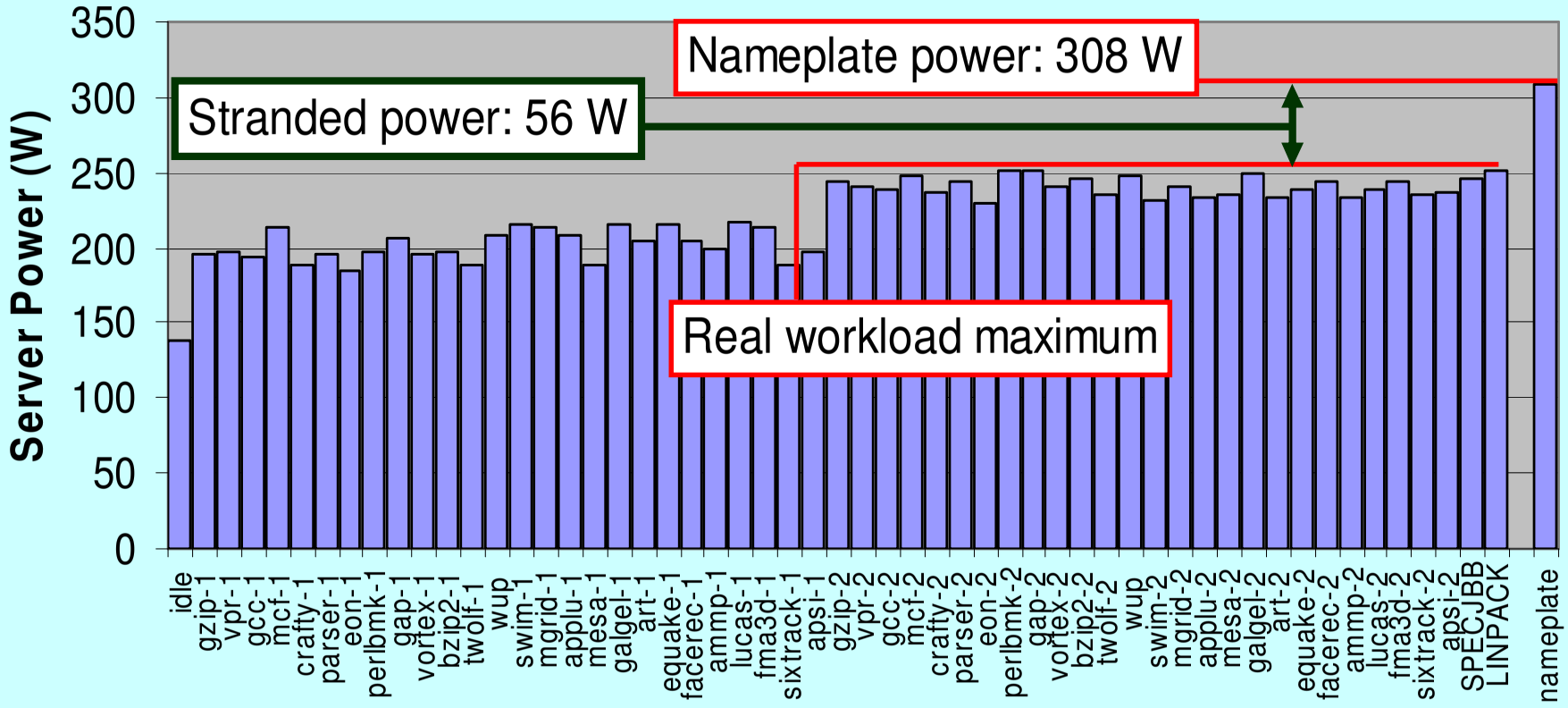
Reference: Tschudi, et al., "Measuring and Managing Data Center Energy Use", 2006

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)

Efficiency Loss: Stranded Power

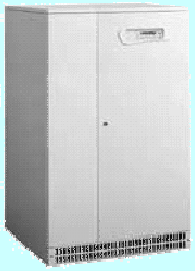
- Data center must wire to nameplate power
- However, real workloads do not use that much power
- Result: available power is ***stranded*** and cannot be used
- Example: IBM HS20 blade server – nameplate power is 56 W above real workloads.

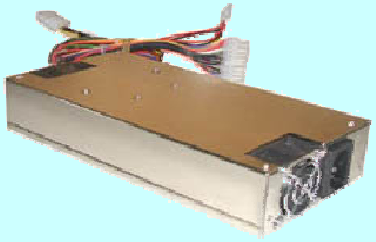


Source: Lefurgy, IBM

Efficiency Loss: Power Conversion

UPS[1]	Power Distribution[2]	Power Supply[3,4]	DC/DC[5]
88 - 92%	98 - 99%	55 - 90%	78% - 93%






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

(2) N. Rasmussen. "Electrical Efficiency Modeling for Data Centers", APC White Paper, 2007

(3) http://hightech.lbl.gov/documents/PS/Sample_Server_PSTest.pdf

(4) "ENERGY STAR® Server Specification Discussion Document", October 31, 2007.

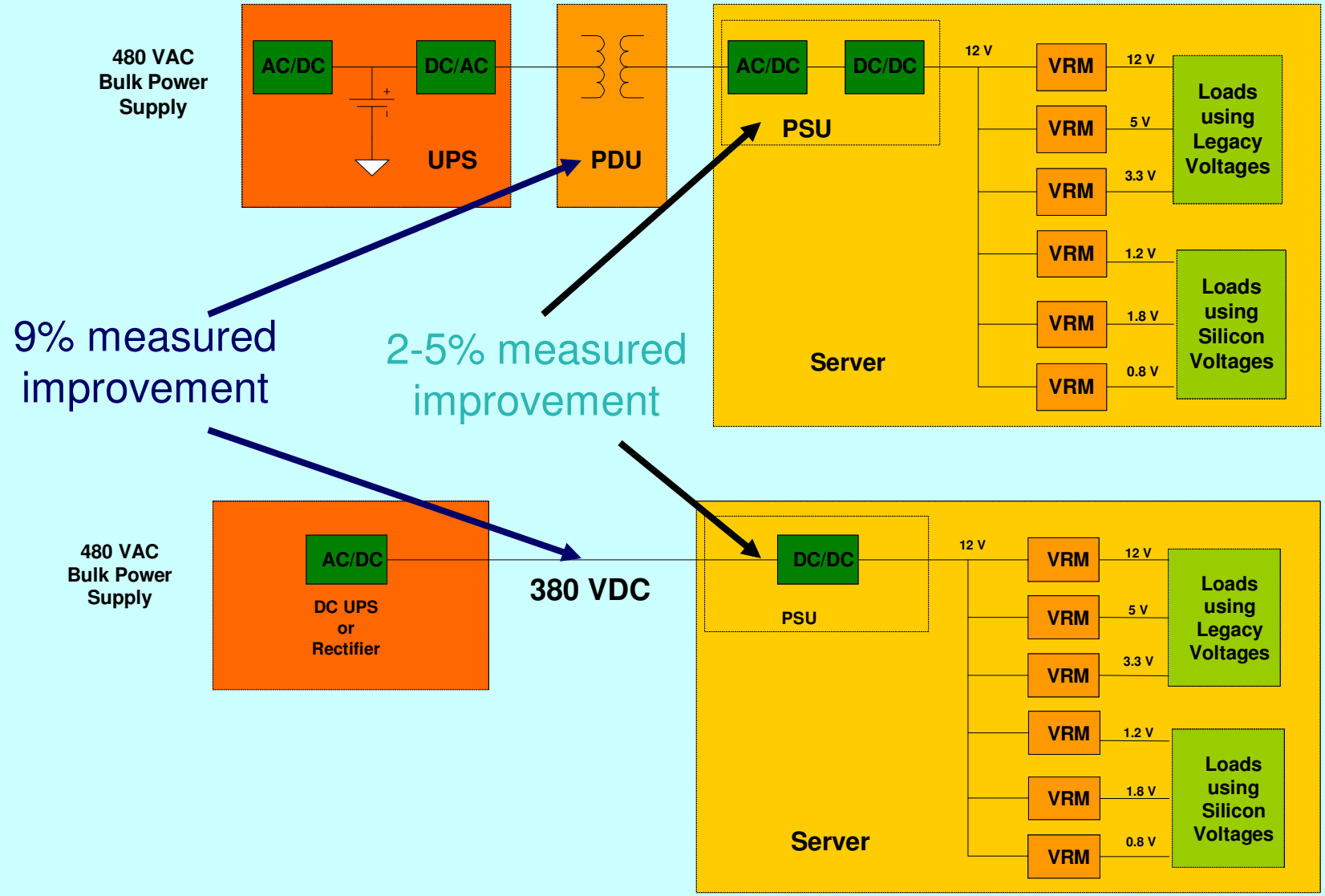
(5) IBM internal sources

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/

AC System Losses Compared to DC



Typical Industry Solutions in 2008: Power Consumption

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

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)

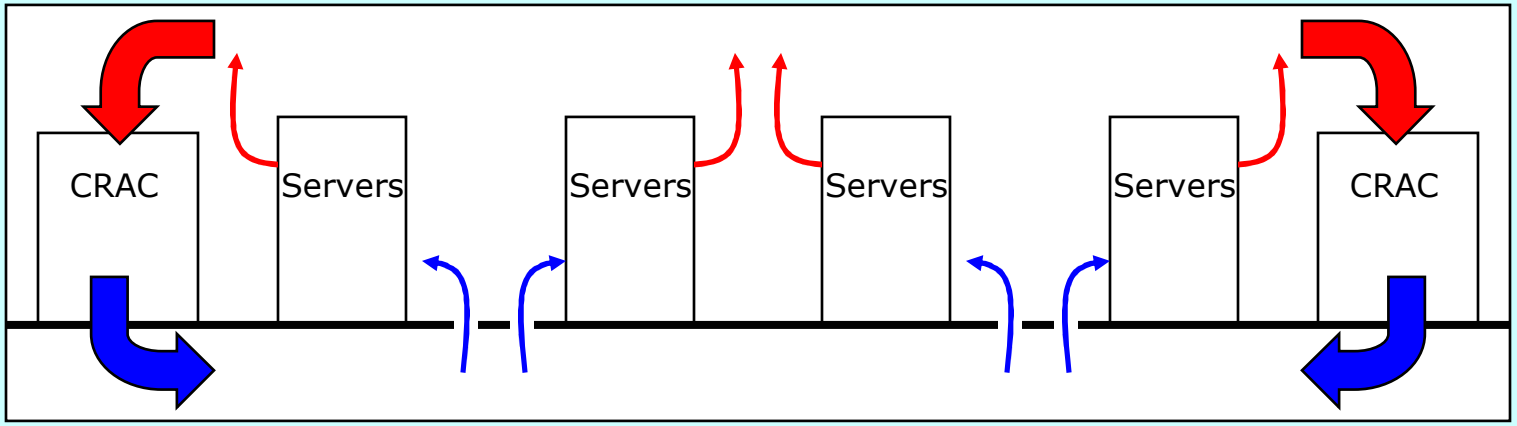
Computer Room Air Conditioning (CRAC) units

Water pumps

Water chillers

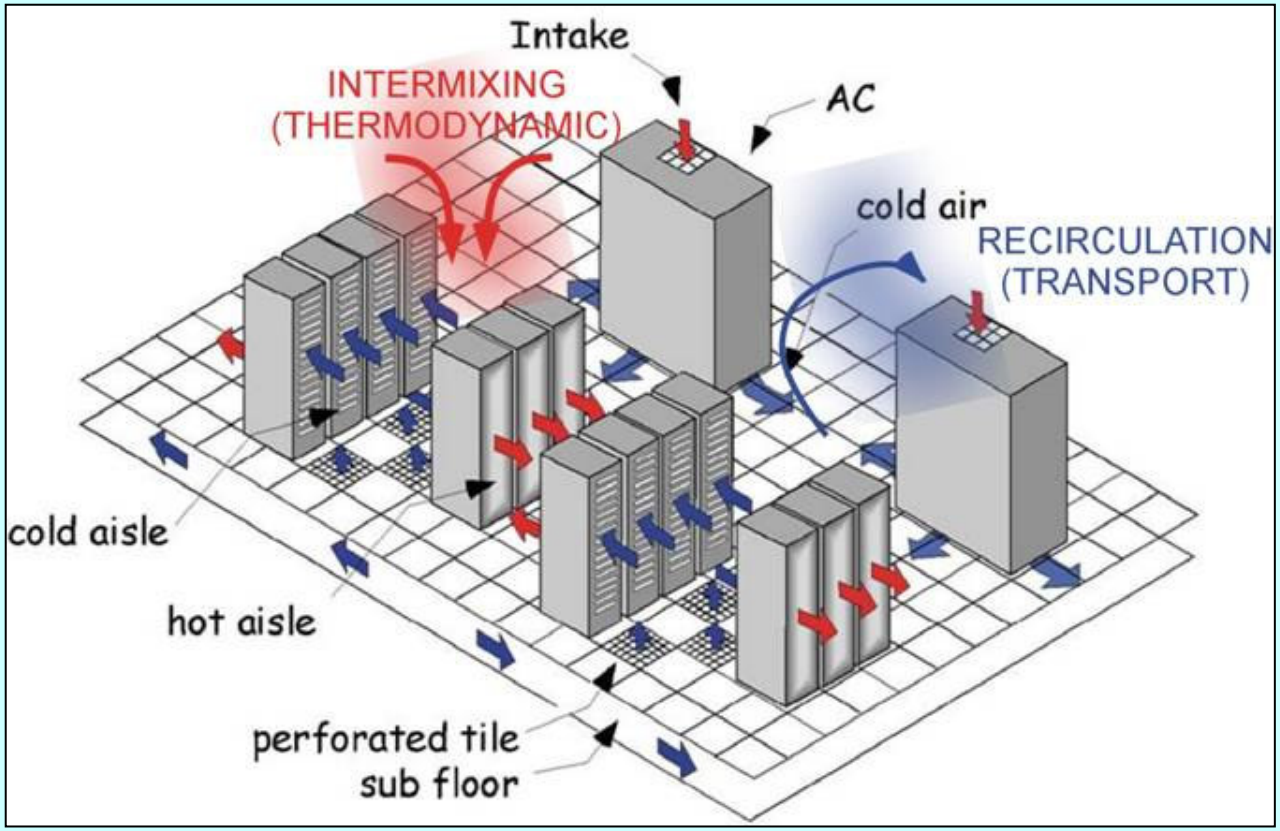
Cooling towers

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 ºF)

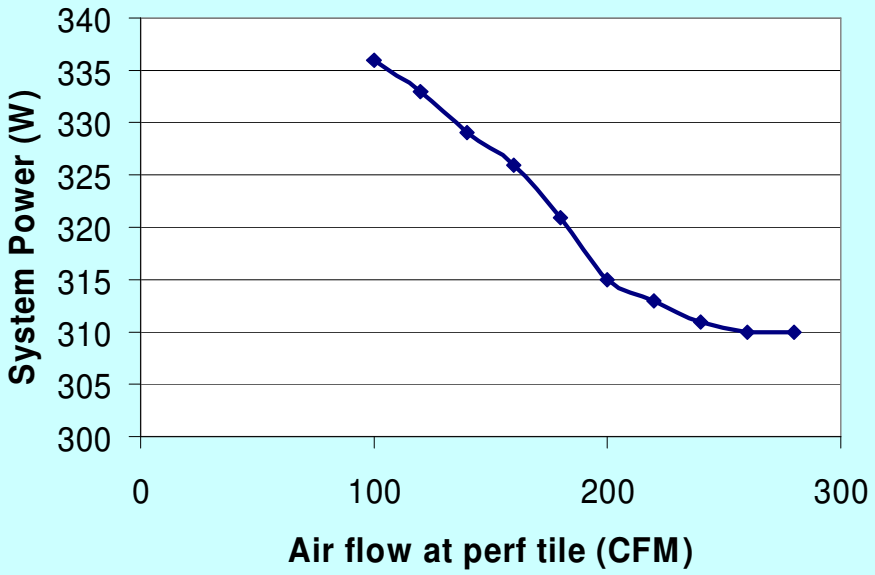
Transport part of cooling:
Low CRAC utilization impacts CRAC blower efficiency (~3 kW/CRAC)

Source: Hendrik Hamann, IBM

Impact of Raised Floor Air Flow on Server Power

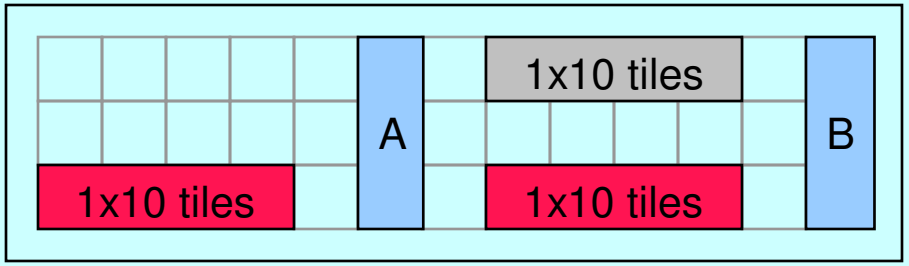
- When there is not enough cold air coming from the perforated tiles
  - Servers fans need to work harder to get cold air across its components.
  - Additionally, increase in rack airflow may cause hot air to overflow into the cold aisle.
- Basic experiment
  - Create enclosed micro-system – rack, 2 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.

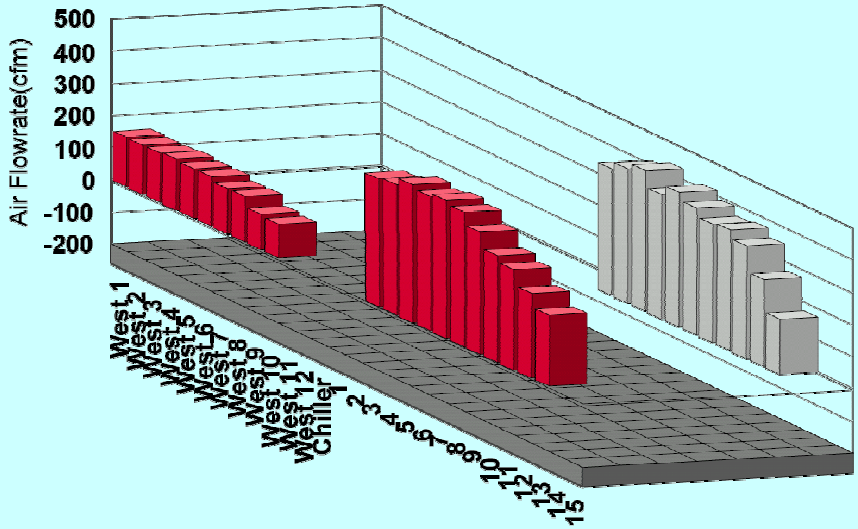
Source: J. Rubio, IBM



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





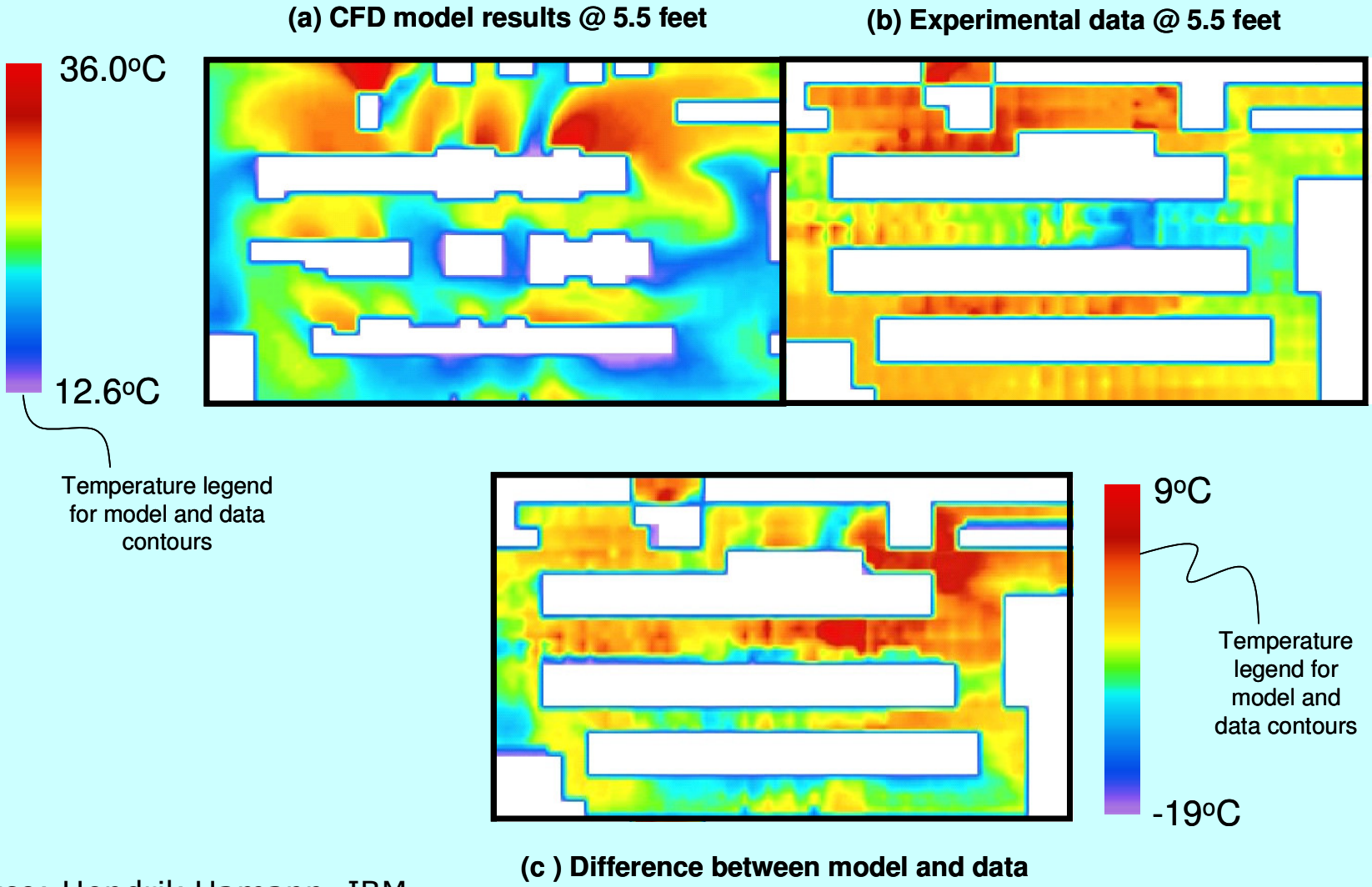
Reference: R. Schmidt, "Data Center Airflow: A Predictive Model", 2000

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

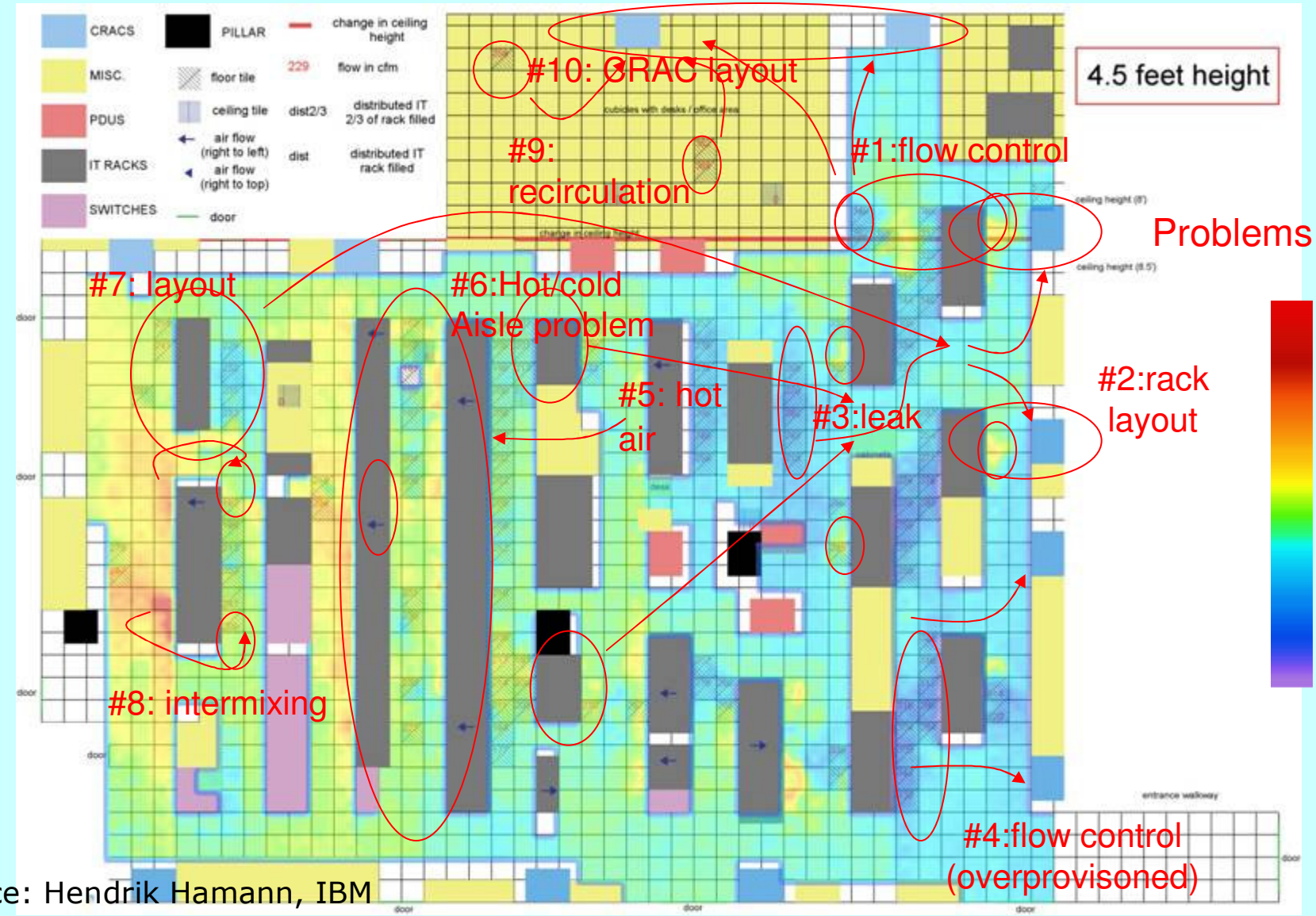
- Measurement-based
  - Find problems in existing data centers
  - Measure the temperature and air flow throughout the data center
  - Highlights differences between actual data center and the ideal data center modeled by CFD

Analysis for Improving Efficiencies, MMT



Source: Hendrik Hamann, IBM

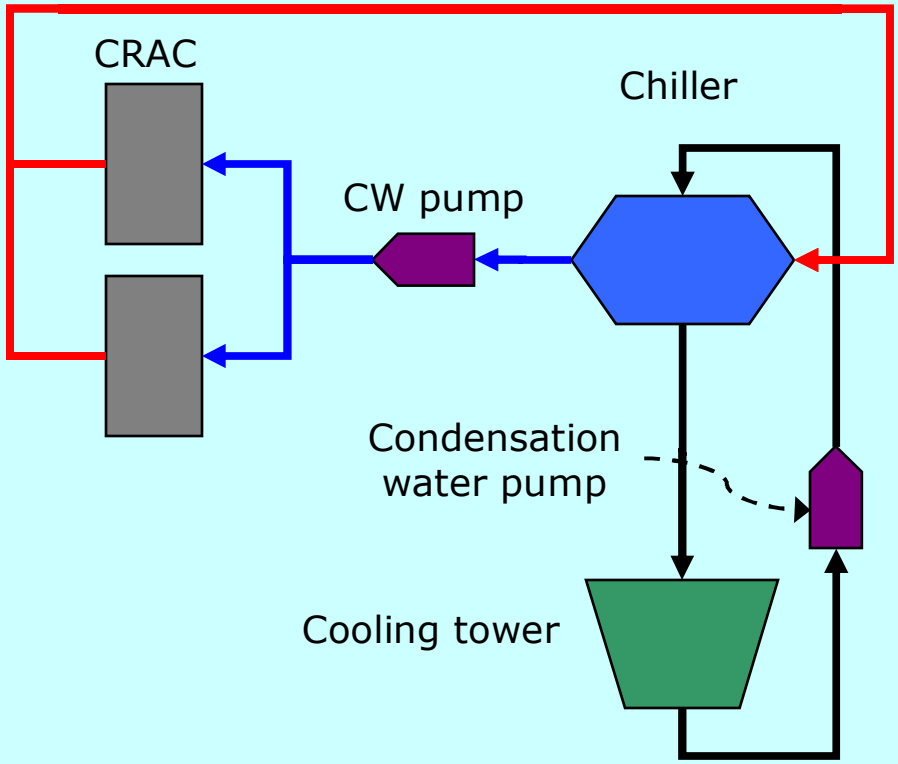
Typical Data Center Raised Floor



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

(1) A. Shehabi, et al. "Data Center Economizer Contamination and Humidity Study", March 13, 2007. http://hightech.lbl.gov/documents/DATA_CENTERS/EconomizerDemoReportMarch13.pdf

(2) Pacific Gas & Electric, "High Performance Data Centers: A Design Guidelines Sourcebook", January 2006. http://hightech.lbl.gov/documents/DATA_CENTERS/06_DataCenters-PGE.pdf

(3) "Cool Thermal Energy Storage", ASHRAE Journal, September 2006

Typical Industry Solutions in 2008: Cooling

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

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

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

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.

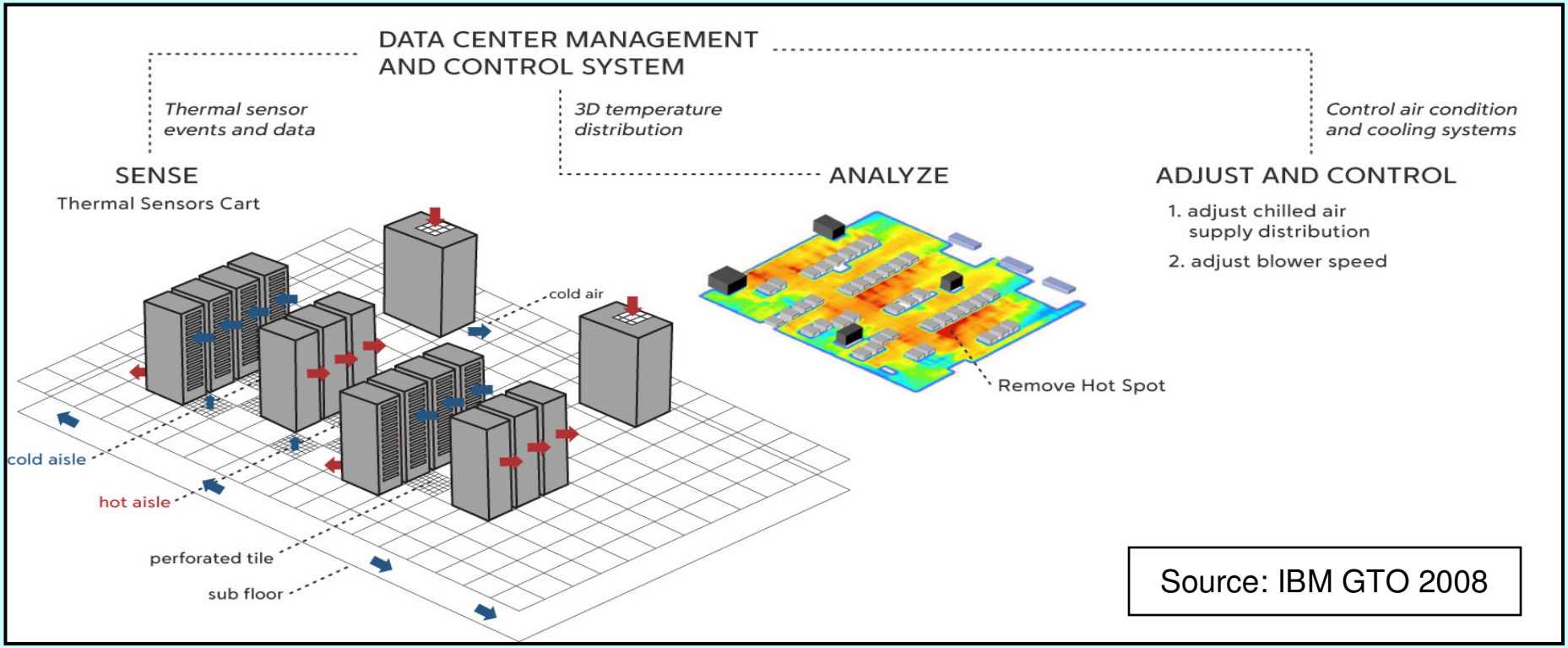
Take Away

- 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

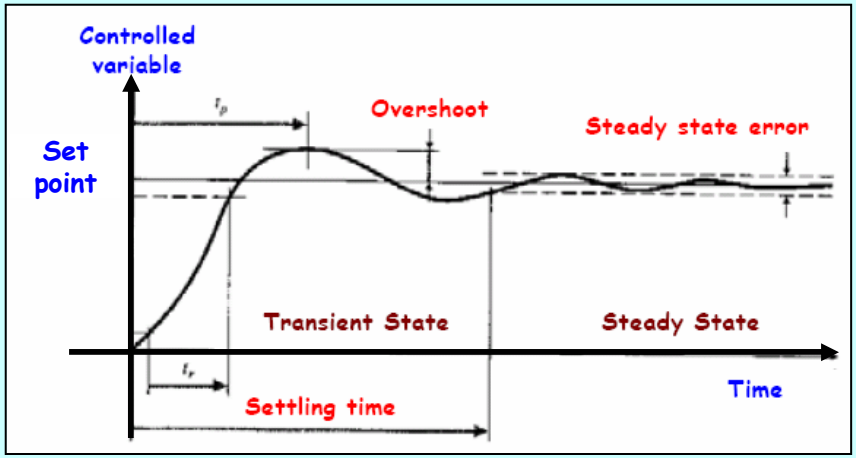
- Monitor power, thermal, performance metrics continuously
- Adapt to changing workloads and input rates
- Meet power budgets, thermal limits, and SLAs
- Adopt closed-loop control techniques



Source: IBM GTO 2008

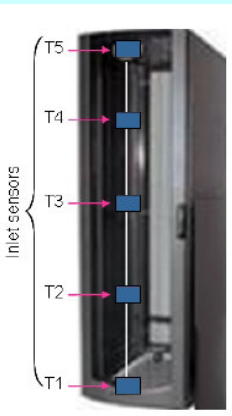
Metrics for closed-loop control

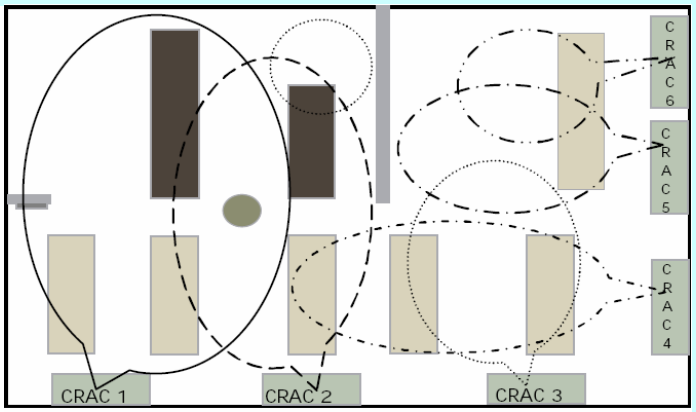
- Accuracy
  - Does the system converge or Is there steady-state error?
- Stability
  - If the input is bounded, then the output should be bounded.
  - Behavior under non-ideal ideal conditions?
- Settling time
  - Short settling time helps reject disturbances
- Overshoot



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

Category	Small (air cooling)	Medium (air and chilled water cooling)	Large (air and chilled water cooling)
Typical size	10K sq ft	30K sq ft	>35K sq ft
Energy savings (% of cooling costs)	40%	30%	15%
Estimated MWh saved	5,300	9,100	10,500

(1) C. Bash, C. Patel, R. Sharma, "Dynamic Thermal Management of Air Cooled Data Centers", HPL-2006-11, 2006
(2) L. Bautista and R. Sharma, "Analysis of Environmental Data in Data Centers", HPL-2007-98, 2007
(3) http://www.hp.com/hpinfo/globalcitizenship/gcreport/energy/casestudies.html

In-depth Example – Power Capping

Function Definition: Operate computer system/sub-system/component within specific power consumption threshold irrespective of load or environmental conditions. Its 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.

IBM Research and U. Tennessee, Knoxville

Server-level Power Control

Charles Lefurgy, Xiaorui Wang, Malcolm Ware

The 4th IEEE International Conference on Autonomic Computing

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

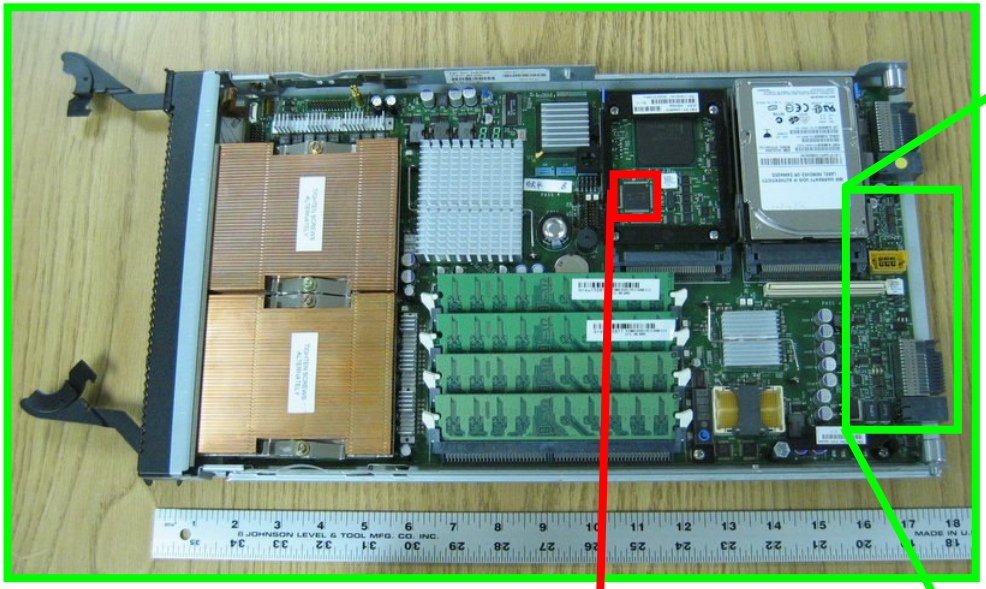
- **Control of peak server-level power (to 0.5 W in 1 second)**
- **Derivation and analysis [see paper]**
  - Guaranteed accuracy and stability
- **Verified on real hardware**
- **Better application performance than previous methods**

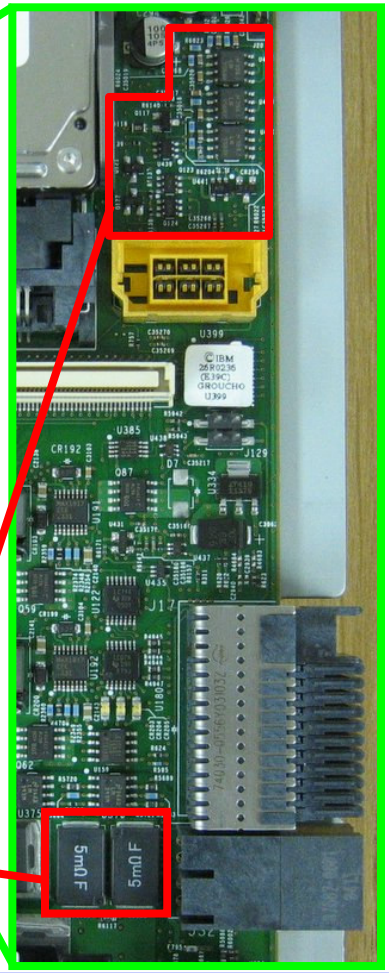
Power measurement

Measure 12V bulk power

0.1 W precision, 2% error

HS20 8843 (Intel Xeon blade)



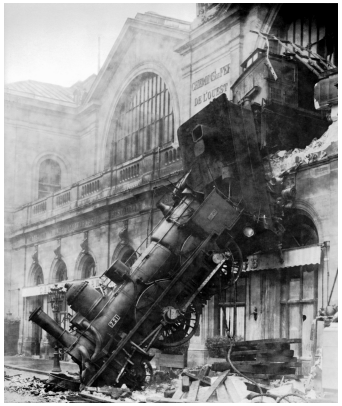


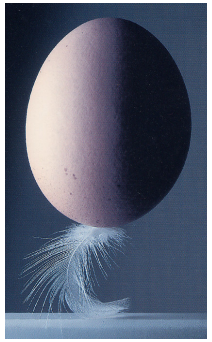
controller firmware on service processor (Renesas H8 2168)

Measurement/calibration circuit

Sense resistors

Options for power control

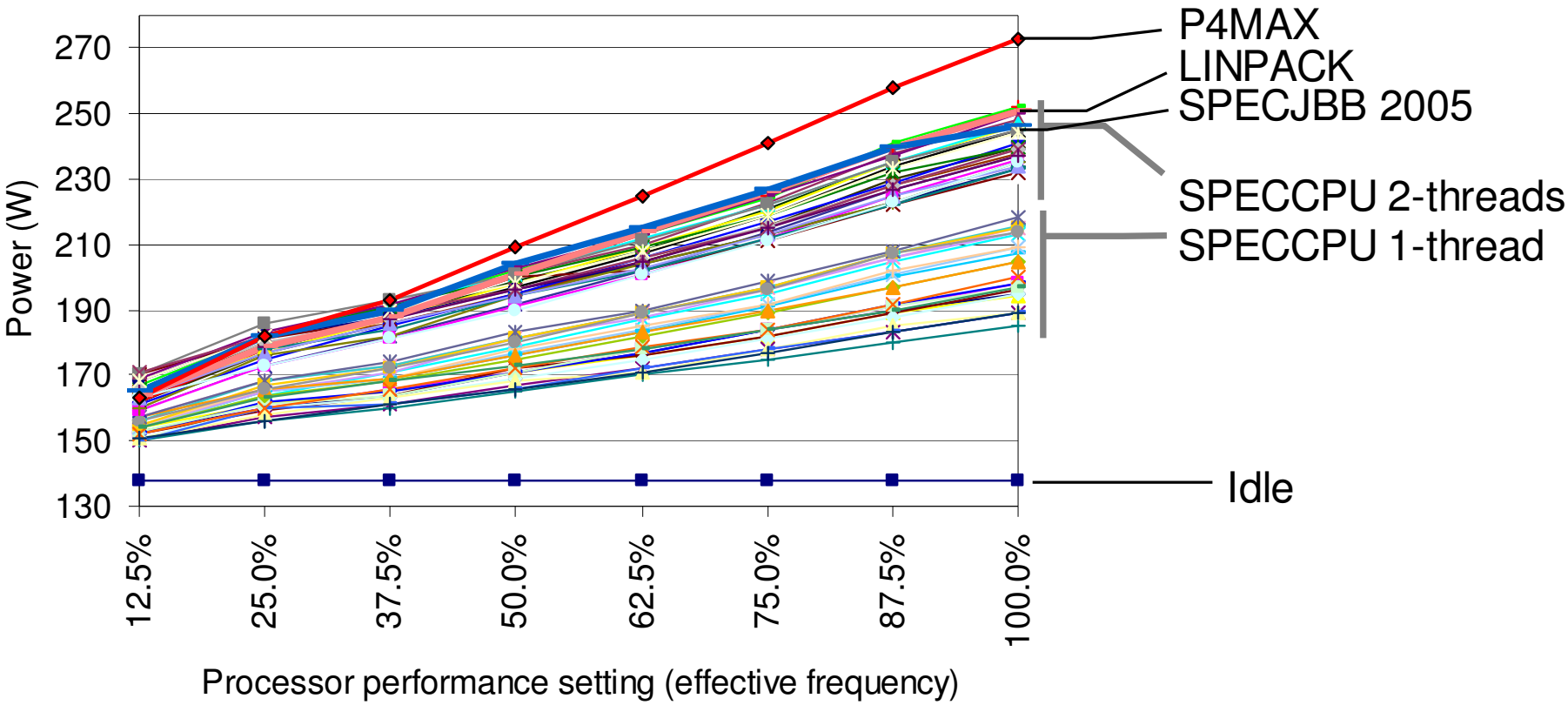




- **Open-loop**
  - No measurement of power
  - Chooses fixed speed for a given power budget
  - Based on most power hungry workload
- **Ad-hoc**
  - Measures power and compares to power budget
  - +1/-1 adjustments to processor clock throttle register
- **Proportional Controller (“P control”)**
  - Designed using control theory
  - Guaranteed controller performance

Open loop design

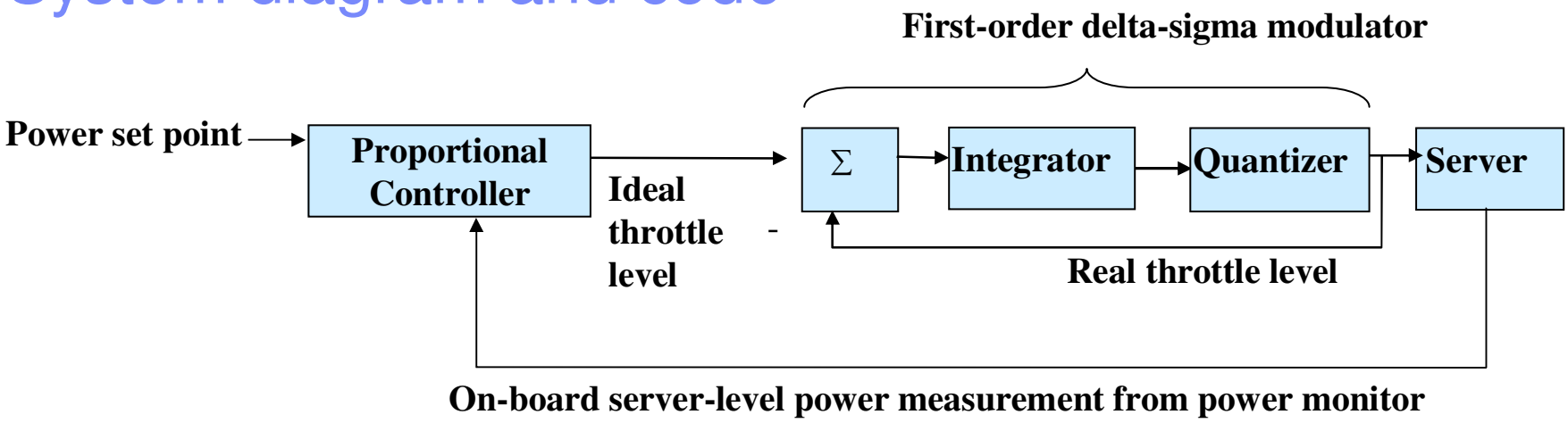
- **P4MAX workload used as basis for open-loop controller**
- **Graph shows maximum 1 second power for workload**



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

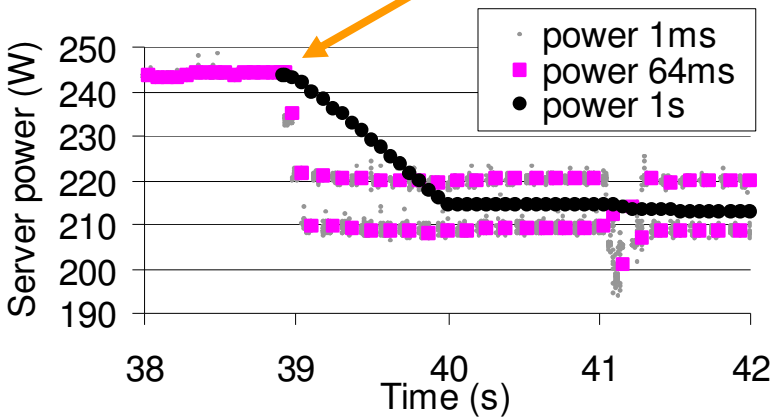


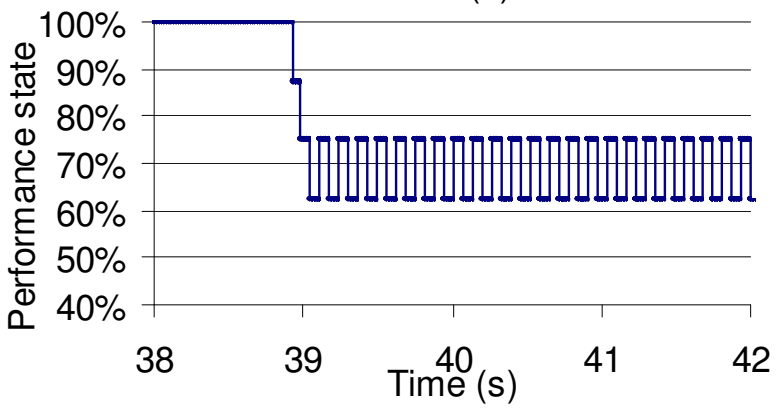

```
// Controller code
error = setpoint – power_measurement;
ideal_throttle = throttle + (1/A) * error;
// Actuator code (First-order delta-sigma modulation)
throttle = truncate(ideal_throttle);
frac = ideal_throttle – throttle;
total_fraction = total_fraction + frac;
if (total_fraction > 1) {
    throttle = 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?

Set point = 211.0 W

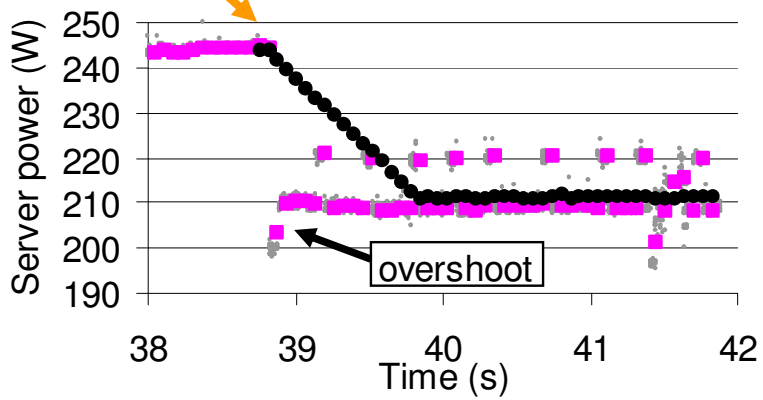
Ad-hoc

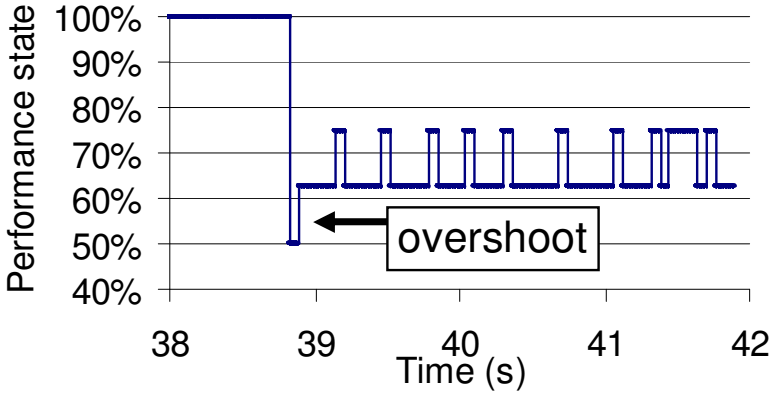




Settles to 216.0 W 5 W Violation
CPU speed: 68.8%

P Controller

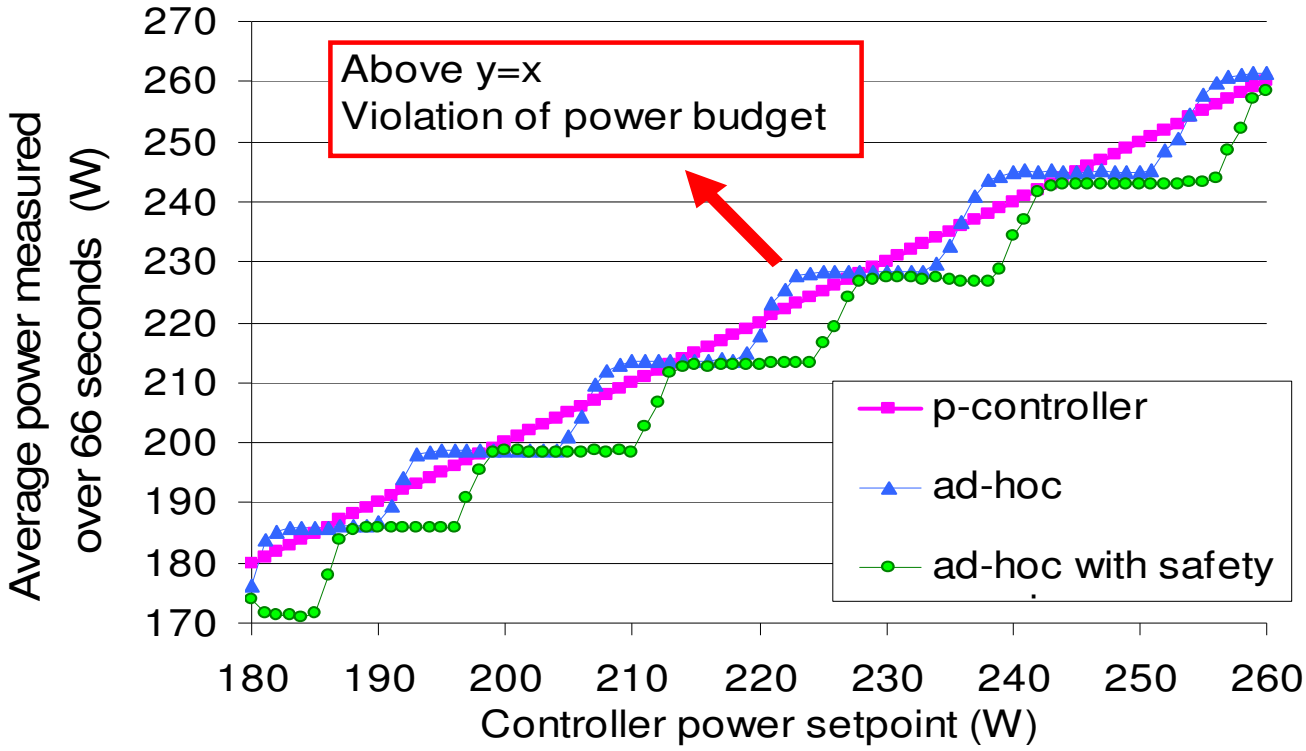




Settles to 211.0 W No violation
CPU speed: 65.8%

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

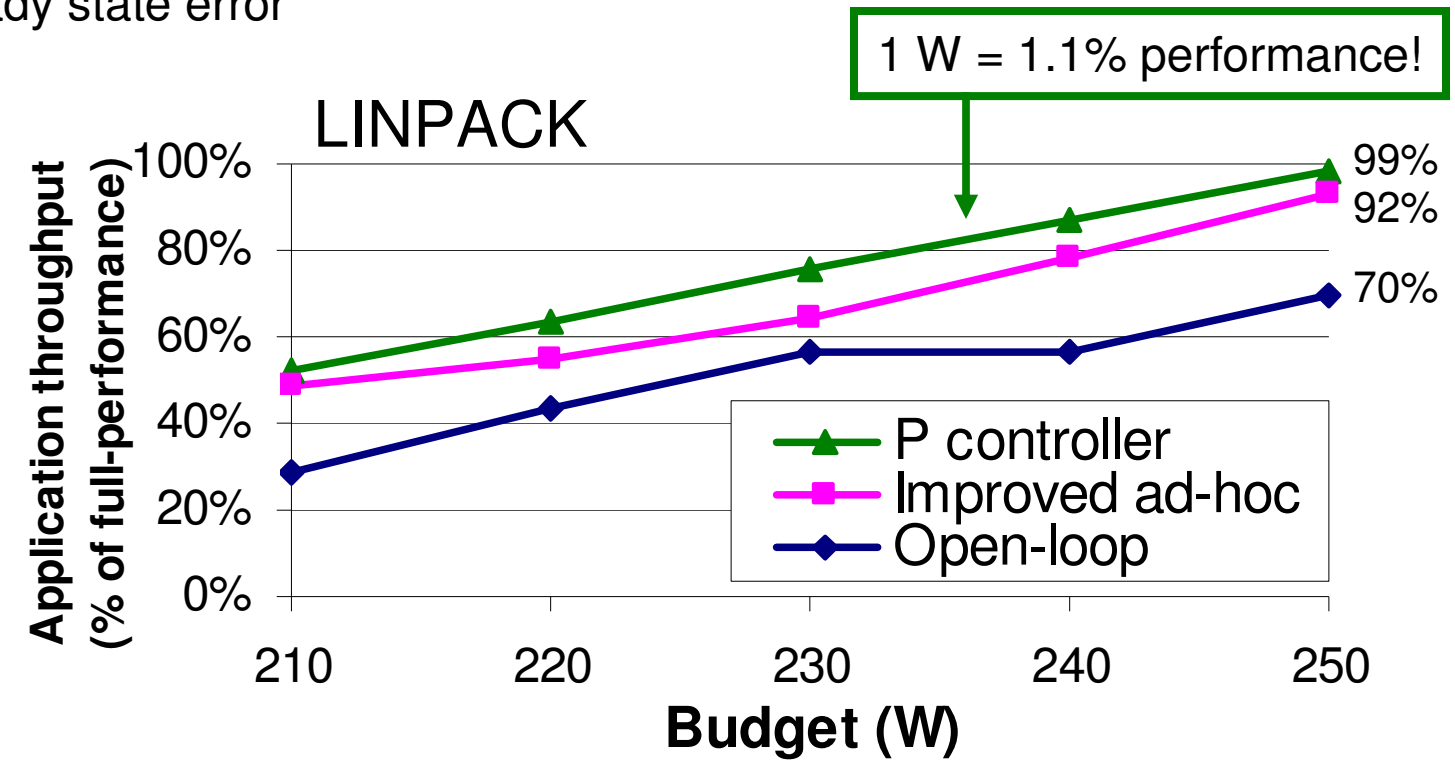
- **Run each controller with 5 power budgets**
- **Compare throughput of workloads**

- **Table shows settings used for each controller**

Power budget	Open-loop processor performance setting	Ad-hoc (with safety margin) set point	P control set point
250 W	75%	238.9 W	245.0 W
240 W	62.5%	229.1 W	235.2 W
230 W	62.5%	219.3 W	225.4 W
220 W	50%	209.5 W	215.6 W
210 W	37.5%	199.7 W	205.8 W

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



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 – 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.