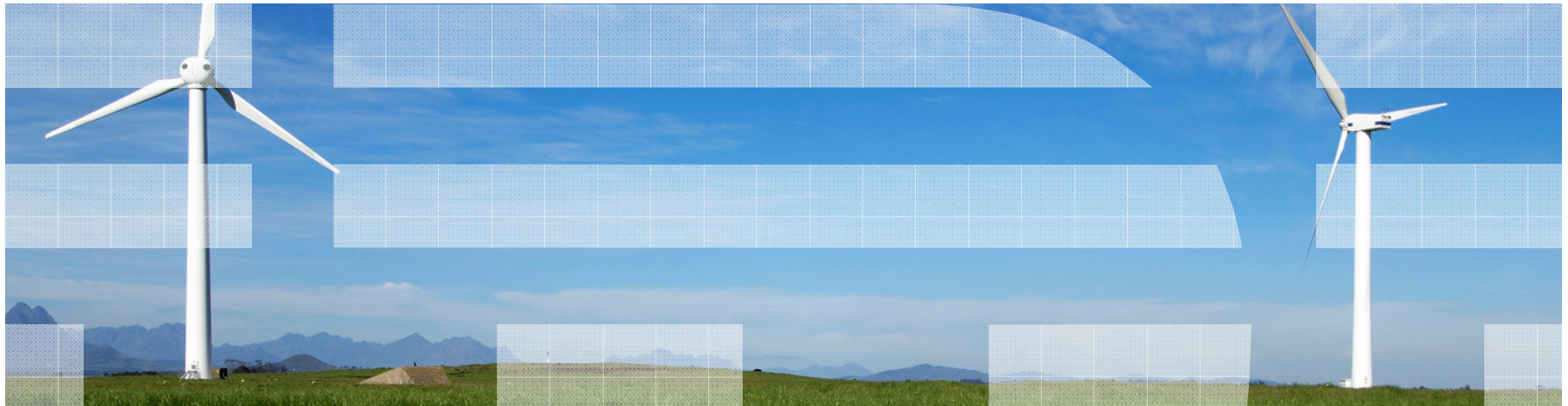


Energy-Efficient Data Centers and Systems

Charles Lefurgy, Malcolm Allen-Ware, John Carter, Wael El-Essawy, Wes Felter, Alexandre Ferreira, Wei Huang, Anthony Hylick, Tom Keller, Karthick Rajamani, Freeman Rawson and Juan Rubio



Data center energy matters

- In 2005, data centers accounted for
 - 1.0% of world-wide energy consumption
 - 1.2% of US energy consumption
 - Consumption doubled between 2000-2005
 - 16% annual growth rate
- Unsustainable



- Drivers of the DC crisis
 - IT demand outpacing energy-efficiency improvements
 - Cloud services
 - Escalating CMOS power density
 - IT refresh is 5x faster than facilities
 - Increasing energy costs

Sources: 1. Koomey, "Worldwide Electricity Used in Data Centers", Environmental Research Letters, 2008
2. Report to Congress on Server and Data Center Energy Efficiency, U.S. Environmental Protection Agency, 2007

Why listen to us?

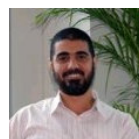
- Technologies
 - 1st power measurement in a server product
 - 1st power capping in a server product
 - Power shifting
 - AMESTER
- Contributions to IBM products
 - POWER6 and POWER7 EnergyScale firmware
 - System x Active Energy Manager firmware
 - IBM Systems Director Active Energy Manager
- Energy optimization of customer data centers
- Patents and publications on power management (2005 – present)
 - 38 US patents awarded
 - 34 peer-reviewed publications



Malcolm Allen-Ware



John Carter



Wael El-Essawy



Wes Felter



Alex Ferreira



Wei Huang



Anthony Hylick



Tom Keller



Charles Lefurgy



Karthick Rajamani



Freeman Rawson



Juan Rubio

Schedule

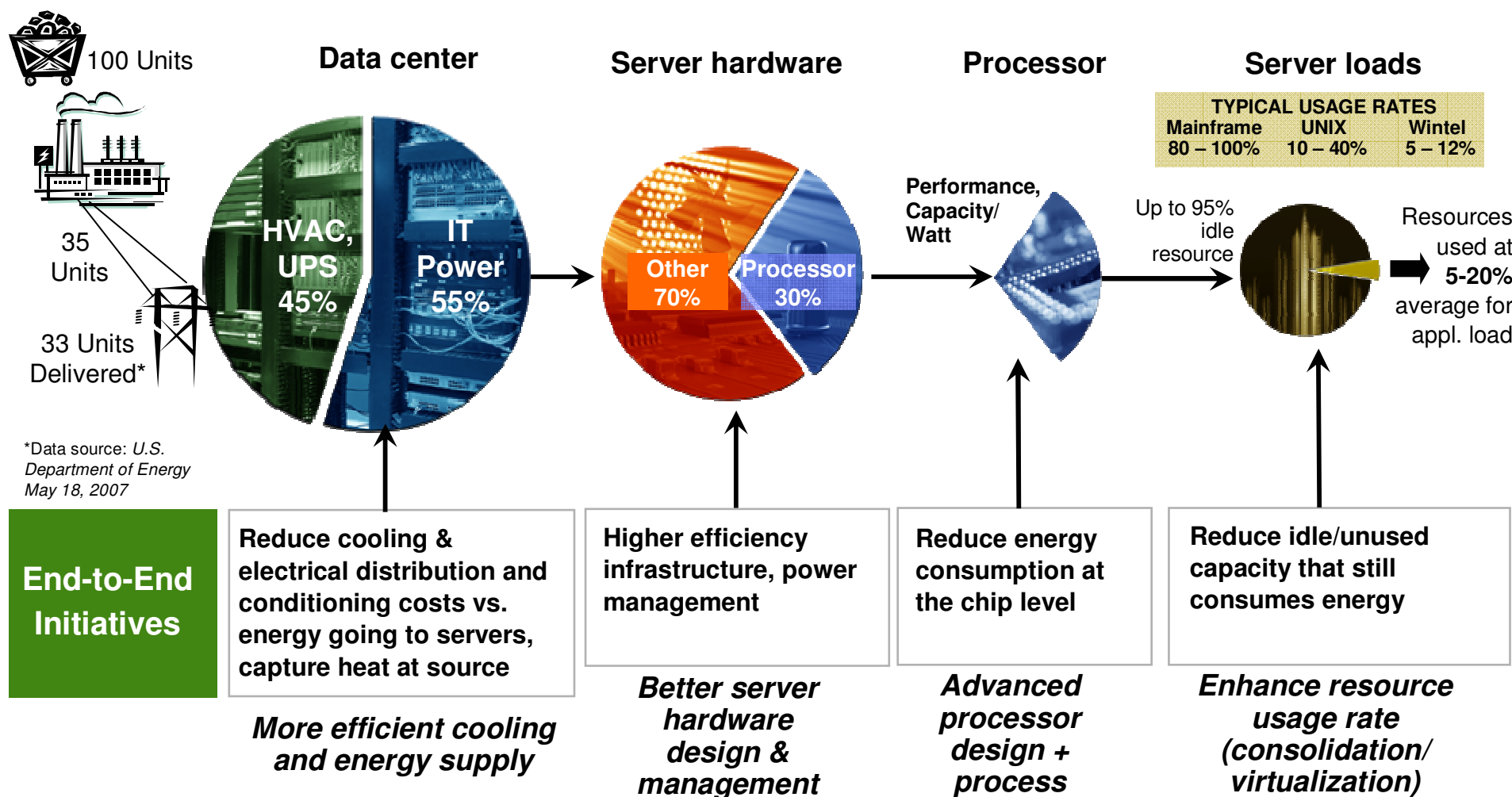
Time	Subject	Presenter
8:00 AM	Fundamentals	Charles Lefurgy
10:00 AM	BREAK	
10:15 AM	Storage	Wes Felter
11:00 AM	Networking	Wes Felter
12:00 PM	LUNCH	
1:00 PM	Cloud	Karthick Rajamani
1:45 PM	Energy-efficient software	Karthick Rajamani
2:15 PM	Modeling	Juan Rubio
3:15 PM	BREAK	
3:30 PM	Emerging technology	Karthick Rajamani
4:30 PM	END	

What is the problem?

There are many dimensions to the problem

1. Very little of the delivered power is converted to useful work
2. Poor allocation of provisioned resources
 - Over cooling – too many air conditioning units are on
 - “Stranded power” – available power is fragmented across circuit breakers
3. Fixed capacity
 - Reaching power and cooling limits of facility
 - Peak capacity cannot always be increased with business growth
4. Power is a first-class design constraint for server design
 - Peak power consumption requirements vary by market
 - Component-level and system-level
 - CMOS technology scaling no longer providing historic trends in energy-efficiency
5. Total cost of ownership
 - Capital expenses (Building data center, buying equipment, etc.)
 - Operational expenses (Energy costs, staffing, etc.)

1. Very little of the delivered power is converted to useful work

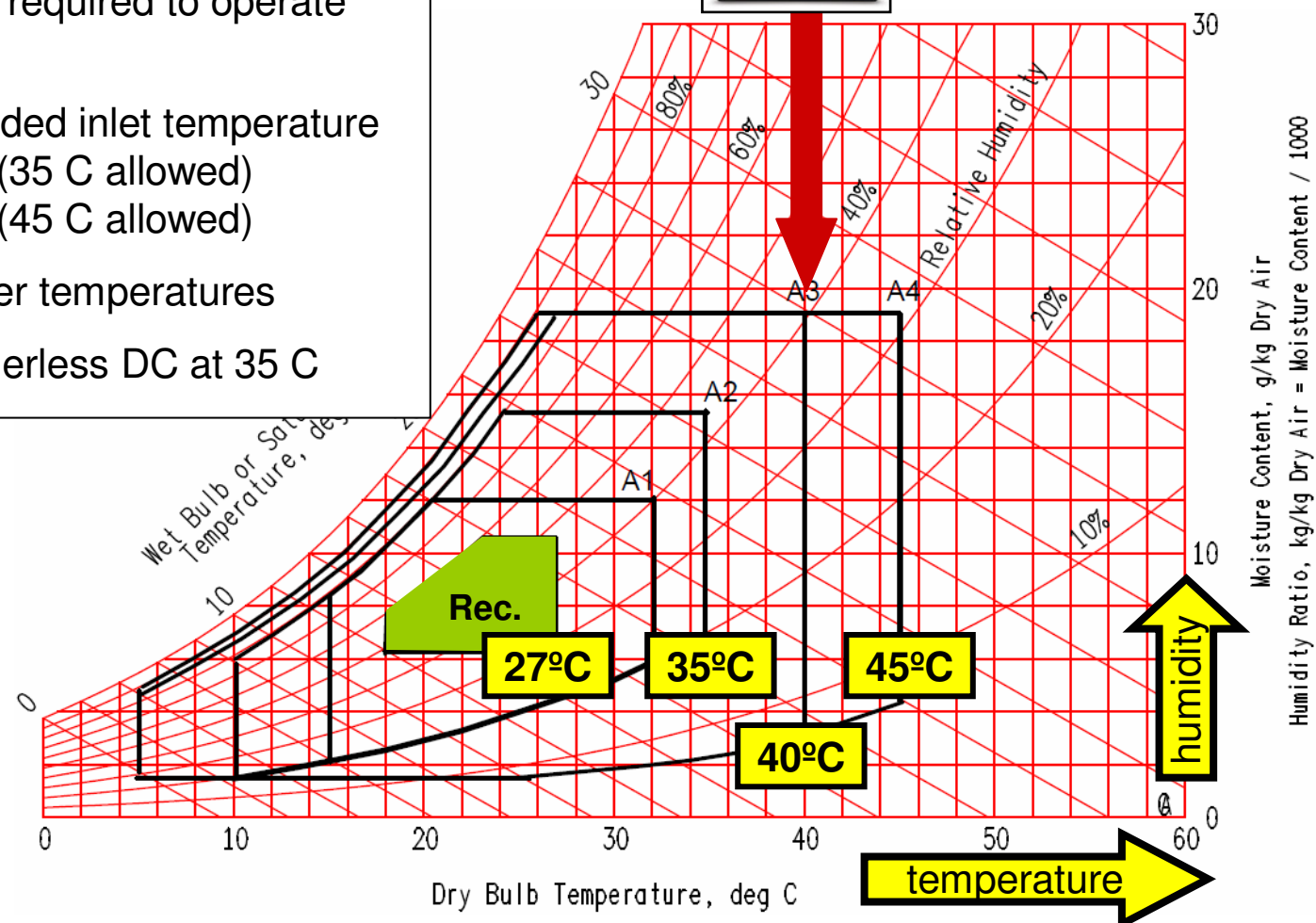


2. Wasting provisioned resources -- Overcooling



**Rackable
CloudRack C2
2009**

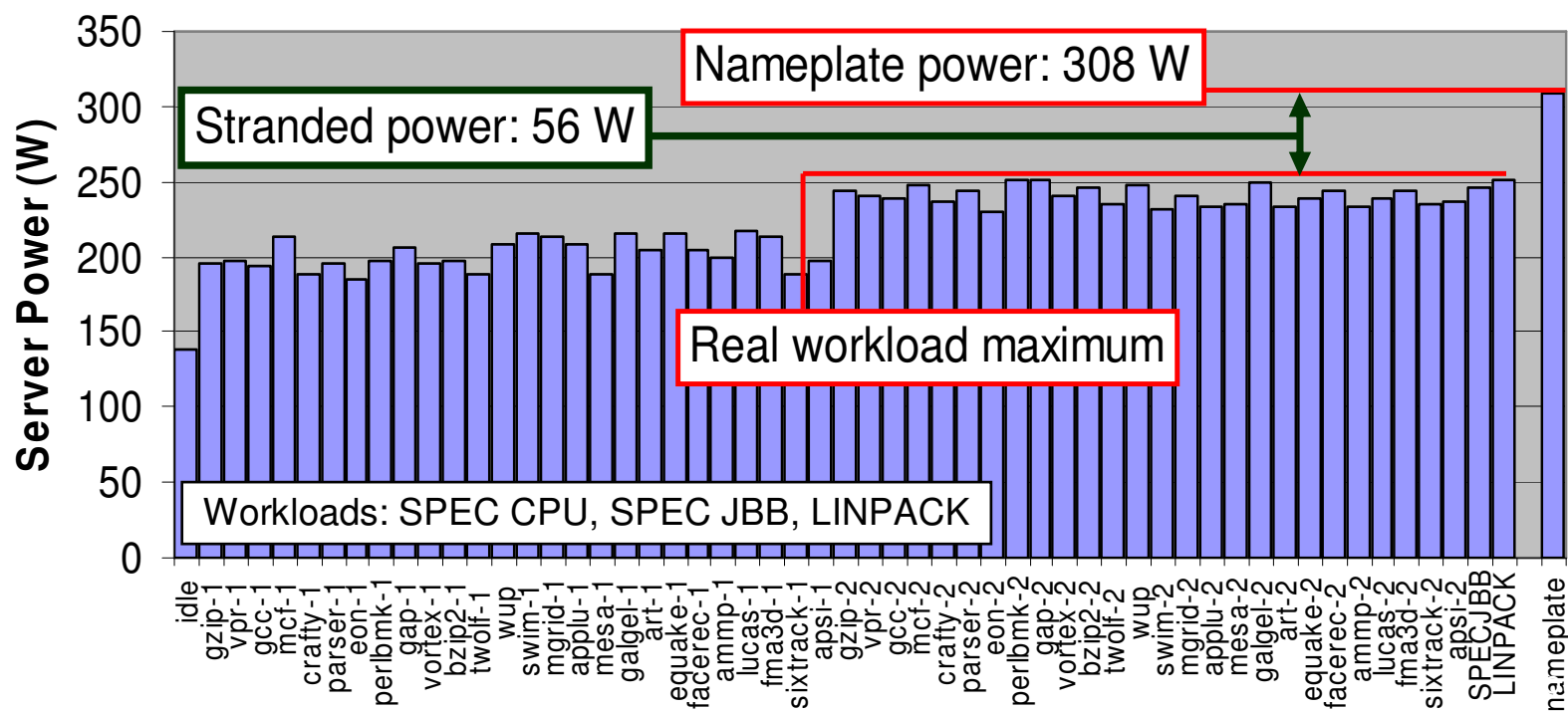
- Spending more than required to operate servers
- ASHRAE recommended inlet temperature
 - 2004: 25 C max (35 C allowed)
 - 2011: 27 C max (45 C allowed)
- Vendors test at higher temperatures
- Microsoft Dublin chillerless DC at 35 C



Source: 2011 Thermal Guidelines for Data Processing Environments – Expanded Data Center Classes and Usage Guidance, American Society of Heating, Refrigerating and Air-Conditioning Engineers

Wasting provisioned resources -- Stranded power

- Using **nameplate power** (worst-case) to allocate power on circuit breakers
 - However, real workloads do not use that much power
 - Result: available power is **stranded** and cannot be used
- Stranded power is a problem at all levels of the data center
- Example: IBM HS20 blade server – nameplate power is 56 W above real workloads.



What are the consequences of wasted capacity?

- Run out of data center capacity
 - Data center is considered “virtually full”
 - Unnecessary expansion of DC results in large capital expense
 - Disruption to business operations
- Maintaining under-utilized data centers is expensive
 - Electric losses are higher at lower utilizations
 - Inefficiency of power and cooling equipment at low loads

3. Fixed capacity

- Peak Rack power
 - Air-cooled rack peak power is roughly 30 kW
 - Older DC cooling infrastructure may not support such high load
 - Common to see empty slots in rack
- Peak DC power is a problem in some geographies
 - Example: New York City
 - Too expensive to expand power delivery
 - Limits business growth

Power supply still a vexation for the NSA

“The spy agency has delayed the deployment of some new data-processing equipment because it is short on power and space. Outages have shut down some offices in NSA headquarters for up to half a day...Some of the rooms that house the NSA's enormous computer systems were not designed to handle newer computers that generate considerably more heat and draw far more electricity than their predecessors.”

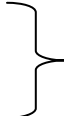
-- Baltimore Sun, June 2007

Snafus forced Twitter datacenter move

“A new, custom-built facility in Utah meant to house computers that power the popular messaging service by the end of 2010 has been plagued with everything from leaky roofs to insufficient power capacity, people familiar with the plans told Reuters.”

-- Reuters, April 1, 2011

4. Power is a first-class design constraint for servers

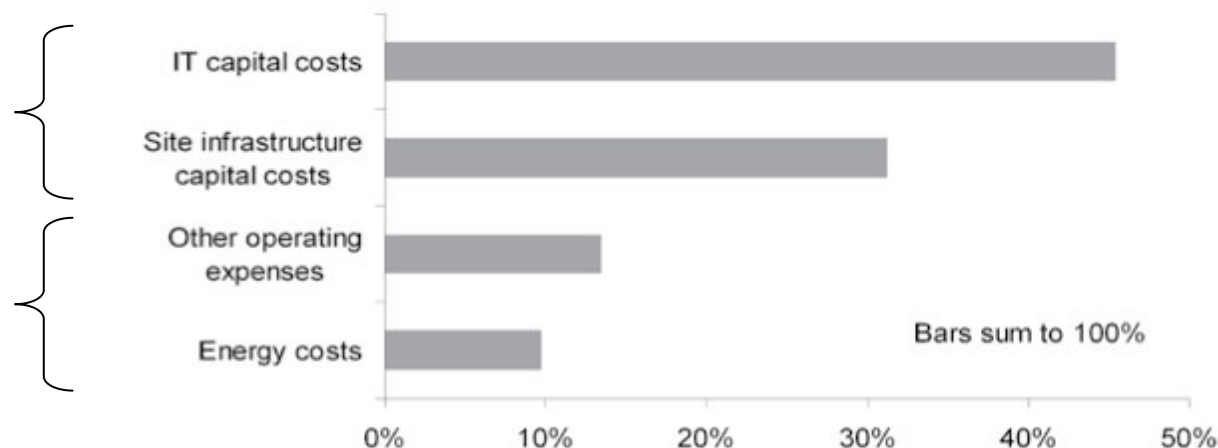
- Power and performance are top design parameters for microprocessors and servers
- Power constraints exist across all classes of computing equipment
 - Laptop (30 – 90 W)
 - Desktop (100s W)
 - Server (200W - 5kW)
 - Data center (1-20 MW)
- Peak power constraint means high performance requires high energy-efficiency
- Server components
 - CPU socket power
 - Memory power
 - Electrical cord limits
 - Power supply limits (physical size, reuse of standard designs) Air cooling limits and fan noise consideration

5. Total cost of ownership

- Tier-3 HPC data center for financial analytics in 2007
 - 4.4 MW capacity (IT + cooling)
 - \$100M USD installed cost
 - 1 W is spent on cooling and UPS for every 1 W for IT equipment)
- Capital costs dominate operating cost
- Opportunity: better energy-efficiency can reduce facility capital costs
 - Pack more revenue producing servers into existing facility
 - Delay building new data center

$\frac{3}{4}$ capital costs

$\frac{1}{4}$ operating costs



Annualized cost by component as a fraction of the total

Source: Koomey et al., *A Simple Model for Determining True Total Cost of Ownership for Data Centers*, Version 2.1, Whitepaper, Uptime Institute, 2008

Metrics

“If you can’t measure it, you can’t manage it”

Metrics and benchmark overview

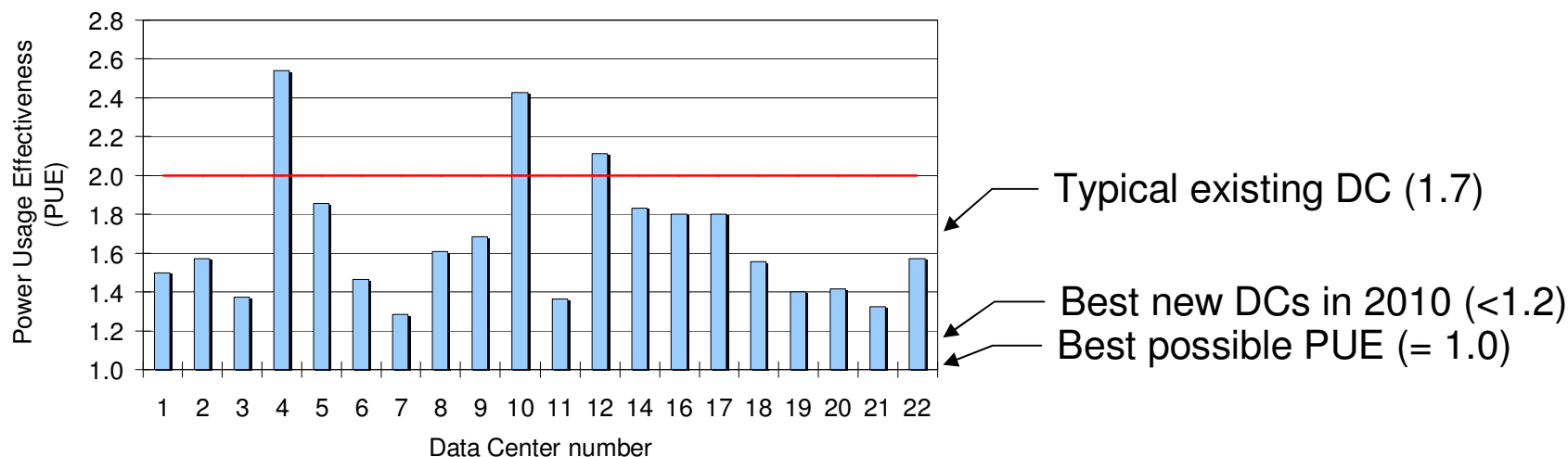
- Data center metrics
 - PUE
 - Green500
- Server benchmarks that require power measurement
 - SPECPower_ssj2008 (2007)
 - SPECweb2009 (2009)
 - SAP server power benchmark (2009)
 - ENERGY STAR Computer Server (2009)
- Server benchmarks with optional power measurement
 - SPECvirt_sc2010
 - TPC-C
 - TPC-E
 - TPC-H
- Storage benchmarks ([in storage section](#))
 - Storage Performance Council: SPC-1/E; SPC-1C/E (2009)
 - SNIA (in development)
 - ENERGY STAR Storage (in development)
- System benchmark
 - SAP system power benchmark (in development)

PUE: Power Usage Effectiveness

- Indicates energy efficiency of entire facility:

$$PUE = \frac{\text{Total facility power}}{\text{IT equipment power}}$$

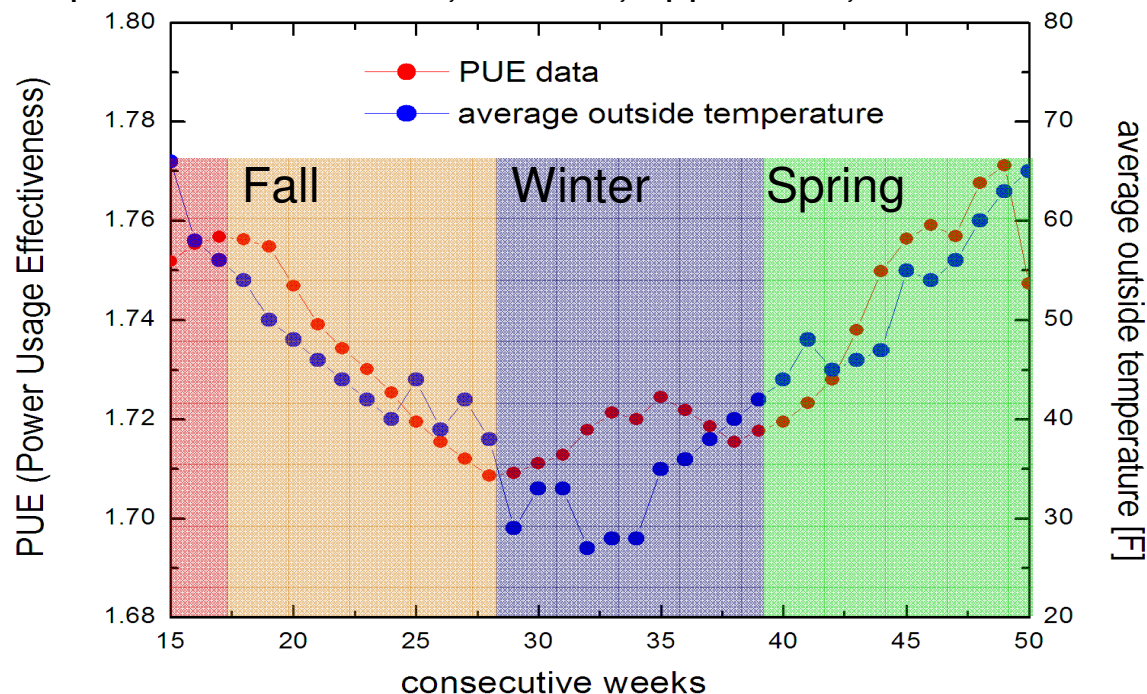
- Only data center metric widely recognized across industry
- Metric created by The Green Grid



Source: Tschudi et al., "Measuring and Managing Energy Use in Data Centers." HPAC Engineering, LBNL/PUB-945, 2005.

Problems with PUE

- Appropriate metering is often not in place
- Score is dependent on weather, location, application, and tier level

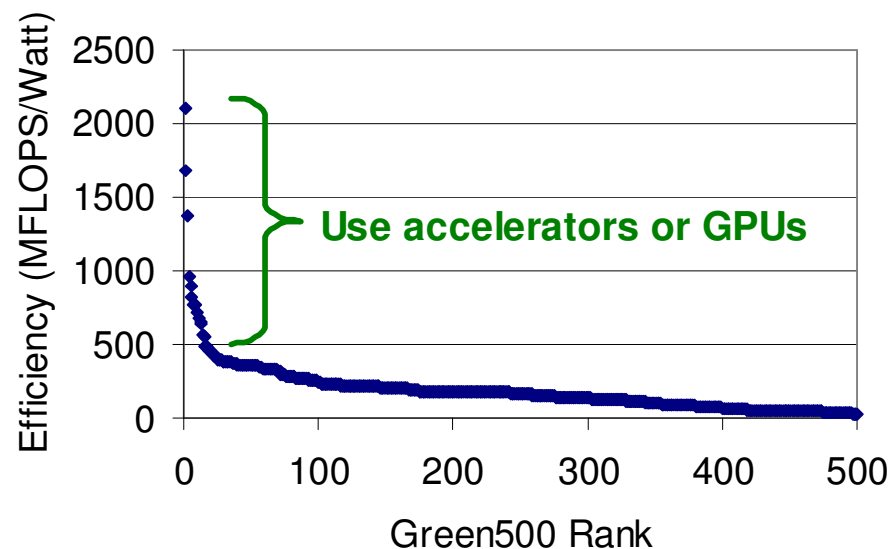


Source: Hendrik Hamman, IBM

- Rewards inefficiency in the server (e.g. poor AC/DC conversion, fan power is included)
- PUE is insufficient for “proving” and managing energy efficiency

The Green500 list

- Energy-efficiency for High Performance Computing
 - Large clusters are costly to operate
(ASC Purple @ 4.5 MW, \$0.12/kWh → \$4.7M/year)
 - Site must be designed to supply power
- Green500 list reorders the Supercomputing TOP500 list for energy-efficiency
 - Metric: LINPACK performance / Power for computer
 - Does not include computer room cooling



Green500 as of June 2011 (source: <http://www.green500.org>)

Rank	MFLOPS/W	Site	Computer	Power (kW)	TOP500 rank
1	2097	IBM – Watson Research Center	NNSA/SC Blue Gene/Q Prototype 2	41.0	109
2	1684	IBM – Watson Research Center	NNSA/SC Blue Gene/Q Prototype 1	38.8	165
3	1375	Nagasaki U. – self-made	Intel i5, ATI Radeon GPU, Infinibad QDR	34.24	430
4	958	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670 Nvidia GPU	1243.8	5
5	891	CINECA/SCS – SuperComputing Solution	IBM iDataPlex DX360ME, Xeon 2.4, nVidia GPU, Infiniband	160	54

Other proposed metrics from The Green Grid

- DCeP (Data Center energy Productivity)
 - Hard to define a workload that can be run well across data centers

$$DCeP = \frac{\text{Useful Work Produced}}{\text{Total Data Center Energy Consumed}}$$

- ERE (Energy Reuse Effectiveness)
 - Fix PUE to show benefit for reusing energy
 - Example: Using hot water from DC to heat nearby building

$$ERE = \frac{\text{Cooling} + \text{PowerDistribution} + \text{Lighting} + \text{IT} - \text{Reuse}}{\text{IT}}$$

- CUE (Carbon Usage Effectiveness)
 - Measure sustainability
 - In addition to PUE

$$CUE = \frac{\text{Total CO}_2 \text{ emissions caused by the Total Data Center Energy}}{\text{IT Equipment Energy}}$$

- WUE (Water Usage Effectiveness)
 - Measure water use

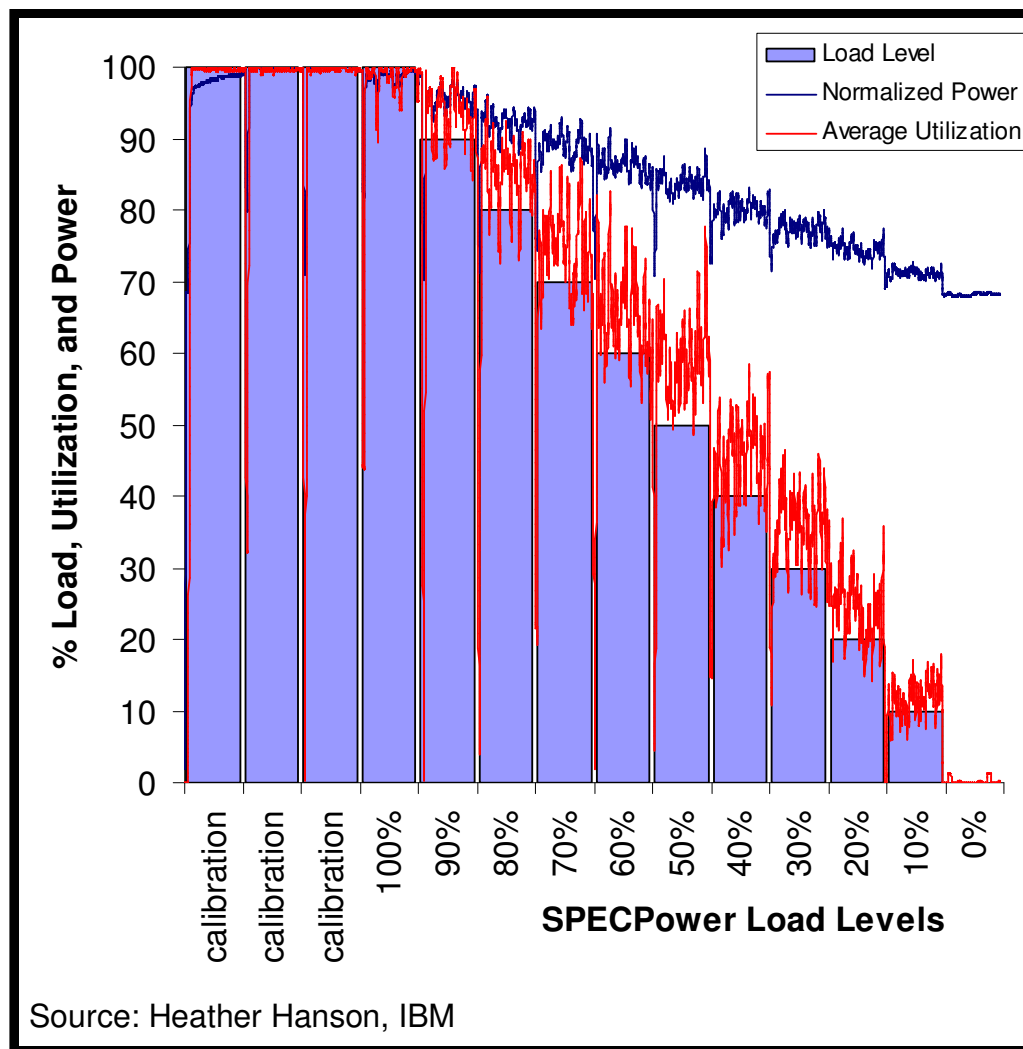
$$WUE = \frac{\text{Annual Site Water Usage}}{\text{IT Equipment Energy}}$$

SPECPower_ssj_2008



- Measure transaction-oriented servers
- Based on SPECjbb, a Java performance benchmark.
- 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

$$\frac{\sum_{idle}^{100\%} \text{Throughput per level}}{\sum_{idle}^{100\%} \text{power per level}}$$



Energy Star



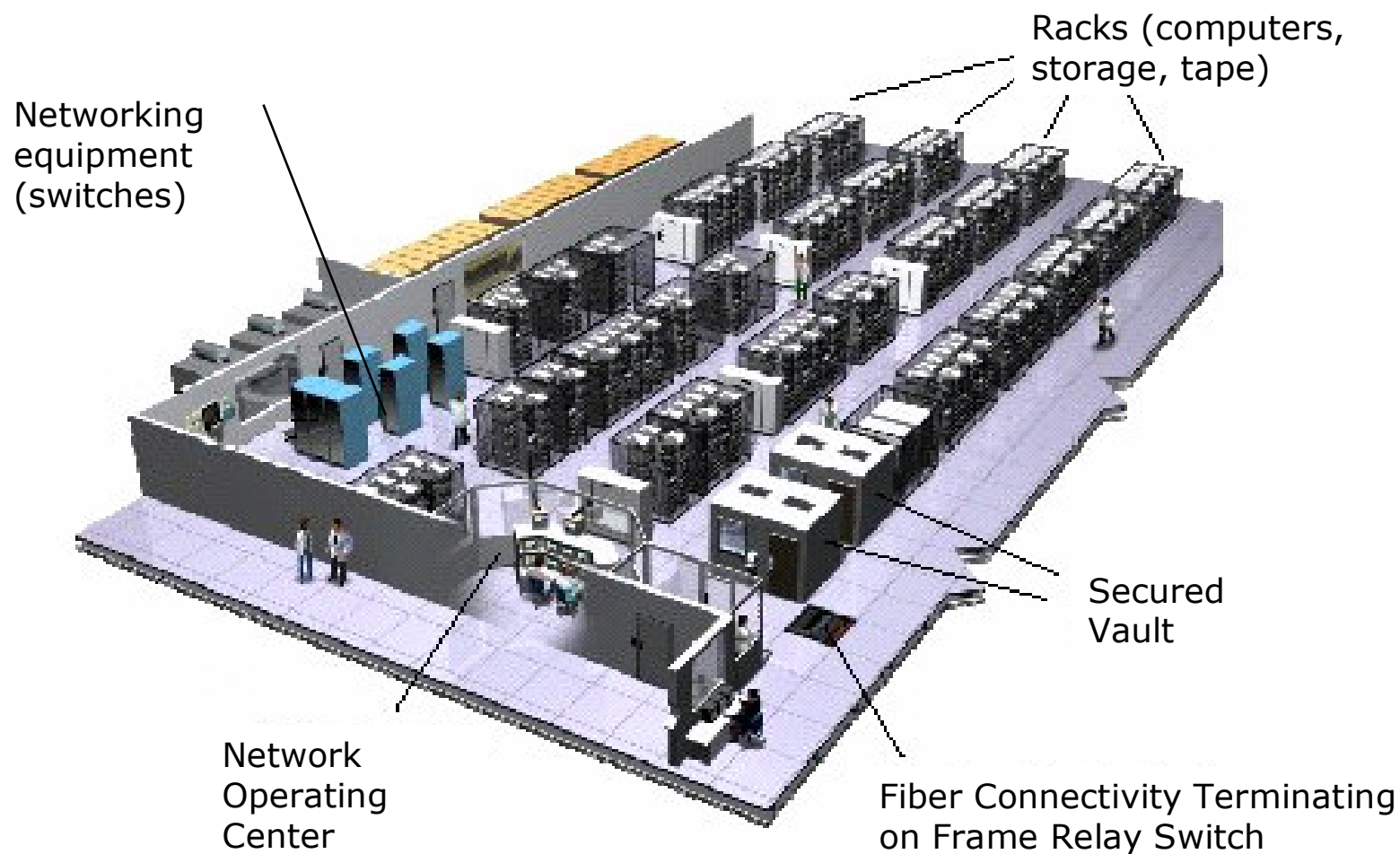
- United States Environmental Protection Agency
- Not a benchmark, but a government specification for compliance allowing a manufacturer to use the Energy Star mark to help customers identify energy-efficient products.
- Version 1 for Servers (2009)
 - Power supply efficiency requirements under loading of 10%, 20%, 50%, 100%, (varies by power supply capacity)
 - Idle power limits, depending on configuration
 - Allowances made for extra components
 - 8 W per additional hard drive
- Version 2 for Servers (in development)
 - Expected to report workload energy efficiency over a range of utilization levels
- Expanding to storage, UPS, and data center (in development)

Issues with server benchmarks

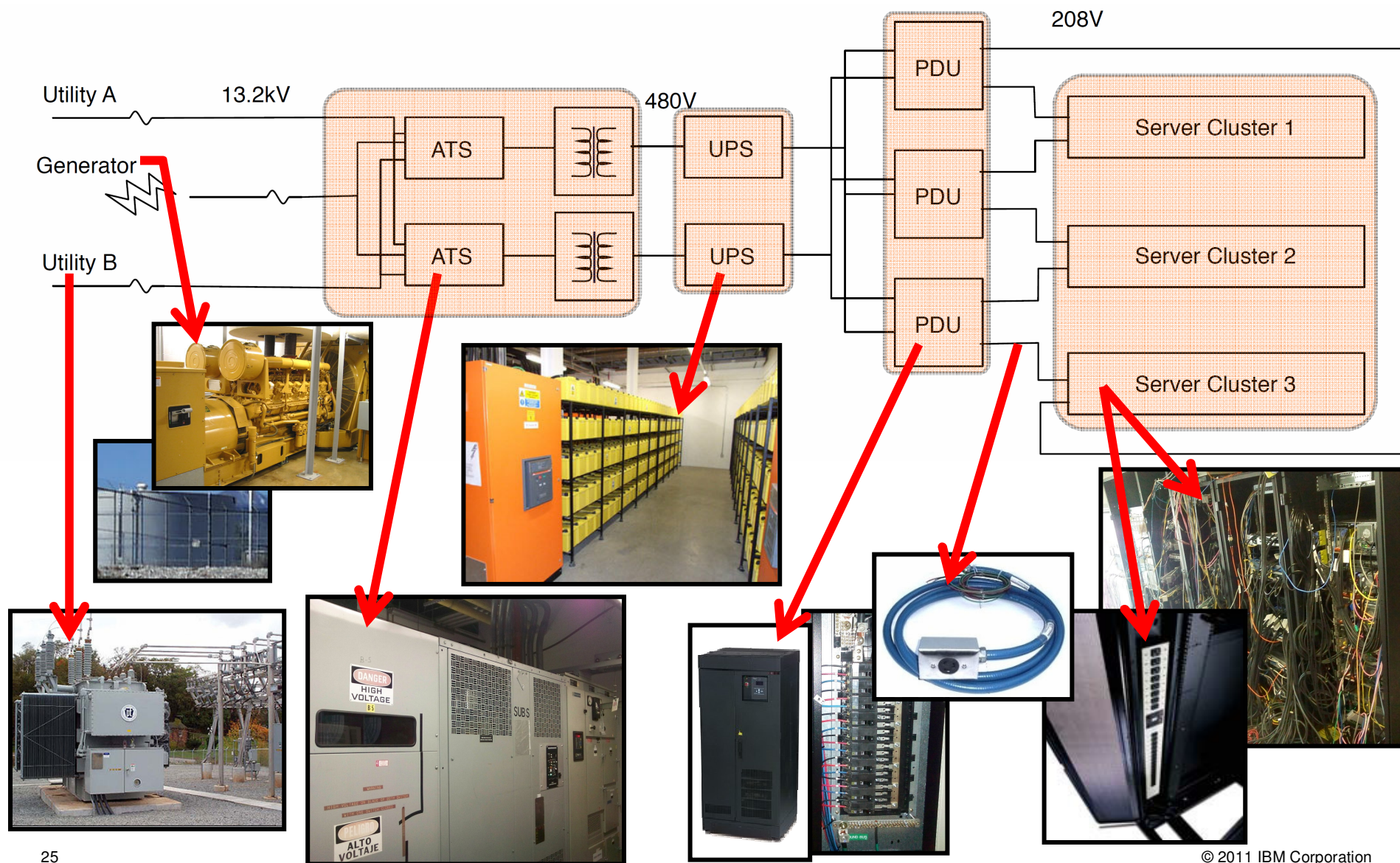
- Lack of realism
 - Do not include network and remote storage loads
 - SAP System Power benchmark will include network and storage
 - No task switching
 - Very strong affinity
- Coverage of server classes
 - Best SPECPower score likely on 1- and 2-socket servers with limited memory
 - Robust (redundant) configurations are penalized
 - Example: Today, dual power supplies reduce conversion efficiency

Data center facilities

A Typical Data Center Raised Floor

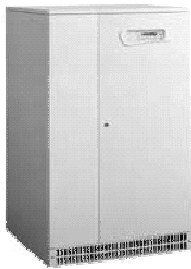


Power delivery infrastructure for a typical large data center



Data center power conversion efficiencies

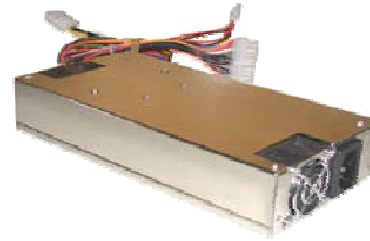
UPS⁽¹⁾
88 - 92%



Power Distribution⁽²⁾
98 - 99%



Power Supply^(3,4)
55 - 90%



DC/DC⁽⁵⁾
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

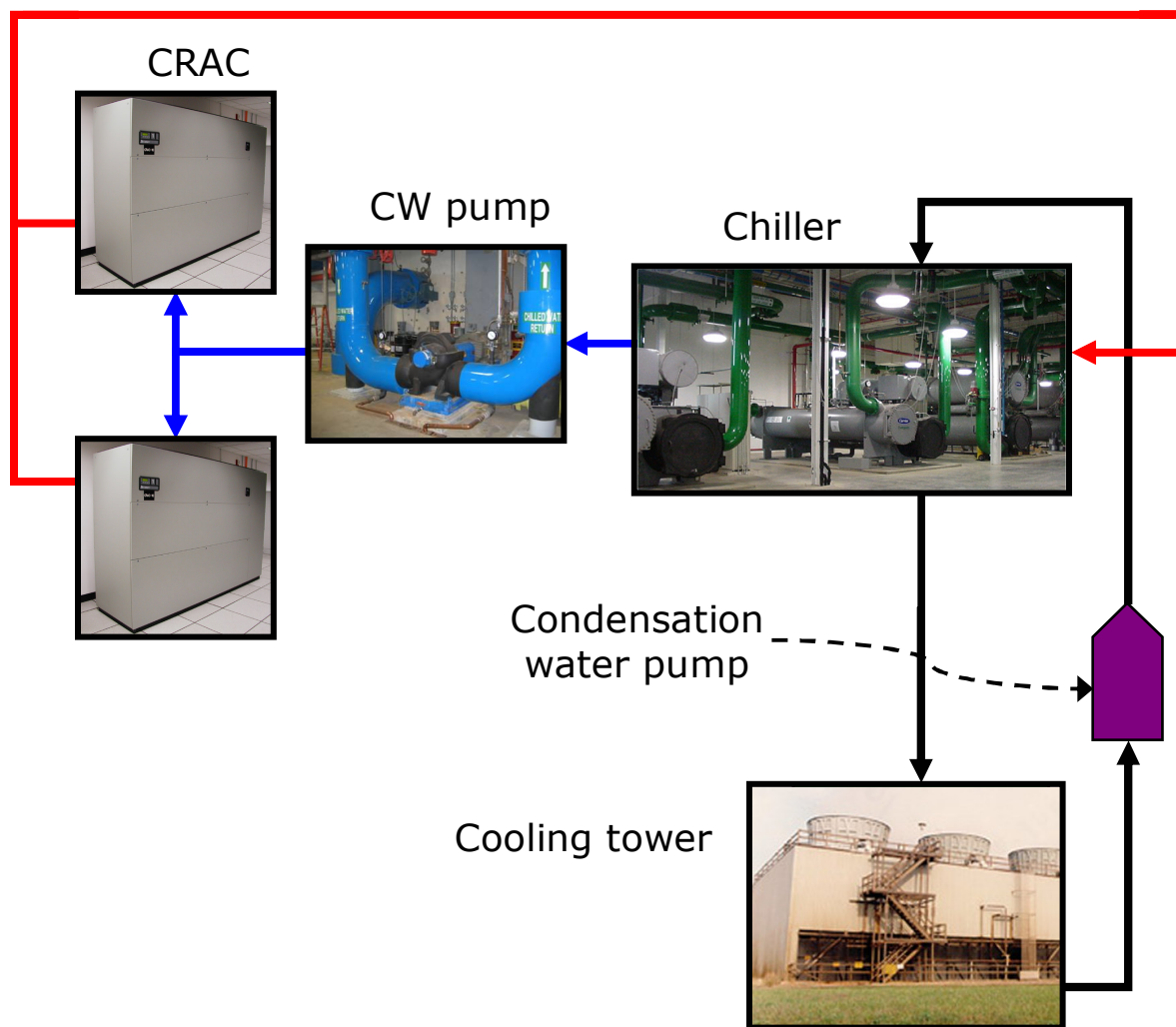
(4) "ENERGY STAR® Server Specification Discussion Document", October 31, 2007.

(5) IBM internal sources

Cooling infrastructure for a typical large data center

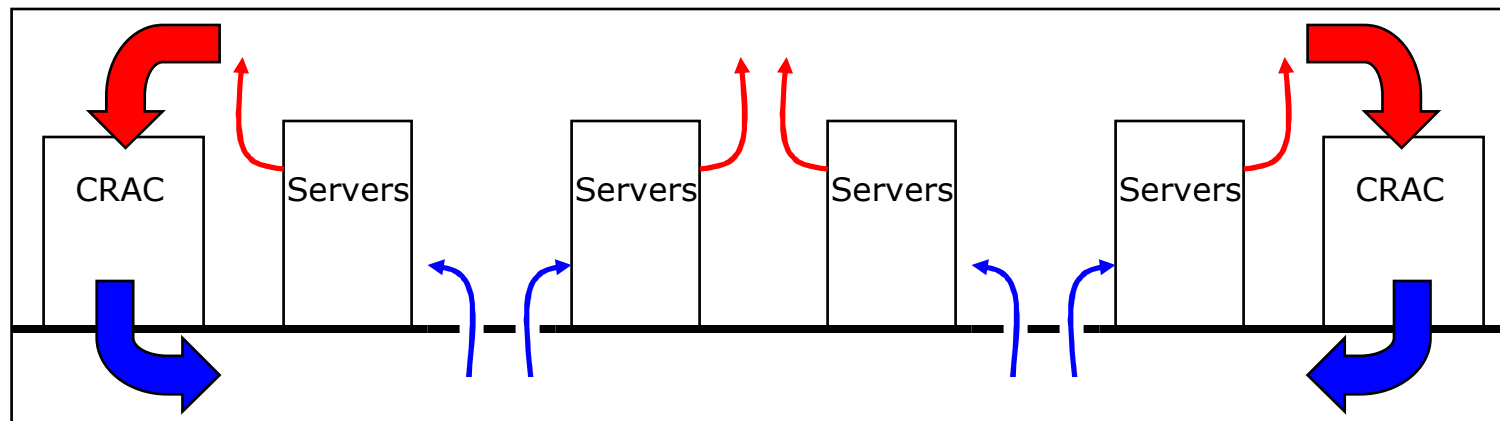
- Two water loops
- Chilled water (CW) loop
 - Chiller(s) cool water which is used by CRAC(s) to cool down the air
 - Chilled water usually arrives to the CRACs near 10°C (50°F)
- Condensation water loop
 - Usually ends in a cooling tower
 - Needed to remove heat out of the facilities

Sample chilled water circuit



Raised floor cooling

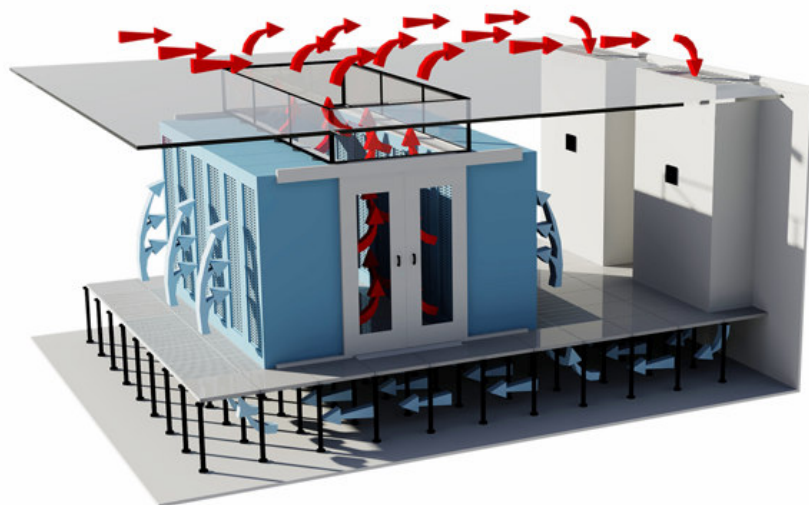
- Racks
 - Arranged in a hot-aisle cold-aisle configuration
- Computer room air conditioning (CRAC) units
 - Located in raised-floor room or right outside of raised-floor room
 - Blower moves air across the raised floor and across cooling element
 - Most common type in large data centers uses chilled 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



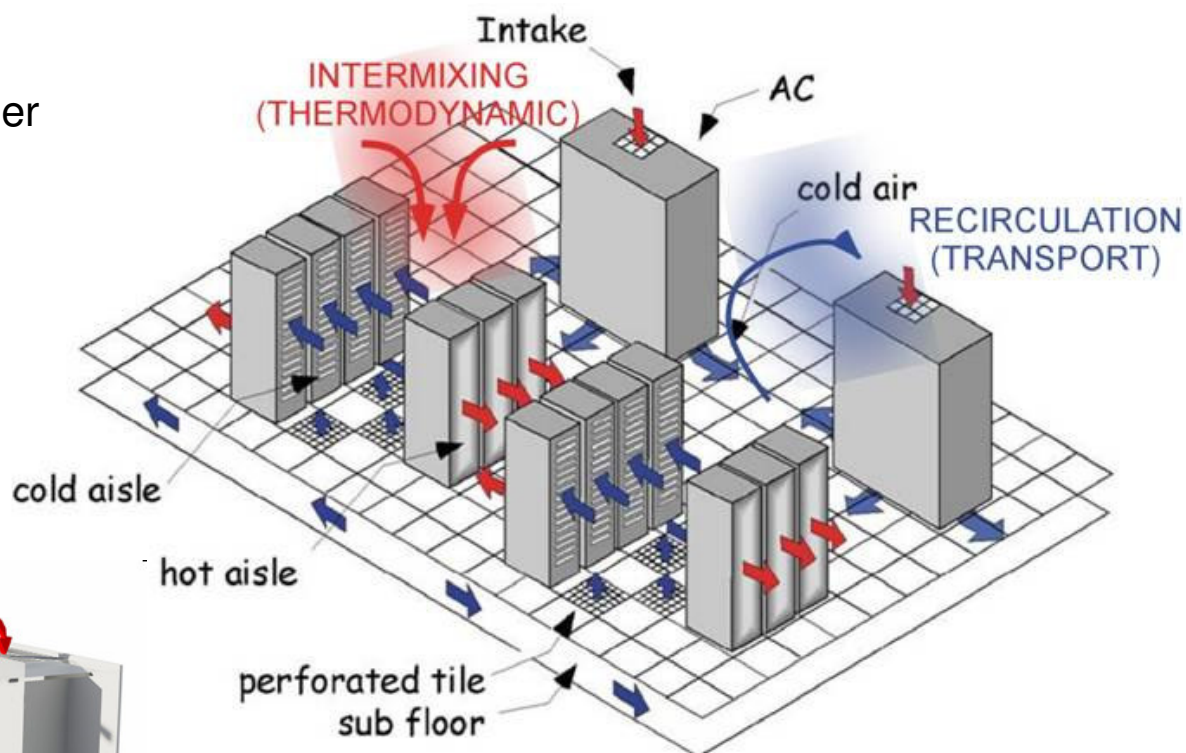
Raised floor cooling

- Hot-cold intermixing
 - Cold air at inlet of air conditioner
 - Hot air at inlet of IT equipment
- Best practice is to separate hot and cold air

Hot-aisle containment

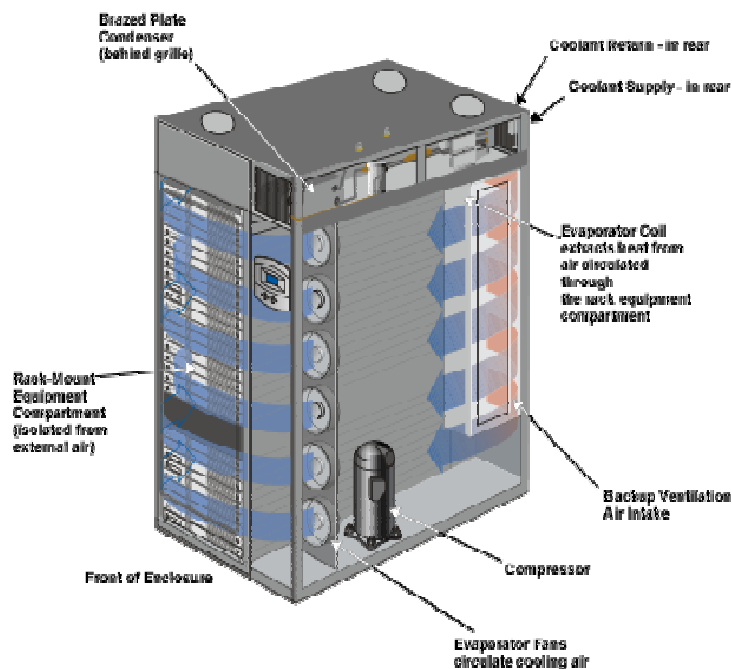


Conventional data center



Commercial liquid cooling solutions for racks

- Purpose
 - Reduce localized hotspots
 - Allow higher power density in older facilities
 - Optimize cooling by rack
 - No raised floor required
- 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



Liebert XDF™
Enclosure (1)



APC InfraStruXure
InRow RP (2)

“Liebert XDF™ High Heat-Density Enclosure with Integrated Cooling”,

http://www.liebert.com/product_pages/ProductDocumentation.aspx?id=40

“APC InfraStruXure InRow RP Chilled Water”,

http://www.apcc.com/resource/include/techspec_index.cfm?base_sku=ACRP501

Container data centers

- Modular data center design
 - IT equipment container: servers, storage, network switch
 - Physical infrastructure container: chiller, UPS/batteries, etc.
 - Site provides power, chilled water, and network
 - Pre-assembled with multi-vendor equipment
- Benefits
 - Pay as you grow
 - No raised floor required, just a concrete slab
 - Cheaper and quicker than retrofitting a data center
 - Rapid delivery (2-3 months from order)
 - Support high power density (34 kW/rack)

Microsoft's Chicago data center (2009)
Calculated annual PUE of 1.22

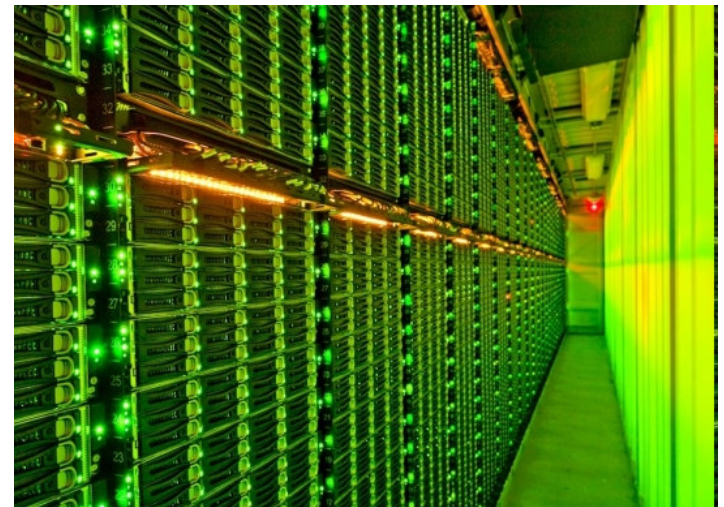
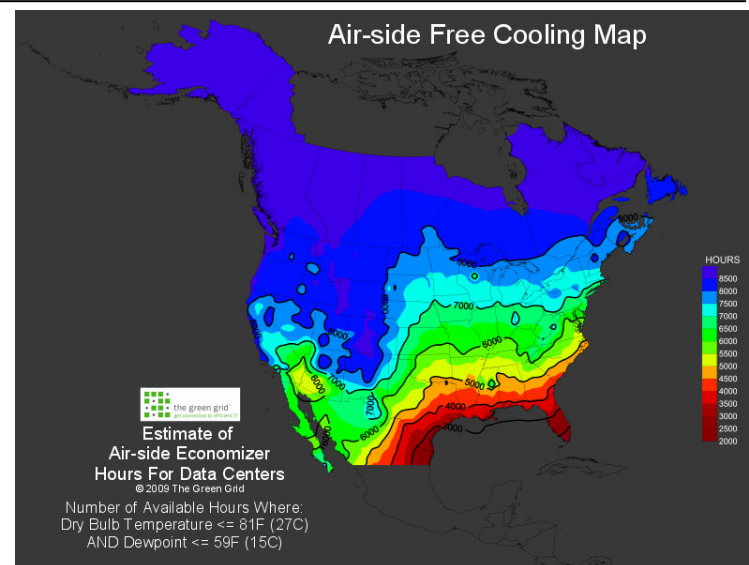


Photo source: CNET.com

Outside air cooling

- Many geographies can use outside air for cooling
 - Reduce or eliminate mechanical chillers
 - Moderate filtration recommended
- Yahoo Compute Coop data center (2010)
 - PUE 1.08 (with evaporative cooling)
 - Oriented for prevailing winds
 - 100% outside air cooling (no chillers)
 - Server inlet air typically 23 C
 - Use evaporative cooling above 26 C
 - Servers reach 26 – 30 C for 34 hours/year

Yahoo Compute Coop in Lockport, NY



Source: Chris Page, "Air & Water Economization & Alternative Cooling Solutions
–Customer Presented Case StudiesData Center Efficiency Summit, 2010

Open Compute Project



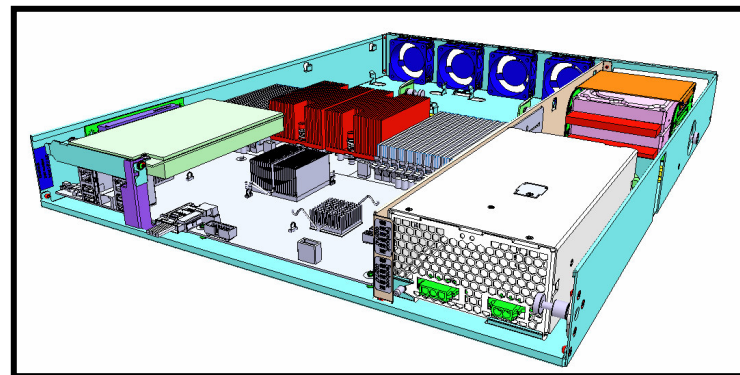
- Facebook released specifications and mechanical designs for data center and server (April 2011)

- Data center
 - Electrical
 - Mechanical
 - Racks
 - Battery cabinet

- Server
 - Chassis
 - Motherboard
 - Power supply

- PUE 1.07 (Oregon, December 2010)

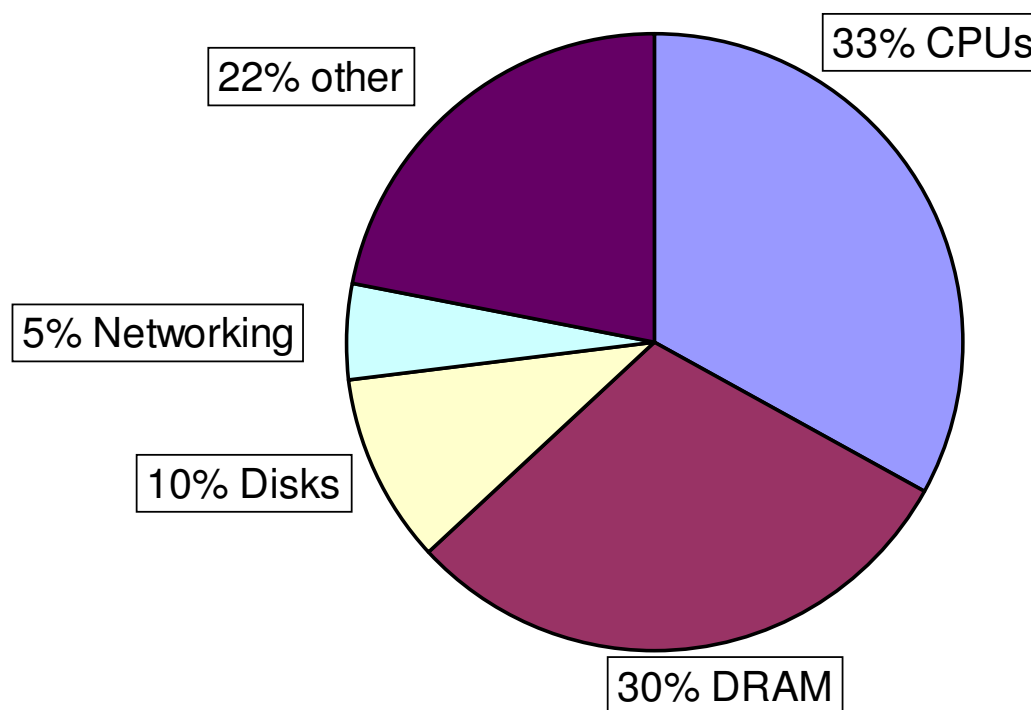
- <http://opencompute.org>



IT equipment

IT equipment

- Servers
- Storage
- Network

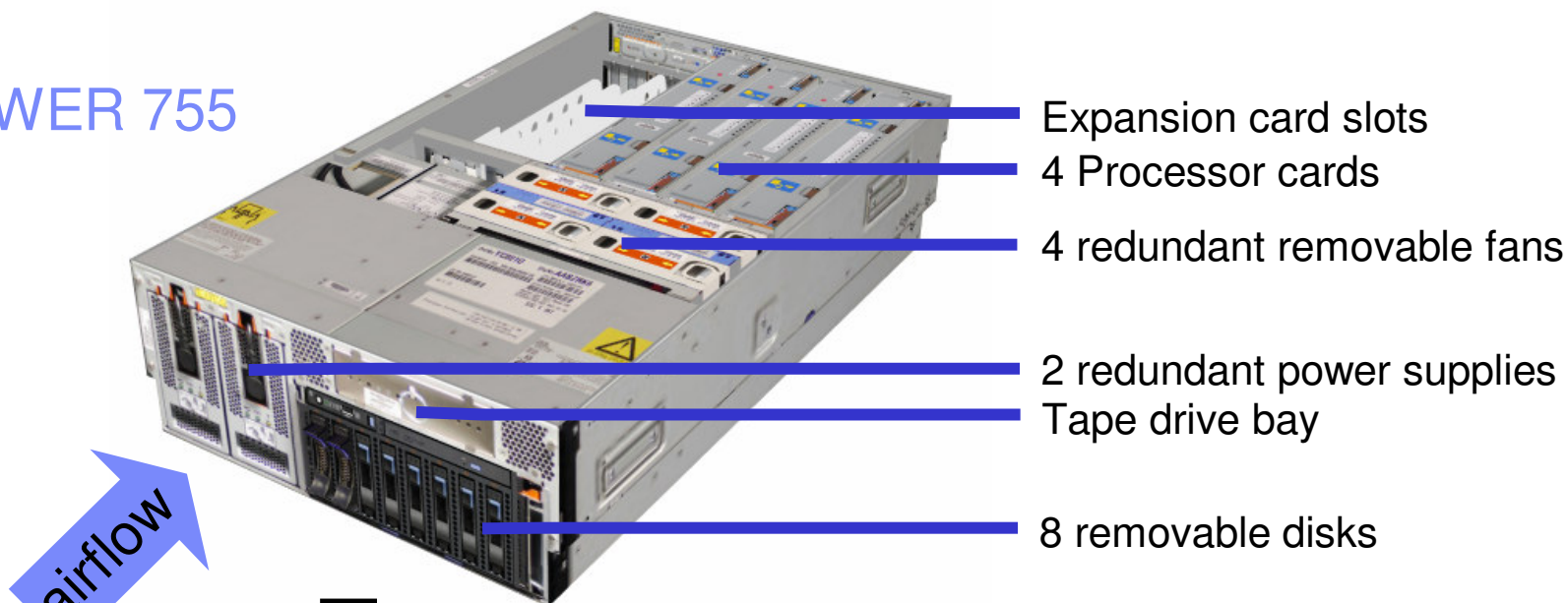


Server peak power by hardware component from a Google data center (2007)

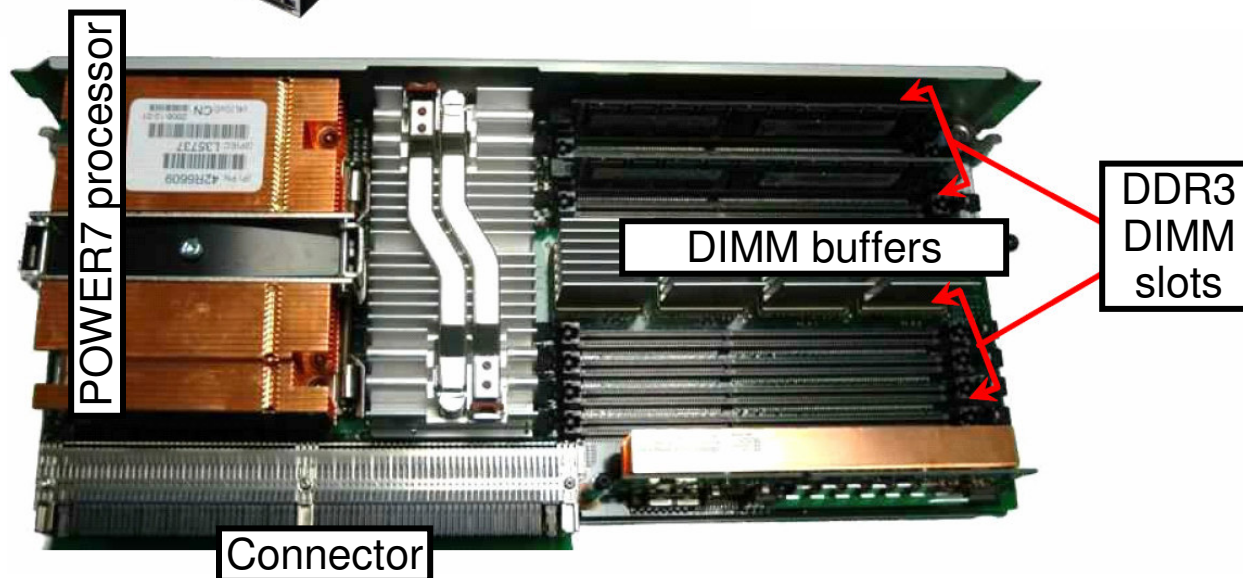
Source: Luiz André Barroso and Urs Hölzle, The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Morgan & Claypool, 2009.

Server components

IBM POWER 755

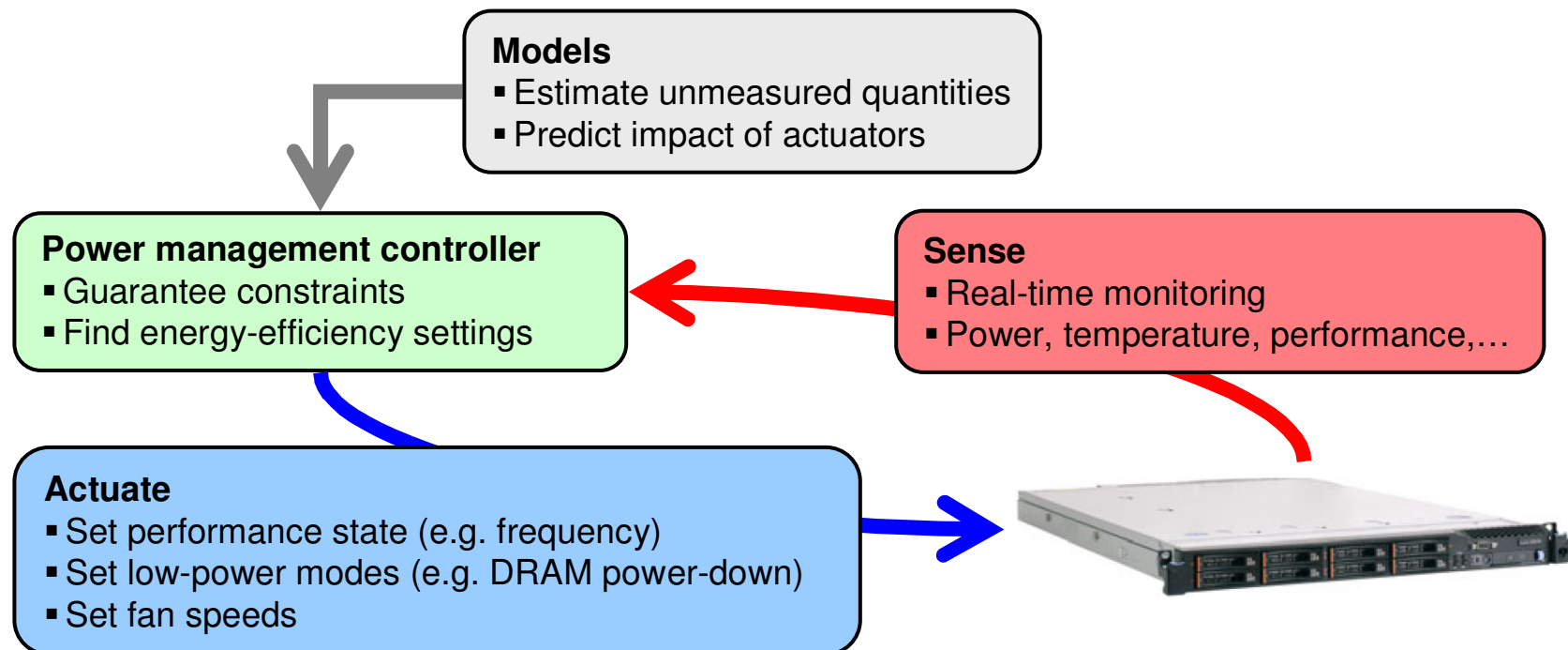


Processor card



Address variability in hardware and operating environment

- Complex environment
 - Installed component count, ambient temperature, component variability, etc.
 - How to guarantee power management constraints across all possibilities?
- Feedback-driven control
 - Capability to adapt to environment, workload, varying user requirements
 - Regulate to desired constraints even with imperfect information



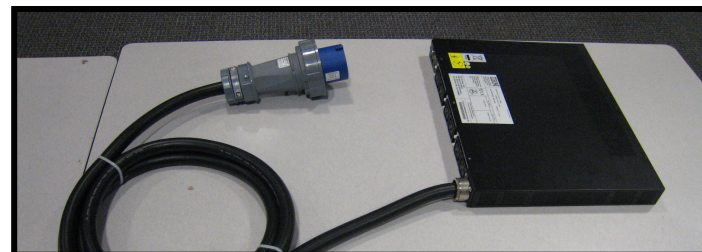
Sensors: temperature

- 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 PECL, 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.
 - Example: POWER7 processor has 44 digital thermal sensors per chip
 - Example: Nehalem EX has 9 digital thermal sensors per chip
 - Example: DDR3 specification has thermal sensor on each DIMM

Sensors: power

- AC power
 - External components – Intelligent PDU, SmartWatt
 - Intelligent power supplies – PMBus standard
 - Instrumented power supplies
- DC power
 - Most laptops – battery discharge rate
 - IBM Active Energy Manager – system power
 - Measure at VRM
- Within a chip (core-level)
 - Power proxy (model using performance counters)
- Sensor must suit the application:
 - Access rate (second, ms, us)
 - Accuracy
 - Precision
 - Accessibility (I2C, Ethernet)

Example: IBM DPI PDU+



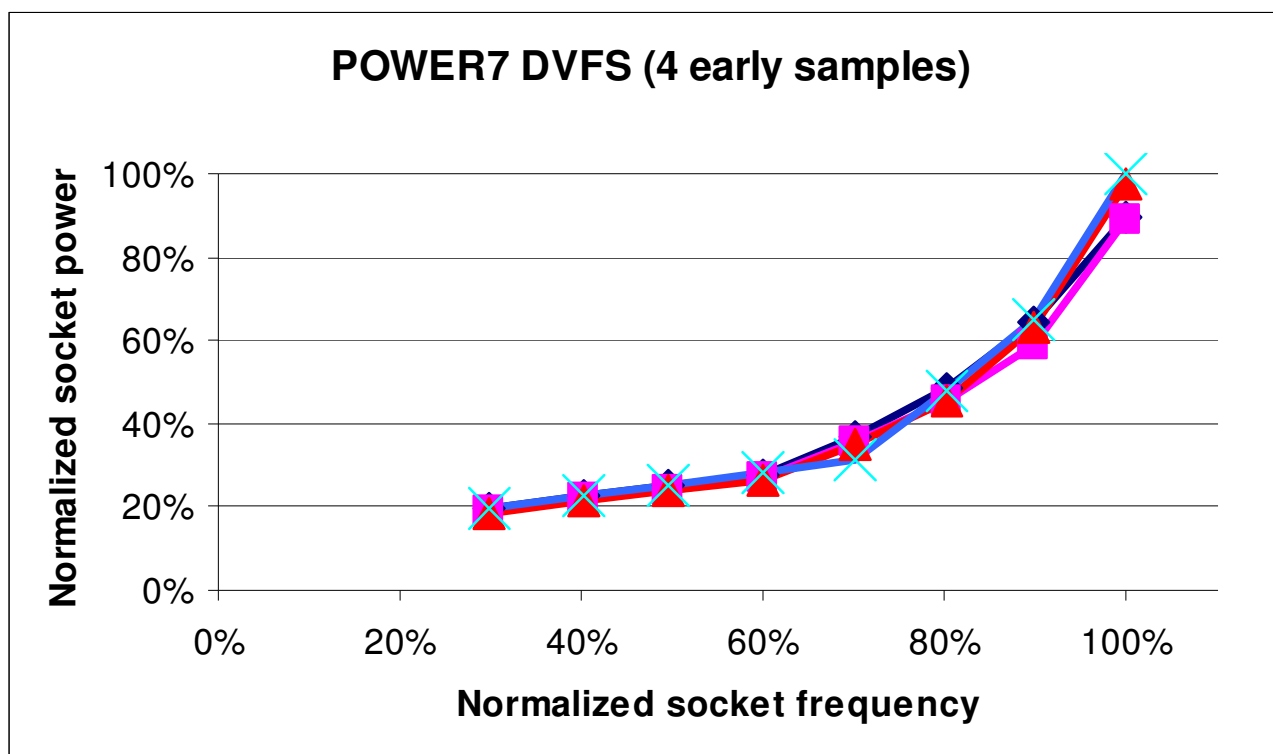
J1	J2	J3	J4	J5	J6	1A	3A	5A
201.3	201.3	198.9	198.9	199.7	199.7	Output Voltage(V)		
Min	Min	Min	Min	0.2	Min	Output Current(A)		
Min	Min	Min	Min	Min	Min	Output Power(W)		
0	0	0	0	0	0	PDU Watt Hour Usage		
0	0	0	0	1	0	Cumulative KW-Hrs		

Sensors: activity and performance

- '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
 - Example: Instructions fetched per cycle
 - Example: Non-halted cycles
- 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
 - Example: Transactions per Watt.

Actuators: microprocessor active states

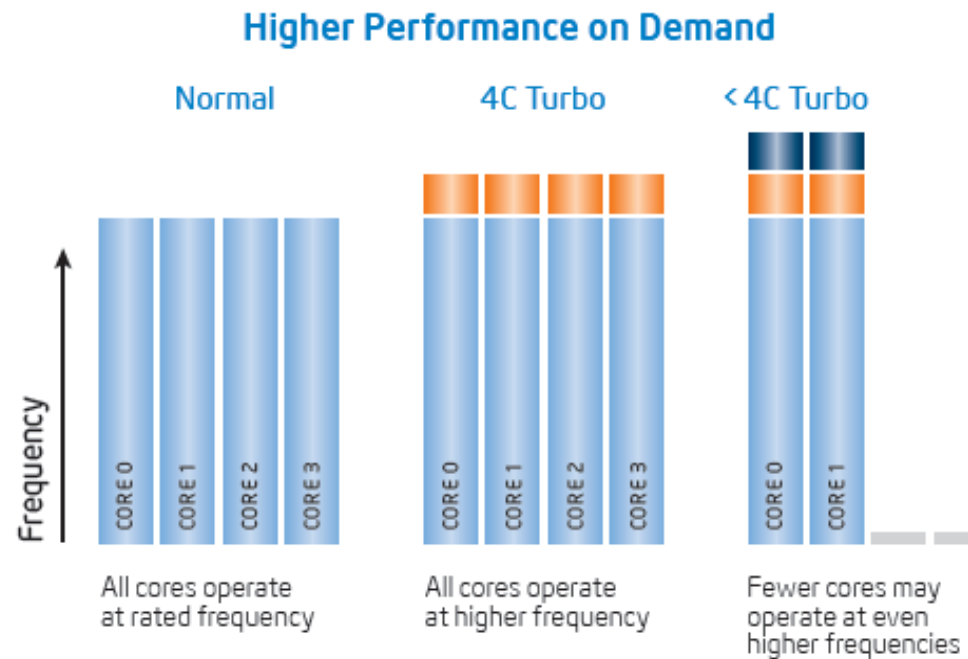
- Dynamic voltage and frequency scaling (DVFS) in modern server processors.
 - Called “P-states” (performance states)
- Today, voltage typically shared across cores on a chip. Cores have independent clocking and may use different frequencies.



Source: Karthick Rajamani, IBM

Actuators: turbo mode

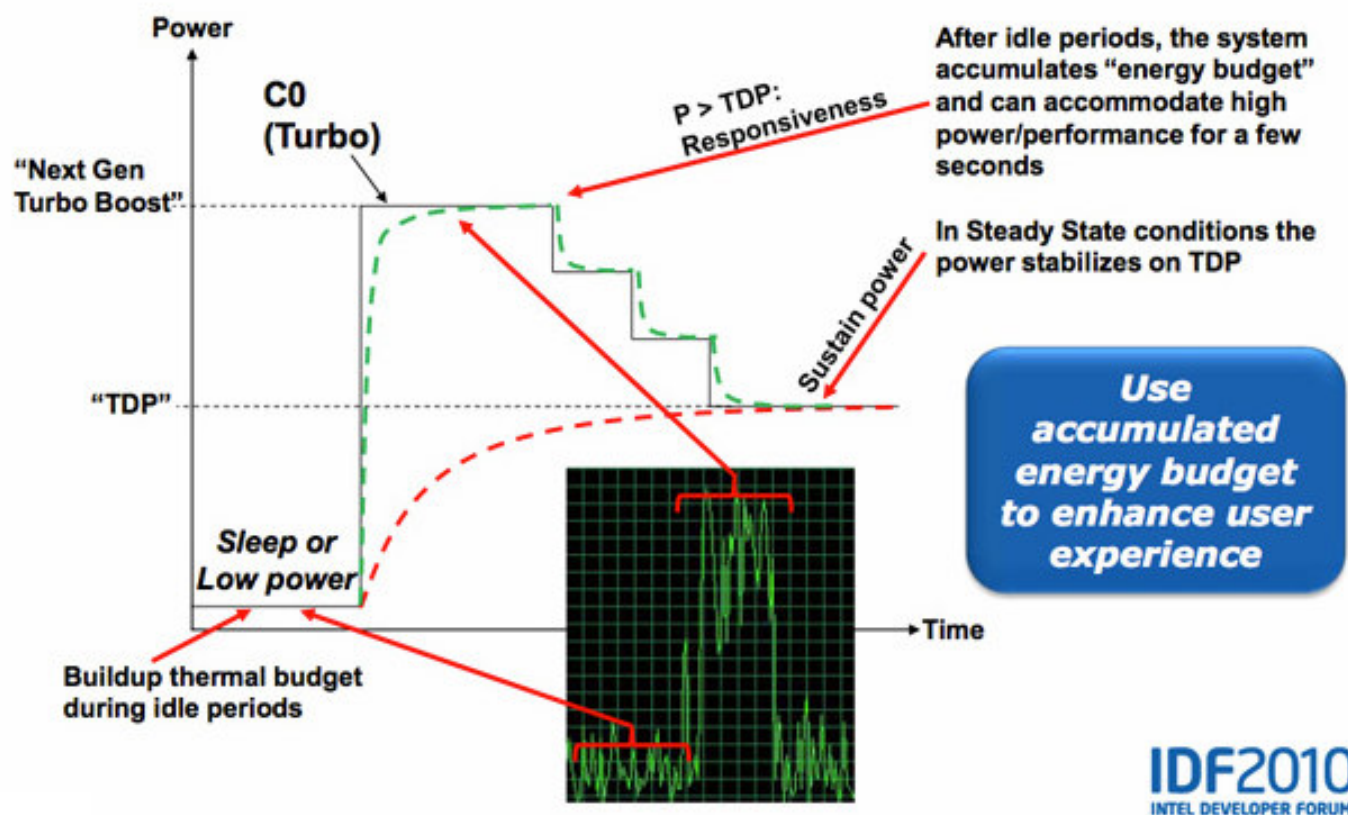
- Turbo frequencies are available on Intel and IBM microprocessors
 - Opportunistic performance boosting beyond nominal (guaranteed level)
 - When power and thermal headroom is available
- Example: Intel Turbo Boost when there is power headroom



Source: Intel Design Forum, 2010

Actuators: turbo mode

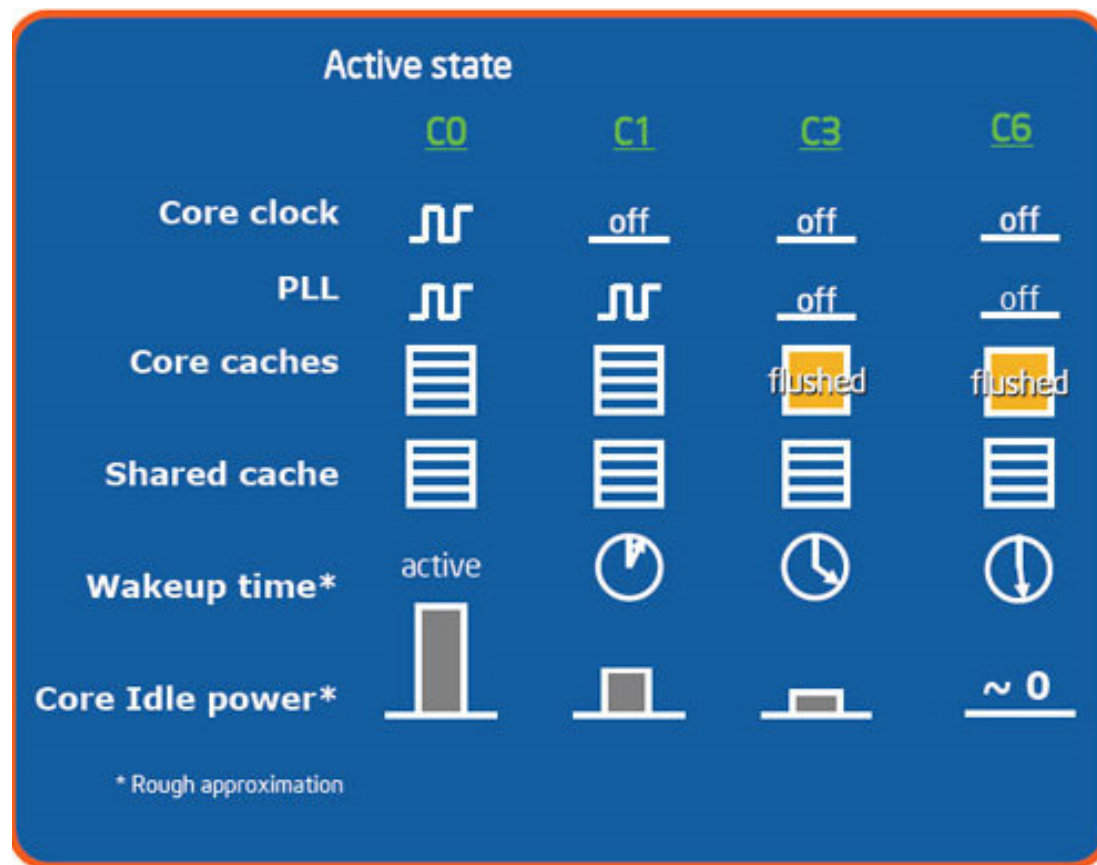
- Example of turbo boost when there is thermal headroom



Source: Intel Design Forum, 2010

Actuators: microprocessor idle states

- Use energy-efficient idle states when OS cannot schedule work
 - Waiting on IO
 - Waiting for server transactions
 - Request batching
- Trade-off power reduction with latency to invoke and wakeup
- Core-level and chip-level states
 - Once all cores are in a low state, chip can go to next lowest state
 - Example: Chip can reduce voltage to retention level, only when all cores are idle.

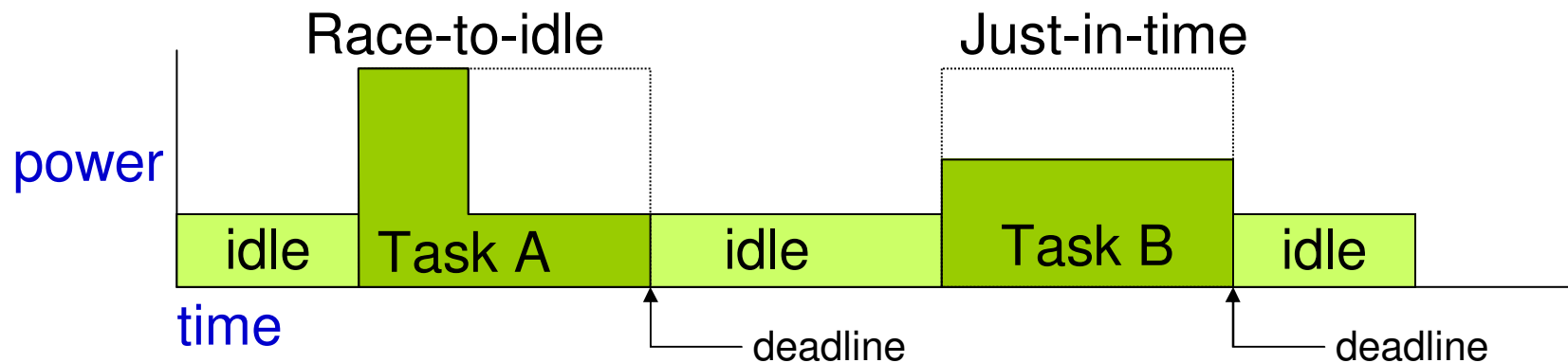


Intel Nehalem idle states

Source: Intel Design Forum, 2008

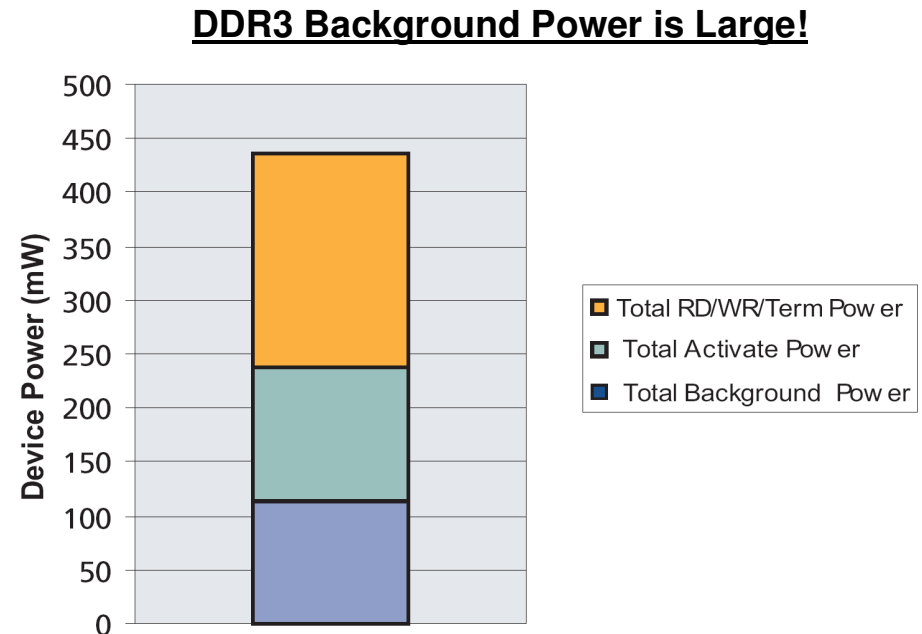
Race-to-idle vs. Just-in-time

- Two strategies to save energy. Both are useful.
- Race-to-idle
 - Complete work as fast as possible and go into lowest power idle state (or off)
 - Concern: wake-up time can be long (minutes to boot server and reload cache)
 - Opportunity: more granular idle modes with different wake-up times
 - Example: OS idle loop using idle states
- Just-in-time
 - Complete work slowly to just meet deadlines (or service level agreement)
 - Useful if running CPU faster does not complete work faster (memory-bound task)
 - Example: Use DVFS to CPU speed to memory bandwidth
 - Useful when idle modes are not available or wake-up time is too long



Actuators: DRAM

- Memory consumes power as soon as plugged in
 - Idle power no longer negligible
 - Idle power increases with DIMM size
- Power increases with access rate to memory
- Power modes
 - Powerdown
 - Self-refresh
- Power down
 - Power is ~10-20% less than active standby
 - Wake-up penalty (7-50ns)
 - Power down groups of ranks within a DIMM
 - DRAM Idle, IO Circuits off, Internal Clock Off, DLL (Dynamic Loop Lock) Frozen
 - Needs Refresh
- Self Refresh
 - Power ~60%-70% less than active standby
 - Wake-up penalty: 600 – 1300 ns
 - Put whole DIMM into self-timed refresh
 - Hub chip also in lower-power state
 - DRAM Idle, IO Circuits off, Internal Clock Off, DLL Off
 - Needs No Refresh



* Fig. from Micron TN-41-01

Source: Kenneth Wright (IBM) et al., “Emerging Challenges in Memory System Design”, tutorial, 17th IEEE International Symposium on High Performance Computer Architecture, 2011.

Thermal constraints

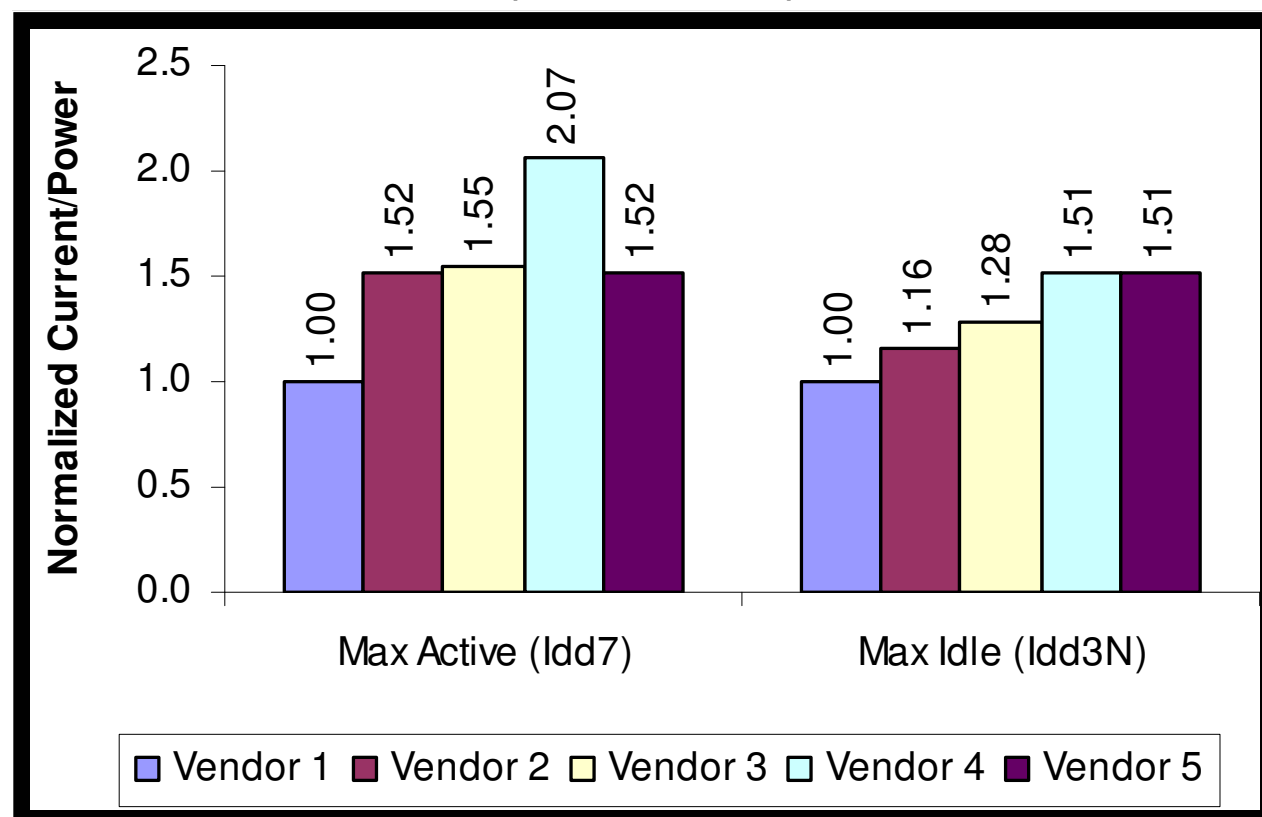
- JEDEC defines max allowable temperature for DRAM
 - Source: Micron, TN-00-08, *Thermal Applications*
 - Functional limit: 95 C (industrial applications)
 - Reliability limit: 110 C (prevent permanent damage)
- Inputs:
 - System thermals (air flow, inlet air, fan ramp, preheat)
 - DRAM current / power (self heat)
- Constraints / Worst case:
 - Fan fail (20-25 C increase)
 - DIMM position (depending on air flow, etc.)
 - Inlet air (Processor speed, processor heat sink)
- Solutions:
 - Use low power modes
 - Throttling (Limit max bandwidth → limits active current)
 - Double refresh ($\geq 85\text{C}$)

Source: Kenneth Wright (IBM) et al., “Emerging Challenges in Memory System Design”, tutorial, 17th IEEE International Symposium on High Performance Computer Architecture, 2011.

Power variability

- Power can vary due to manufacturing of same part (processor leakage power)
- Power consumption is different across vendors

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

Important concepts for energy management

Energy proportional computing

- Popularized by Google (2007)
- Definition: Energy consumed is proportional with work completed
 - Ideal: When no work is performed, consume zero power
 - Reality: When servers are idle, their power consumption is significant
- Today, it is still not uncommon for idle servers to consume 50% of their peak power.
 - Many components (memory, disks) do not have a wide range for active states.
 - Power supplies are not highly efficient at every utilization level.
- Can apply to data centers, servers, and components

Energy proportional computing

“The Case for
Energy-Proportional
Computing,”
Luiz André Barroso,
Urs Hölzle,
IEEE Computer
December 2007

Energy Efficiency =
 $\text{Utilization} / \text{Power}$

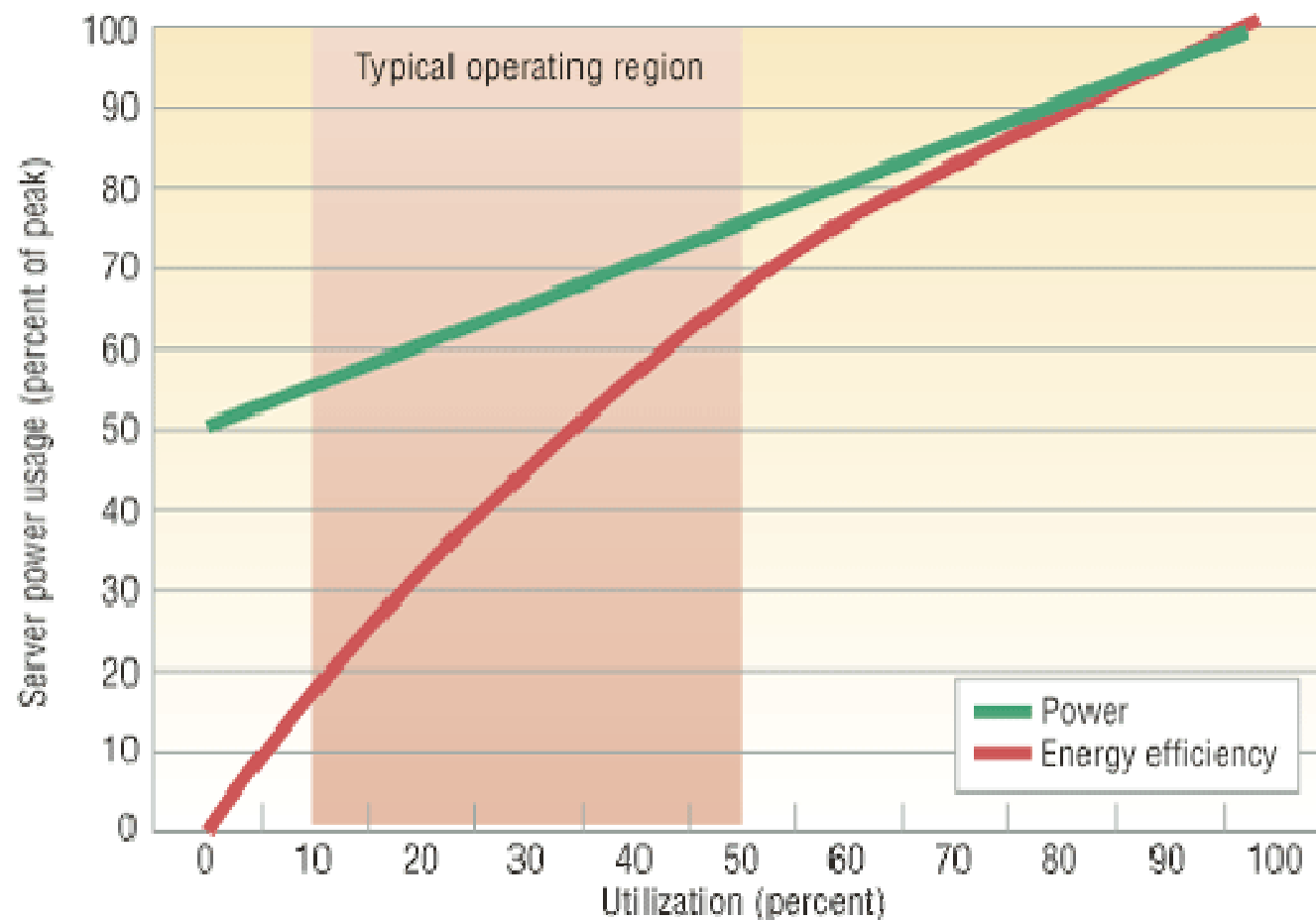


Figure 2. Server power usage and energy efficiency at varying utilization levels, from idle to peak performance. Even an energy-efficient server still consumes about half its full power when doing virtually no work.

Energy proportional computing

"The Case for
Energy-Proportional
Computing,"
Luiz André Barroso,
Urs Hölzle,
IEEE Computer
December 2007

Averages
found
outside
the cloud

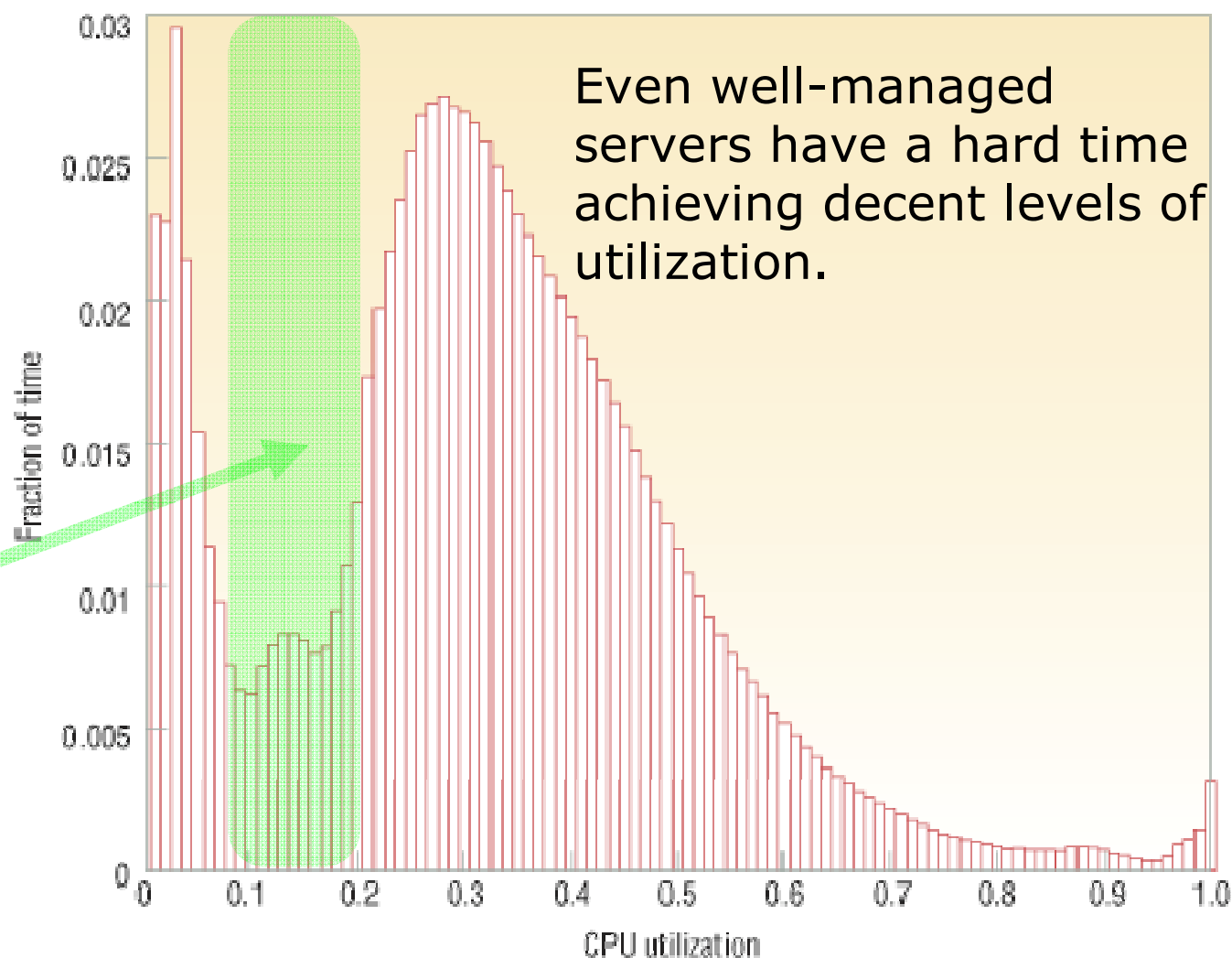


Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum

Virtualization – opportunities for power reduction

- Virtualization enables more effective resource utilization by consolidating multiple low-utilization OS images onto 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 that multiple, less energy-efficient, under-utilized systems can be replaced with fewer, more energy-efficient, higher performance systems for
 - A net reduction in energy costs.
 - Lower infrastructure costs for power delivery and cooling.
- More in [cloud section](#)



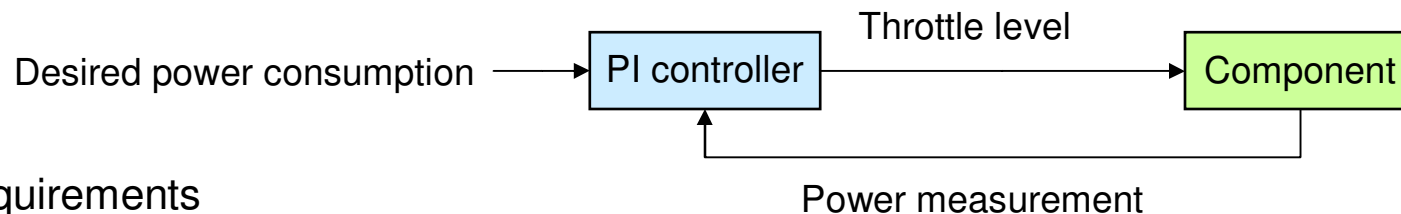
Workload-optimized systems

- Specialize hardware/software to run a workload efficiently
 - Remove components required for general purpose computing
 - Size components (memory/processor/network) to workload
 - ASIC
 - FPGA
 - Custom instruction sets,
 - Accelerators (hardware + SW libraries)

- Examples in [emerging technologies section](#)

Power capping

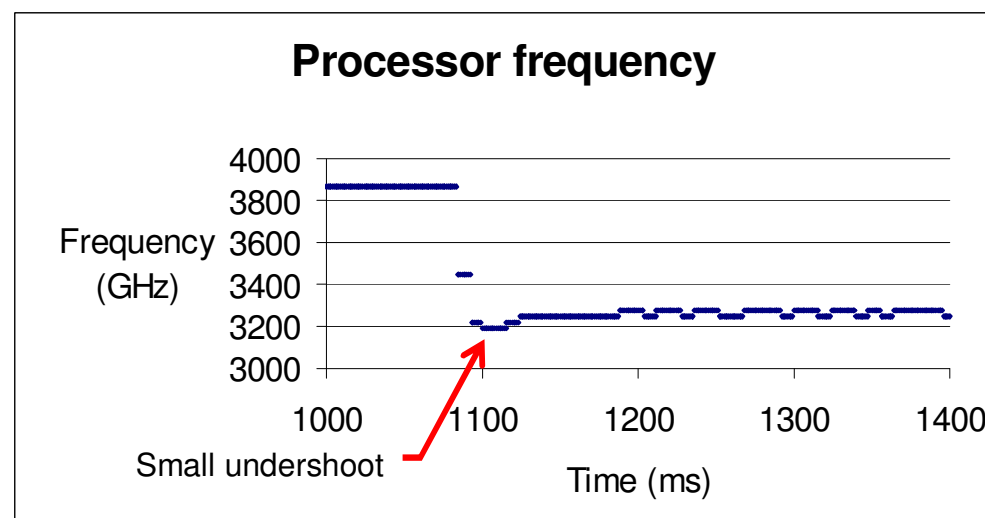
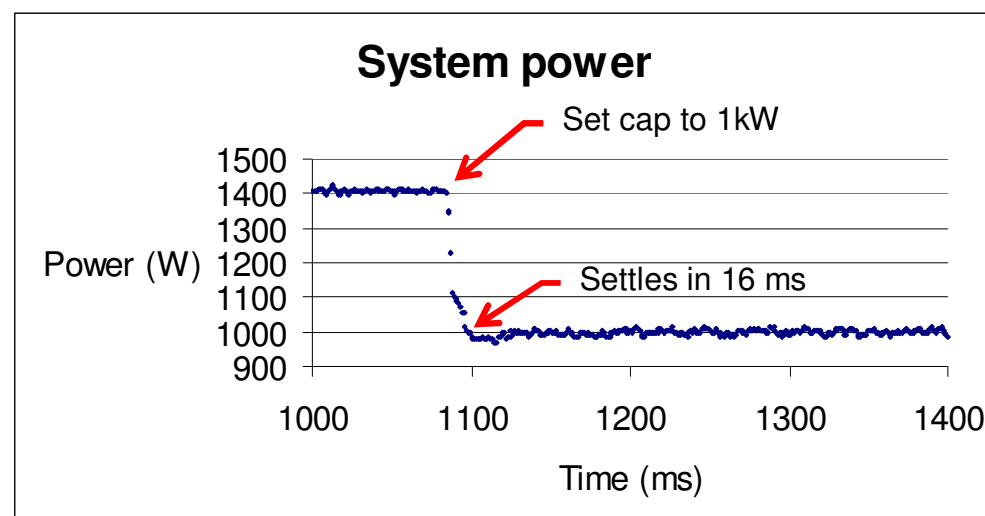
- Power capping is a method to control peak power consumption
 - High power use → slow down; Low power use → speed up
 - Can be applied at many levels: components, servers, racks, data center
 - Control-theoretic approach



- Requirements
 - Precision measurement of power
 - Measurement error translates to lost performance
 - Components with multiple power-performance states
 - Example: microprocessor voltage and frequency scaling
- Impact
 - Power capping provides safety @ worst-case power consumption
 - Allows IT equipment to oversubscribe available power (better performance @ typical-case power)
 - Stranded power is reduced (lowering cost)
 - Power delivery is designed for typical-case power (lowering cost)

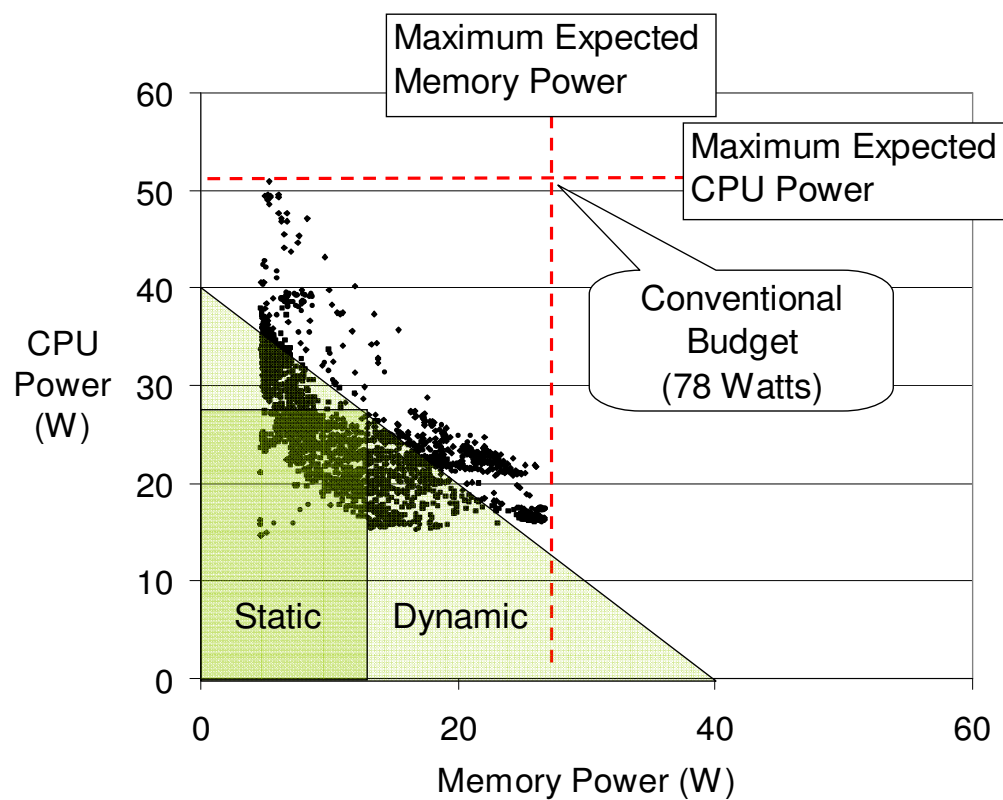
Power capping – Example

- IBM Power 750 Express server
- Caps when redundant power supply fails or customer sets power cap target
- Every 8 ms, measure system power and adjust processor voltage and frequency to meet power cap
- Drop power cap from 2 kW to 1 kW
- Settles to 2% of target in 16 ms (2 control intervals)



Power shifting

- Set a power budget on every component to control aggregate power
- Shifting = dynamically adjusting power budgets to improve aggregate performance
 - Requires performance measurement from components

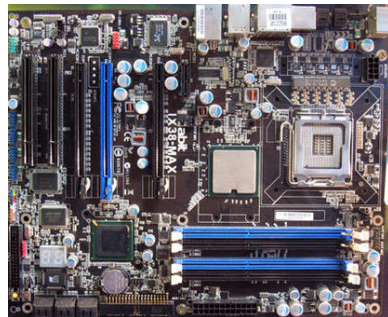
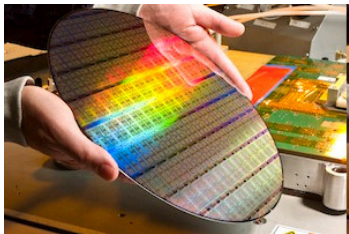


- Example: power shifting across CPU and memory for PPC 970 computer
- Points show execution intervals of many workloads with no limit on power budget
- “Static” encloses unthrottled intervals for 40 W budget: 27 W CPU, 13 W Memory
- “Dynamic” encloses unthrottled intervals for 40 W budget and power shifting
 - Better performance than static design for 40 W
 - Lower cost than conventional 78 W power supply design

Shift power to where it is consumed efficiently

■ Opportunities

- Intra-chip: shift between function units, cores, cores \leftrightarrow caches
- Intra-node: shift between processors, processors \leftrightarrow DRAM, leakage/fans
- Intra-rack: shift between nodes, storage \leftrightarrow compute, disaggregated DRAM
- Intra-data center: cross-node optimization (placement, migration, consolidation)
- Across data centers: time shifting, power arbitrage, enhanced reliability



Take away

- Data center and server power management addresses many problems
 - Huge costs for cooling and power delivery
 - Constraints on achievable performance
 - Constraints on data center capacities limiting IT growth
- Governments, data center operators, and IT vendors are all engaged
 - New benchmarks and metrics (PUE, SPECPower)
 - New standards under development
- In last 10 years, we have seen considerable innovation in data center design
 - Containers, outside air, etc.
- Data centers and servers are becoming more instrumented
 - Many sensors and actuators allow for adaptive, flexible behavior
 - Power capping, shifting
 - Virtualization and dynamic consolidation
- Many techniques for managing power and cooling
 - Consolidation, workload-optimized systems, capping, shifting, etc.

Selected Reading

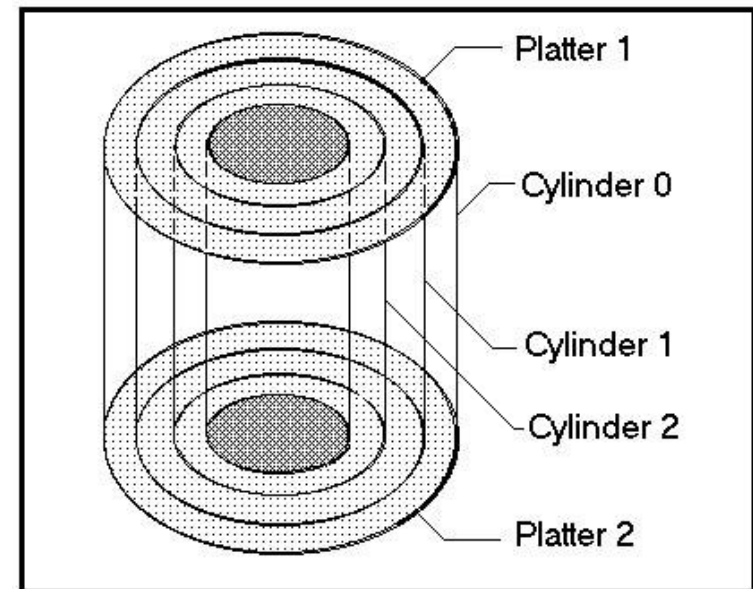
- What is the problem?
 - US EPA, Report to Congress on Server and Data Center Energy Efficiency, August 2007.
http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf
 - American Society of Heating, Refrigerating and Air-Conditioning Engineers <http://www.ashrae.org>
 - James Hamilton's blog <http://perspectives.mvdirona.com/>
- Metrics
 - The Green Grid <http://www.thegreengrid.org>
 - W. Feng and K. Cameron, "The Green500 List: Encouraging Sustainable Supercomputing", IEEE Computer, December 2007. <http://www.computer.org/portal/web/csdl/doi/10.1109/MC.2007.445>
- Data center and servers
 - Luiz André Barroso and Urs Hölzle, The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Morgan & Claypool, 2009.
 - Michael Floyd et al., "Introducing the Adaptive Energy Management Features of the POWER7 Chip", IEEE Micro, March/April, 2011.
- Energy proportional computing
 - Luiz André Barroso, and Urs Hölzle, "The Case for Energy-Proportional Computing," IEEE Computer, December 2007.
- Power capping and shifting
 - Charles Lefurgy, Xiaorui Wang, and Malcolm Ware, "Power capping: a prelude to power shifting", Cluster Computing, Springer Netherlands, November 2007.
 - W. Felter, K. Rajamani, T. Keller, C. Rusu, "A Performance Conserving Approach for Reducing Peak Power in Server Systems", ICS 2005.

Storage

Storage power problem looming right behind memory power (and it's already dominant in some environments)

- Problem: Many enterprise data centers spend upwards of 40% of their IT power on storage

- SAS (15K, FC, ...) drives are fastest, but highest power and cost
- Optimizing performance can drive low resource utilization (spread data across many spindles)
- Other parts of the data center are becoming more energy proportional but storage is not
- Optimizing for better energy efficiency can potentially reduce performance



- Standards bodies are including power and energy metrics along with performance
 - Storage Performance Council (SPC)
 - SNIA
 - EPA



More customer attention to power and energy saving features at purchase time

Storage power problem looming right behind memory power (and it's already dominant in some environments)

- Opportunities:
 - Move away from high-cost, high-power enterprise SAS/FC drives
 - Consolidation (fewer spinning disks = less energy, but less throughput)
 - Hybrid Configurations (Tiering/Caching): Replace power-hungry SAS with SATA (for capacity) and flash/PCM (for IOPS)
 - SATA consumes ~60% lower energy per byte
 - Flash can deliver over 10X SAS performance for random accesses
 - Flash has lower active energy and enables replacing SAS with SATA and spindown (by absorbing I/O activity)
 - Issue: But what data should be placed in what storage technology and when?
 - Opportunistic spindown
 - Write offloading
 - Deduplication/Compression
- Challenge: IO-intensive applications have strict latency and bandwidth requirements



Storage Background – What Customers Want

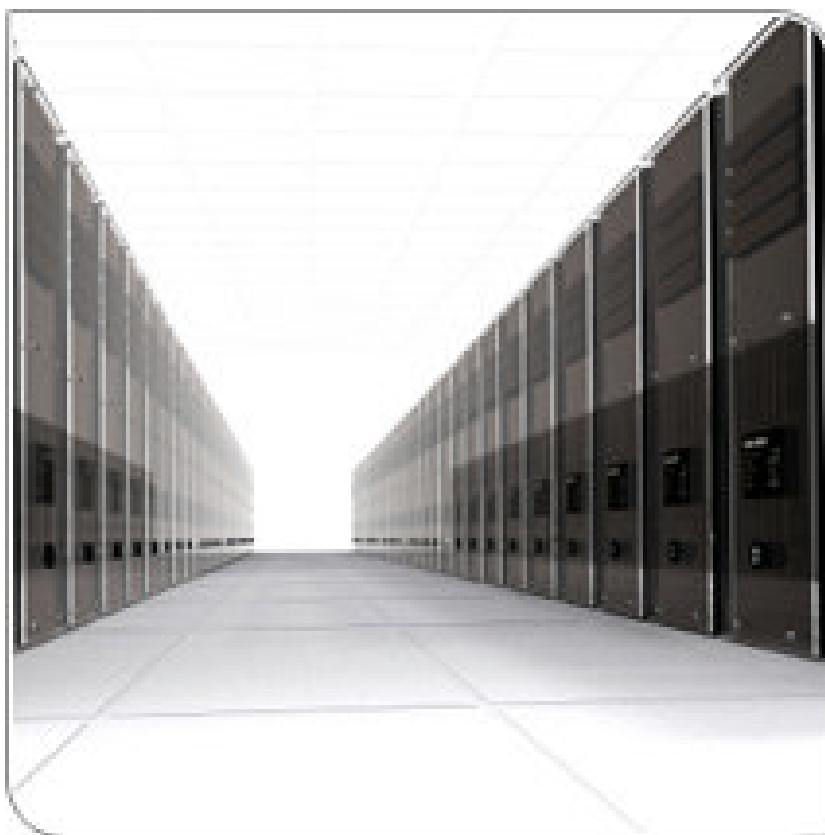
Availability

Reliability

Redundancy

Accessibility

Fault-tolerance



Low-Latency

Throughput

Capacity

Feature-rich
(extensible management
options via simple software)

All at low cost, low power, and low energy!!!

We're moving in the right direction

Major trend: High-speed enterprise drives → high-capacity SATA drives and SSDs

- SATA drives have significantly lower price per GB, and also Watt per GB

Yesterday

Excessive Power:

- No Spin-down
- Fans and controllers

Costly:

- 15K RPM SAS
- Wasted capacity

Wasted Capacity:

- RAID Configuration
- Short-stroking



Today

Reduced Power:

- Aggressive Spindown
- System power mgm't
- Flash to absorb I/O

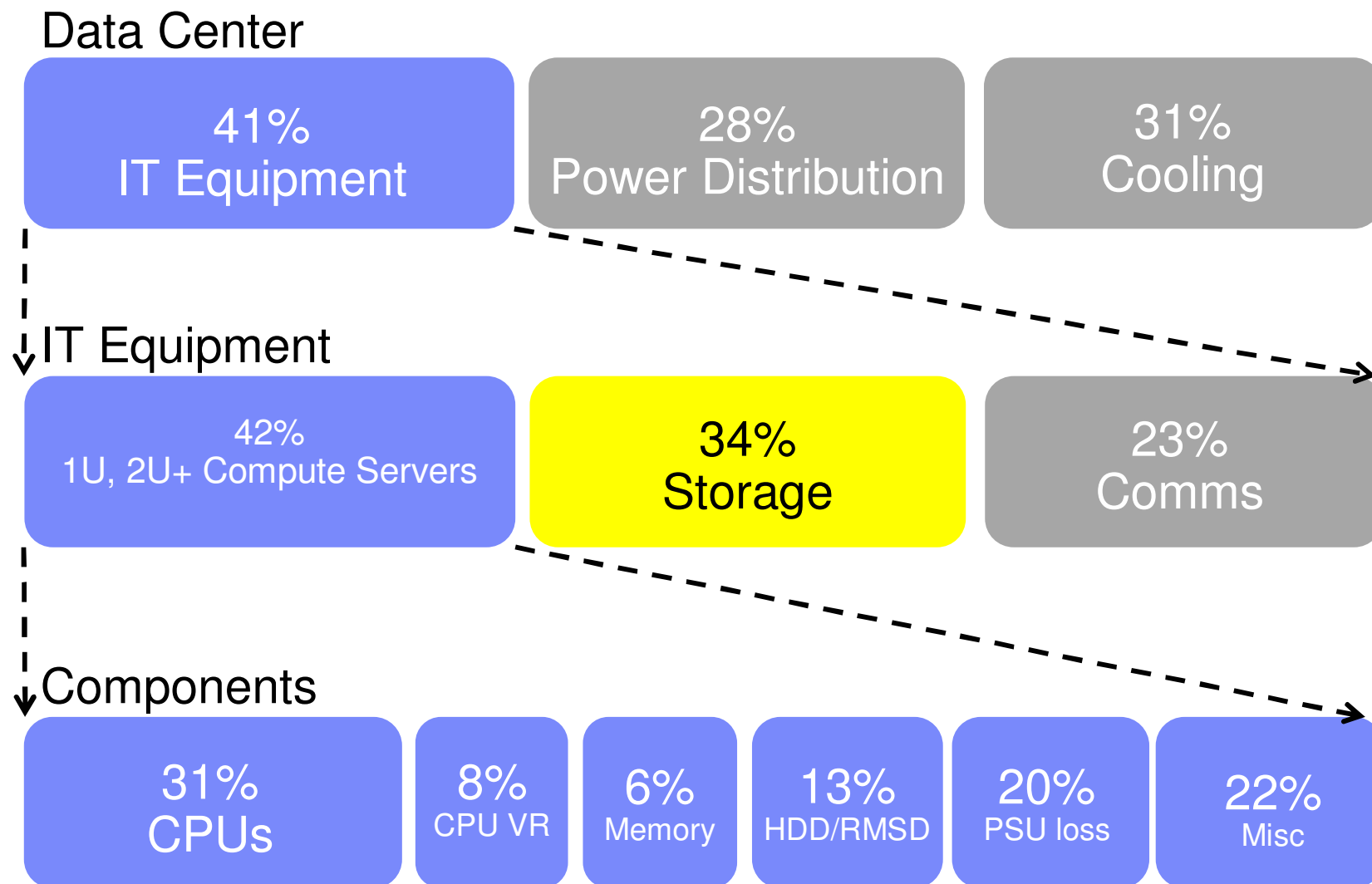
Reduced Cost:

- SATA + SSD
- Storage virtualization

Increased Capacity:

- Dense SATA drives
- No short-stroking
- Deduplication

Where does the power go?



Source: Dell Measurements, presented at VMWorld 2007

Storage Growth

“...Total disk storage systems capacity shipped reach 3,645 petabytes, growing 54.6% year over year.”

– International Data Corporation (IDC) about 2Q10

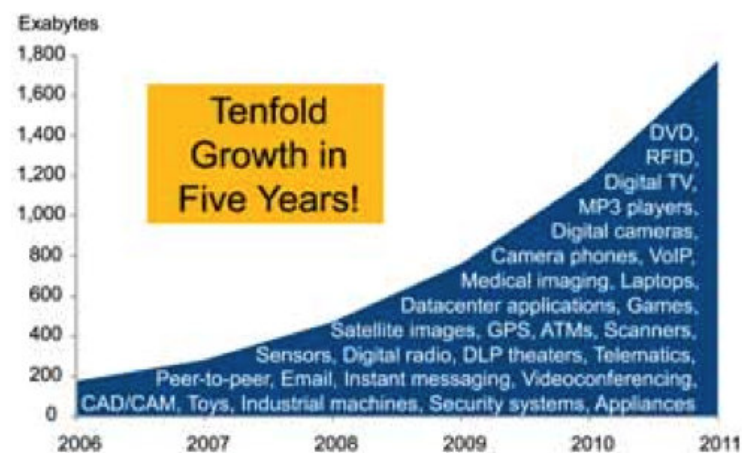
source (press release):

<http://www.idc.com/about/viewpressrelease.jsp?containerId=prUS22481410§ionId=null&elementId=null&pageType=SYNOPSIS>



Figure 1

Digital Information Created, Captured, Replicated Worldwide



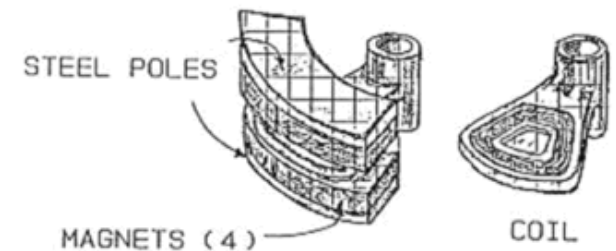
Source: IDC, 2008

<http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>

© 2011 IBM Corporation

Storage Power Capping

- Goal: Add more storage for the same power capacity
- Massive Arrays of Idle Disks (MAID)
 - Increase the number of spindles for better throughput but manage for power and energy
 - Much academic work and included in some products
- Disk Acoustic Modes (i.e., slowing down the seek arm)
 - Seek power is the result of current consumption through the voice coil motor during positioning of the read/write heads
 - Reducing the current through the voice coil motor reduces the actuator arm speed
 - Increases seek times (lower performance)
- Throttling of I/O
 - Reducing the amount of work sent to the storage system to keep drives in idle states
- Lower Power States during Idle periods
 - Placing drives in standby mode when possible



(a) FLAT COIL



Storage Energy Saving

- Goal: Reduce energy consumption and save money on energy costs
- Massive Arrays of Idle Disks (MAID)
 - Much academic work and included in products
- Replace Storage Media and/or hybrid storage configurations
 - Removing mechanical disks altogether or placing them in lower power states



all-SAS
100% energy



all-SATA
45% energy



all-SATA + SSD
16% energy

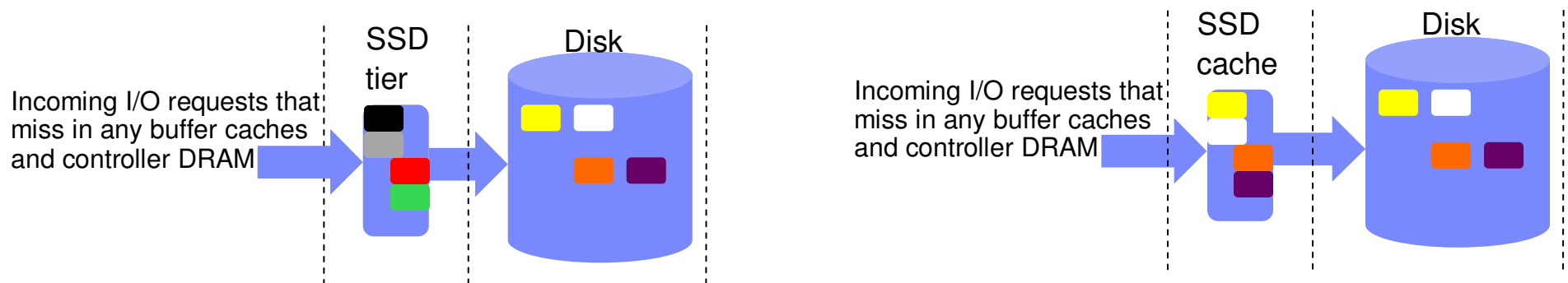
Tiering vs. Caching

■ Tiering

- Tiers represent the trade-off between performance, capacity, and power
- Data lives in either one tier or the other, but never simultaneously occupying both
 - Capacity increases slightly
- Some are proposing more than 2 tiers
 - e.g., flash→15k→10k or flash→disk with no spindown→disk with spindown
- Decisions about where data lives are made at time intervals from 30 minutes to hours typically
 - Due to this timescale, tiering can't keep up with fast changes in workloads

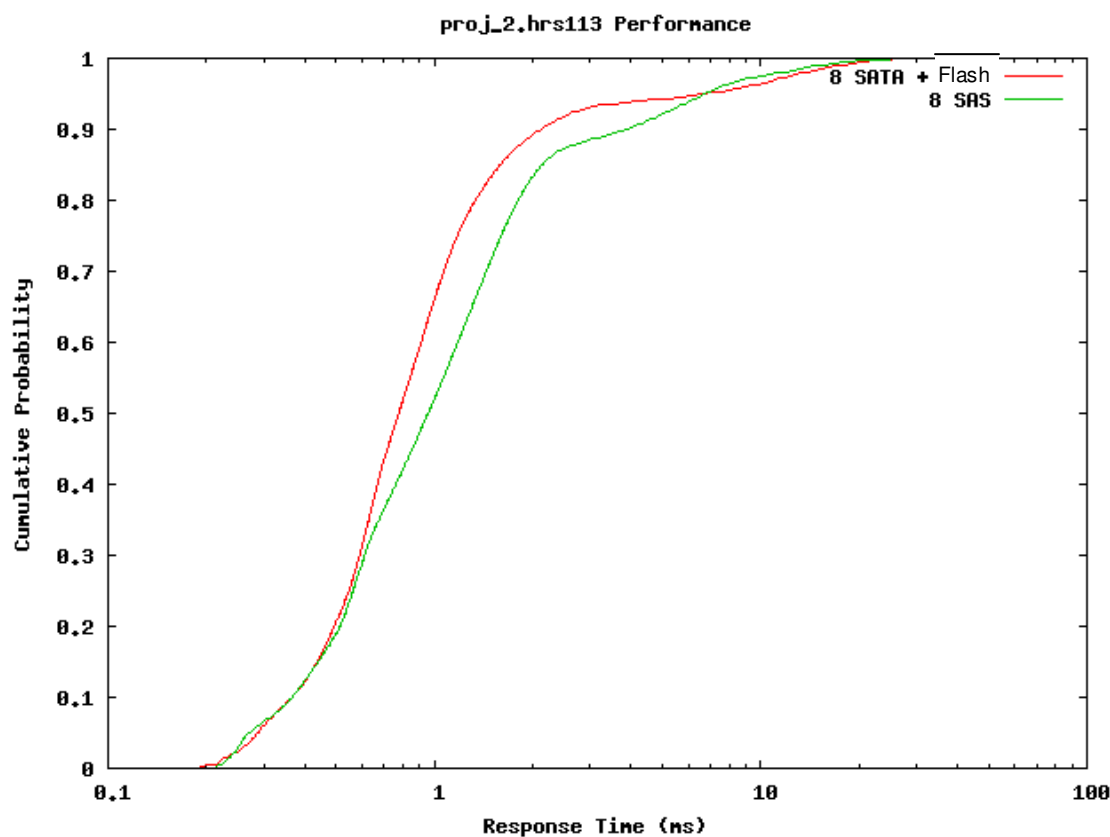
■ Caching

- Cache's purpose is to increase performance
- Data in the cache is a copy of data in the other tier or (in the case of a write) is about to be copied back to the other tier
- Decisions about where data lives are made at each access (read or write)
 - Responsive to immediate changes in workload



Performance Need Not Suffer

Empirical experiments 8-disk RAID-6 SAS array vs. 8-disk RAID-6 SATA array + SSD cache




For the medium-duty workloads we used, flash...

- offers more spindown opportunity
- matches or exceeds the performance of SAS

To Spin Down or Not to Spin Down...(workloads matter)

- Some environments/workloads are already tuned with the right spindown approach
 - e.g., backup/archive (MAID)
- Some environments/workloads cannot tolerate any multi-second latency from a spinup delay
 - e.g., OLTP
- Spindown is appropriate for medium-duty workload environments
 - e.g., email, virtualization, filers



Caution:
not to exceed rated
spindown cycles

Software technologies

- Thin-provisioning
 - Software tools to report and advise on data management/usage for energy- and capital-conserving provisioning
- Deduplication
 - Either at the file or block level
 - Ensures only one copy of data is stored on disk (e.g., replicate copies are turned to pointers to the original)
- Storage virtualization
 - Allows for more storage systems to be hidden behind and controlled by a central controller that can more efficiently manage the different storage systems
 - Abstract physical devices to allow for more functionality
 - Enables powerful volume management

Emerging Technologies

- Phase-change memory (PCM)
- STTRAM, MRAM
- And other Storage Class Memories (SCM)

Metrics and Benchmarks

- Storage Performance Council (SPC)
 - SPC-1C/E (2009)
 - For smaller storage component configurations (no larger than 4U with 48 drives)

<http://www.storageperformance.org/home>

- SPC Benchmark 1/Energy

- Idle Test
- IOPS/Watt
- Extends SPC-1C/E to include more complex storage configurations

The logo for the Storage Performance Council, featuring the words "Storage Performance Council" in a bold, sans-serif font. "Storage" is in blue, "Performance" is in orange, and "Council" is in blue.

- SNIA Emerald (in development)
 - Idle power
 - Maximum response time
 - Performance (IOPS) per Watt
 - <http://www.snia.org/home/>



- EPA (EnergyStar Data Center Storage Product Specification) (in development)
 - Power Supply Efficiency
 - Active and Idle State
 - Power Management Requirements
 - Not up to date due to confidentiality until released
 - http://www.energystar.gov/index.cfm?c=new_specs.enterprise_storage



© 2011 IBM Corporation

Emerging Industry Solutions

- IBM Storwize V7000
 - SAN Volume Controller Software (SVC)
 - Virtualizes SANs and eases storage and volume management
 - EasyTier Software
 - For hybrid storage systems (flash-based SSDs and mechanical disks), dynamically moves data between SSD tier and disk tier for best performance
- IBM DS8800
 - SVC
 - EasyTier
- EMC VNX
 - tiering + spin-down support
- Sun Oracle Exadata Database Storage Machine
 - Flash caching
 - Volume management for easy scalability

Take-Aways and References

- Data is growing and the storage systems to satisfy the capacity demand are not energy-proportional
- More of the data center is becoming more energy-efficient, and storage energy consumption is becoming dominant
- Storage Power Capping and Energy Saving Techniques
 - Hybrid storage architectures
- Metrics and benchmarks adopted by the industry are beginning to drive this issue and the importance of focus in this area
- References:
 - Dennis Colarelli and Dirk Grunwald. Massive arrays of idle disks for storage archives. pages 1–11. In Proceedings of the 2002 ACM/IEEE International Conference on Supercomputing, 2002.
 - Charles Weddle, Mathew Oldham, Jin Qian, An-I Andy Wang, Peter Reiher, and Geoff Kuenning. PARAID: a gear-shifting power-aware raid. *ACM Transactions on Storage*, 3(3):33, October 2007.
 - D. Chen, G. Goldberg, R. Kahn, R. I. Kat, K. Meth, and D. Sotnikov, Leveraging disk drive acoustic modes for power management. In Proceedings of the 26th IEEE Conference on Mass Storage Systems and Technologies (MSST), 2010.
 - Wes Felter and Anthony Hylick and John Carter. Reliability-aware energy management for hybrid storage systems. To Appear in the Proceedings of the 27th IEEE Symposium on Massive Storage Systems and Technologies (MSST), 2011.

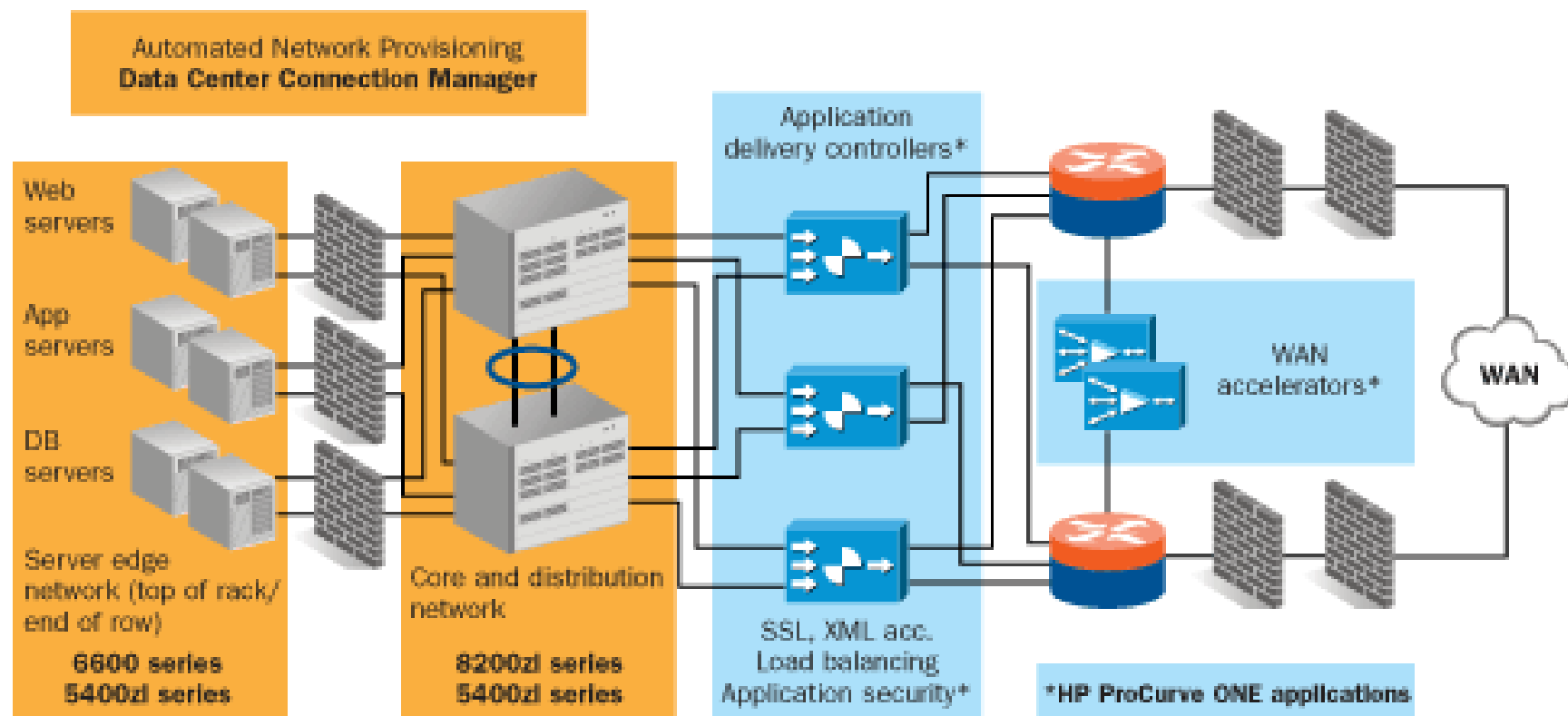


Networking



What's in a network?

Solution architecture



Source: HP

Network power in the data center — Overview

- Includes Ethernet LANs and Fibre Channel SANs
 - Virtually no published work on SAN power modeling/management!
- Most power is in switches
 - Network interface cards (NICs) & appliances (firewalls, etc.) are counted as servers
 - Router power is usually small due to lower number of routers than switches
- Small (10-20% of data center IT power) but growing
 - Replacement of direct-attached storage with networked storage (often driven by virtualization)
 - More dynamic environments (e.g. VM migration) demand more bandwidth
 - Emerging bandwidth-intensive analytics workloads
- Currently not energy-proportional
 - Switch power is proportional to number of active ports (if you're lucky)
 - Reduces the overall proportionality of the data center
- Energy is a small fraction of total network cost
 - Few modifications can be justified based on energy savings

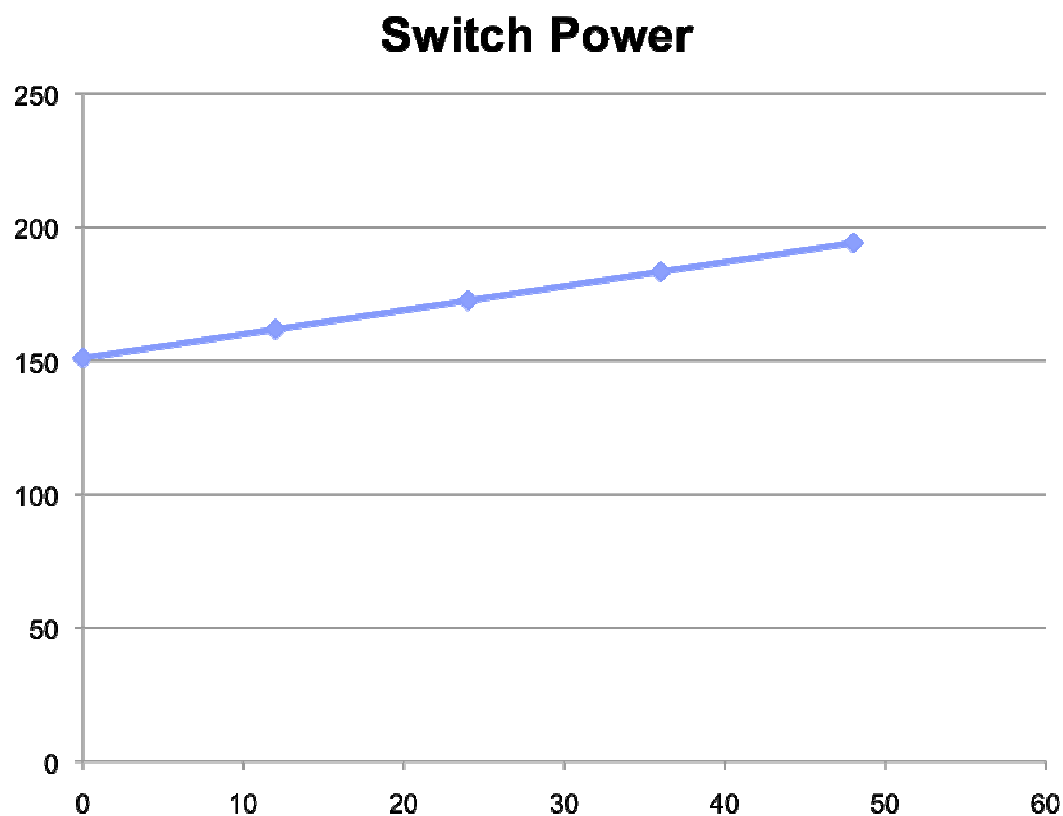
Switch Power Model

- $\text{switch_power} = \text{base_power} + \text{active_ports} * (\text{MAC_power} + \text{PHY_power})$
- Base_power includes crossbar, packet buffers, control plane, etc.
- $\text{MAC_power} = \text{MAC_base} + \text{MAC_activity_factor} * \text{utilization}$
 - Media Access Control (MAC) layer deals with protocol processing
 - MAC_activity_factor ~ 0 in many switches (depends on degree of clock gating)
- PHY_power depends on speed, media (copper or optical), and distance
 - Physical layer (PHY) performs data coding/decoding
- For chassis switches, use fixed power for the chassis and power model for each line card
- Optimization: Minimize total length of cabling to minimize power
 - Use top-of-rack switches instead of home-run cabling

Source: Priya Mahadevan, Sujata Banerjee, Puneet Sharma: Energy Proportionality of an Enterprise Network. Green Networking Workshop 2010

Switch Power Example

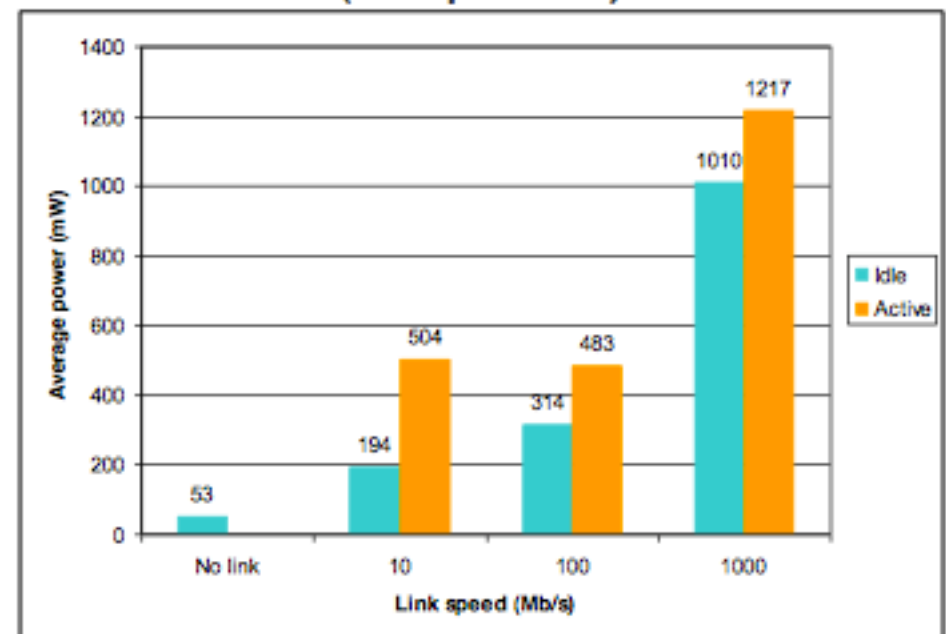
- 48-port 1000BASE-T switch
- base_power = 151 W
- MAC_power = 0.68 W/port
- PHY_power = 0.22 W/port



Port Power Example

- Intel 1000Base-T copper gigabit NIC
 - MAC_power varies from 53-270 mW depending on utilization
 - Slightly energy proportional
 - PHY_power varies from 150-1000 mW depending on link speed
 - 1000 Mb/s is 3x the power of 100 Mb/s,
 - 5x the power of 10 Mb/s
 - Does your laptop really need 1000 Mb/s?
 - Unplugging the cable saves significant power
-
- NICs and switches use the same PHYs
 - Switches may have higher MAC power per port

Single-port PCIe 10/100/1000 Mb/s controller (MAC plus PHY)



Source: Intel, Intel® 82573L Gigabit Ethernet Controller, 130 nm

"Idle" = no traffic

"Active" = bi-directional, line-rate traffic

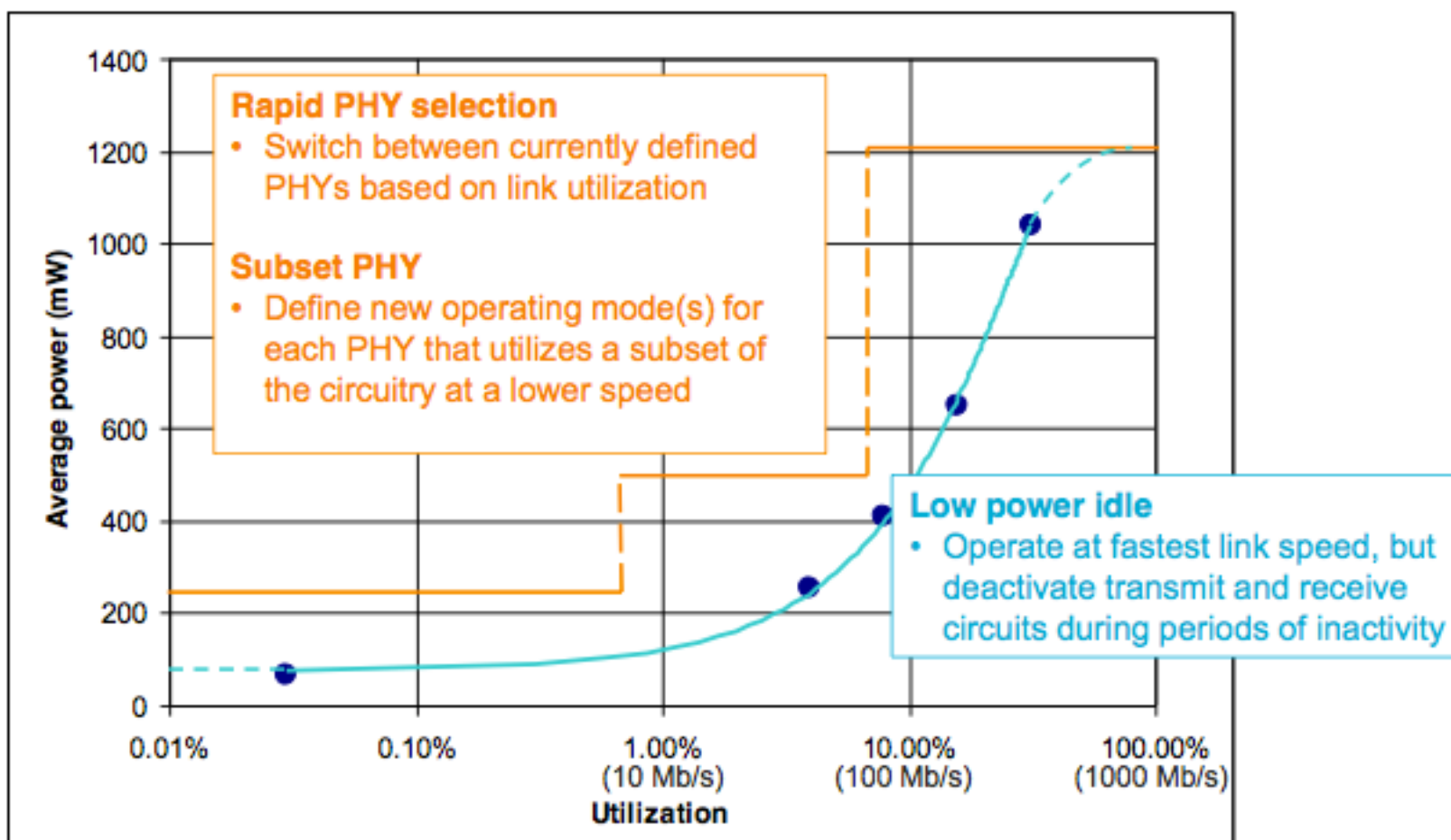
IEEE 802.3az Energy Efficient Ethernet

- Final standard ratified in October 2010
- Products just now appearing
- Free power savings – should be no performance loss and no additional cost
- Low-power idle (LPI) allows PHY Tx side to shut off when no packet is being transmitted
 - Saves ~400mW per 1 Gbps port
- Rx side remains on continuously
- Makes Ethernet more energy-proportional

Existing Actuators and Near-Term Power Reduction Policies

- What's possible with the equipment you have today?
- Enable 802.3az where possible
- Find unplugged ports and turn them off
 - Requires administrative action to turn port back on if needed
 - Can save ~1W/port
 - Mostly benefits older switches; newer switches power off unplugged ports automatically
- Consolidate onto fewer switches and turn off unused switches
- Manual port rate adaptation (only for copper)
 - Gather utilization data on switch ports
 - Run 1Gbps ports at 100Mbps when possible
 - Can save ~0.5W/port (out of ~2-4W/port)
 - Future: Run 10Gbps at 1Gbps or slower (save 4W?)
 - Requires administrative action; could be automated via SNMP

Rate Adaptation vs. Low Power Idle



Source: Intel, Configuration: Traffic profile = "Trace_VOIP_*.txt", low-power idle initialization wait = 10 ms, sleep time = 1 ms, wake time = 10 ms

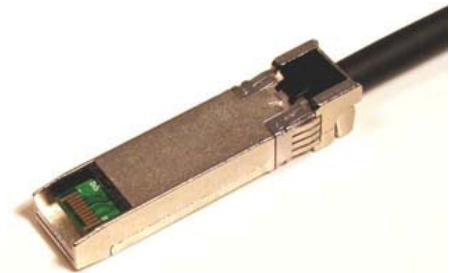
Simulation C-code and traffic profiles available at <http://www.ieee802.org/3/az/public/tools/index.html>

Architectural Considerations for Future Data Centers

- Tradeoffs between server, storage, and network power
 - Shared storage may reduce storage power and increase network power
 - Virtualization may reduce server power and increase network power (due to higher bandwidth)
- Converge LAN and storage networks
 - Reduce number of switches
 - Reduce number of active ports
- Replace high-end chassis switches with scale-out topology of small switches
 - 2X the ports -> over 2X the power
 - Feature-rich switches are higher power

10GbE Copper Dilemma

- 10GBASE-T (Cat6 cable, RJ-45 connector)
 - High power (>5 W/port)
 - More expensive
 - Short range (100 m)
 - Power proportional to range, but always higher than -CR
 - Backwards compatible with gigabit (RJ-45 connector)
 - Can reduce speed and power
 - Compatible with existing data center cable plant
 - Extra latency (+2 us) due to error correction
- 10GBASE-CR (aka SFP+ direct attach twinax)
 - Low power (<1 W/port)
 - Lower cost (despite costlier cables)
 - Very short range (<10 m)
 - Not backwards compatible with gigabit
 - Fixed 10 Gb/s speed
 - Very low latency (~ 0.1 us)



References

- Priya Mahadevan, Sujata Banerjee, Puneet Sharma: Energy Proportionality of an Enterprise Network. Green Networking Workshop 2010
 - Empirical study from HP Labs
 - Switch power model, rate adaptation, and throwing away equipment
- Sergiu Nedevschi, Lucian Popa, Gianluca Iannaccone, Sylvia Ratnasamy, David Wetherall: Reducing Network Energy Consumption via Sleeping and Rate-Adaptation, NSDI 2008
 - Switch power model, buffer-and-burst, rate adaptation
- Brandon Heller, Srinivasan Seetharaman, Priya Mahadevan, Yiannis Yiakoumis, Puneet Sharma, Sujata Banerjee, Nick McKeown: ElasticTree: Saving Energy in Data Center Networks. NSDI 2010
- Dennis Abts, Mike Marty, Philip Wells, Peter Klausler, Hong Liu: Energy Proportional Datacenter Networks. ISCA 2010
 - Shut off some links and switches during low network load
 - Requires multipathed topology (not yet common)
- Nathan Farrington, Erik Rubow, and Amin Vahdat: Data Center Switch Architecture in the Age of Merchant Silicon. Hot Interconnects 2009
 - Considers cost and power tradeoffs of different switch designs

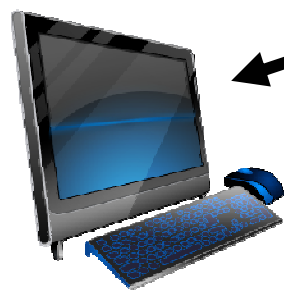
Cloud Computing: Virtualized Resources and Energy Management

Cloud Computing

User's view of a cloud

- Pay-as-you-use cost model
- Rapidly grow (shrink) compute capacity to match need
- Ubiquitous access

Cloud:
Computing infrastructure designed for dynamic provisioning of resources for computing tasks.



The growth of cloud-based computing means an increased use of data centers for computing needs

Provider's view of a cloud

- Shared resources, consolidation driving up utilization, efficiency.
- Leverage economies of scale for optimizations.
- Increased flexibility in sourcing equipment, components, pricing.

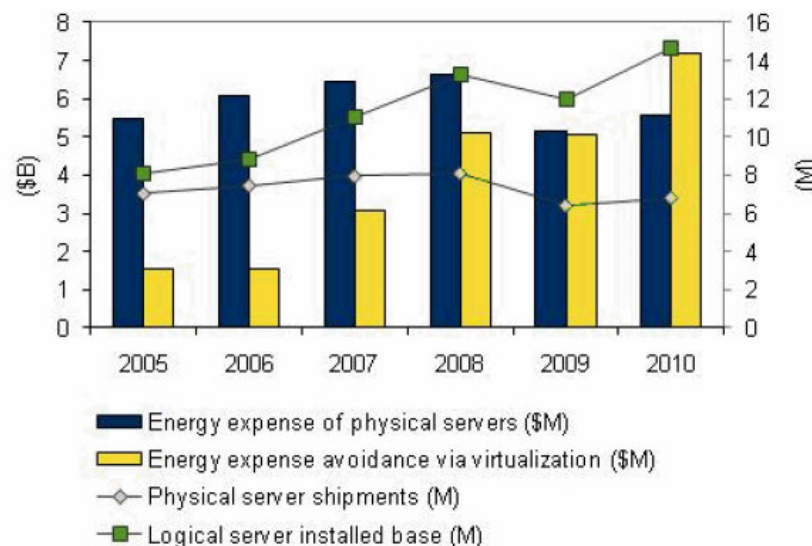
Energy-efficiency, Virtualization and Cloud Computing

- “..the energy expense associated with powering and cooling the worldwide server installed base increased 31.2% over the past five years ... In 2009, the server energy expense represented \$32.6 billion, while the server market generated \$43.2 billion.”*

- “..customers have avoided \$23.5 billion in server energy expense over the past six years from virtualizing servers.”*

*
Datacenter Energy Management: How Rising Costs, High Density, and Virtualization Are Making Energy Management a Requirement for IT Availability (IDC Insight Doc # 223004), Apr 2010

Worldwide Server Virtualization Installed Base and Energy Expense Savings, 2005–2010



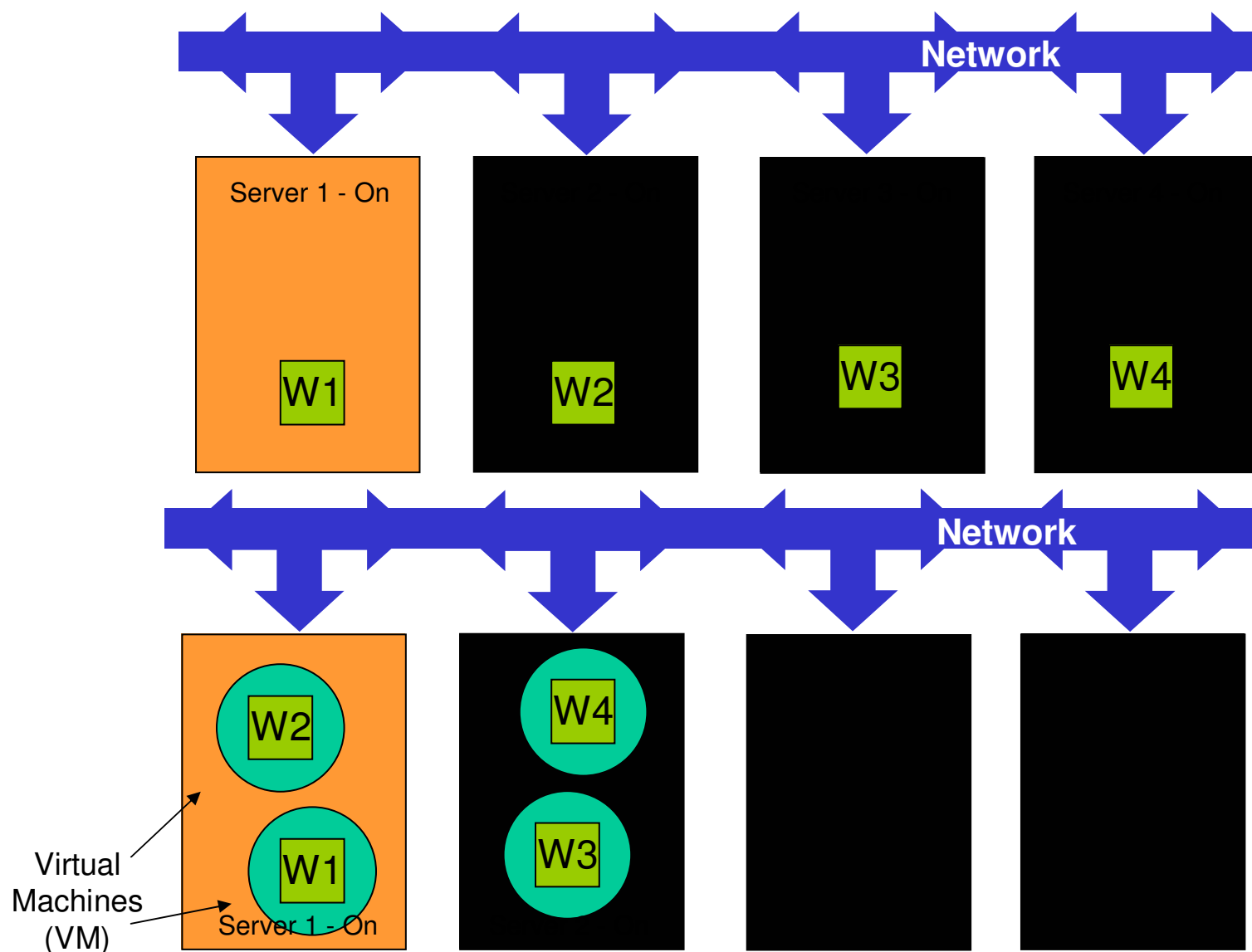
Source: IDC, 2010

- “Cloud will grow from a \$3.8 billion opportunity in 2010, representing over 600,000 units, to a \$6.4 billion market in 2014, with over 1.3 million units.”. In Worldwide Enterprise Server Cloud Computing 2010–2014 Forecast Abstract (IDC Market Analysis Doc # 223118), Apr 2010.
- “cloud computing to reduce data centers energy consumption from 201.8TWh of electricity in 2010 to 139.8 TWh in 2020, a reduction of 31%”, in Pike Research Report on “Cloud Computing Energy Efficiency”, as reported in Clean Technology Business Review, December, 2010.

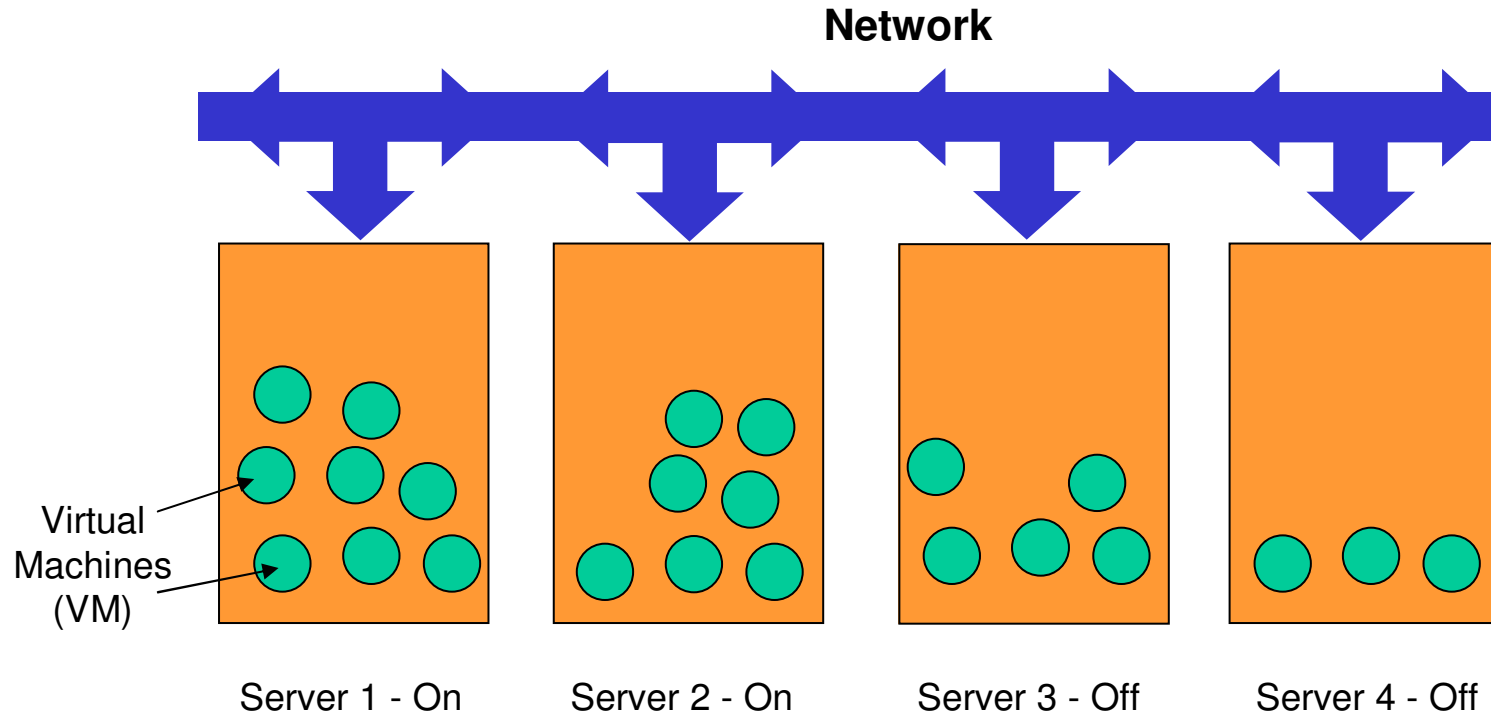
Efficiency from Cloud Model for Computing

- **Better utilization of systems drives increased efficiency**
 - Increased sharing of resources – lower instance of unused resources.
 - Less variability in aggregate load for larger population of workloads – better sizing of infrastructure to total load.
- **Computing on a large-scale saves materials, energy**
 - Study shows savings through less materials for larger cooling and UPS units.
 - Similar savings also possible in IT equipment
- **Economies of scale fund newer technologies**
 - Favor exploitation of newer (riskier), cheaper cooling technologies because of scaled up benefits.
 - Favor re-design of IT equipment with greater modularity, homogeneity with efficiency as a driving concern.

Virtualization as Cloud Computing Enabler



Dynamic Consolidation with Live Migration



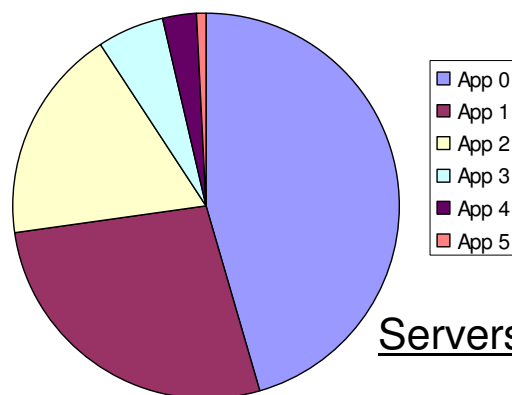
- Need support in platform and VMM/hypervisor for live migration/partition mobility.
- Network connectivity between hosts.
- Server power-on/-off capability (typically managed through service processor connections).

Case Study: Consolidation Benefit Analysis in an Enterprise Data Center*

- Explore opportunities for power savings through consolidation of server in data centers under realistic constraints
- Clusters are formed based on physical proximity, ownership by different lines of business, and instruction set architecture family

Example Consolidation – 6 server Cluster

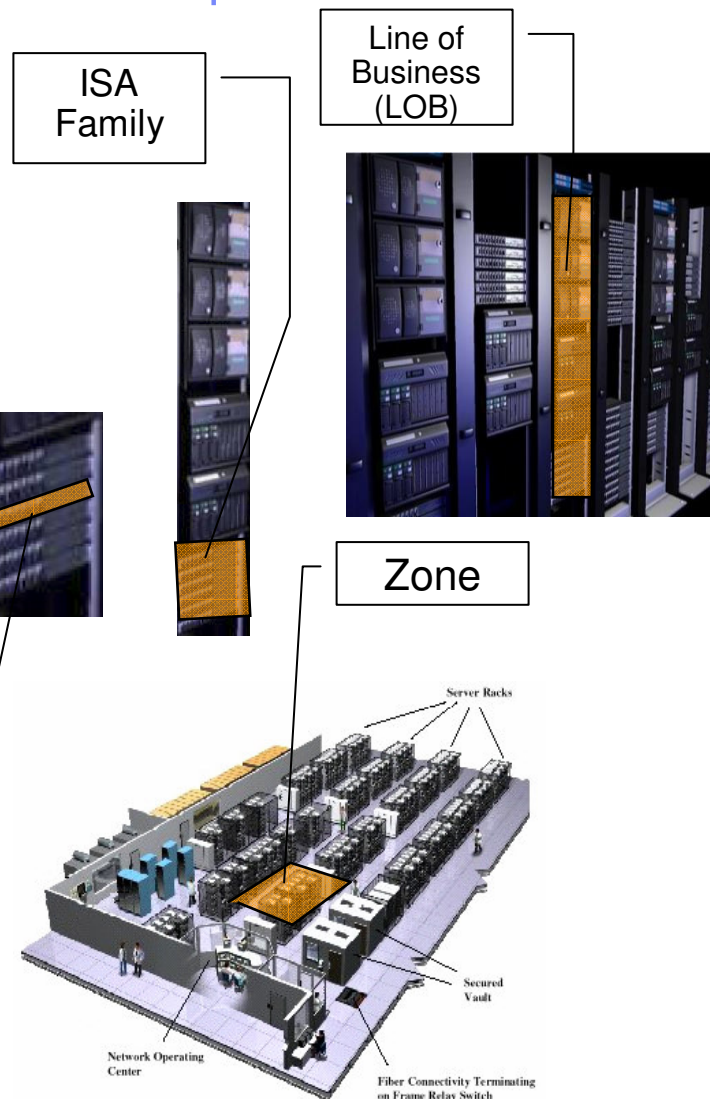
Server0 utilization pie chart



Servers1-5 Off

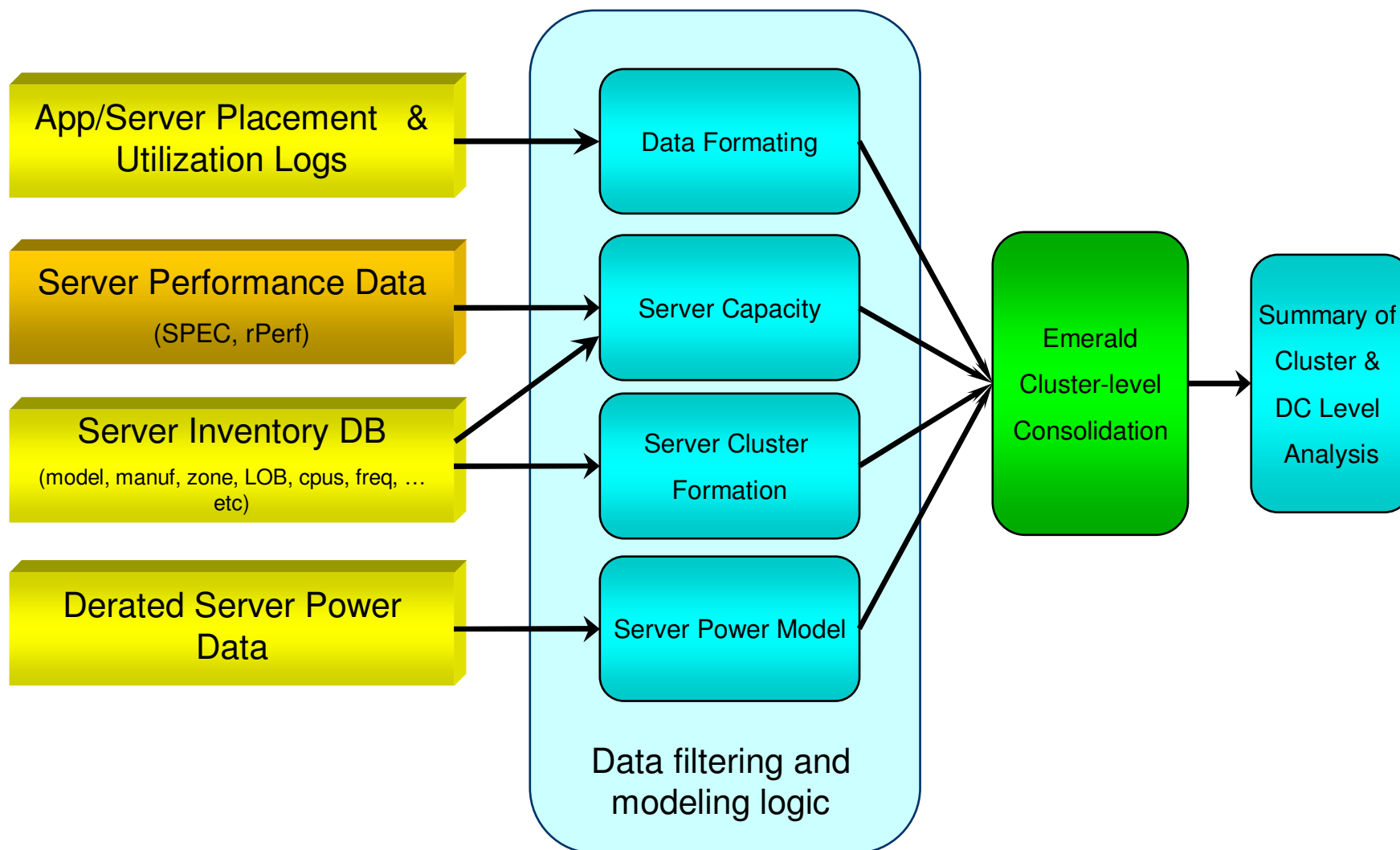
- 70 Zones
- Over 2000 servers
- 5 days performance data

Cluster
(Zone, LOB, ISA)



*Study conducted by Wael El-essawy, Karthick Rajamani, Juan Rubio, Tom Keller

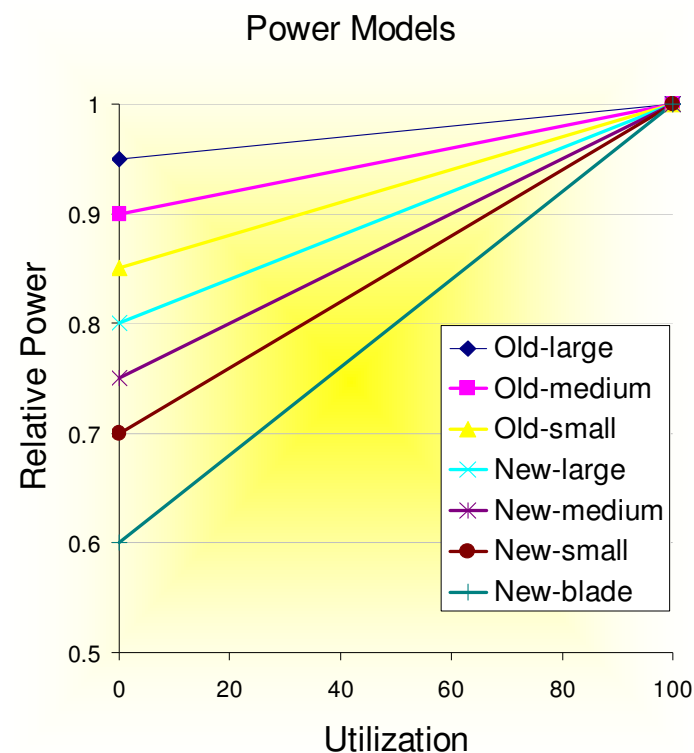
Case Study: Consolidation Methodology



Case Study: Performance and Power Models

- Performance logs sample *Average CPU Utilization* every 15/60 minutes
- Take a common performance metric to compare servers with different architectures
- Capacity:
 - **$CINT\ Rate \times Freq_{scale} \times \#CPUs_{scale}$**
 - *Other Performance metrics can be used (SPEC Web, rPerf, etc.)*
- Utilization logs for a server are relative to its capacity
- Start with non-virtualized servers

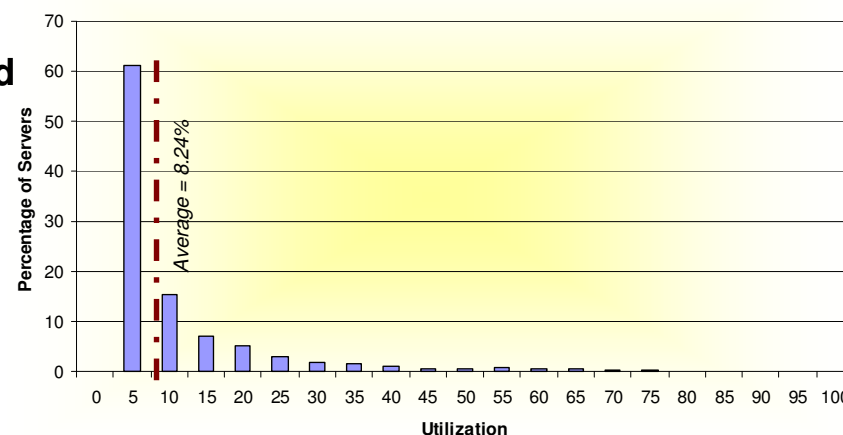
- Assume *linear* power model for each server between:
 - P_{max} (Power at 100% utilization)
 - P_{idle} (Idle Power)
- P_{max} :
 - Determined by the server derated power
 - Represents how much power is allocated
 - Usually less than nameplate power
 - Maximum configuration
- P_{idle}
 - Determined by the server age and model
 - Assumes no DVFS



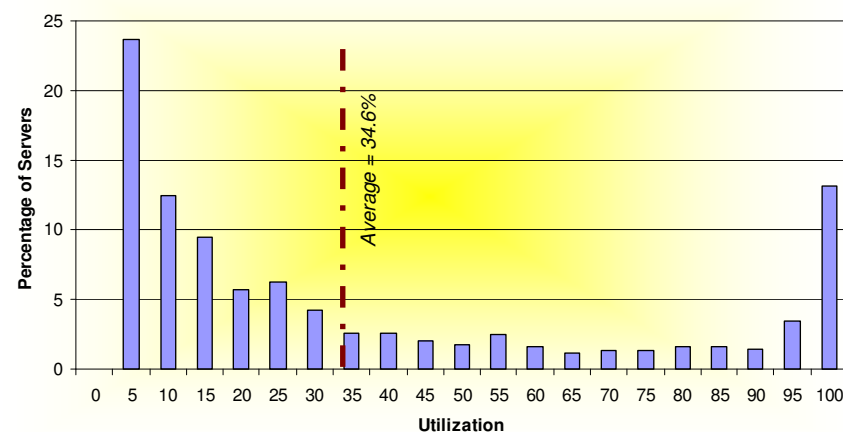
Case Study: Consolidation Results

- 2,292 servers out of 2,977 total servers are idled
 - 23% servers remain active, the rest can be turned off**
 - Per cluster, on average:
 - Active servers: 1.25,
 - Idle servers: 4.2
 - Average Jobs per server after consolidation: 4.3
- Data Center Servers are mostly underutilized
 - 8% Overall average CPU utilization
 - 76% of the servers are less than 10% utilized
 - Only 2% of the servers are more than 50% utilized
- Cluster-level consolidation significantly raised server utilization
 - 35% Average consolidated Utilization
- Cluster-level consolidation significantly lowered aggregate server power
 - 74% reduction in total server power**

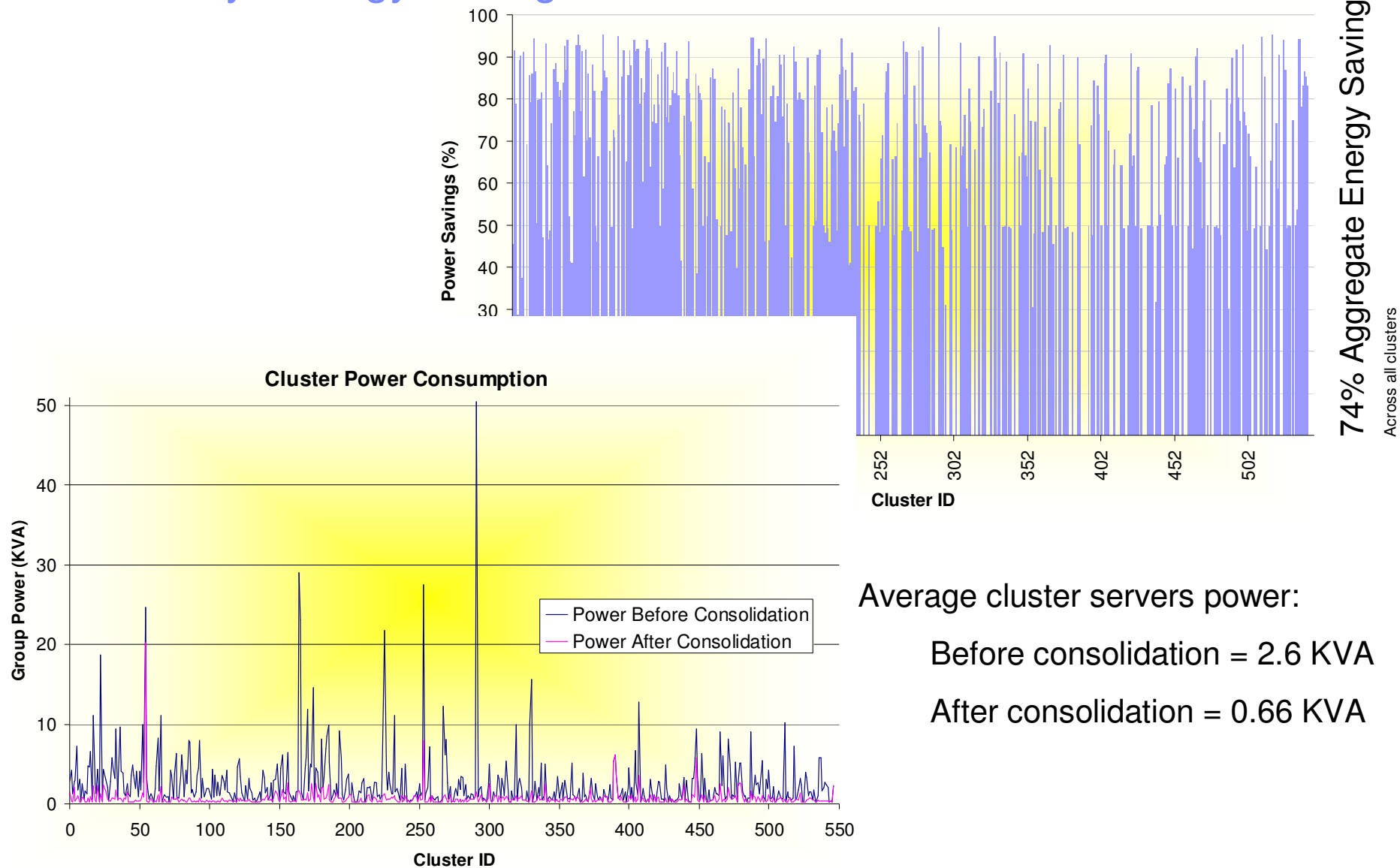
Input Utilization Histogram



Consolidated Utilization Histogram



Case Study: Energy Savings



Dynamic Provisioning and Consolidation

- **Virtual Machine (VM) consolidation as a bin-packing problem.**
 - Bin size: Server capacity
 - Object size: Historical VM CPU utilization summary
- **Extensions for practical solutions**
 - Limit packing to a fixed percentage of server capacity, avoid resource saturation.
 - Accommodate requirements for other resources such as VM memory needs.
 - Provision VM resources using prior characterization with expected workload.
 - Factor in SLAs and/or adopt runtime performance monitoring.
- **Possible additional optimizations/considerations**
 - Techniques for better prediction of future load characteristics.
 - Factor in multiple optimization concerns with utility-function based frameworks.
 - Factor in server/cluster power limits and power consumption for placement.
 - Adopt energy-aware placement strategies in heterogeneous server-workload environments.
 - Factor in VM migration characteristics and cost.
 - Factor in server on/off temporal characteristics.
 - Understand and address impact of other shared resources such as processor caches, networks, I/O.

Consolidation and Other Techniques

- Dynamic Voltage and Frequency Scaling
 - Often evaluated as a competing solution.
 - Provides better responsiveness to load changes, with potentially lower energy savings.
 - Applicable to non-virtualized and virtualized environments without VM migration support.
 - Can be transparently leveraged as complementary solution.
 - Should be explicitly leveraged in conjunction for superior optimization.

- Thermal Management
 - Consolidation can increase the diversity in data center thermal distribution
 - Thermal-aware consolidation/task placement strategies to mitigate thermal impact of consolidation.
 - Modular cooling infrastructure controls are an important complement to consolidation solutions to reduce overall datacenter energy consumption.
 - Integration of energy-aware task placement/consolidation with thermal management solutions can be a successful approach to full Data center energy optimization.

Considerations When Evaluating/Optimizing Cloud for Energy Efficiency

- **Energy cost of transport (network) to and from Cloud**
 - Volume of data transported between Cloud and local network impact which applications are more energy-efficient with a Cloud
 - Network topology and components can have a big impact on overall efficiency.
 - However, most of the networking energy cost is often hidden/transparent to a user.
- **Performance impact of shared resource usage**
 - Higher response times for tasks can render a Cloud solution infeasible; a hybrid computing solution is a likely compromise.
 - Lowered performance can imply lowered energy-efficiency
- **Modularity, responsiveness of the infrastructure supporting the Cloud**
 - Higher consolidation can exacerbate cooling issues in a non-modular, cooling-constrained facility.
 - Cooling infrastructure not tunable to changes can be a source of inefficiency in consolidated environments.
 - Servers with slow on/off times can limit exploitation of dynamic consolidation.
 - Networking within the cloud can be a factor for dynamic consolidation benefits.

Energy Proportionality versus Dynamic Consolidation

- Energy proportional components
 - Consume power/energy in proportion to their utilization
 - Ideally, no energy is consumed if no load and energy consumption scales in proportion to load.
- Server consolidation provides energy proportionality with non-ideal system components
 - Just enough servers are kept active to service the consolidated load allowing the rest to be powered off.
 - The granularity for scaling energy to load is in energy for entire servers.

Can increasing energy proportionality in server component designs render server consolidation solutions obsolete ?

- **Ideal energy proportionality is still far from reality, so continue with server consolidation.**
- Clusters of servers heterogeneous in their efficiencies would continue to benefit from energy-aware task placement/consolidation.
- Cooling solutions without good tuning options can interact sub-optimally with energy proportional hardware requiring intelligent task consolidation/placement to improve overall datacenter efficiency.

Energy Accounting

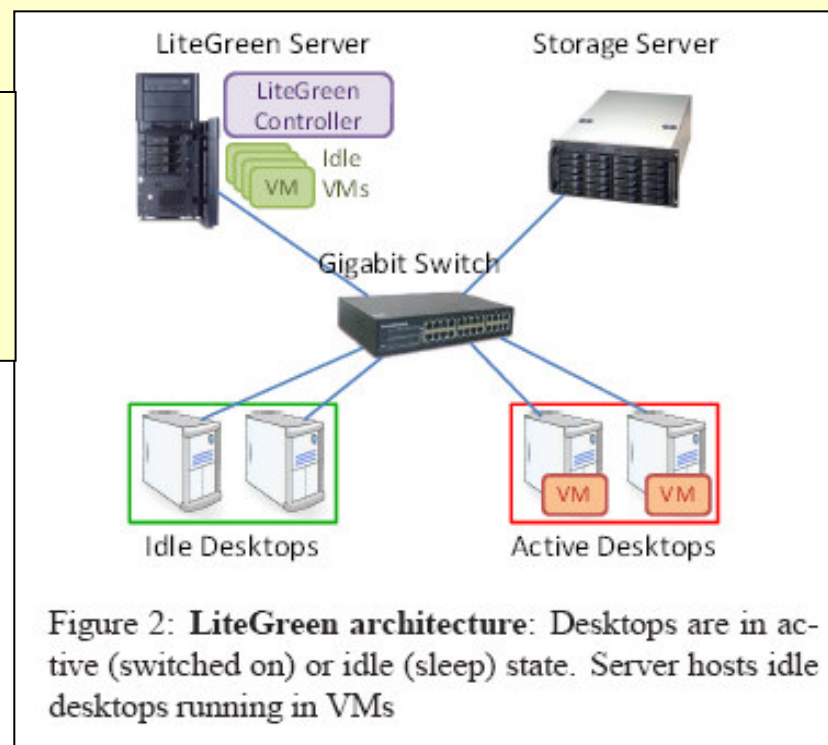
- Basic Motivation
 - Charging, incentivizing customers to allow better infrastructure utilization.
 - Identify in-efficiencies and unanticipated consumption.
 - Adapt resource provisioning and allocation with energy-usage information for more efficient operation.
 - Energy profiling of software to guide more efficient execution.
- Different approaches and challenges
 - **Activity-based approaches (Modeling)**
 - Accuracy of models
 - Inability to capture power variation with environment and manufacturing.
 - **Power-measurement based approach (Measurement)**
 - Synchronizing measurement with resource ownership changes.
 - Granularity of measurements versus resource ownership/usage.
 - **Common Challenges**
 - State changes with power management
 - Fairness considerations

Reducing power of Networked PCs

- **Problem:** Networked PCs always on to provide remote access capability even when mostly idle, wastes power.
- **Solutions:**
 - Using special NIC hardware to keep limited networking active even while allowing the PC to sleep.
 - Set up a *Sleep Proxy*. Sleep proxies maintain the *network presence* for the sleeping PC and wake it up when needed, sleep proxies themselves could themselves be special virtual machines¹.
- Virtualize the PC. Migrate the PC Virtual Machine to a designated holding server for such VMs, power down the PC till its resources are needed².

²LiteGreen: Saving Energy in Networked Desktops Using Virtualization, Tathagata Das, Pradeep Padala, Venkata N. Padmanabhan, Ramachandran Ramjee, Kang G. Shin, USENIX 2010.

¹SleepServer: A Software-Only Approach for Reducing the Energy Consumption of PCs within Enterprise Environments, Yuvraj Agarwal Stefan Savage Rajesh Gupta, USENIX 2010.



Benchmarks: SPECvirt_sc2010

- **SPEC's first virtualized environment benchmark**
- **Reporting in performance-only and performance-per-watt categories**
 - Two efficiency categories:
 1. Full system (server + storage) performance-per-watt
 2. Server performance-per-watt
- **Workload organized in sets of VMs, called *tiles*.**
 - Six VMs per tile running three applications, applications are modified versions of SPECweb2005, SPECjAppServer2004, and SPECmail2008.
 - Acceptable performance criteria set for each application within a tile.
- **Measures**
 - Performance measure is arithmetic mean of normalized performance measure for each of the three applications expressed as <Performance>@<number of VMs>.
 - Allows for fractional tiles.
 - Peak performance/Peak power is the measure for performance-per-watt categories
- **Still in early adoption stage**
 - Released 2010
 - Eighteen results in performance category and one in each performance-per-watt category.

Takeaway

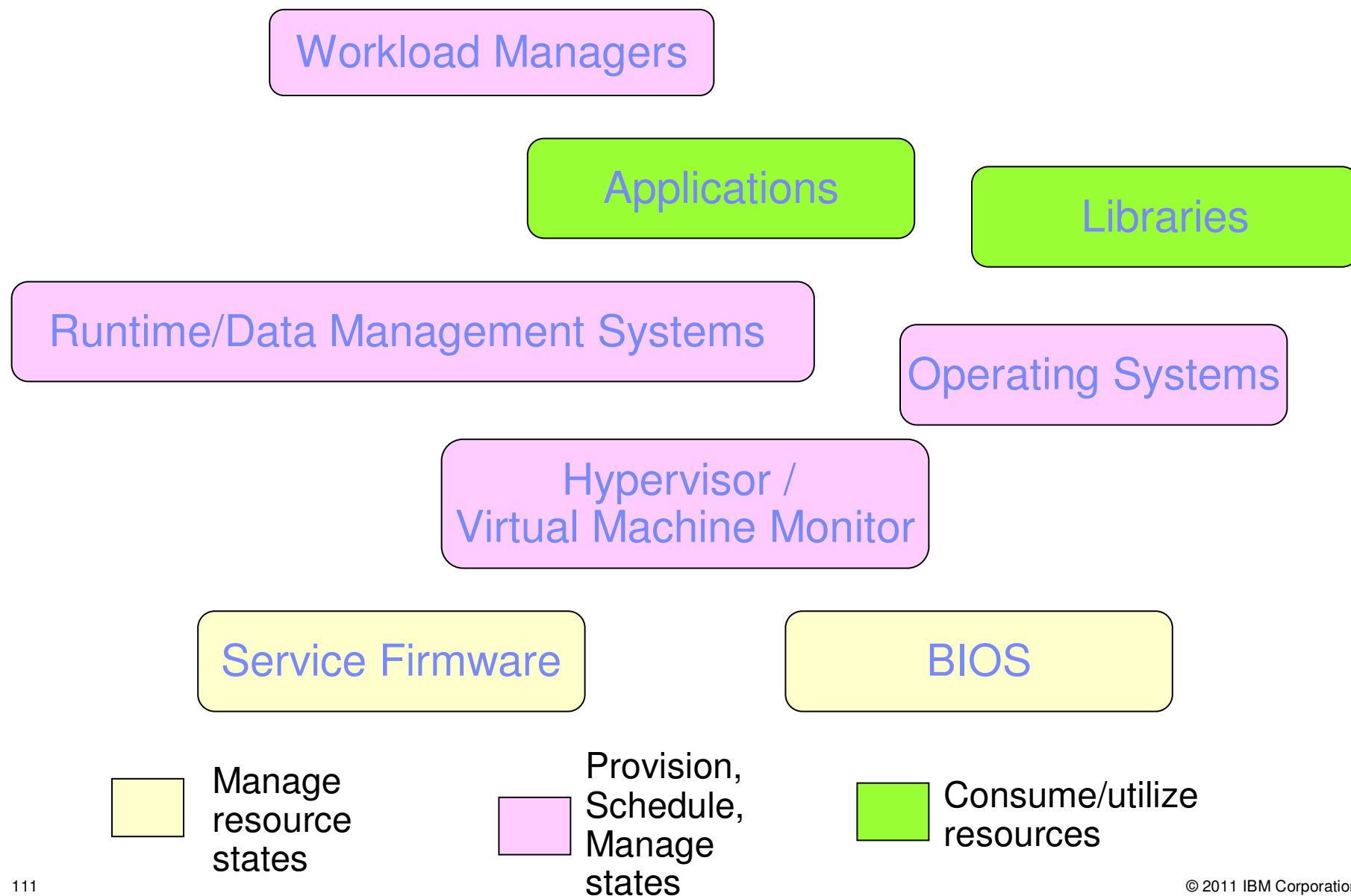
- Cloud is an attractive computing infrastructure model with rapid growth because of its on-demand resource provisioning feature.
 - Growth of cloud computing would lead to growth of large data centers. Large-scale computing in turn enables increased energy efficiency and overall cost efficiency.
- The business models (cloud provider) around clouds incentivize energy efficiency optimizations creating a big consumer for energy-efficiency research.
 - The cloud's transparent physical resource usage model facilitates sharing and efficiency improvements through virtualization and consolidation.
 - Energy-proportionality and consolidation need to co-exist to drive Cloud energy-efficiency
 - End-to-end (total DC optimization) design and operations' optimization for efficiency will also find a ready customer in Cloud Computing.
- Efficiency optimization while guaranteeing SLAs will continue to drive research directions in the Cloud.

References

1. Using Virtualization to Improve Datacenter Efficiency, Version 1, Richard Talaber, Tom Brey, Larry Lamers, Green Grid White Paper #19, January, 2009..
2. Quantifying the Environmental Advantages of Large-Scale Computing, Vlasia Anagnostopoulou, Heba Saadeldeen, Frederic T. Chong, International Conference on Green Computing, August, 2010. (material and operational cost reduction).
3. Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport, Jayant Baliga, Robert W A Ayre, Kerry Hinton, and Rodney S Tucker, Proceedings of the IEEE 99(1), January, 2011.
4. pMapper: power and migration cost aware application placement in virtualized systems, A. Verma, P. Ahuja, and A. Neogi, in Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware, 2008.
5. Energy Aware Consolidation for Cloud Computing, Shekhar Srikantaiah, Aman Kansal, Feng Zhao, HotPower 2008.
6. Performance and Power Management for Cloud Infrastructures, Hien Nguyen Van, Frédéric Dang Tran and Jean-Marc Menaudy, 3rd IEEE International Conference on Cloud Computing, 2010.
7. Mistral: Dynamically Managing Power, Performance, and Adaptation Cost in Cloud Infrastructures, Gueyoung Jung, Matti A. Hiltunen, Kaustubh R. Joshi, Richard D. Schlichting, Calton Pu, ICDCS 2010.
8. vGreen: A System for Energy Efficient Computing in Virtualized Environments, Gaurav Dhiman, Giacomo Marchetti, Tajana Rosing, ISLPED 2009.
9. Temperature-Aware Dynamic Resource Provisioning in a Power-Optimized Datacenter, Ehsan Pakbaznia, Mohammad Ghasemazar, and Massoud Pedram, DATE 2010.
10. Trends and Effects of Energy Proportionality on Server Provisioning in Data Centers, Georgios Varsamopoulos, Zahra Abbasi, and Sankeep K. S. Gupta, International Conference on High Performance Computing (HiPC), December, 2010.
11. Virtual Machine Power Metering and Provisioning, Aman Kansal, Feng Zhao, Jie Liu, Nupur Kothari, Arka A. Bhattacharya, ACM SOCC 2010.
12. VMeter: Power Modelling for Virtualized Clouds, Ata E Husain Bohra and Vipin Chaudhary, IPDPS 2010.
13. VM Power Metering: Feasibility and Challenges, Bhavani Krishnan, Hrishikesh Amur, Ada Gavrilovska, Karsten Schwan, GreenMetrics 2010, in conjunction with SIGMETRICS'10), New York, June 2010 (Best Student Paper).
14. LiteGreen: Saving Energy in Networked Desktops Using Virtualization, Tathagata Das, Pradeep Padala, Venkata N. Padmanabhan, Ramachandran Ramjee, Kang G. Shin, USENIX 2010.
15. SleepServer: A Software-Only Approach for Reducing the Energy Consumption of PCs within Enterprise Environments, Yuvraj Agarwal Stefan Savage Rajesh Gupta, USENIX 2010.
16. Somniloquy: Augmenting Network Interfaces to Reduce PC Energy Usage, Y. Agarwal, S. Hodges, R. Chandra, J. Scott, P. Bahl, and R. Gupta, NSDI'09, Berkeley, CA, USA, 2009

Energy-efficient Software

Software Components and Compute Resources



The Many Roles of Software in Energy-efficient Computing

- **Exploiting lower energy states and lower power operating modes**
 - Support all hardware modes e.g. S3/S4, P-states in virtualized environments
 - Detect and/or create idleness to exploit modes.
 - Software stack optimizations to reduce mode entry/exit/transition overheads.
- **Energy-aware resource management**
 - Understand and exploit energy vs performance trade-offs e.g. Just-in-time vs Race-to-idle
 - Avoid resource waste (bloat) that leads to wasted energy
 - Adopt energy-conscious resource management methods e.g. polling vs interrupt, synchronizations.
- **Energy-aware data management**
 - Understand and exploit energy vs performance trade-offs e.g. usage of compression
 - Energy-aware optimizations for data layout and access methods e.g. spread data vs consolidate disks, inner tracks vs outer tracks
 - Energy-aware processing methods e.g. database query plan optimization
- **Energy-aware software productivity**
 - Understand and limit energy costs of modularity and flexibility
 - Target/eliminate resource bloat in all forms
 - Develop resource-conscious modular software architectures
- **Enabling hardware with lower energy consumption**
 - Parallelization to support lower power multi-core designs
 - Compiler and Runtime system enhancements to help accelerator-based designs

Processor and System State Management

- ACPI states for OSPM, Intel (Enhanced) SpeedStep, AMD PowerNow
 - Encounter incomplete Chipset/BIOS support and/or lack of enablement by user.

Global (G) State	Sleep (S) State	Processor Core (C) State	Processor State	System Clocks	Description
G0	S0	C0	Full On	On	Full On
G0	S0	C1/C1E	Auto-Halt	On	Auto-Halt
G0	S0	C3	Deep Sleep	On	Deep Sleep
G0	S0	C6	Deep Power Down	On	Deep Power Down
G1	S3	Power off	Power off	Off, except RTC	Suspend to RAM
G1	S4	Power off	Power off	Off, except RTC	Suspend to Disk
G2	S5	Power off	Power off	Off, except RTC	Soft Off
G3	NA	Power off	Power off	Power off	Hard off

Source:

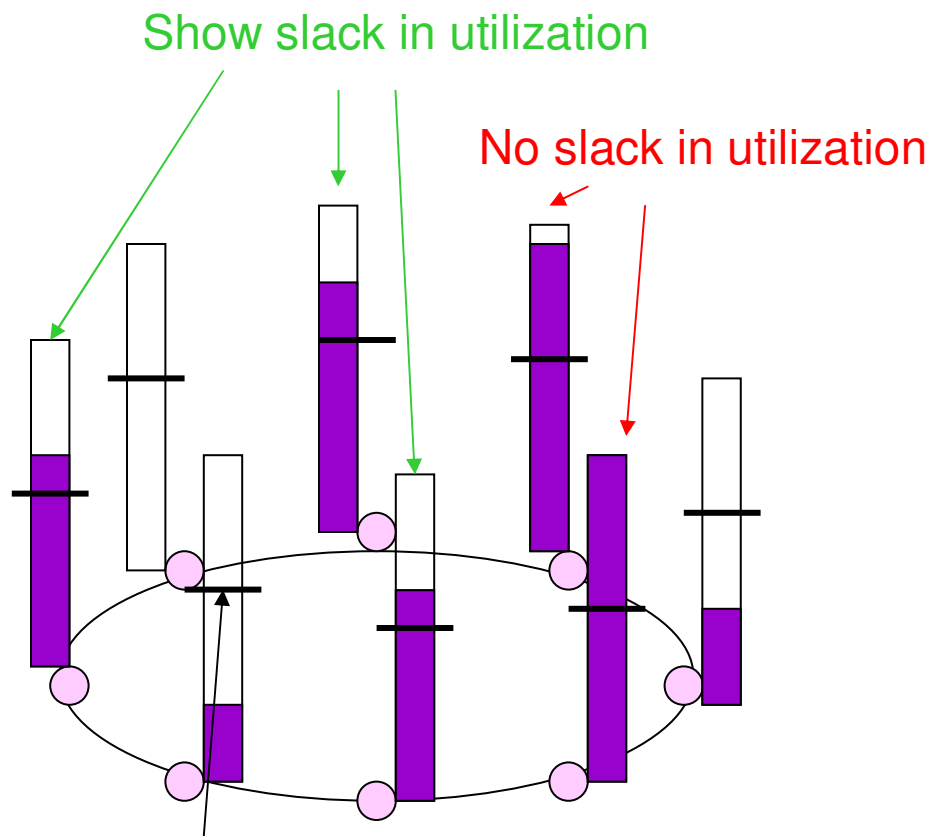
Mondira (Mandy) Pant, Intel

Presentation at GLSVLSI, May 2010

- Linux *governors* for user-level power management.
- Folding on IBM POWER platforms.
- Managing states in virtualized environments via service firmware, aggregate utilization, hypervisor with OS hints.
- Increasing idle exploitation opportunities – tickless kernels and timer/interrupt-service migrations.
- Coordinating voltage-frequency scaling and idle state management.

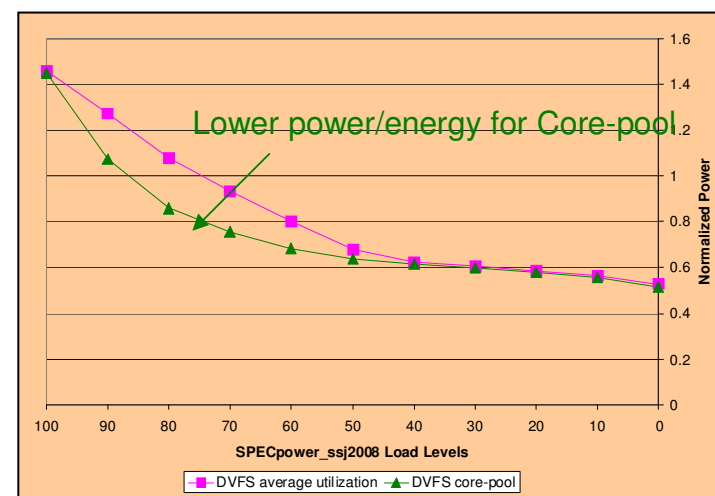
Exploiting Dynamic Voltage and Frequency Scaling: A New Approach

Core-pool algorithm* for slack detection in multi-core, multi-threaded environments

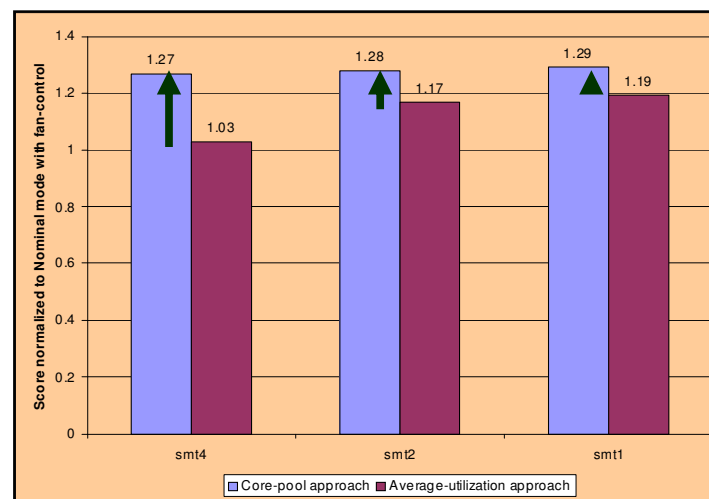


Minimum Utilization Threshold

Power reduction for SPECpower_ssj2008



Improvement greater for increased SMT



*Power-performance Management on an IBM POWER7 Server,
114 Rajamani et al, ISLPED 2010

Memory Sub-system Power Management

- **Idle memory power**

- Large memory systems can have a greater fraction of memory power in idle devices.
- Exploiting DRAM idle power modes critical to energy-efficiency.
- Power-aware virtual memory, coordinated processor scheduling and memory-state management.

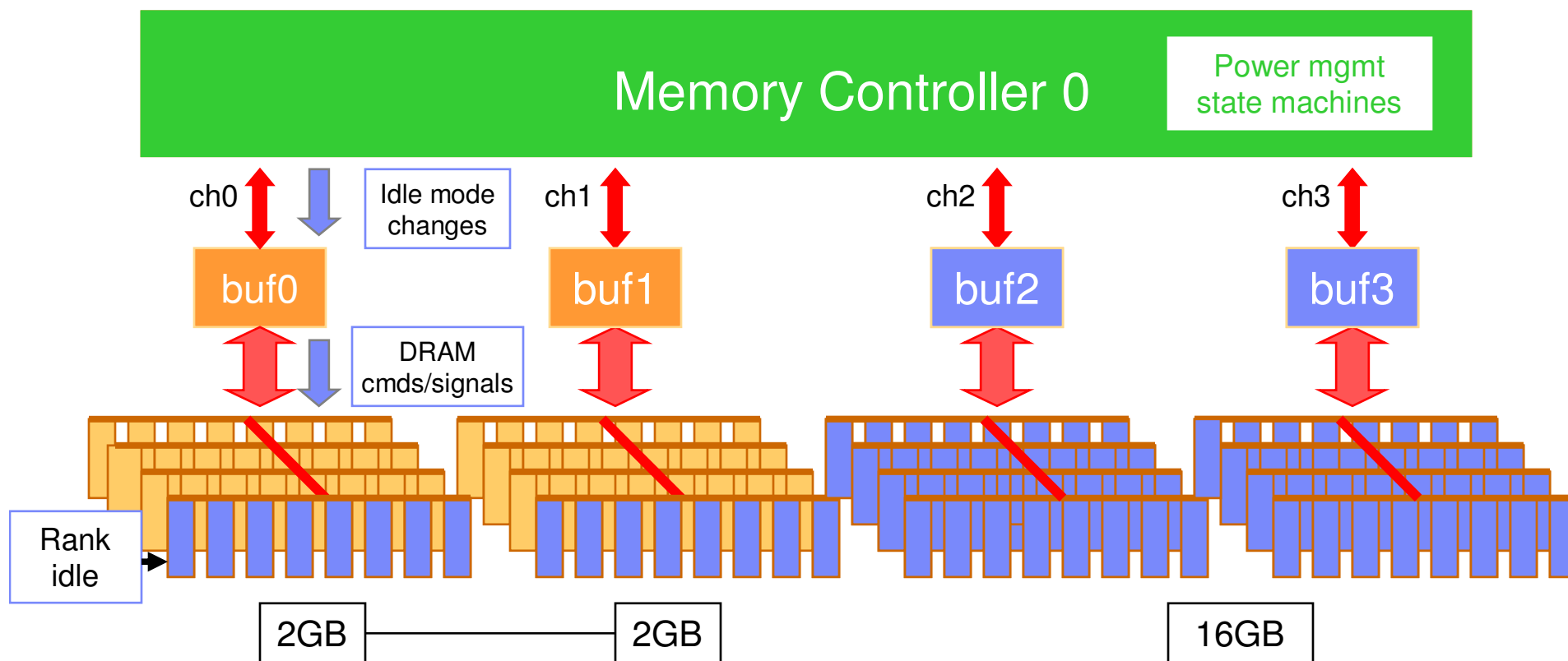
- **Active memory power**

- DRAM device active power is not reducing fast enough to keep pace with bandwidth growth demands.
- Providing adequate power for DRAM accesses can be critical to system performance.
- Power shifting between processor and memory - regulating power consumption for maximizing performance.

- **Support in today's servers**

- Transparent to systems software and applications
- System-state driven e.g. S3 state entry can place DRAM in self-refresh mode.
- Idle-detect driven - DRAM power-down (e.g. Nehalem EX, POWER6) and self-refresh (e.g. POWER7) triggered when memory controller detects adequate idleness.

Dynamic Memory Idle State Management



- Large regions of memory need to be idle before lower power mode can be used.
- Higher savings/latency mode ($O(\mu s)$) needs even larger regions to be idle, infeasible.
- Granularity needs worsen with larger capacity devices/DIMMs, i.e., can be worse than shown..
- ***Hypervisor and system software involvement to consolidate data in fewest power domains can maximize idle opportunities***

Ideas we are exploring for software assisted memory power management

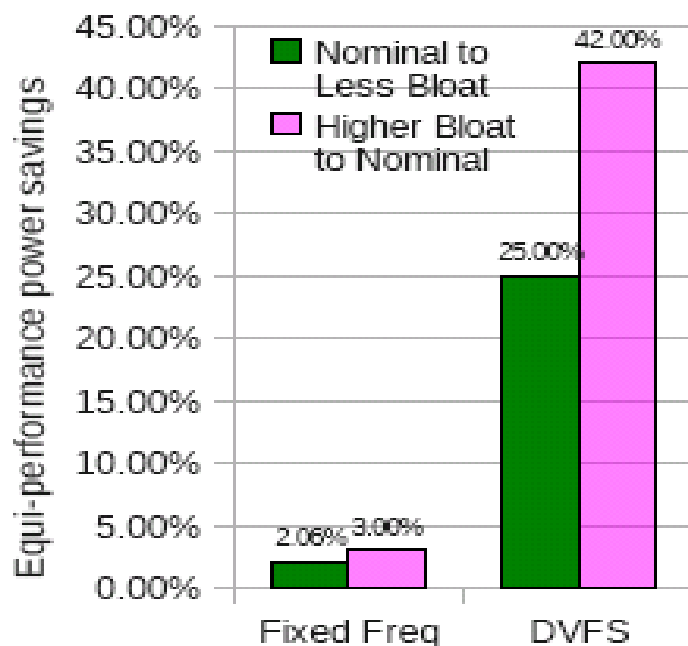
- **Energy-aware virtual memory re-sizing by hypervisor/operating systems**
 - Improve system memory utilization and lower physical memory occupancy by idle data to reduce associated energy.

- **Affinity-aware placement and memory allocation limiting device occupancy**
 - Lowering memory access cost (active power) and memory occupancy cost (standby power) .

- **Software assisted *tiered* main memory architecture**
 - Facilitate incorporation of new memory technologies and/or aggressive exploitation of low power (but higher latency) modes for energy-efficient capacity expansion.

Software Bloat

- Modularity and flexibility for software development can have performance and energy-efficiency overheads.
 - Temporary object bloat scenarios for SPECpower_ssj2008, data measured on POWER750.
 - Shows **equi-performance power** (and consequently energy) for different levels of bloat.
 - Primary source of inefficiency is lower performance due to cache pollution and memory bandwidth impact from higher incidence of temporary objects.



Nominal – Original, unmodified code

Higher Bloat – Disabled explicit object reuse at one code site

Less Bloat – Introduced object reuse at another code site

Use of DVFS enabled big power reductions when bloat is reduced

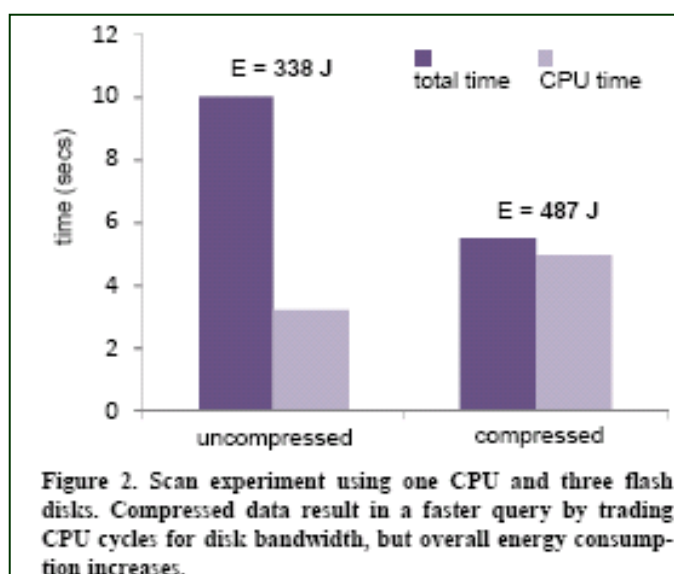
Source:

The interplay of Software Bloat, Hardware Energy Proportionality and System Bottlenecks
– HotPower 2011.

Energy-Performance Trade-offs

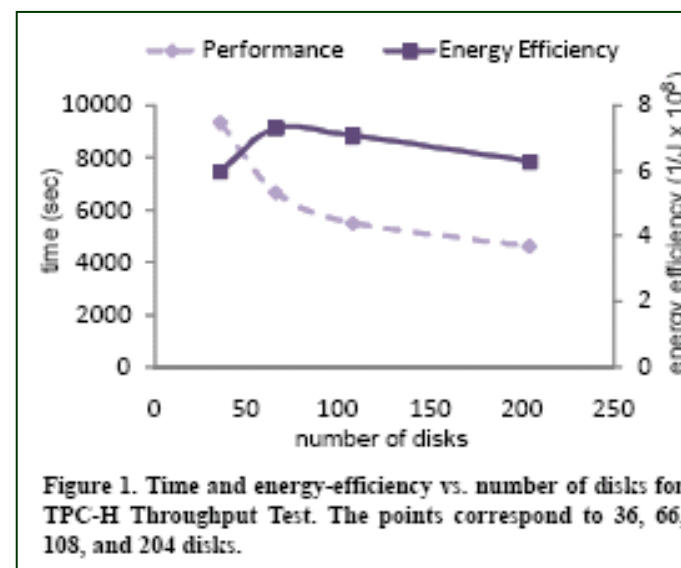
Optimizing for performance need not always optimize energy-efficiency. Examples

- Race-to-idle optimizes performance, but inefficient if workload is memory bound.
- Usage of compression to improve performance under limited storage access bandwidth.



- Usage of disk parallelism to address limited storage bandwidth.

Source for figures: Energy Efficiency: The New Holy Grail of Data Management Systems Research, S Harizoupoulos, M A Shah, J Meza, P Ranganathan, CIDR Perspectives 2009.



Cluster, Parallel and High-performance Computing Applications

- **Energy-aware server pool sizing for multi-server applications.**
 - Incorporate energy-aware optimizers in workload management systems to choose the number/type of servers for multi-tier workloads based on real time site traffic.
- **Cluster resource sizing and power-mode usage based on load.**
 - Load-balancer for cluster can utilize energy considerations to shut down additional servers not required for SLA compliance.
- **Coordinating processor/system state management and job scheduling**
 - Job schedulers for Supercomputer clusters can adapt performance states of servers to nature of workload launched on specific servers.
- **Energy-aware parallel application runtimes/libraries**
 - Exploiting processor idle states at synchronization points
 - Exploiting network link states based on communication patterns

Potential Optimizations in Data Management

- **Query optimizers in database systems**
 - Factor in performance implications of accessing disk-resident/memory-resident data in formulating a query plan, incorporate energy considerations.
- **Group/batch processing of queries**
 - Both throughput and energy-efficiency can be improved by (delayed) batch processing of related queries trading of higher latencies for individual (early) queries.
- **Enabling adoption of energy-efficient media**
 - Optimize software stack for usage of newer media like Flash with better energy-efficiency for random I/O, enable tiered storage.
- **Data layout and energy optimizations**
 - Coordinate data accesses and disk idle-mode change commands based on knowledge of data layout on disks to lower disk energy.
- **Energy-efficient data node management**
 - Adopt energy-aware data replication/placement strategies in multi-node/multi-replica environments.

Downloadable tools for energy-awareness

PowerTOP

– <http://www.linuxpowertop.org/>

```

File Edit View Terminal Go Help
PowerTOP version 1.8 (C) 2007 Intel Corporation

Cn          Avg residency      P-states (frequencies)
C0 (cpu running) (12.9%)      1.71 Ghz  9.8%
C1          0.0ms ( 0.0%)      1200 Mhz  0.3%
C2          10.7ms (87.1%)      800 Mhz   0.5%
C3          0.0ms ( 0.0%)      600 Mhz   89.4%
C4          0.0ms ( 0.0%)

Wakeups-from-idle per second : 81.2 interval: 15.0s
Power usage (ACPI estimate): 14.1W (6.6 hours) (long term: 136.4W,/0.7h)

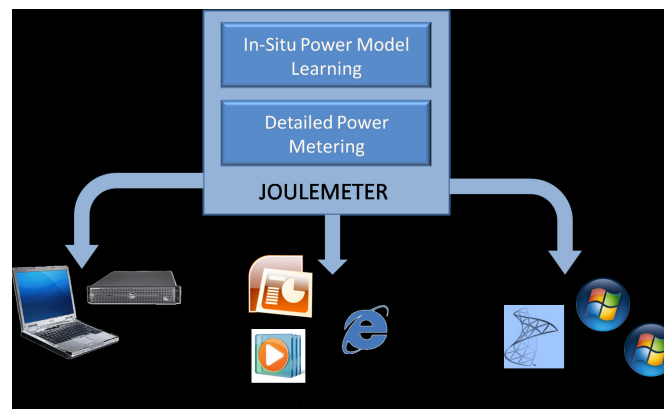
Top causes for wakeups:
34.4% ( 31.9) <interrupt> : ipw2200, Intel 82801DB-ICH4, Intel 82801DB-I
19.4% ( 18.0) firefox-bin : futex_wait (hrtimer_wakeup)
15.5% ( 14.4) X : do_setitimer (it_real_fn)
11.5% ( 10.7) evolution : schedule_timeout (process_timeout)
4.3% ( 4.0) <kernel module> : usb_hcd_poll_rh_status (rh_timer_func)
3.9% ( 3.6) <interrupt> : libata
1.8% ( 1.7) <kernel core> : sk_reset_timer (tcp_delack_timer)
1.2% ( 1.1) X : schedule_timeout (process_timeout)
1.1% ( 1.0) Terminal : schedule_timeout (process_timeout)
1.1% ( 1.0) xfce4-panel : schedule_timeout (process_timeout)
0.6% ( 0.5) <kernel module> : neigh_table_init_no_netlink (neigh_periodic
0.5% ( 0.5) spamd : schedule_timeout (process_timeout)
0.5% ( 0.5) events/0 : ipw_gather_stats (delayed_work_timer_fn)
0.4% ( 0.3) xfdesktop : schedule_timeout (process_timeout)
0.4% ( 0.3) firefox-bin : sk_reset_timer (tcp_write_timer)
0.3% ( 0.3) nsd : futex_wait (hrtimer_wakeup)
0.2% ( 0.2) xscreensaver : schedule_timeout (process_timeout)
0.2% ( 0.2) ksnapshot : schedule_timeout (process_timeout)

Suggestion: Disable the unused bluetooth interface with the following command:
hciconfig hci0 down ; rmmod hci_usb
Bluetooth is a radio and consumes quite some power, and keeps USB busy as well.
Q - Quit R - Refresh B - Turn Bluetooth off

```

JouleMeter

– <http://research.microsoft.com/en-us/projects/joulemeter/default.aspx>



Take Away

- Energy-aware software is integral to energy-efficient computing
- Intelligent resource provisioning and management is key at all levels of resource management.
- Appropriately managing component low-power modes requires architecting software for dynamic power-performance trade-off management, idle opportunity detection and creation.
- It is important to realize flexibility in software development without exacerbating resource waste leading to lowered performance and inefficiency.
- Data placement and access have important implications on resource usage and consequently energy-efficiency.

References

1. Power-performance Management on an IBM POWER7 Server, Karthick Rajamani, Malcolm Ware, Freeman Rawson, Malcolm Ware, Heather Hanson, John Carter, Todd Rosedahl, Andrew Geissler, Guillermo Silva, Hong Hua, 2010 IEEE/ACM International Symposium on Low-power Electronics and Design (ISLPED 2010).
2. Energy Reduction in Consolidated Servers through Memory-Aware Virtual Machine Scheduling, Jae-Wan Jang, Myeongjae Jeon, Hyo-Sil Kim, Heeseung Jo, Jin-Soo Kim, Member, and Seungryoul Maeng, IEEE Transactions on Computers 60(4), April 2011.
3. The New Holy Grail of Data Management Systems Research, S Harizoupoulos, M A Shah, J Meza, P Ranganathan, CIDR Perspectives 2009
4. The Thrifty Barrier: Energy-efficient Synchronization in Shared-memory Multiprocessors, J. Li, J.F. Martínez, and M.C. Huang, In International Symposium on High Performance Computer Architecture (HPCA), February 2004.
5. On Evaluating Request-Distribution Schemes for Saving Energy in Server Clusters, Karthick Rajamani and Charles Lefurgy, ISPASS 2003.
6. Towards Eco-friendly Database Management Systems, Willis Lang and Jignesh M Patel, 4th Biennial Conference on Innovative Data Systems Research, Jan 2009.
7. Exploring Power-performance Trade-offs in Database Systems, Zichen Xu, Yi-Cheng Tu, Xiaorui Wang, 26th IEEE International Conference on Data Engineering, March, 2010.
8. Robust and Flexible Power-Proportional Storage, Hrishikesh Amur, James Cipar, Varun Gupta, Gregory R. Ganger, Michael A. Kozuch, Karsten Schwan, ACM Symposium on Cloud Computing (SoCC 2010), Indianapolis, June 2010.
9. Evaluation and Analysis of GreenHDFS: A Self-Adaptive, Energy-Conserving Variant of the Hadoop Distributed File System, Rini T. Kaushik, Milind Bhandarkar, Klara Nahrstedt, 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010.
10. Compiler-directed Energy Optimization for Parallel Disk Based Systems, S. W. Son, G. Chen, O. Ozturk, M. Kandemir, A. Choudhary, IEEE Transactions on Parallel and Distributed Systems (TPDS) 18(9), September 2007.

Energy Modeling

Section Outline

- Modeling of energy-efficient data centers
 - Principles used in data center modeling tools
 - State of the art in data center modeling
 - Future research topics (model integration, off-line vs. real-time modeling)

Modeling Goals and Process

- Goal can be:
 - Estimate variables that are hard to measure (e.g., total energy spend in power conversion)
 - Understand impact of changes to the scenario (e.g., introduce new server, change temperature set point, failure of cooling unit)
 - Optimize a scenario (e.g., determine best location for a new server, reduce number of applications that fail after a cooling unit shuts down)
- To evaluate the energy efficiency of computer systems, it is necessary to model both workloads and their physical environment
- Researchers have produced several tools to model parts of the system:
 - Workloads in computer systems (e.g., SimpleScalar, SimOS, Simics) and large scale systems (e.g., MDSim)
 - Physical properties such as current drawn, power dissipation and heat transport

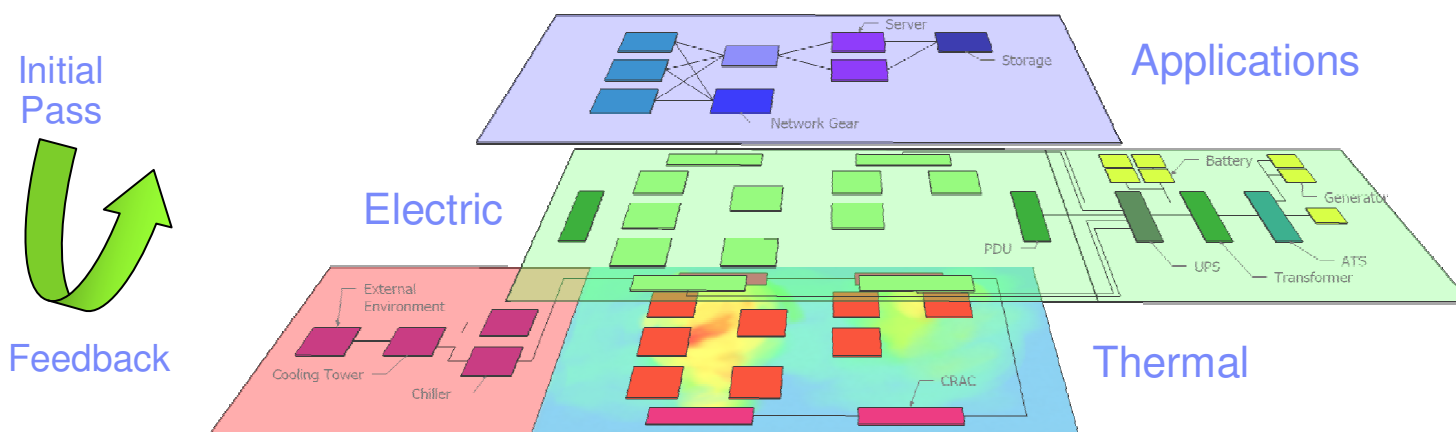
Modeling Workflow

- A full system simulation requires:
 - Modeling of the application workloads
 - Use those to drive the power load models
 - Use power loads to drive the electrical network
 - Use power loads as heat loads to drive the thermal models
 - Use the thermal transfers to evaluate the facility cooling system
- Feedback from later stages is needed to improve accuracy:
 - Ambient temperature affects cooling within the server → impacts power consumed by fan, and leakage power of processor
 - Power management of server → can impact performance of workload
 - Failure in one domain can propagate to other domains

Step
 μs to ms

ms to sec

sec to min



Assessing Modeling Solutions

- Multiple tools exist, each modeling different aspects of the problem
 - Selecting the right ones is key!
- Modeling domain focus:
 - Application: performance, utilization
 - Electric: server power, data center current distribution, energy consumption
 - Thermal: room air temperature, heat transport, mechanical plant
 - Reliability: thermal cycles, electric quality
 - Cost: operational expenses, capital expenses, return-on-the-investment (ROI)
- Data:
 - Measurement-based: use real workloads or systems, and sensors
 - Analytical: use models of system to estimate state variables
- Execution:
 - Real-time
 - Off-line

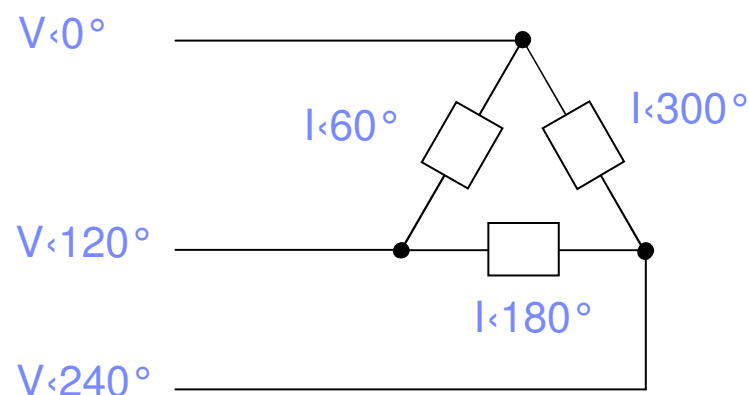
Electrical/Power Modeling

- The first stage in developing a comprehensive data center power model is to estimate power for the individual components
- Power (in the AC domain) is the complex product of the current and voltage
- Power Factor (pf) is the portion of the virtual power that is actually consumed
 - Most, but not all translates into heat
- Most systems' power supplies have power factor correction
- So, a power model can estimate power as the sum of those products
- Caveat: data center power distribution is usually done in 3-phases
 - The current is not in phase with the voltage
 - Furthermore, with unbalanced phases (unequal loads on each branch), result in changes to the angle of the current
 - Requires a calculation of the resulting power factors in situations with unbalanced phases
 - Usually, the power network is AC, which requires a complex number solver
- Tools usually have models to represent the efficiency of power delivery components
 - Transformer or cable power losses, etc.

$$\vec{P} = \vec{V} \times \vec{I}$$

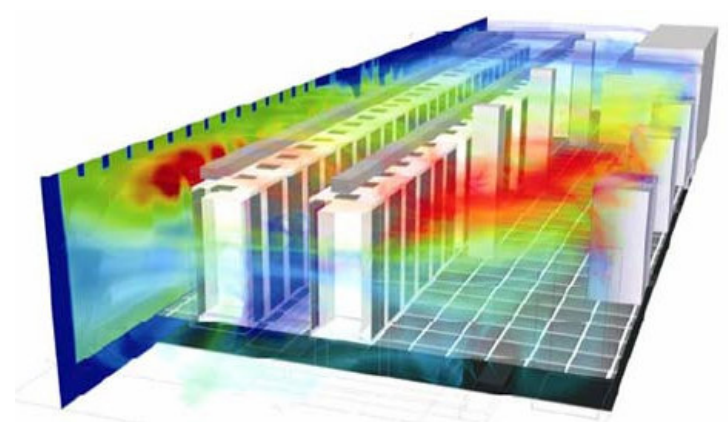
$$P_{VA} = V_{RMS} \times I_{RMS} \quad (\text{in Volt - Ampere})$$

$$P_{REAL} = pf \times V_{RMS} \times I_{RMS} \quad (\text{in Watts})$$



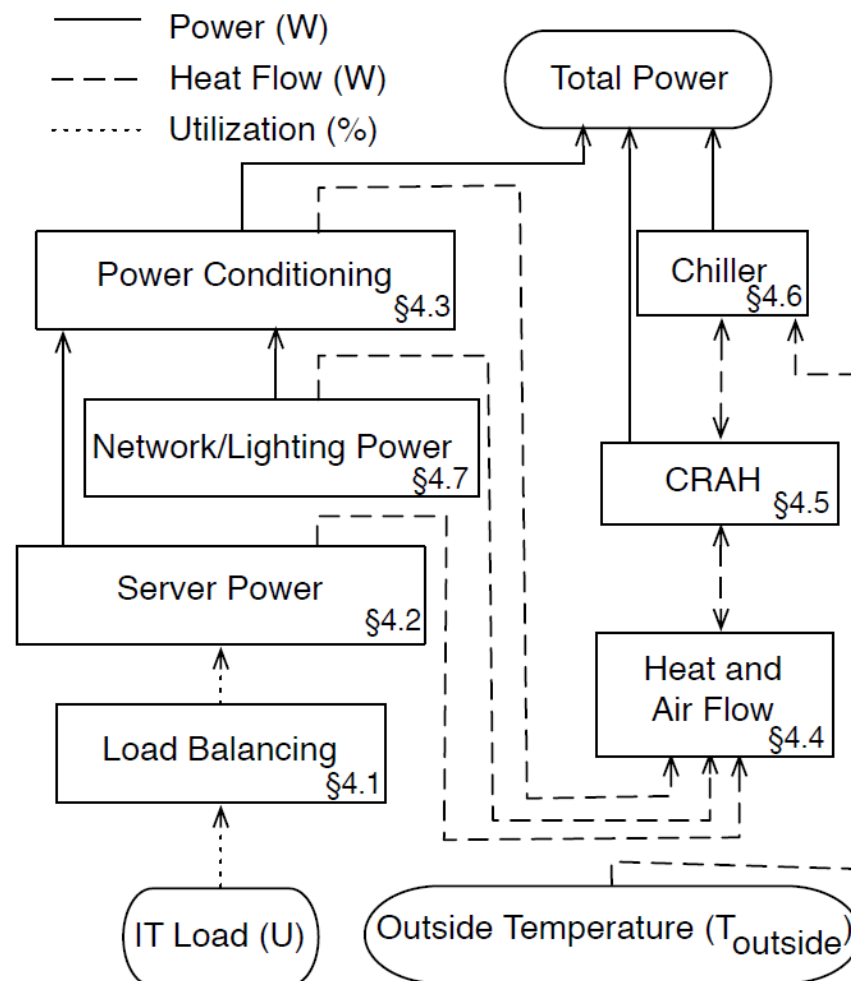
Thermal Modeling

- There are multiple methods, which use varying degrees of basic principles and system characterization
- Computational Fluid Dynamics (CFD)
 - Determine heat loads (power of system), transports (air or liquid flows), topology and boundary conditions
 - Apply finite element (FE) mathematics on system
 - Use transport and thermal equations (Navier-Stoke)
- System characterization:
 - Perform experiments on system
 - Build polynomial models for components
 - Obtain “steady-state” by solving system of equations
- Energy balance equations
 - Arithmetic tabulation of power loads and heat removal capabilities of equipment



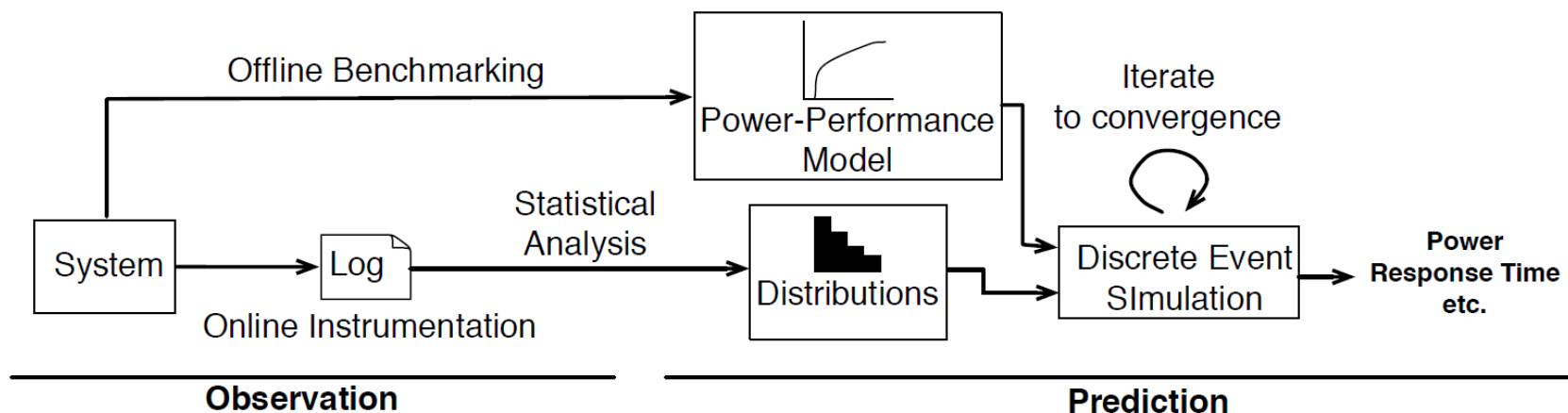
Data Center Power Model

- Source:
 - University of Michigan
 - “Understanding and Abstracting Total Data Center Power”**, Workshop on Energy-Efficient Design (WEED), held in conjunction with ISCA 2009.
- Focus:
 - Electric power of data centers
- Approach:
 - Power is a function of equipment utilization and ambient outside temperature
- Good for fast exploration of high-level what-if scenarios, with simplified models.



Stochastic Queuing Simulation

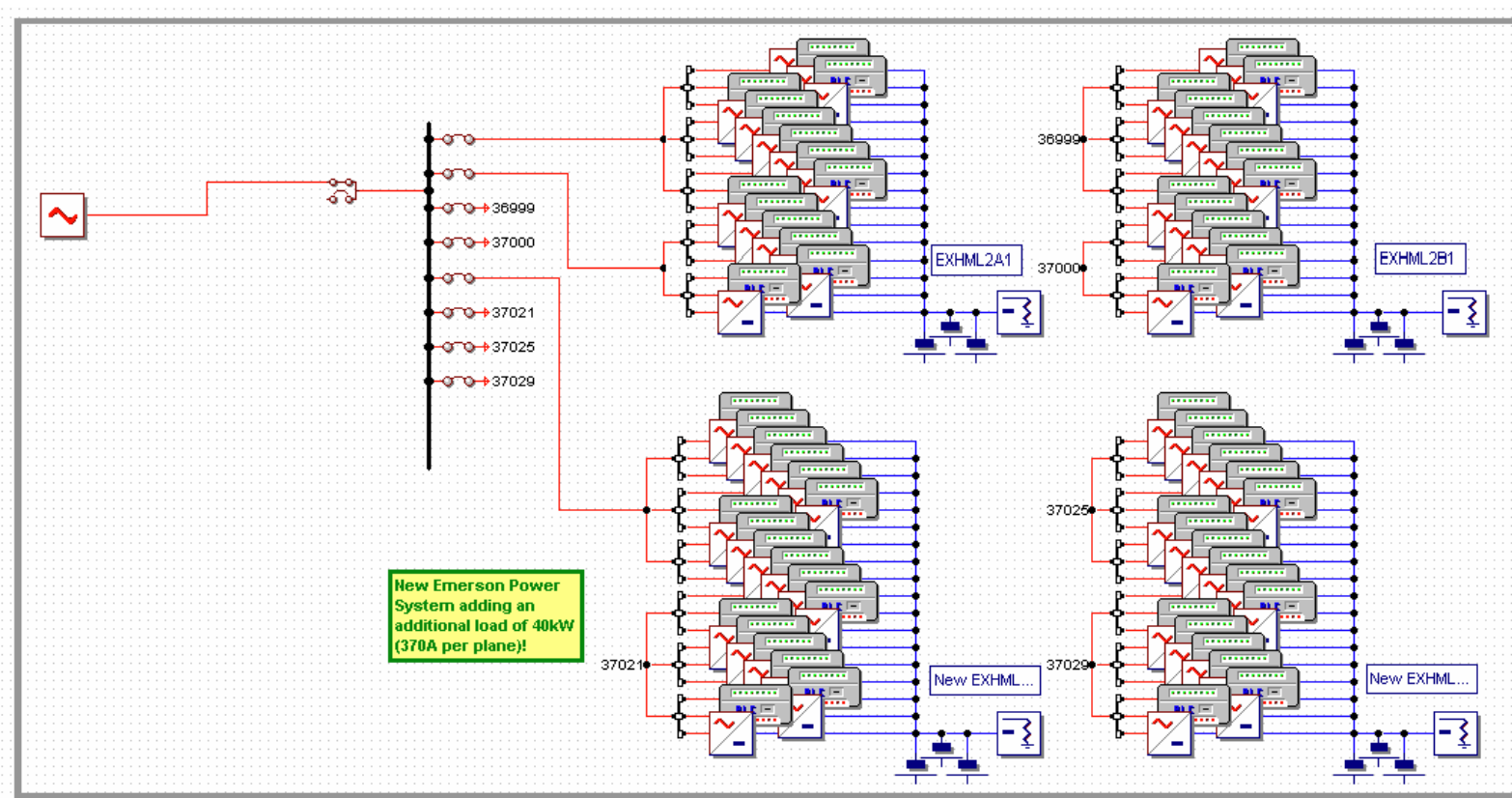
- Source:
 - University of Michigan
 - “**Stochastic Queuing Simulation for Data Center Workloads**”, Workshop on Exascale Evaluation and Research Techniques (EXERT), 2010.
- Focus:
 - Integrate workload characteristics in a data center power model
- Approach:
 - Characterize equipment
 - Characterize workloads → build distributions
- Suited for data center design, and “what-if” modeling, not for runtime management



Data Center Power Delivery Reliability Model

- Source:
 - Frank Bodi, “***Super Models in Mission Critical Facilities***”, INTELEC 2010
- Focus:
 - Electrical and mechanical modeling of power distribution and its use to detect failures
- Approach:
 - Develop an electrical model for each component
 - Models are connected according to topology of data center
- Virtual stress test
 - Monte-Carlo simulation of failures: loss of main power, loss of redundant power, switching sequences
 - Determine mean-time-between-failure (MTBF) → useful to determine equipment deficiencies

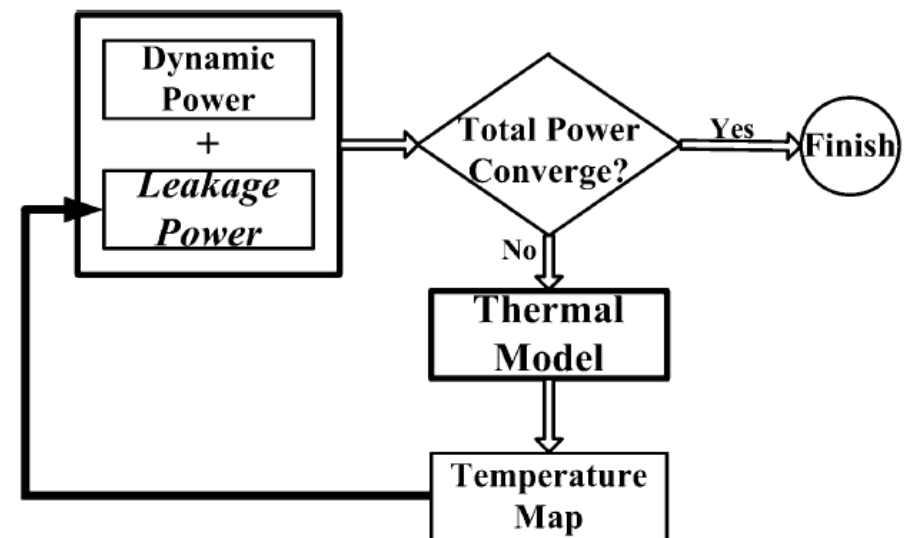
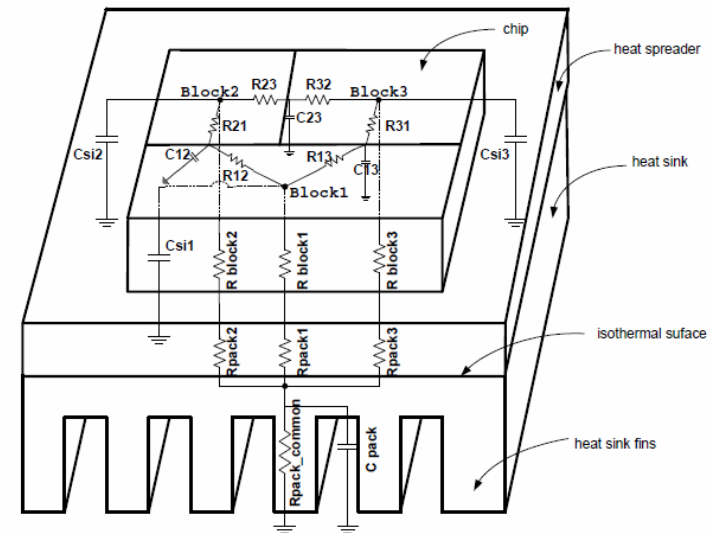
Data Center Power Delivery Reliability Model



- Components modeled:
 - Incoming AC grid, standby power generator, transfer switches, USP, transformers, power panels, breakers, computer and air-conditioning loads
- Data center is represented as a “net-list” of components

HotSpot: a Compact Thermal Model at the Processor-Architecture Level

- Source:
 - Univ. Virginia
 - “**HotSpot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design**”, Transactions of VLSI, 2006
 - <http://lava.cs.virginia.edu/HotSpot/>
- Focus:
 - Thermal modeling of microprocessors
 - Integration with power simulation
- Approach:
 - All thermal interfaces (heat sink, heat spreader, silicon) are represented as resistors or capacitors
 - Values are obtain out of “basic principles”
 - RC-network for the microprocessor and package is iteratively solved



Thermal Faults Modeling Using an RC model

- Source:
 - Univ. of Pittsburgh
 - “**Thermal Faults Modeling Using a RC Model with an Application to Web Farms**”, ECRTS, 2007
- Focus:
 - Thermal modeling of servers
- Approach:
 - Abstracts properties of the system and develops network inspired by electrical components
 - Current sources \rightarrow heat producers
 - Voltage \rightarrow temperature
 - Easy to develop hierarchical models:
 - Components \rightarrow servers \rightarrow cluster

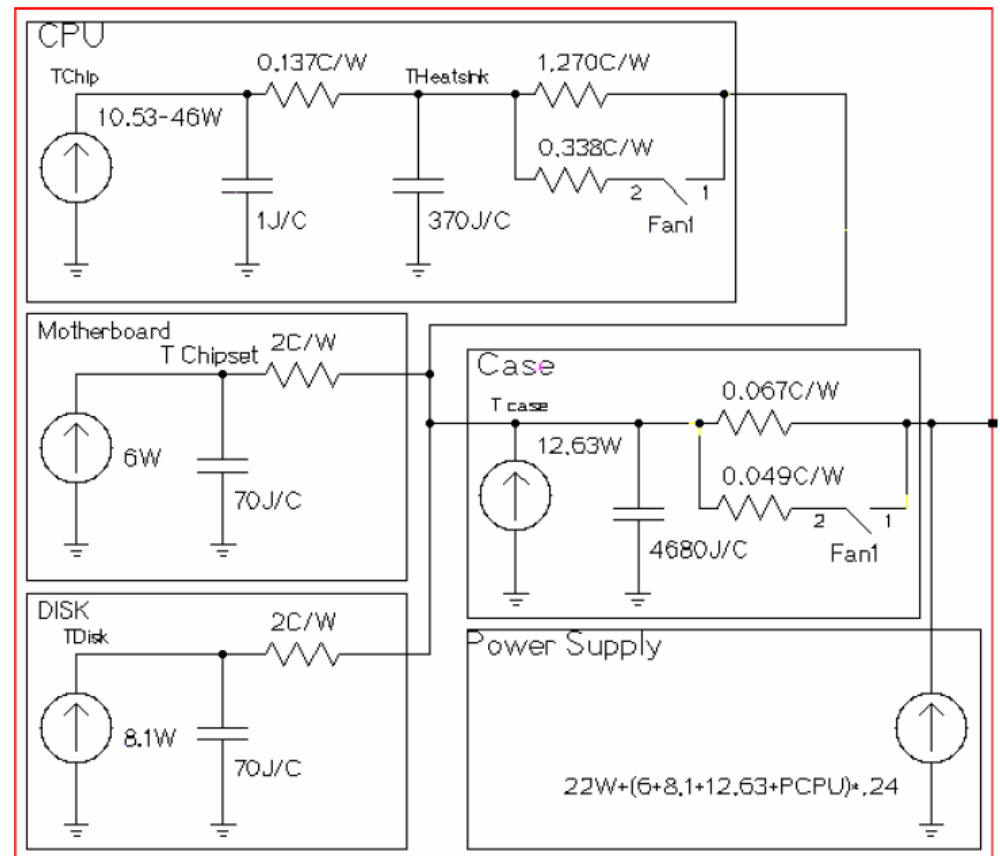
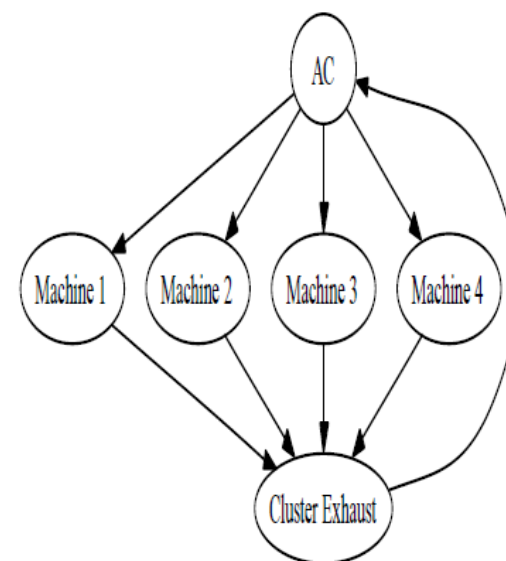
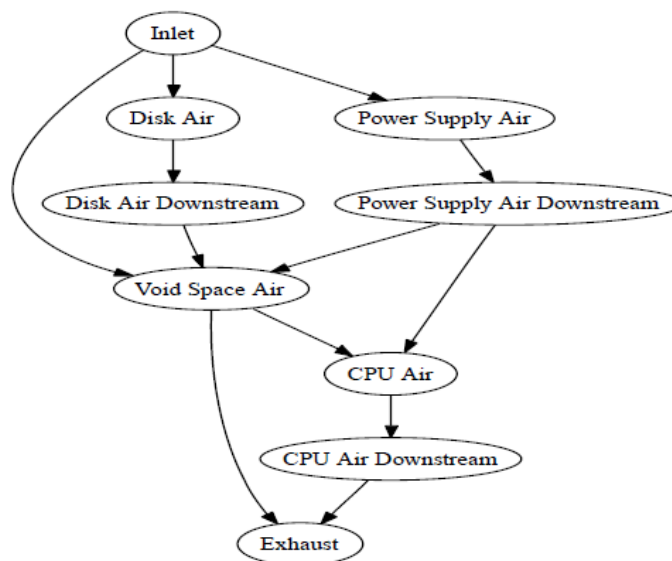
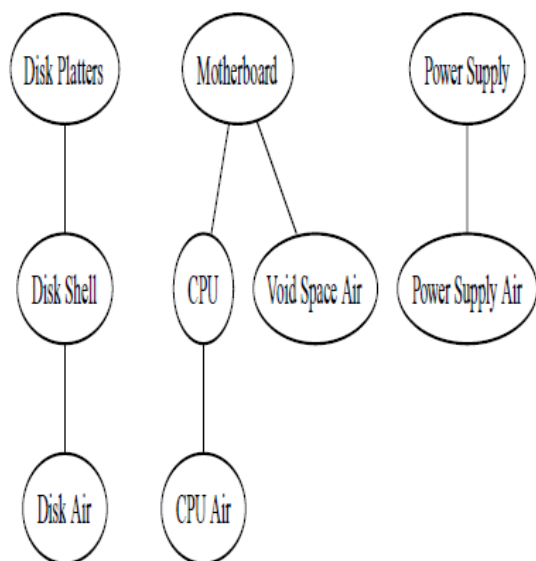


Figure 4. Thermal model for a server (in a mid-tower case).

Mercury Suite: Server-level Thermal Model

- Source:
 - Rutgers University
 - “**Mercury and Freon: Temperature Emulation and Management for Server Systems**”, ASPLOS, 2006.
- Focus:
 - Thermal modeling of servers
- Approach:
 - Build graphs to represent heat transfer paths and air flow paths
 - Based on conservation of energy laws
- Provides a good link between data center level and chip level thermal models



Mercury Suite: Server-level Thermal Model

- Basic principles:
 - Conservation of energy:

$$Q_{gained} = Q_{transfer} + Q_{component}$$

- Newton's Law of heat transfer. Heat transfer from region of highest to lowest temperature according to:

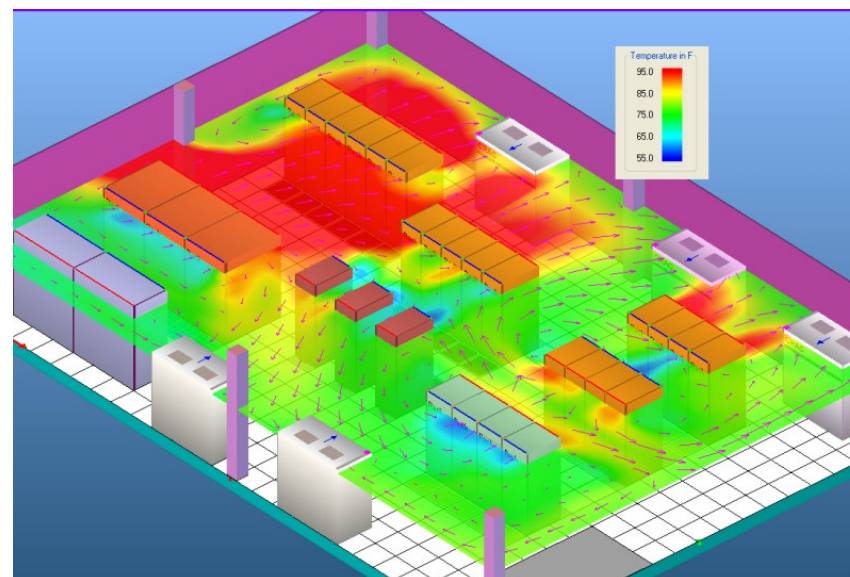
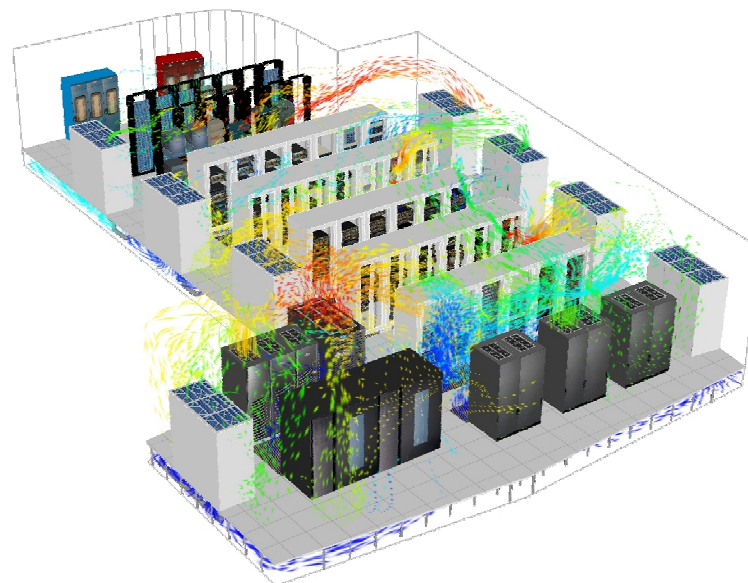
$$Q_{transfer,1 \rightarrow 2} = k \times (T_1 - T_2) \times time$$

- 'k' is a function of the heat capacity of the material, the area exposed to airflow, the speed of the air, its moisture content and pressure
- In this work, 'k' is assumed constant and computed empirically.

Component	Property	Value	Unit
Disk Platters	Mass	0.336	kg
	Specific Heat Capacity	896	$\frac{J}{K kg}$
	(Min, Max) Power	(9, 14)	Watts
Disk Shell	Mass	0.505	kg
	Specific Heat Capacity	896	$\frac{J}{K kg}$
CPU	Mass	0.151	kg
	Specific Heat Capacity	896	$\frac{J}{K kg}$
	(Min, Max) Power	(7, 31)	Watts
Power Supply	Mass	1.643	kg
	Specific Heat Capacity	896	$\frac{J}{K kg}$
	(Min, Max) Power	(40, 40)	Watts
Motherboard	Mass	0.718	kg
	Specific Heat Capacity	1245	$\frac{J}{K kg}$
	(Min, Max) Power	(4, 4)	Watts
Inlet	Temperature	21.6	Celsius
	Fan Speed	38.6	ft^3/min

Other Commercial Thermal Tools

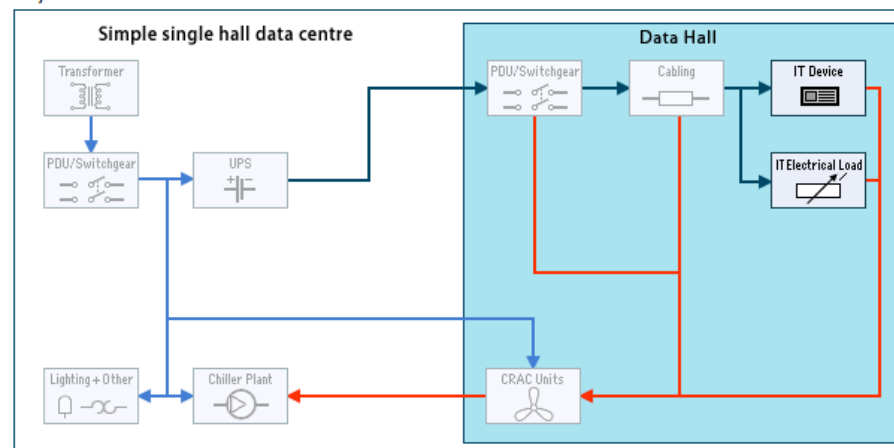
- General purpose simulation tools
 - Accurate and high resolution CFD (computational fluid dynamics)
 - High performance and parallel solvers
- Popular tools:
 - ANSYS: Fluent
 - Mentor Graphics: FloTherm
 - SolidWorks: FloWorks
- Challenges:
 - Not particularly addressing data centers
 - Need accurate input data (dimensions of equipment, thermal properties, etc.)
 - Steep learning curves
 - Can be expensive
- New breed of tools suited for data center modeling
 - ANSYS: CoolSim
 - Future Facilities: 6Sigma
 - Innovative Research: TileFlow
 - Mentor Graphics: FloVent



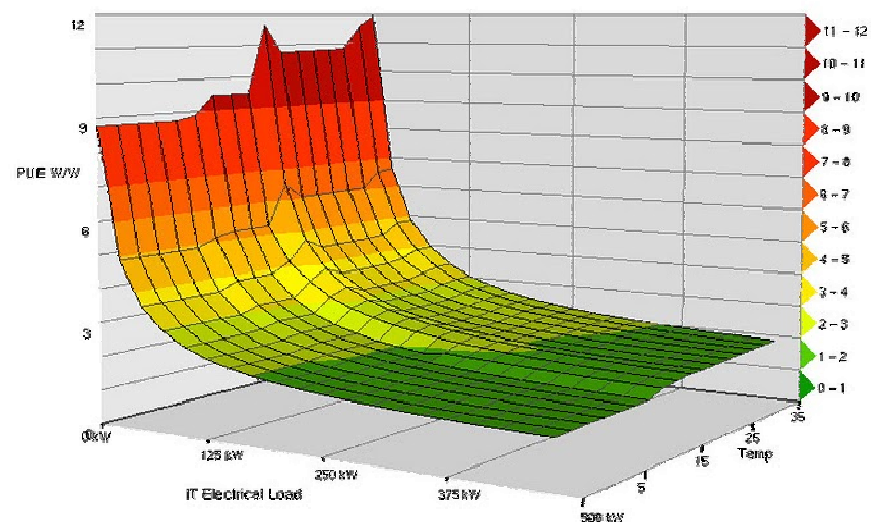
BCS (British Computer Society) Data Centre Simulator

- Source:
 - <http://dcsg.bcs.org/welcome-dcsg-simulator>
 - <http://www.romonet.com/content/prognose>
- Focus:
 - High-level data center cost and energy simulator
- Approach:
 - Use component efficiency curves from manufacturers
 - Build electrical and thermal “topology” of data center
 - Energy balance of components
- Reports:
 - Detailed information and classification of energy consumption
 - Simulation across seasons
 - Cost

Layout:

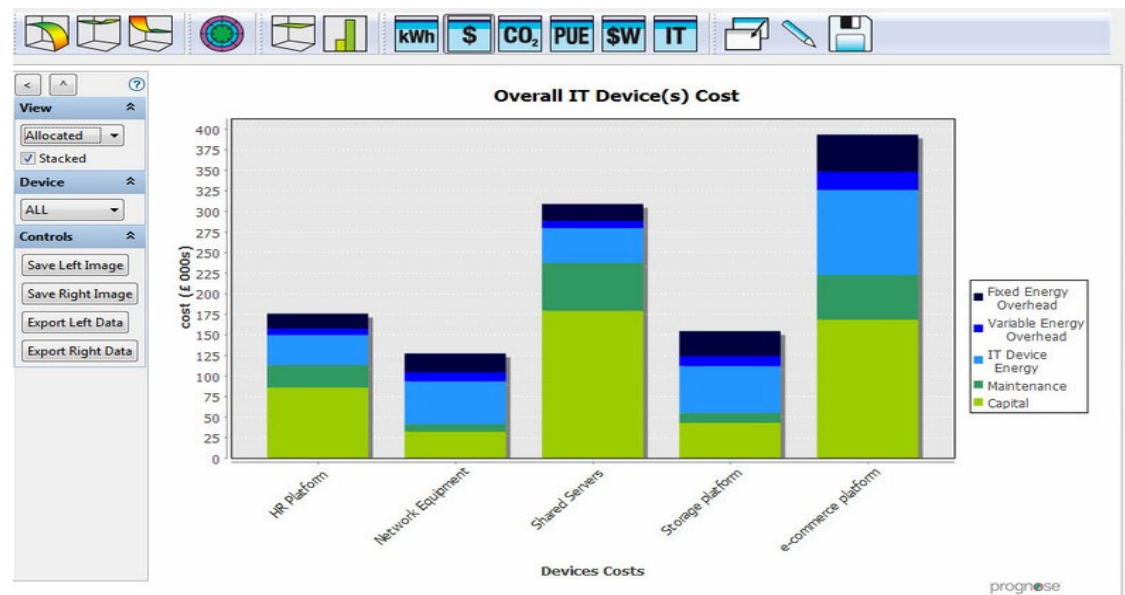
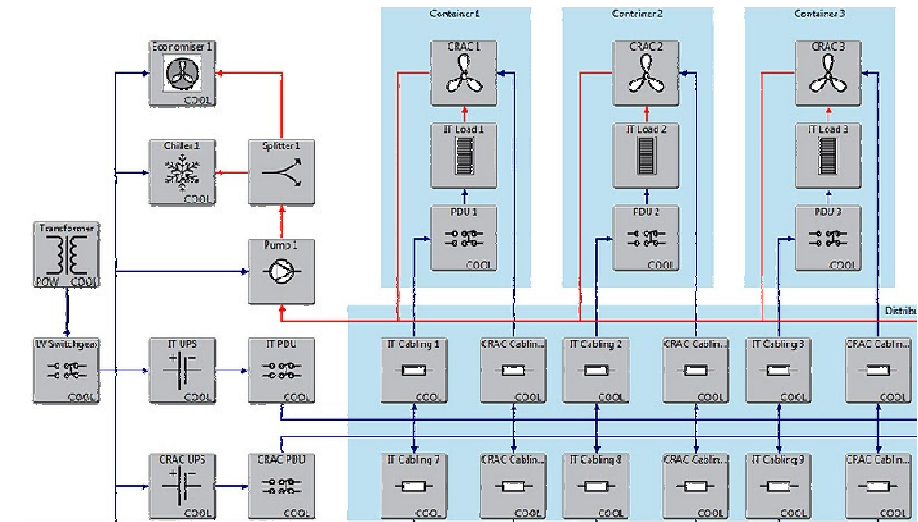


PUE Chart - Sample data centre



BCS (British Computer Society) Data Centre Simulator

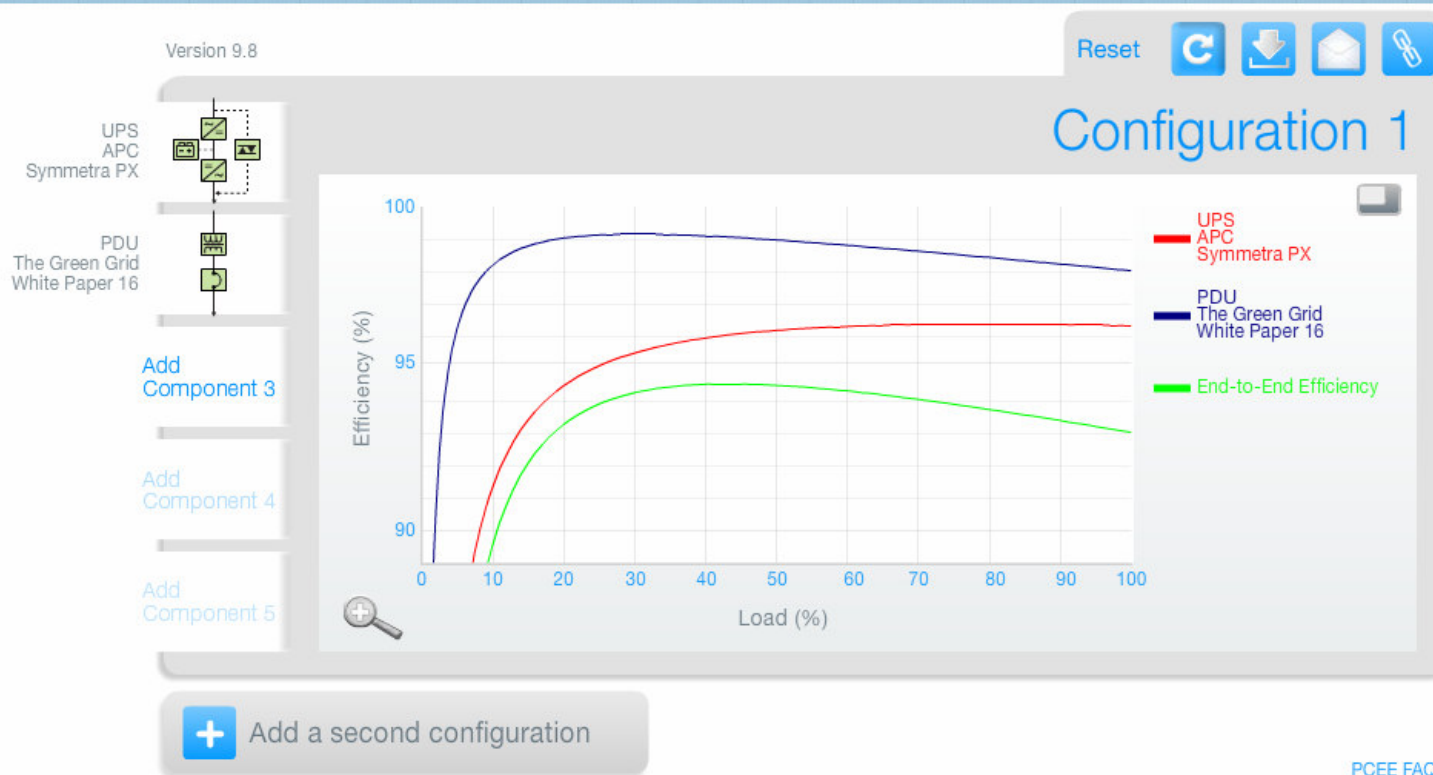
- Tool has great topology input capabilities that allow modeling of complex systems
- Reports:
 - Detailed information and classification of energy consumption
 - Simulation across seasons
 - Cost



GreenGrid Data Center Modeling Work

- Source:
 - The Green Grid consortium
 - <http://thegreengrid.org/library-and-tools.aspx?category=All&type=Tool>
- Focus:
 - Address multiple aspects of data center energy efficiency
- Approach:
 - High level tools, useful for planning or rough estimation
- Power Usage Effectiveness Estimator
 - <http://estimator.thegreengrid.org/puee>
 - <http://estimator.thegreengrid.org/pcee>
- Free-cooling Estimated Savings
 - For US http://cooling.thegreengrid.org/namerica/WEB_APP/calc_index.html
 - For Europe http://cooling.thegreengrid.org/europe/WEB_APP/calc_index_EU.html
 - For Japan http://cooling.thegreengrid.org/japan/WEB_APP/calc_index_jp.html

Power Configuration Efficiency Estimator



[Legal Disclaimer](#)

[PCEE FAQ](#)

[PCEE Help](#)

If you have any questions or comments, please contact powerdata@lists.thegreengrid.org.

Power Usage Effectiveness Estimator

Version 6.0

Facility Name

PUE 1.3	Facility Power Total	57.2 kW	Module Chain
			Number of Modules: 0
DCIE 0.769	IT Equipment Power Total	44 kW	Module Total
			0kW
			Module IT Equipment Power Total
			0kW

[What is the core?](#)

Core
pPUE: 1.3

Core Power Total	57.2 kW
IT Equipment Power Total	44 kW

[What is a module?](#)

[Legal Disclaimer](#)

[PUEE Help](#) [PUEE FAQ](#)

Core [Add Description](#)

Infrastructure Infrastructure Total: 13.2 kW

Type	Sub-Type	Dissipation ?	Qty.	Notes
HVAC	Computer Room Air Condit	3.4 kW	3	
Power	Power Distribution Units (P	1 kW	2	
Power	Lighting	0.1 kW	10	

[+ Add a Component](#)

IT Equipment IT Equipment Total: 44 kW


Type	Sub-Type	Dissipation ?	Qty.	Notes
Compute Devices	Servers	1.1 kW	40	

[+ Add a Component](#)

[Reset](#)

If you have any questions or comments, please contact puee-comments@lists.thegreengrid.org.

© 2011 IBM Corporation


the green grid™
 get connected to efficient IT

Free-Cooling Estimated Savings

US/CANADA LOCATION (ZIP CODE):

DEGREES IN: ☒ FAHRENHEIT ☐ CELSIUS

ALLOW MIXING OF SUPPLY AND RETURN AIR ☒

ALLOW HUMIDIFICATION ☒

MAX LIMIT

MIN LIMIT

DRYBULB TEMP THRESHOLD (DEG):

DEWPOINT TEMP THRESHOLD (DEG):

REL. HUMIDITY THRESHOLD (%):

DESIRED CHILLED WATER TEMP (DEG):

COOLING SYSTEM APPROACH TEMP (DEG):

[Comment Now](#)

DATA CENTER IT POWER (kW):

POWER USAGE EFFECTIVENESS (PUE):

TOTAL FACILITY POWER (kW):

OVERHEAD POWER (kW):

PERCENT OF OVERHEAD POWER FOR COOLING SYSTEM (%): % kW

PERCENT OF COOLING SYSTEM POWER FOR CHILLER (%): % kW

PERCENT OF COOLING SYSTEM POWER FOR TOWER (%): % kW

PERCENT OF COOLING SYSTEM POWER FOR PUMPS/FANS (%): % kW

PERCENT OF OVERHEAD POWER FOR POWER LOSSES and LIGHTING (%): % kW


ELECTRIC COST (\$ per kWh)

HOURS MEETING CRITERIA FOR FREE-AIR COOLING:

ESTIMATED SAVINGS USING FREE-AIR COOLING:

HOURS MEETING CRITERIA FOR WATER SIDE ECONOMIZER:

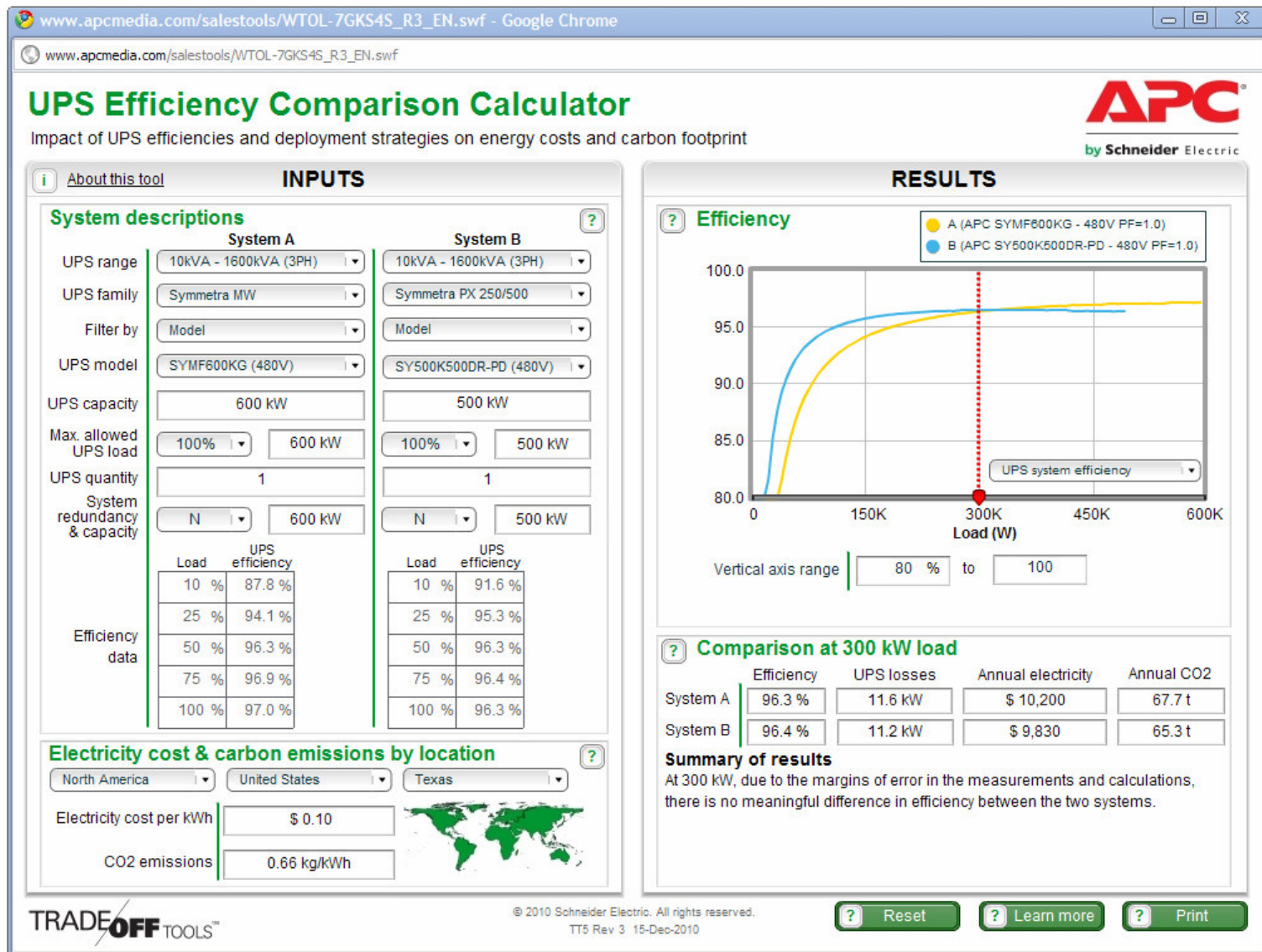
ESTIMATED SAVINGS USING WATER SIDE ECONOMIZER:

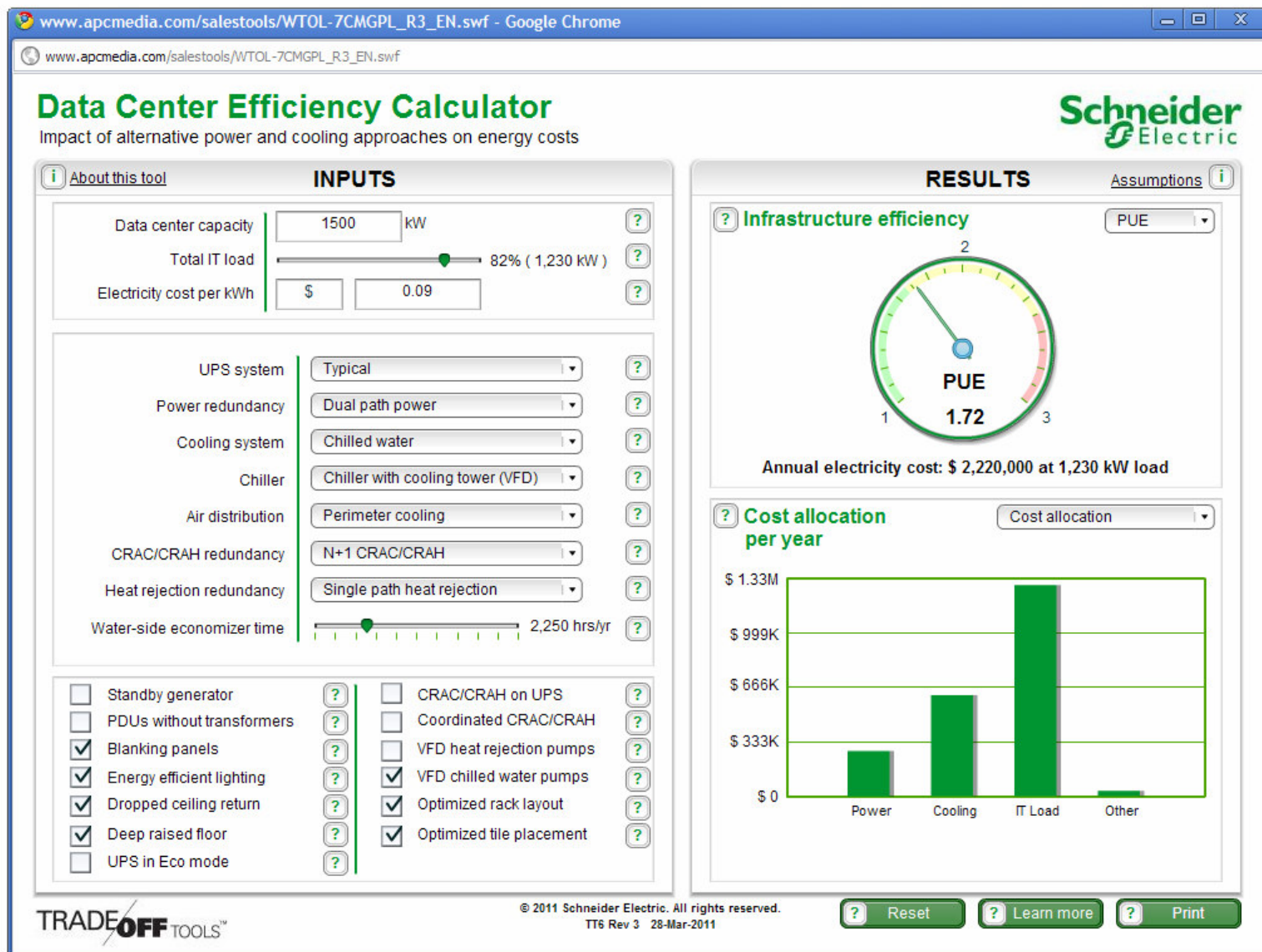
APC Models to Estimate Cost of Ownership

- Source:
 - APC (American Power Conversion), manufacturer of data center infrastructure equipment
 - <http://tools.apc.com/>
 - <http://www.apc.com/tools/isx/tco/>
- Focus:
 - Model energy efficiency of APC's InfraStruxure™ components to show their TCO (total cost of ownership)
- Approach:
 - Use energy efficiency curves for components
 - “Arithmetic” tabulation of energy consumption and cost

APC Models – UPS Efficiency Comparison Calculator



APC Models – PUE Calculator



APC Models – Cost of Ownership

InfraStruxure™

InfraStruxure™

InfraStruxure™ Total Cost of Ownership

Please provide your information below, then click "Get Results." Use the help button next to each variable name for further explanation.

Basic Input

Design Power Rating (Watts)	<input type="text" value="120000"/>
2N Design?	<input type="radio"/> Yes <input checked="" type="radio"/> No
Force APC Solution to use a Raised Floor?	<input type="radio"/> Yes <input checked="" type="radio"/> No

Optional Input

Starting Load - % of Design Power	<input type="text" value="10"/>
Ending Load - % of Design Power	<input type="text" value="30"/>
Average Rack Power (Watts/Rack)	<input type="text" value="1500"/>
Infrastructure Design Life (Years)	<input type="text" value="10"/>
Data Center Space per Rack (sq ft/Rack)	<input type="text" value="30"/>
% of Infrastructure Design Life when Ending Load is reached	<input type="text" value="50"/>
Installation Labor Rate	<input type="text" value="Typical"/> ▼
Electricity Cost (\$/kw-hr)	<input type="text" value="0.07"/>

Reload Default Values

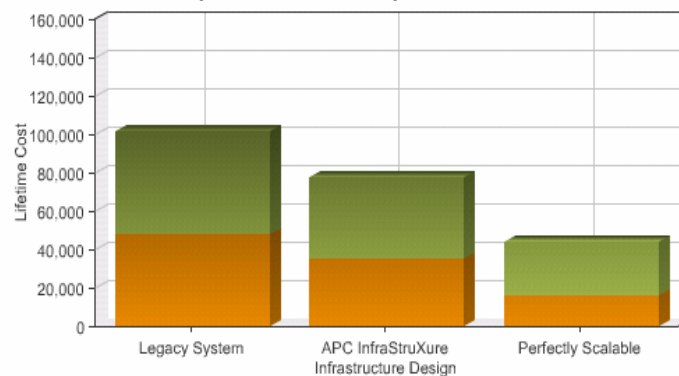
Get ResultsAdvanced Options

APC Models – Cost of Ownership

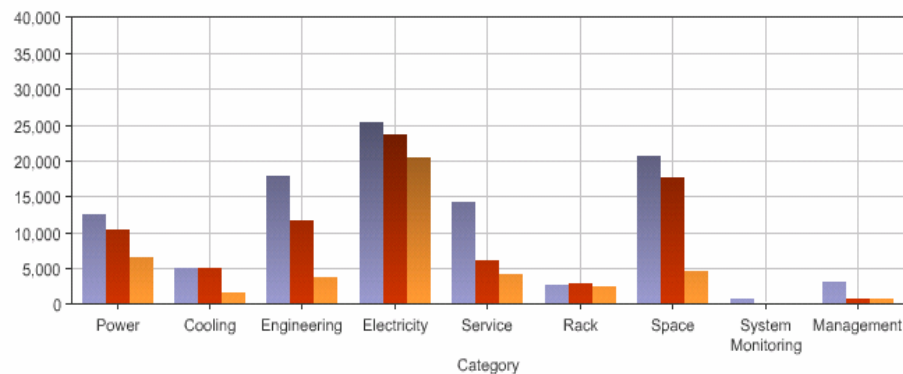
Design Life Costs

	Legacy System	APC InfraStruXure	Perfectly Scalable	APC cost savings over legacy
Total Cost (CAPEX + OPEX)	\$2,043,859.67	\$1,563,670.77	\$887,382.72	23.49%
Cost per Rack	\$102,192.98	\$78,183.54	\$44,369.14	

Comparison of Lifetime Expenses - Per Rack



TCO by Category / Average Racks Used over Lifetime

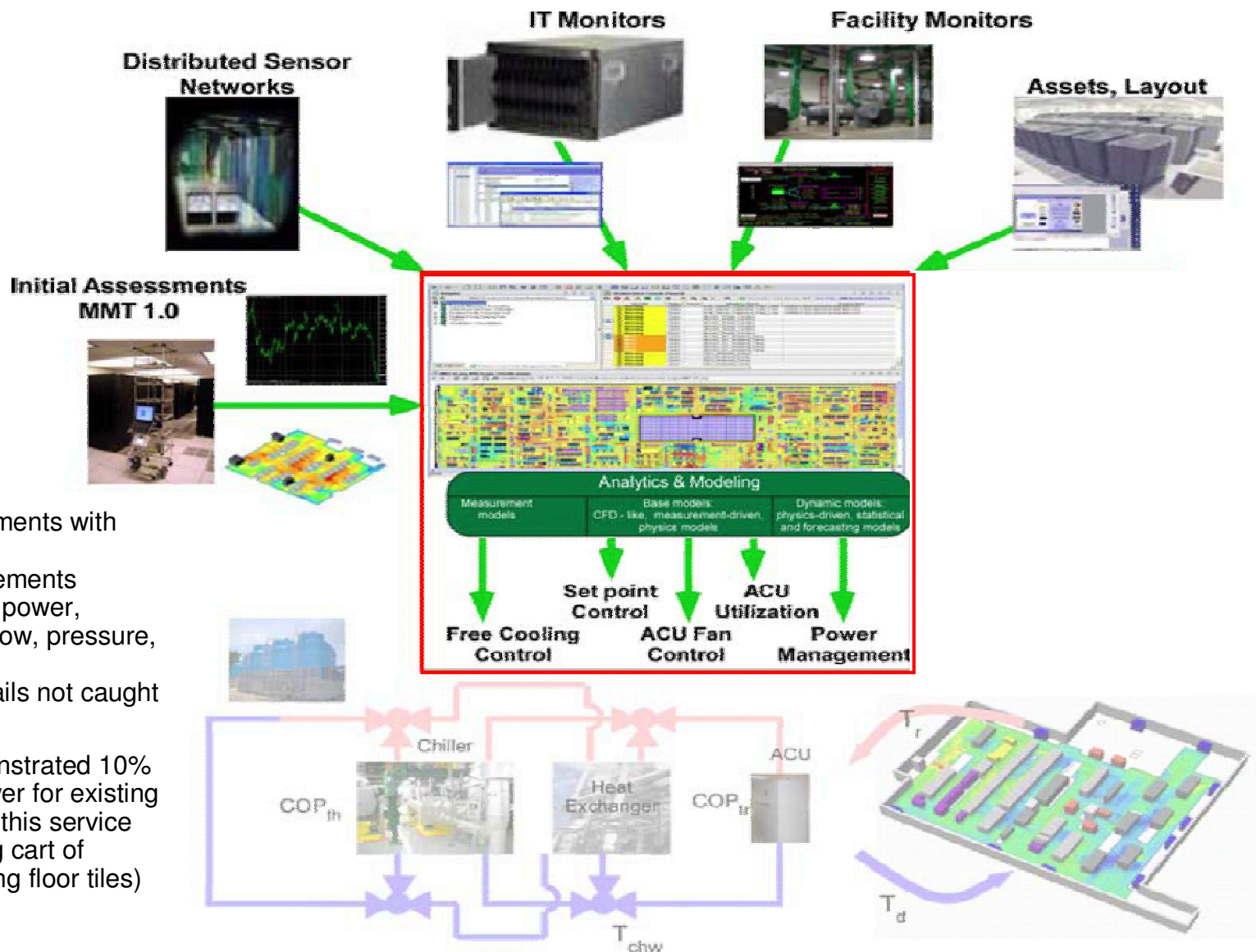


IBM Measurement and Management Technology (MMT)

- Source:
 - IBM Research
 - “***Measurement-based modeling for data centers***”, IThERM, 2010
- Focus:
 - Thermal modeling of data centers for improving energy efficiency
- Approach:
 - Thermal scanning of data center → build thermal model
 - Thermal sensors are strategically placed → allows using simpler version of thermodynamic equations



IBM Measurement and Management Technology (MMT)



- Coupling measurements with models
 - Point measurements (temperature, power, humidity, air flow, pressure, etc)
 - Model for details not caught by sensors
- On average, demonstrated 10% overall cooling power for existing data centers using this service (with only a moving cart of sensors, and moving floor tiles)

Research Opportunities

▪ **Integration of all models:**

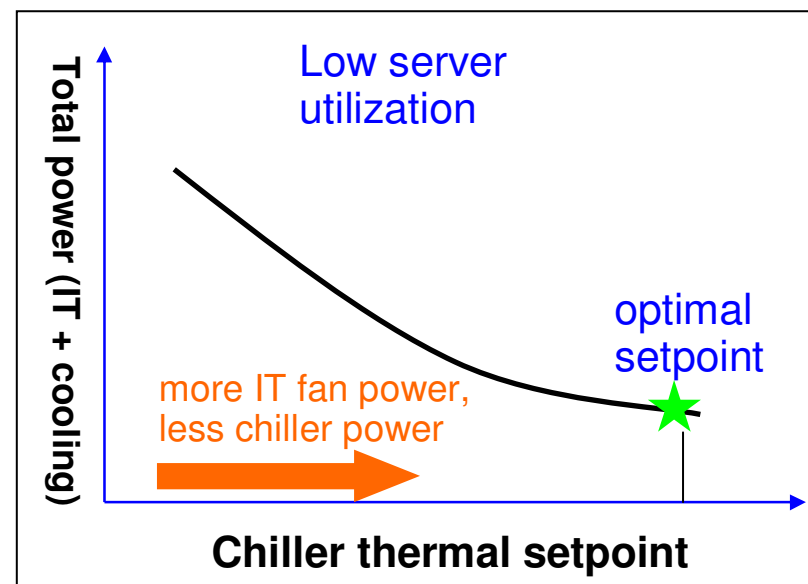
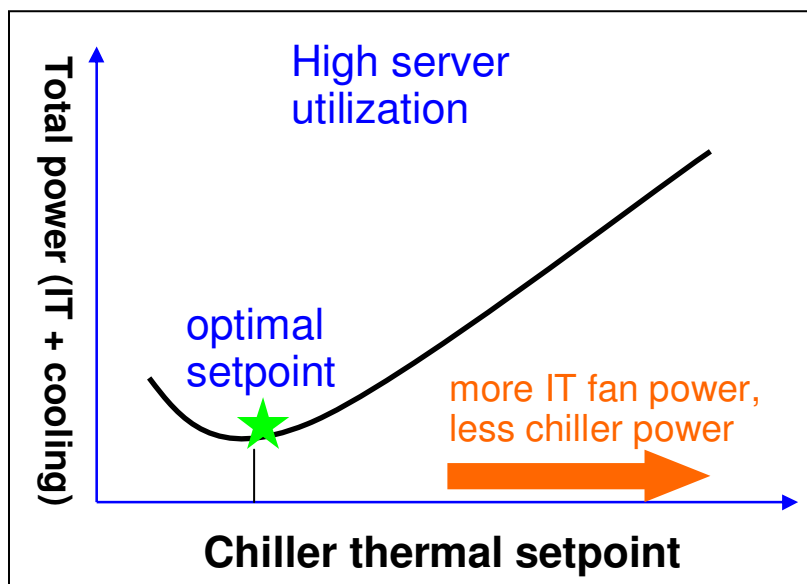
- Different purposes (performance, IT and cooling power, cost, reliability)
 - Is there a need for a tightly-coupled “supermodel”?
 - What is the right interface between multiple simulation domains?
- Time scales in different models (ns, us, ms, sec, min, hour)
 - What is the right time granularity to keep for each domain?
- Spatial scales (chips → servers → racks → data centers)
 - How to scale the simulation to large data centers?
- Advantage:
 - What do we gain by integrating those models?

▪ **Hybrid models:**

- Off-line models are for planning and design
- Real-time models requires sensor data

Thermal-Aware Power Optimization (TAPO) – Optimizing Total Power

- Tradeoff between data center cooling power and IT/server fan power
 - Higher IT/server inlet temperature → less CRAH power, higher server fan power
 - Server fan is limited in form factor, can't use large, power-efficient fans.
 - Fan power is quadratic/cubic to cooling capability
- Total power is a strong relationship with IT utilization per cooling zone
 - Low utilization favors warm inlet temperature
 - High utilization favors cool inlet temperature
- Binary control of CRAH setpoint is close to optimal
 - >10% total Data Center power reduction
- Wei Huang, et al., IGCC 2011, best paper



Reliability

Reliability-Aware Power Performance Optimization

- Reliability has a power cost
- Power management can affect reliability
- Reliability considerations
 - Data center tier classifications
 - Redundant branch circuits increase power delivery infrastructure costs
 - Redundant power delivery components in servers increase power delivery infrastructure costs
 - Chip aging leads to higher operational voltages, reducing energy efficiency
- Power management considerations
 - Thermal cycling of chips and early failures of packages
 - Disk drive failure rates due to spin downs to save power
 - Fan failure rates due to power cycling fans

History of Reliability in Data Centers and the Concept of Tiers

- IT customers expect availability of “Five Nines” or 99.999%
- Although hardware and software platforms may meet Five Nines, the complementary site infrastructure can’t support these availability goals
- Uptime Institute’s Tier Performance Standards established in 1995 has become the default standard for the uninterruptible uptime industry
 - See “Tier Classifications Define Site Infrastructure Performance” white paper
- Actual measured site availability ranging from 99.67% to 99.99%
- Substantially less than Five Nines: conclusion is site availability limits overall IT availability
- Highest tier, Tier IV, first appeared in 1995 with UPS Windward data center project working with IBM and other vendors
 - Requires at least two completely independent electrical systems connected to two redundant power supplies in all IT equipment
 - Last point of electrical redundancy is between UPS and actual IT equipment
 - Human factors are important because 70% or more of all site failures involve people
 - 4 hours for IT to recover from a failure leads to 1 failure every 5 years for 99.995% availability

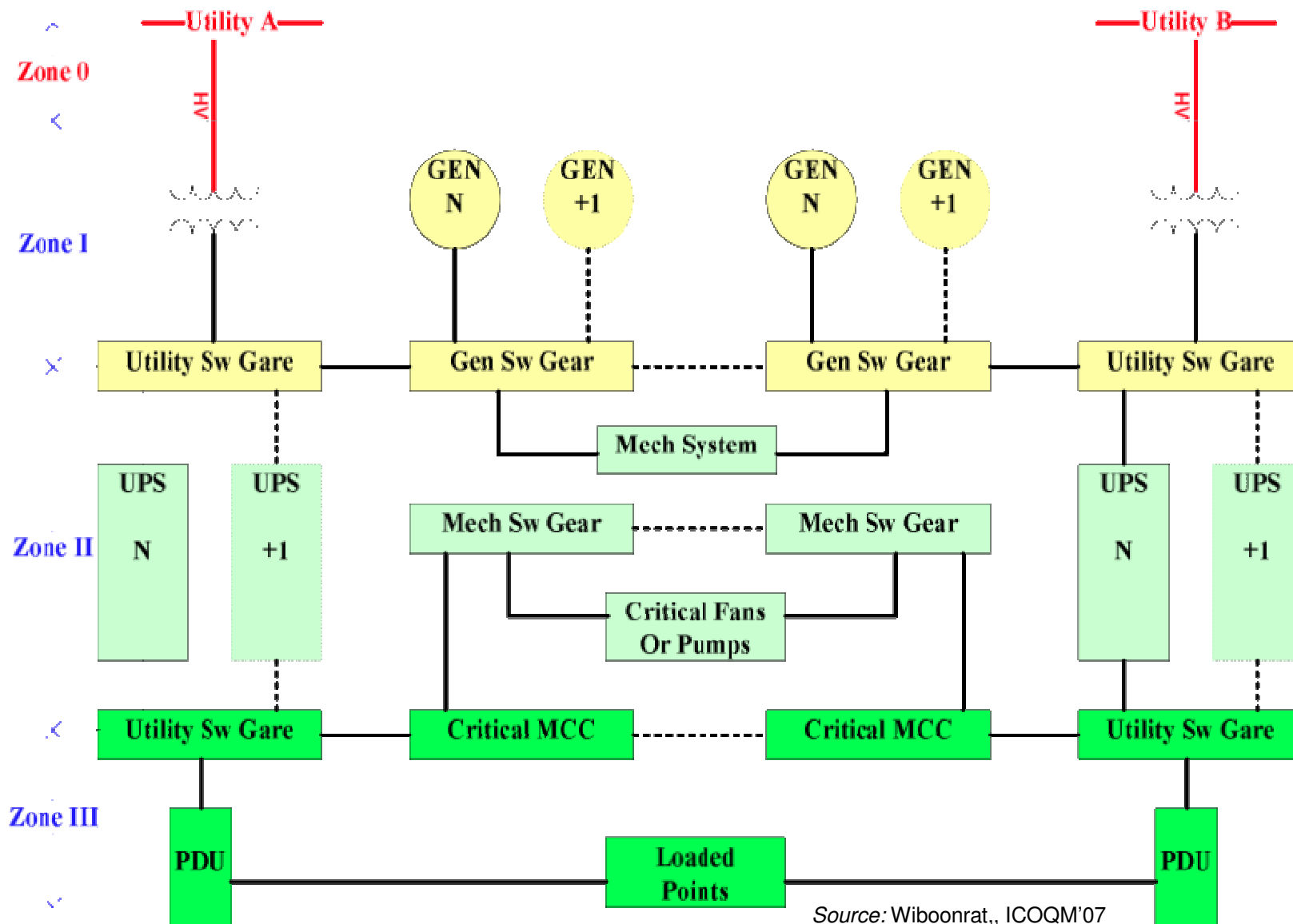
Data Center Tier Descriptions Based on Power, Cost, Reliability

	Tier I	Tier II	Tier III	Tier IV
Utility Voltage (typical)	208, 480	208, 480	12-15kV	12-15kV
Single Points-of-Failure	Many + human error	Many + human error	Some + human error	None + Fire and EPO
Annual Site Caused IT Downtime (actual field data)	28.8 hours	22.0 hours	1.6 hours	0.8 hours
Representative Site Availability	99.67%	99.75%	99.98%	99.99%
Typical Months to Implement	3	3-6	15-20	15-20
Year first deployed	1965	1970	1985	1995
Construction Cost: Raised Floor Usable UPS Output	\$220/sq ft \$10,000/kW	\$220/sq ft \$11,000/kW	\$220/sq ft \$20,000/kW	\$220/sq ft \$22,000/kW

Costs based on 2005 estimates

Source: Uptime Institute white paper: "Tier Classifications Define Site Infrastructure Performance"

Data Center Up-Time Institute Tier 4 Reliability Support

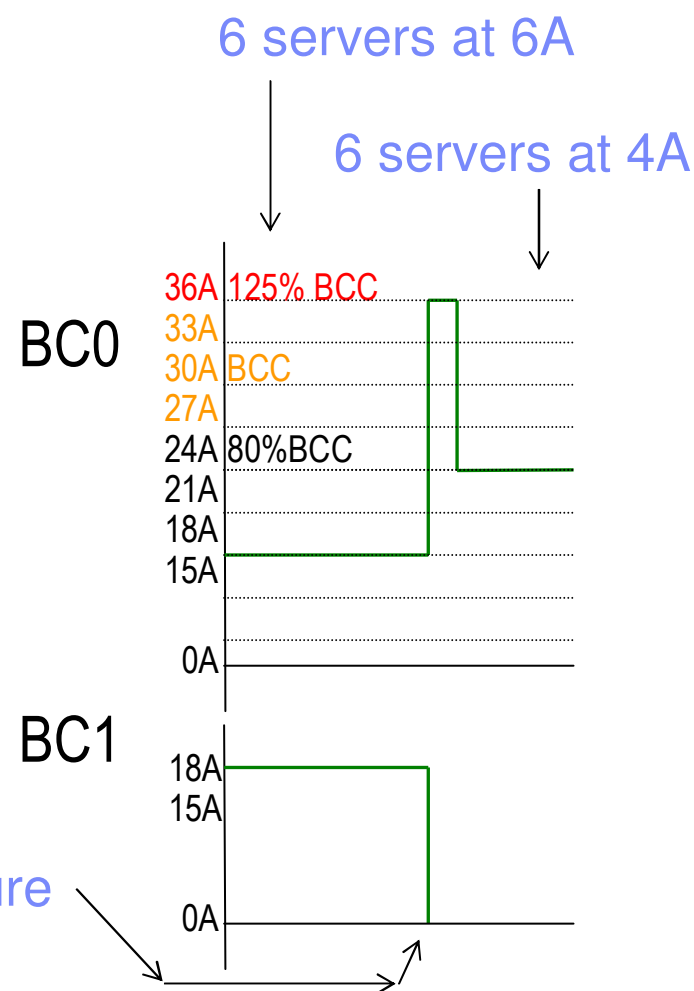
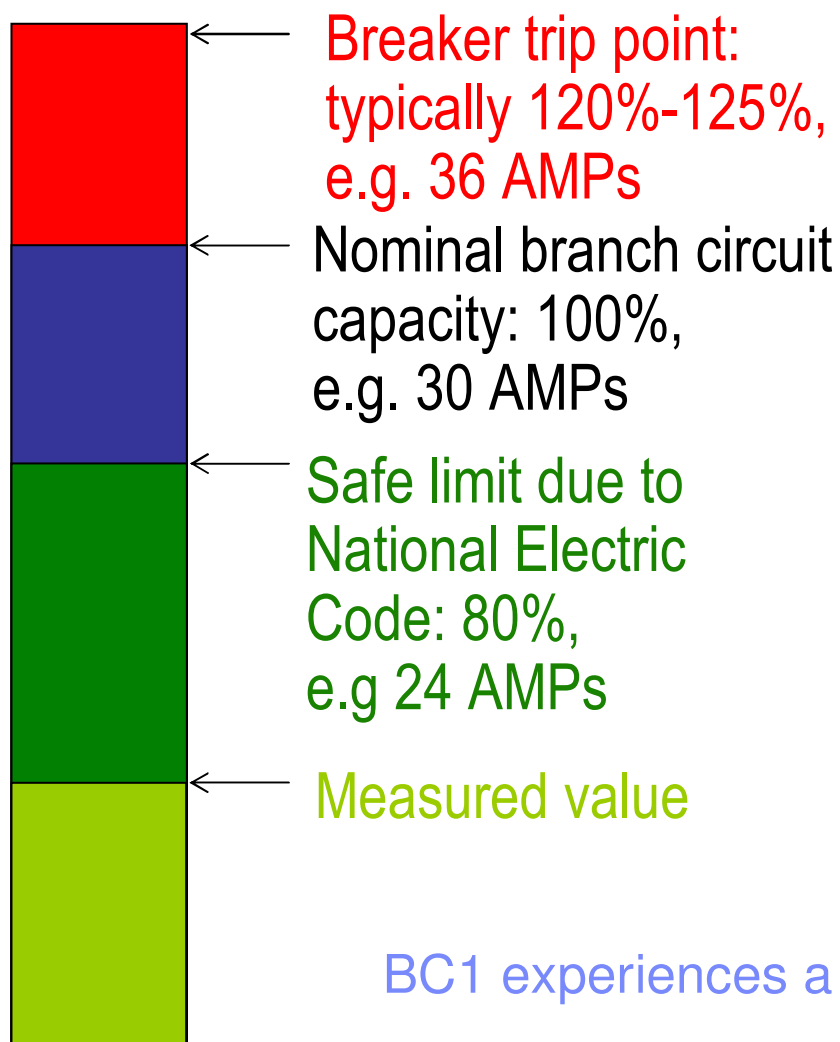


Source: Wiboonrat,, ICOQM'07

Oversubscription of Branch Circuits to Increase Equipment Density in a DC

- The cost of power delivery infrastructure is high, especially for Tier IV data centers
 - Motivation here to use that expensive infrastructure investment and pack more than 50% more equipment than is used today without impacting reliability, but some compromise in performance
- Example: branch circuits and racks
 - Consider a typical data center rack has two branch circuits feeding it for redundancy, say BC0 and BC1, with 30 AMPs per phase on each branch circuit
 - Call this current capacity per branch circuit BCC
 - Within a rack, all the servers are fed from two independent power strips, each power strip fed by $0.8 \times \text{BCC}$ AMPs of current
 - Fuse limit is $1.25 \times \text{BCC}$ AMPs, this means each branch circuit could handle this for a short number of seconds
 - This leads to the actual potential amount of density improvement with oversubscription as $(1.25/0.8)=1.5625$ or a 56.25% denser IT equipment with same Tier IV uptime and redundancy support

Branch Circuit Power (example: 30 AMPs)



Component Redundancy and Reducing Guardbands

- Component redundancy reductions are being pursued to reduce power delivery infrastructure costs through oversubscription
- Guardbands are used for reliable operation of chips
 - Definition of guardband for chips: amount of additional margin in a key parameter, e.g. voltage, to assure chip timing never fails under all worst case scenarios including aging and workloads
- Energy efficiency gains are being pursued from guardband reductions

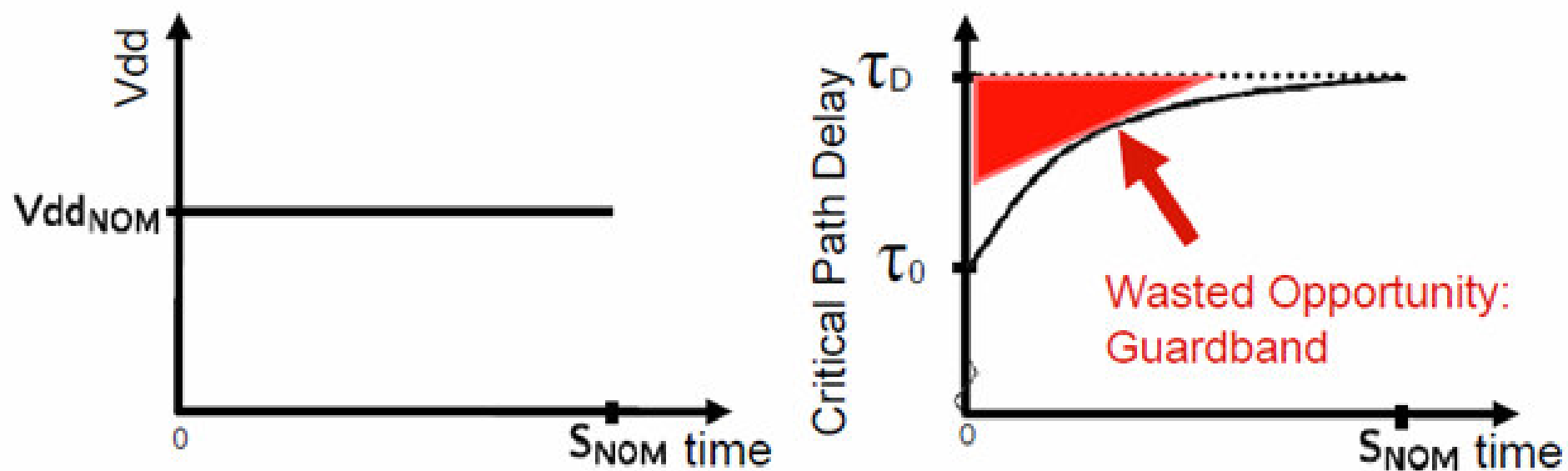
Component Redundancy in Servers Impacts Cost of Power Delivery

- Redundancy in components adds cost to server design
- Voltage Regulators with additional phases for current delivery, fail in place
- Two power supplies, each capable of handling the full load of the server
 - With power capping, new means to “oversubscribe” the supplies so that one power supply can’t handle the full load of the server, but server still continues operation if one of the two supplies fails
 - Oversubscription of power supplies is analogous to the branch circuit oversubscription described earlier

Guardband Reduction at the Chip Level

- Chip aging, workload variability leading to higher voltages over chip lifetimes increasing power consumption
- Josep Torrellas, from the University of Illinois, has a representative research project on this topic
 - Variation-Tolerant Architectures
 - Describes active voltage management techniques to manage aging of chips
 - Reducing voltage guardbands over lifetime of the chip can increase power efficiency

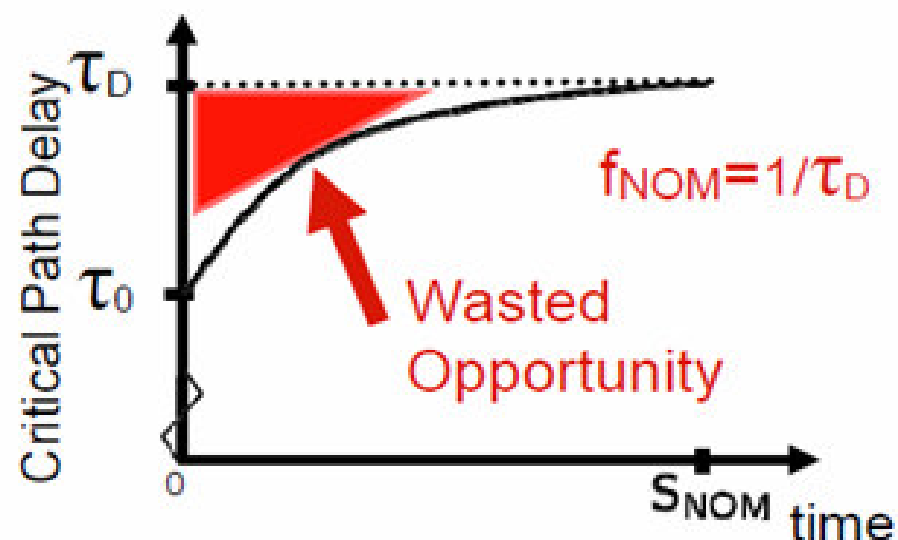
Aging-Induced Degradation



Source: Josep Torrellas, "Variation-Tolerant Architecture"

Managing Aging

Contribution: DVS for Aging Management (DVSAM)

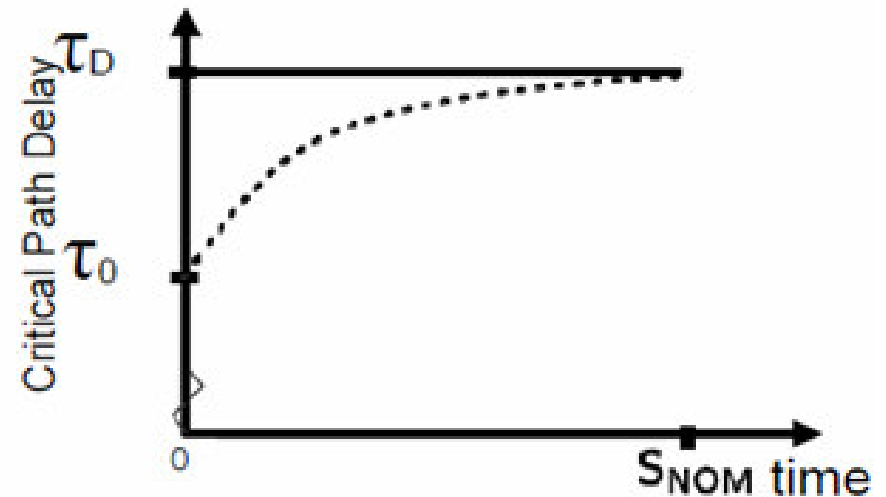
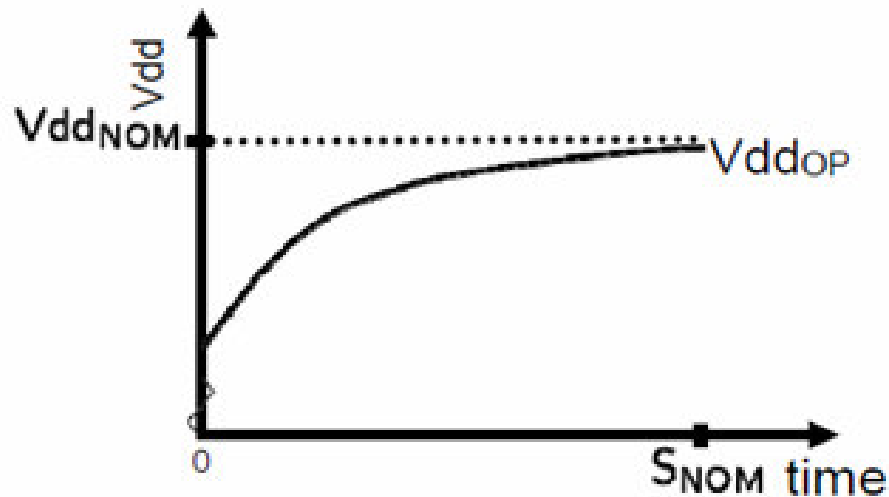


- Continuously change V_{dd} with time to compensate for critical path degradation

- **DVSAM-Pow**: Turn wasted opportunity to power efficiency
- **DVSAM-Perf**: Turn wasted opportunity to higher frequency

Source: Josep Torrellas, "Variation-Tolerant Architecture"

DVSAM-Pow: Power Efficiency



- Start with low V_{dd} and increase slowly
- Critical path delays are kept at τ_D until end: Run at f_{NOM}

Source: Josep Torrellas, "Variation-Tolerant Architecture"

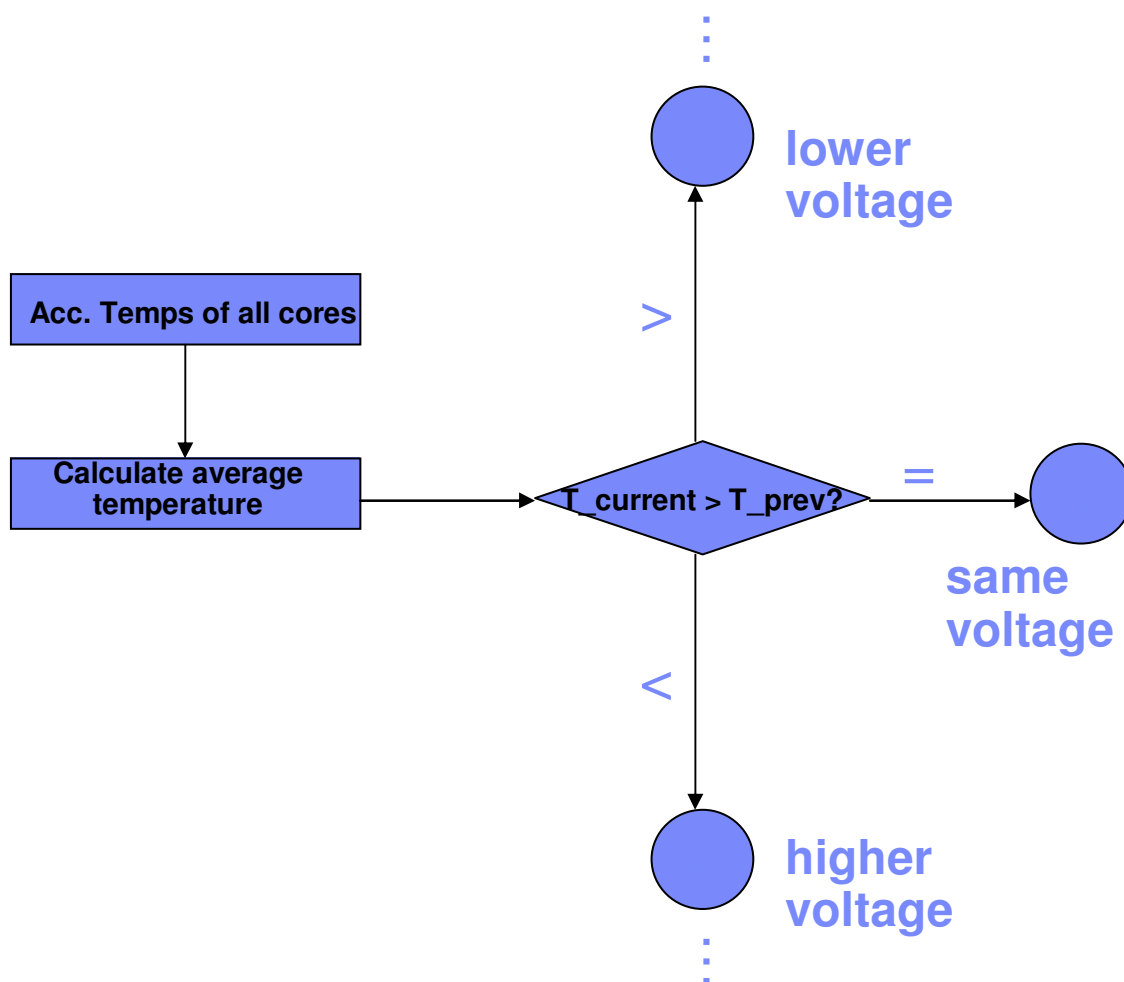
Could Thermal Cycling of Chips Lead to Packaging Failures?

- IBM study of actual customer environments
 - System operation is unique (based on power management policies)
 - Customer applications and workloads on the system
 - Unique data center environment
- IBM Developed a Figure of Merit (FOM)
 - The purpose of the FOM is to have a metric that is related to the frequency and the depth of the thermal cycling going on inside chips
 - Reads and averages a couple of on-chip thermal sensors which are spatially separated and segregated from high power dissipation areas on chip
 - Parse temp data into discrete elements → feed through an algorithm which then normalizes this data to a defined thermal cycle condition → keep a running tab of thermal cycles a given processor experiences in the field
 - FOM is saved on all modules in the field, can be retrieved from returned modules
 - Possible to read FOM values off of machines in the field
 - The larger the FOM, the more likely failures could occur with the packaging
 - A FOM approaching 10,000 over a 7 year lifetime is seen as a problem

-
- The scatter plot displays the relationship between the number of processors (Y-axis, logarithmic scale from 1 to 141) and the Projected FOM Growth (X-axis, linear scale from 0 to 1200). The data points are categorized by color: blue squares, magenta diamonds, and yellow triangles. The blue squares represent the highest number of processors, starting at approximately 141 for low FOM growth and decreasing rapidly as FOM growth increases. The magenta diamonds represent a moderate number of processors, starting around 27 and decreasing to 1. The yellow triangles represent the lowest number of processors, starting around 1 and decreasing to 1. A red circle highlights the worst case FOM value of 1058, which corresponds to a projected FOM growth of approximately 1058 and a number of processors of 1.

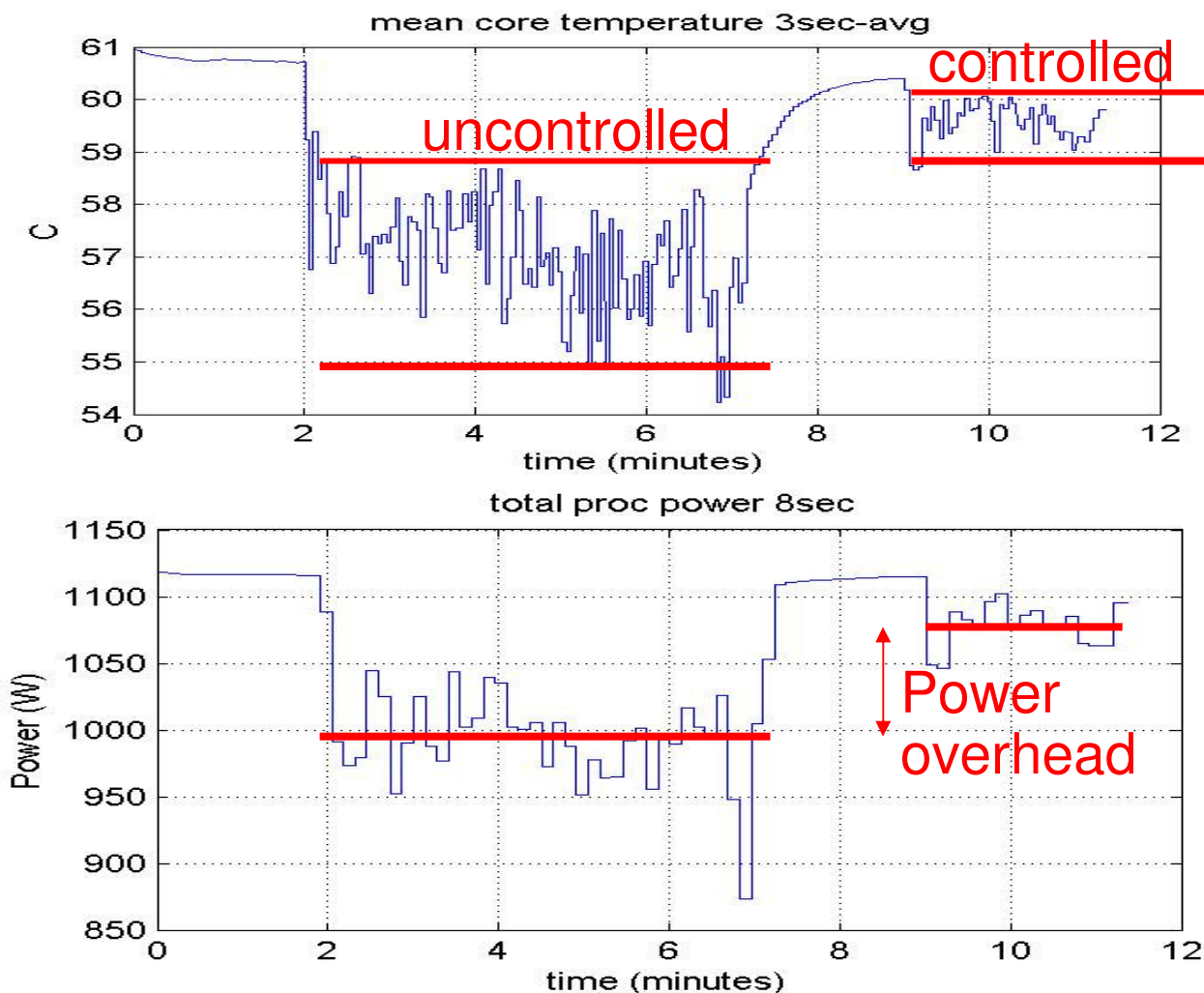
Improvements to Reduce Thermal Cycling

Adjust voltages to control chip temperatures, while leaving performance (frequency) alone



Results from Technique to Mitigate Thermal Cycling

- Average chip temperature shown, with controlled maintaining $\pm 1^\circ\text{C}$ temperature swing, but at higher energy cost
- Technique resulted in 7% reduction in projected FOM growth



Disk Reliability Study by Google

- Many data centers are moving to higher ambient temperatures: is there a risk for disk drives?
- “Failure Trends in a Large Disk Drive Population” FAST’07 paper from Google
 - Over 100,000 hard disk drives studied
 - Examined SMART (Self-Monitoring Analysis and Reporting Technology) parameters from within drives as well as temperatures
 - Found that only at disk temperatures above 40 deg C was there a noticeable correlation to drive failures
 - Some SMART parameters with higher correlation to failures included first scan errors, reallocations, offline reallocations, and probational counts (suspect sectors on probation)
 - Key missing piece from study is extensive power cycling other than reporting that after 3 years, higher power cycle counts can increase failures rate to be over 2%
 - Assumes server class drives are running continuously as the normal mode of operation (little change in power)

Reliability Aware Disk Power Management

- Storage consumes a large fraction (up to 40%) of total IT equipment power
 - High cost of operation (energy cost)
 - Limits capacity of facility
- Disks are not power- or energy-proportional:
 - Consume roughly the same power idle (but spinning) or servicing a request (e.g., 7W idle vs 10W active)
 - Only by spinning disks down (turning off spindle motor) can you achieve very low power state (e.g., sub 1W)
- Spindown or Power Cycling introduces problems:
 - Reliability: Disks are mechanical devices and can only spin-up / spin-down limited number of times over lifetime
 - Latency: It takes 5-10 seconds to spin up a disk and access a block (vs ~10ms when spinning)



Our
focus

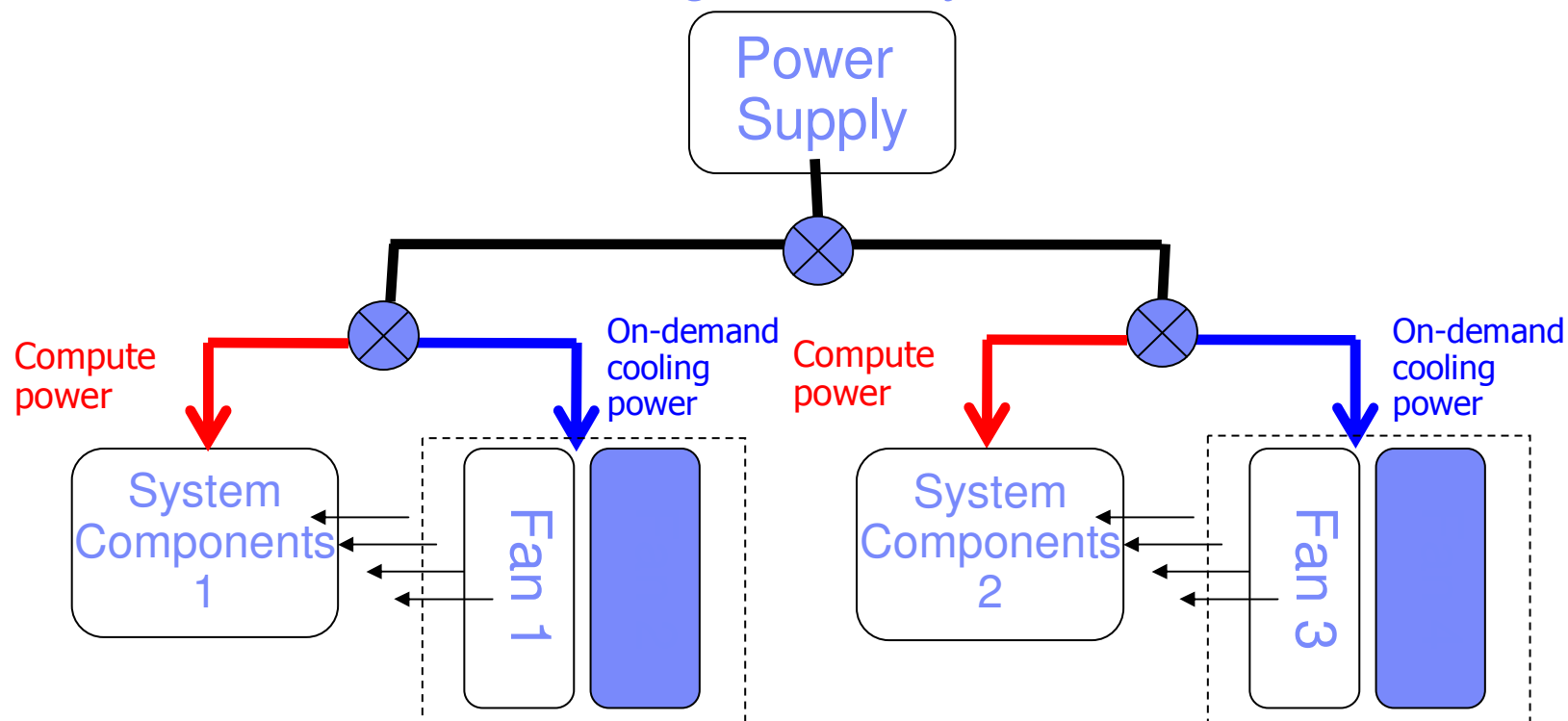
Reliability Aware Approach to Spindown

- Manage idle timeout periods *dynamically*, rather than using a fixed timeout period
 - If disk has been spun down less frequently than conservative rate (e.g., every 15 minutes) in past, can spin down more often in the future
 - Need to limit *lifetime* spindowns to manufacturer specification → maintain lifetime spindown rate
- One approach to controlling spindowns: token bucket
 - Every N minutes (e.g., every 15 minutes), add a token to the “spindown bucket”
 - When energy management policy wishes to spin down disk, must remove token (or defer spindown)

Fan Failure Mechanisms and Impacts on Reliability

- Dynamic fan management for higher ambient conditions for data center PUE: impacts on reliability?
- HP paper: “Cooling Fan Reliability: Failure Criteria, Accelerated Life Testing, Modeling and Qualification”
 - Most common failure mechanism is mechanical due to bearings wearing out
 - Bearings wear out due to loss of lubricant
 - Higher temperatures accelerate MTTF for fans
 - Vendors prefer to quote 25 deg C ambient which is cooler than many of the more efficient Data Center designs with higher ambient temperatures
 - Power cycling may increase failure rates, but powering down fans can maximize energy savings for idle servers

Improvements in Server Cooling Efficiency



- Redundant series fan pairs, for normal mode, only one fan in a set is on (Fan 1 and Fan 3)
- Assign additional cooling (Fan2 or Fan4) on demand
- When one fan fails, the other fan is switched on just-in-time before thermal emergency (a few seconds observed in real system). From then on, use normal mode.
- When a failed fan is replaced, higher performance can be resumed when the utilization requires it.

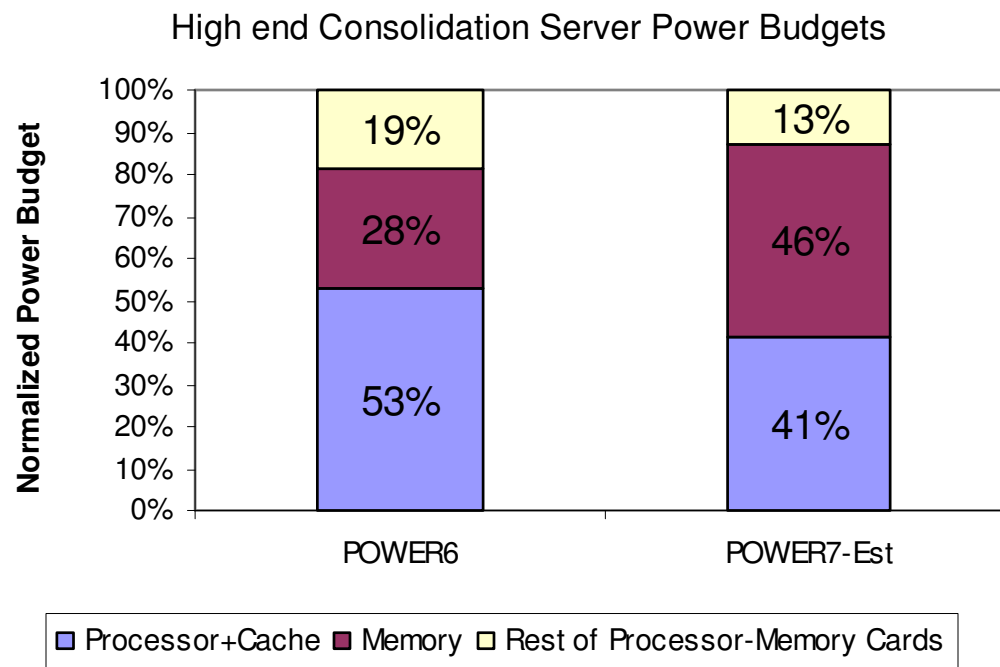
Research in Emerging Technologies and Solutions

Outline

- Storage Class Memories
- Power delivery, cooling and packaging technologies
- Workload optimized systems

Storage Class Memory – Motivation

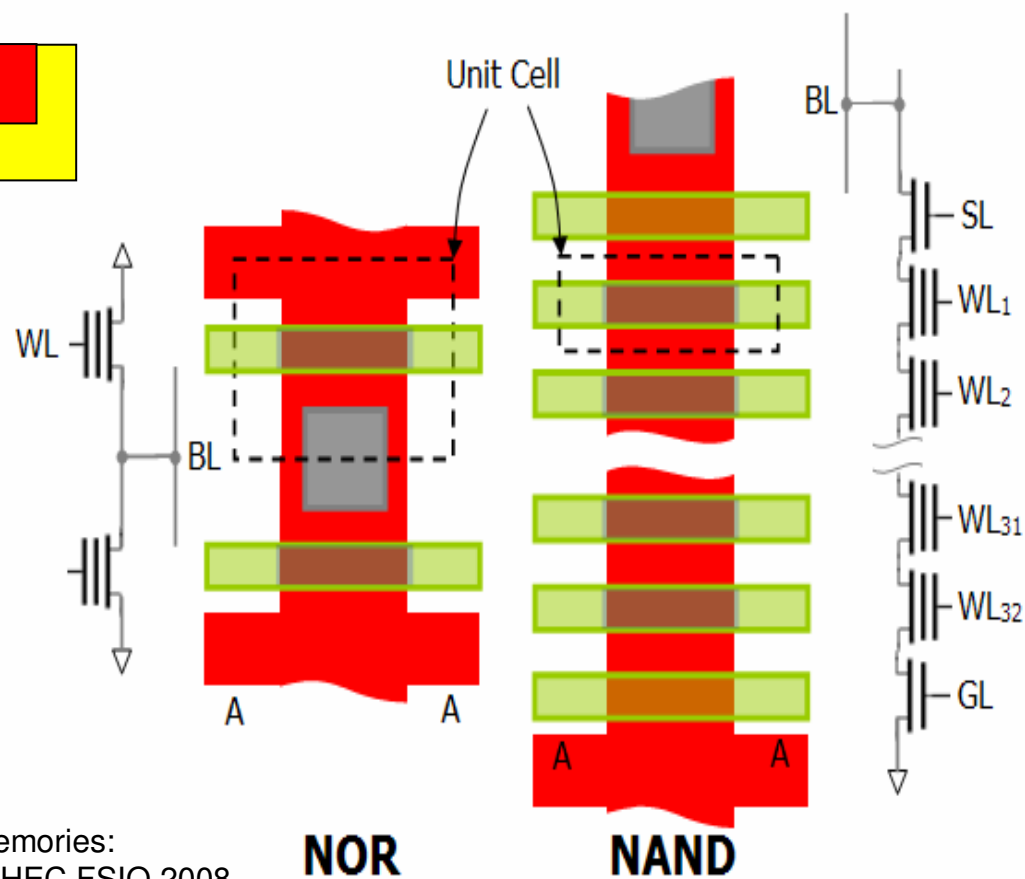
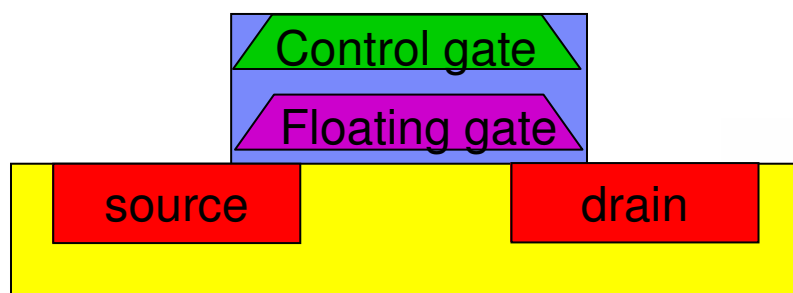
- Memory power is an increasing fraction of server power
- Growing demand for capacity
 - Virtualization
 - Data-intensive applications
 - In-memory databases
- Ramcloud*
 - Latency is a problem, so keep everything in memory
 - Large number of diskless computing nodes.
- Scaling DRAM increasingly difficult (like CMOS scaling of logic chips)



Source: Architecting for Power Management: The IBM POWER7 Approach: Ware, Rajamani, Floyd, Brock, Rubio, Rawson, Carter, HPCA 2010

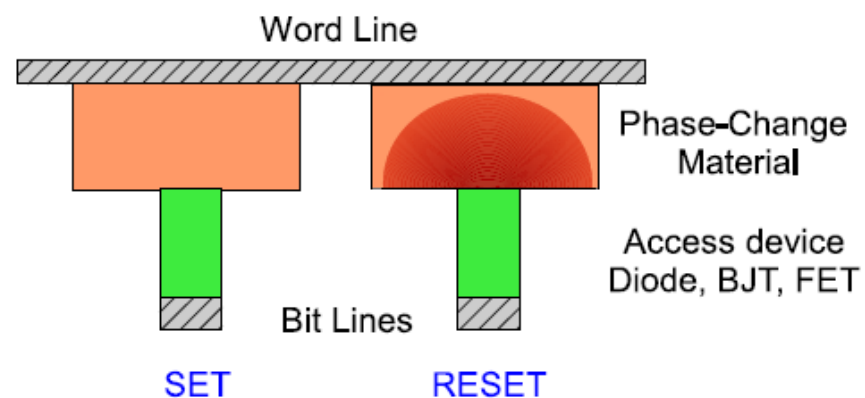
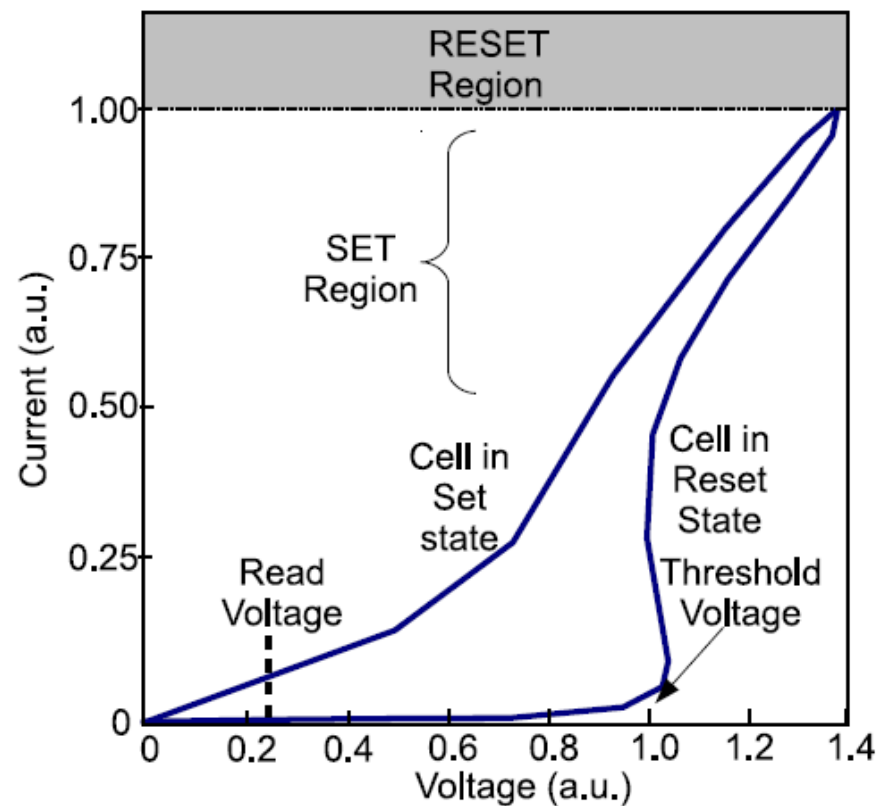
*Ramcloud: <http://fiz.stanford.edu:8081/display/ramcloud/Home>

SCM Technologies – Flash



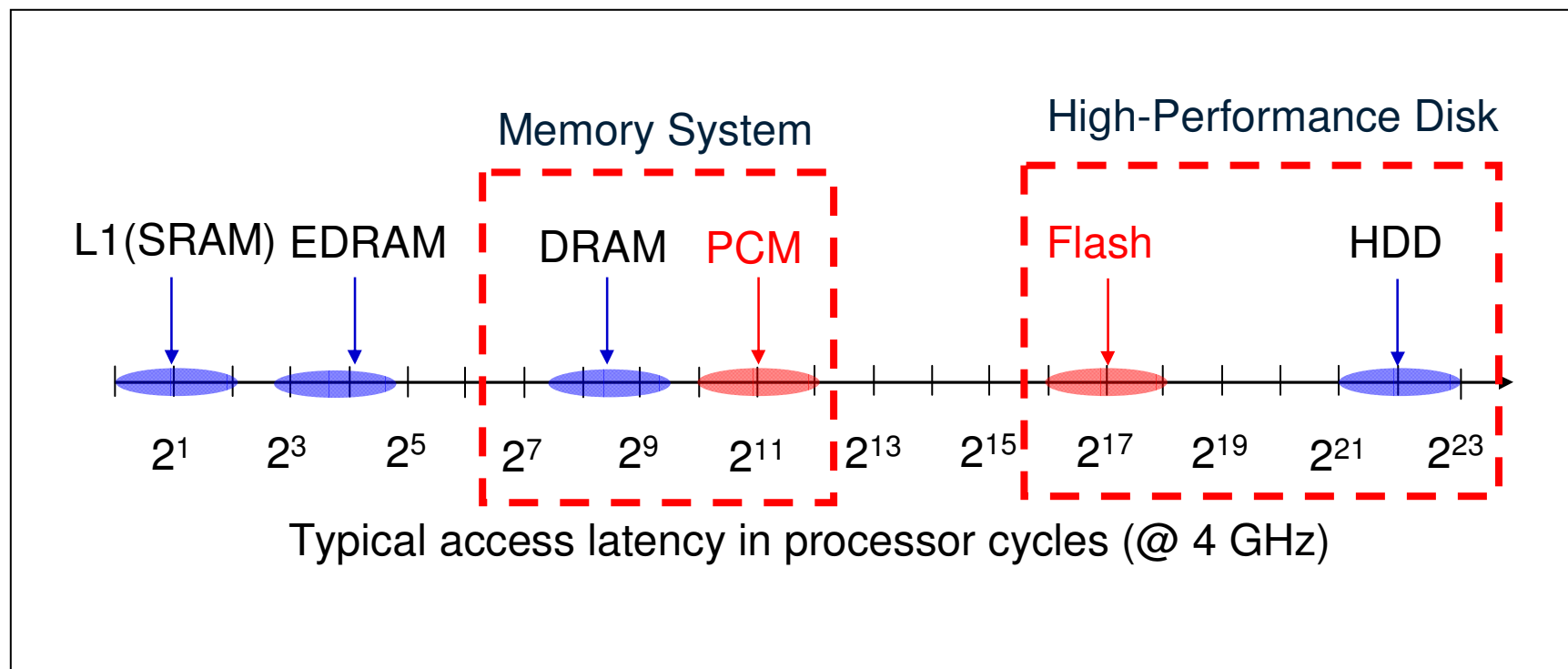
Source: Winfried Wilcke, Flash and Storage Class Memories:
Technology Overview and Systems Impact, Panel at HEC FSIO 2008

SCM Technologies - PCM



Near term use of SCM technologies

Need higher capacity, lower power



Source: Scalable High Performance Main Memory System Using Phase-Change Memory Technology, Moinuddin K. Qureshi, Vijayalakshmi Srinivasan, Jude A. Rivers, ISCA 2009.

PCM, Flash and DRAM: A Quantitative View

Attribute	DRAM	PCM	NAND Flash
Non-Volatile	No	Yes	Yes
Idle Power	100mW/GByte	1 mW/GByte	10 mW/GByte
Erase / Page Size	No / 64Bytes	No / 64Bytes	Yes / 256KB
Write Bandwidth per die	1-6GBytes/s	50-100 MB/s	5-40 MB/s
Page Write Latency	20-50 ns	1 μ s	500 μ s
Page Read Latency	20-50 ns	50 ns	25 μ s
Endurance	10^{16}	10^7	10^5
Maximum Density	4Gbits	4Gbits	64Gbits

Emerging technology trends – Main Memory

- Alternatives
 - PCM - Already commercially available
 - MRAM - Commercially available but limited to 4Mbits
 - STT-RAM - Early prototypes
 - Memristors and dual-gate - single cell or very small prototypes
- All alternatives are persistent
 - Additional power implications (suspend/resume).
 - OS and applications can also use this property.
 - Security
 - In memory data is persistent and can be physically accessed
 - Reliability
 - How to trust information in memory?
- Endurance
- New system architectures

Storage Class Memory as Secondary Storage Alternatives to Magnetic Disks (HDD)

- Flash
 - Lower power
 - 0.5W-2W versus 2W-10W for HDD
 - Lower latency for random I/O
 - Larger number of IOPs: 20K-35K vs 100-300 for HDD
 - Similar sequential access bandwidth
 - Flash has comparable density, but suffers from scalability problem
 - Endurance decreasing – 3K erases
 - Cells more unreliable – More bits dedicated to error-correction
- PCM
 - Less dense than Flash
 - Hybrid designs with Flash
 - Metadata on PCM, data on Flash
 - Reduce write amplification
 - Update in place – PCM re-writable and byte-addressable

Emerging System Architectures with PCM

- PCM + DRAM
 - PCM as main memory, DRAM as large cache
 - Virtual memory managed – IBM Watson
 - Hardware managed – University of Pittsburgh
- 3D architectures
 - 3D chip containing processors and DRAM/PCM
 - IBM, University of Pittsburgh
 - Used to reduce power consumption
 - Diskless (HP Nanostore)

References: Storage Class Memory

1. The Basics of Phase Change Memory Technology :
http://www.numonyx.com/Documents/WhitePapers/PCM_Basics_WP.pdf
2. S. Raoux et al. Phase-change random access memory: A scalable technology. *IBM Journal of R. and D.*, 52(4/5):465–479, 2008.
3. International Technology Roadmap for Semiconductors - 2010 <http://www.itrs.net/Links/2010ITRS/Home2010.htm>
4. Nanostore: Ranganathan, P , From Microprocessors to Nanostores: Rethinking Data-Centric Systems, *IEEE Computer* , vol.44, no.1, pp.39-48, Jan. 2011
5. B. C. Lee et al, Phase Change Technology and the Future of Main Memory, *IEEE Micro, Special Issue: Micro's Top Picks from 2009 Computer Architecture Conferences (MICRO TOP PICKS)*, Vol. 30(1), 2010.
6. Jian-Gang Zhu; , "Magnetoresistive Random Access Memory: The Path to Competitiveness and Scalability," *Proceedings of the IEEE* , vol.96, no.11, pp.1786-1798, Nov. 2008.
7. M. Qureshi et al. Scalable high performance main memory system using phase-change memory technology. *In ISCA '09: Proceedings of the 36th annual international symposium on Computer architecture*, pages 24–33, New York, NY, USA, 2009. ACM.
8. B. C. Lee et al. Architecting phase change memory as a scalable DRAM alternative. *In ISCA '09: Proceedings of the 36th annual international symposium on Computer architecture*, pages 2–13, New York, NY, USA, 2009. ACM.
9. Winfried Wilcke, IBM. Flash and Storage Class Memories: Technology Overview & Systems Impact., *HEC FSIO 2008 Conference*. http://institute.lanl.gov/hec-fsio/workshops/2008/presentations/day3/Wilcke-PanelTalkFlashSCM_fd.pdf

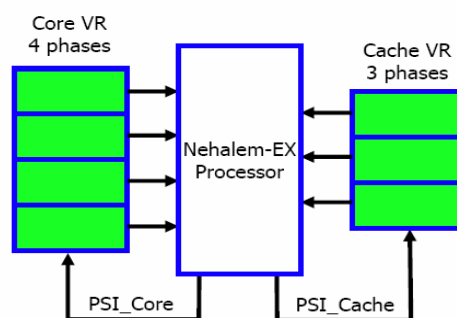
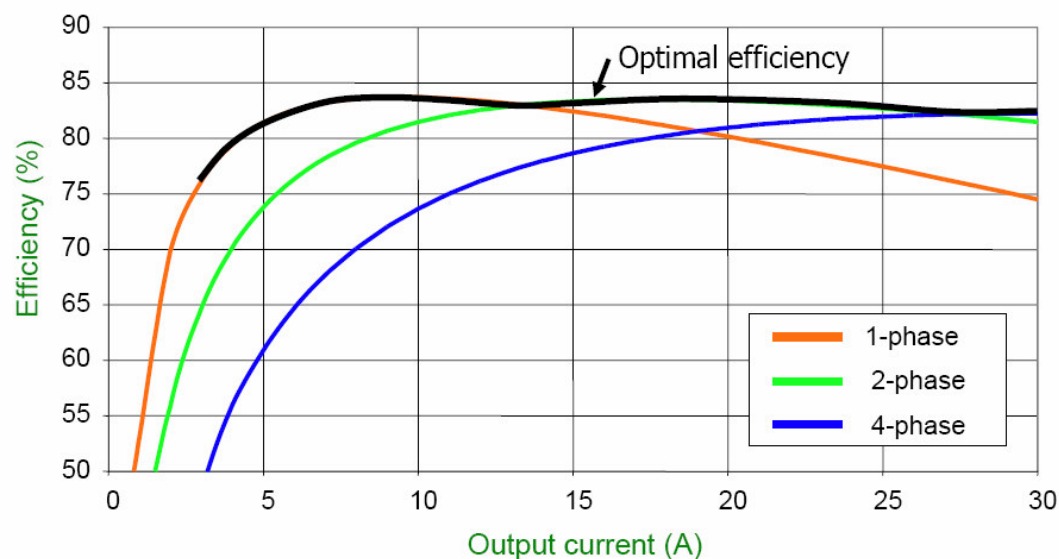
Power Delivery, Cooling and Packaging Technologies

Voltage Regulator Phase Shedding for Increased Efficiency

- Modern processor VRMs are multiphase designs, with the total load split among the phases.
- The efficiency of a multiphase regulator varies with load, with efficiency falling at lower loads.
- When the load is small instead of using all phases, each providing a small current, shut-off some of them and increase the load for the others.
- Nehalem-EX
 - extends the VR phase shut-off to the cache supply,
 - obtains about 2W power reduction per socket in idle mode.

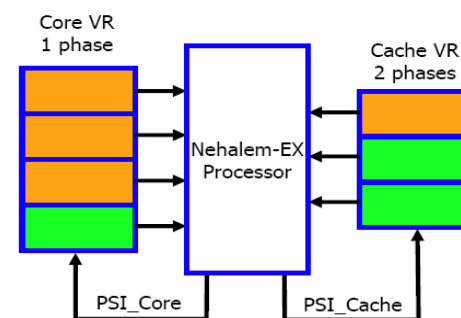
Figures and Data: A 45 nm 8-Core Enterprise Xeon® Processor, S. Rusu et al., IEEE Journal of Solid-state Circuits, 45(1), January 2010.

Load Efficiency of 4-phase Voltage Regulator



Full Load Mode

- All VR phases are enabled
- Maximum VR efficiency



Idle Mode

- Turn off 3 core and 1 cache phases
- Maximum VR efficiency

On-chip Voltage Regulation

- **Benefits**

- Lower distribution losses from higher voltage on-board power distribution.
- Lower energy spent for droop control, as regulation is closer to load.
- Enables fine-grained voltage control (spatial and temporal) leading to better load-matching and improved energy-efficiency
- Reduction in board/system costs, reducing voltage regulation needs on board.

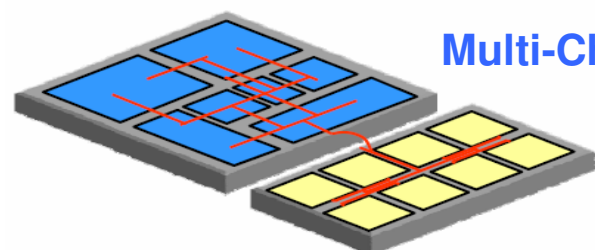
- **Challenges**

- Space overheads on processor chip
- Difficulty realizing good discrete components in same technology as digital circuits.

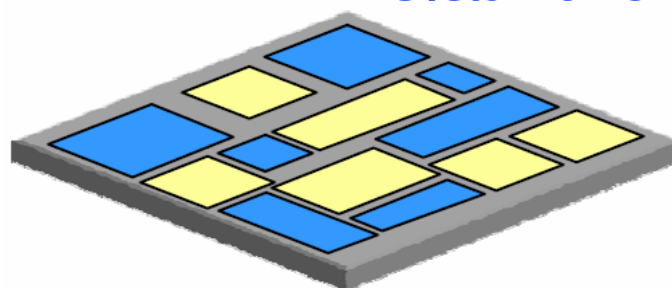
- **Opportunities**

- 3D packaging can help address both challenges above.

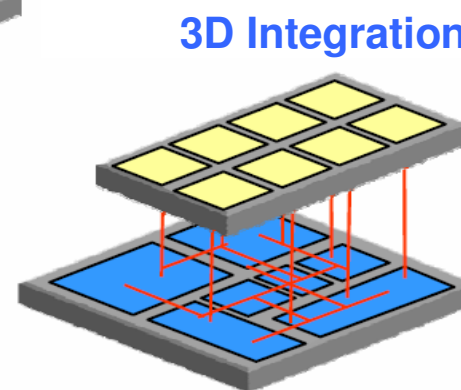
3D Chip Stacks



Multi-Chip Design



System on Chip



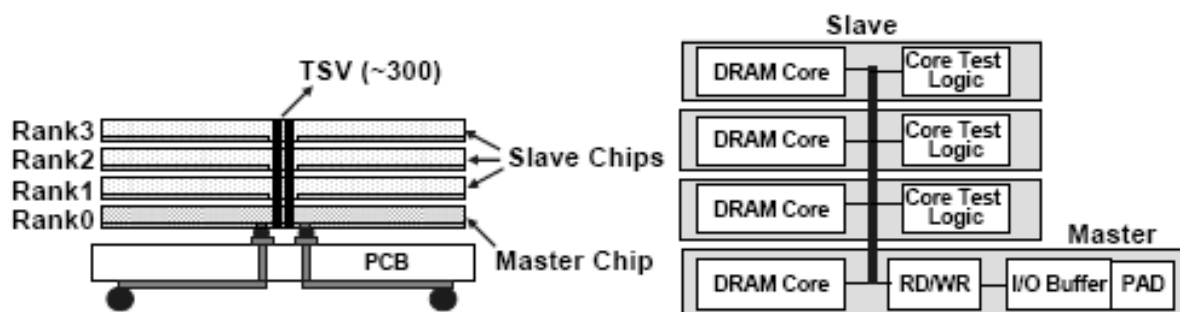
3D Integration

3D Benefits:

- High core-cache bandwidth
- Integration of disparate technologies
- Reduction in wire length
- Reduced interconnect, I/O cost – eliminates off-chip drivers, lower power overheads & faster, higher energy-efficiency

3D

Currently being embraced for DRAM devices e.g. Samsung 8Gb 3D DRAM



Cross-sectional view

Conceptual drawing

❑ **Master chip: Core + Peripheral logic (supports multi-rank operation) + DLL + I/O + Test logic**

❑ **Slave Chip: Core + Test logic**

Power reduction

- Standby (IDD2N): 50%,
- Active (IDD1): 25%

Faster speeds

- Less loading on channel.

Source: 8Gb 3D DDR3 DRAM using Through-Silicon-Via Technology (Samsung), U. Kang et al., IEEE Solid-State Circuits Conference, 2009

Challenges

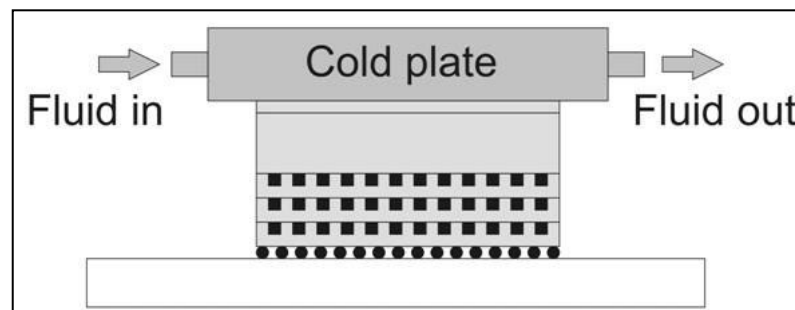
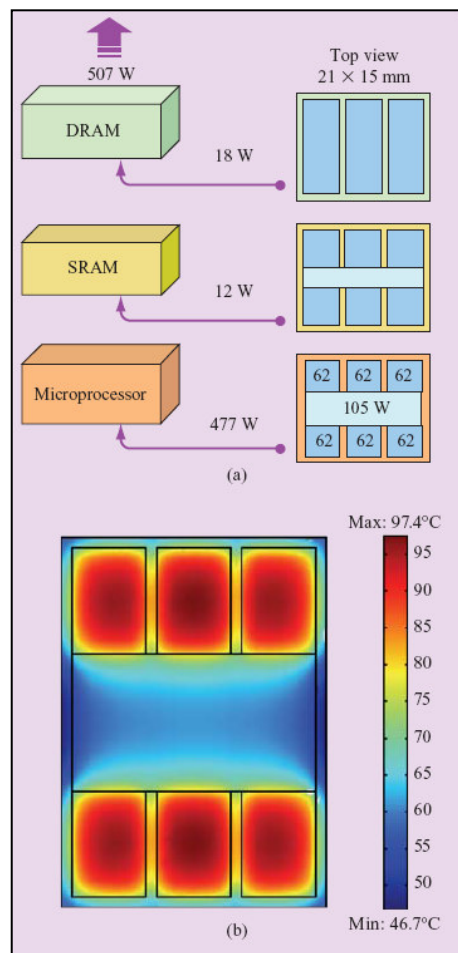
- New technology: initial development costs, tool costs.
- Cooling

CMOSAIC Project

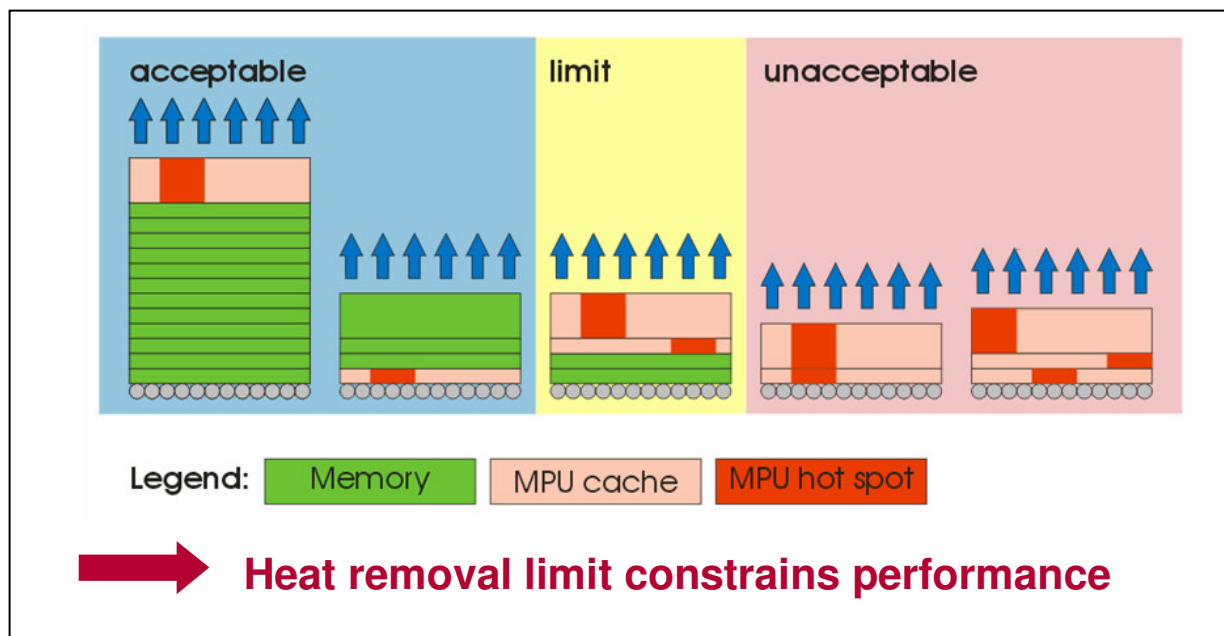
- Ongoing collaborative project between IBM, École Polytechnique Fédérale de Lausanne (EPFL) and the Swiss Federal Institute of Technology Zurich (ETH).
- Evaluate chip cooling techniques to support a 3D chip architecture.
- 3D stack-architecture of multiple cores with a interconnect density from 100 to 10,000 connections per sq. mm.
- Liquid cooling microchannels ~50um in diameter between the active chips.
- Single-phase liquid and two-phase cooling systems using nano-surfaces that pipe coolants—including water and environmentally-friendly refrigerants—within a few millimeters of the chip.
- Two phase cooling
 - Once the liquid leaves the circuit in the form of steam, a condenser returns it to a liquid state, where it is then pumped back into the processor, completing the cycle.

Source: Bruno Michel, IBM

Limits of Traditional Back-side Heat Removal



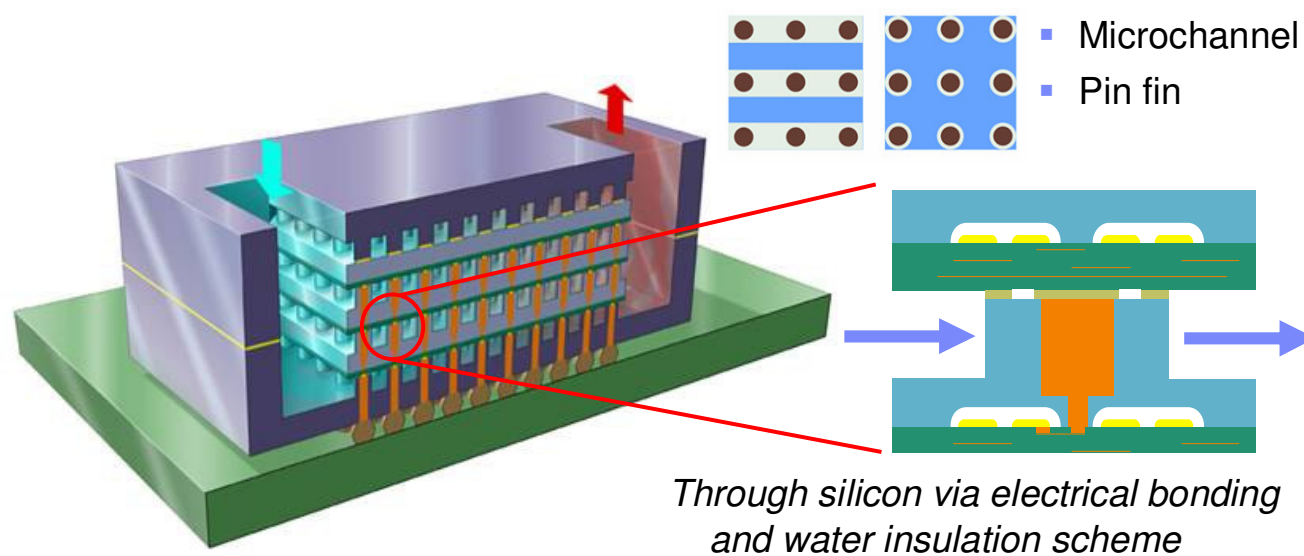
Microchannel back-side heat removal



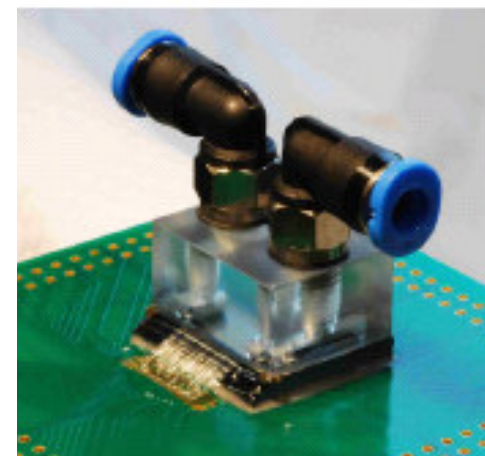
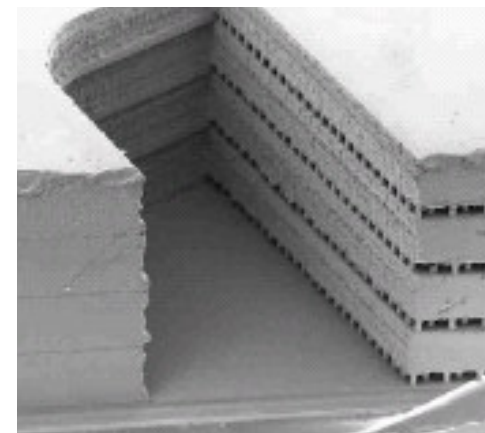
Source: Bruno Michel, IBM

Scalable Heat Removal by Interlayer Cooling

- 3D integration requires (scalable) **interlayer liquid cooling**
- Challenge: isolate electrical interconnects from liquid



cross-section through fluid port and cavities



Test vehicle with fluid manifold and connection

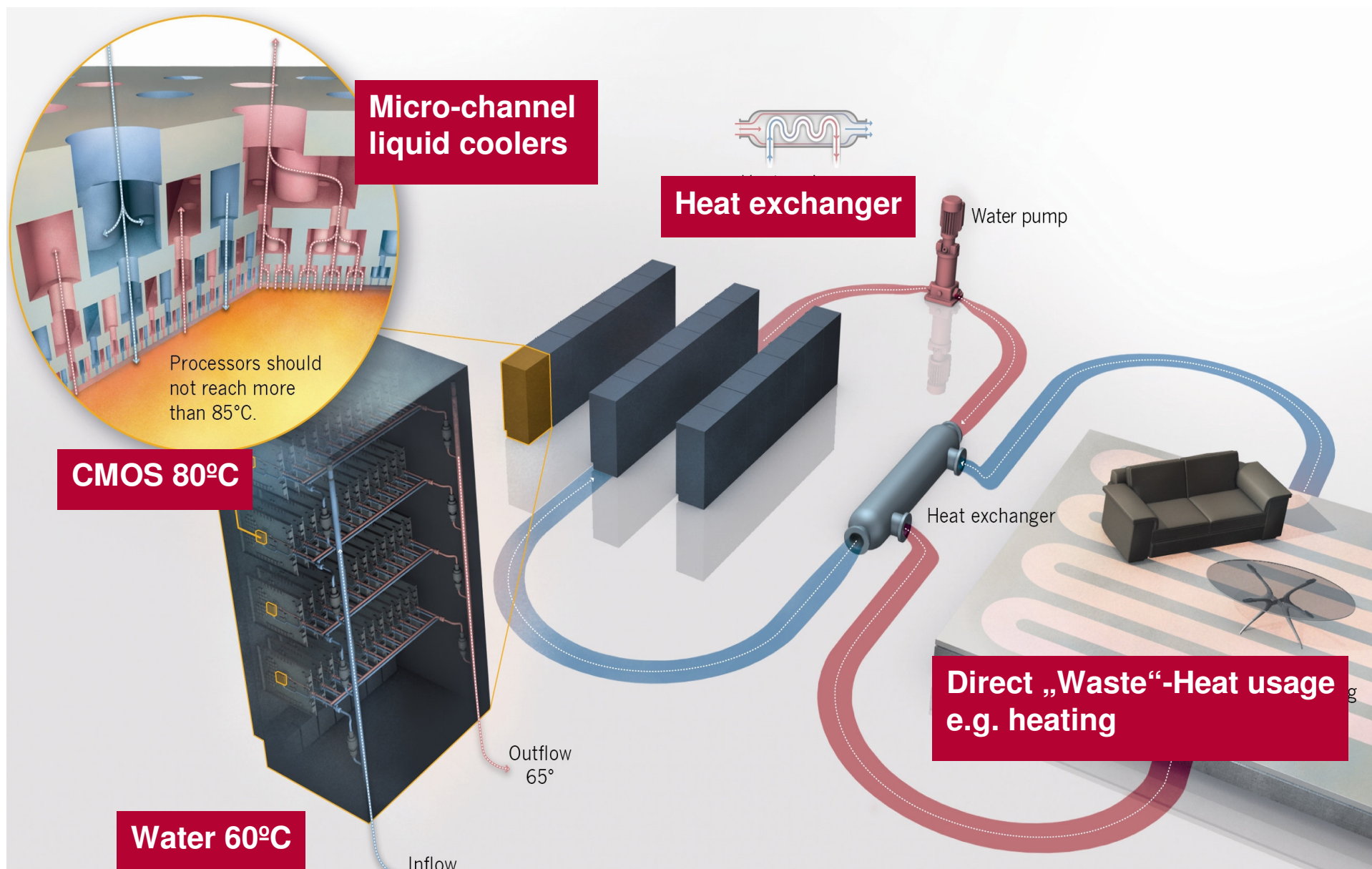
Source: Bruno Michel, IBM

SuperMUC Super Computer (2012)

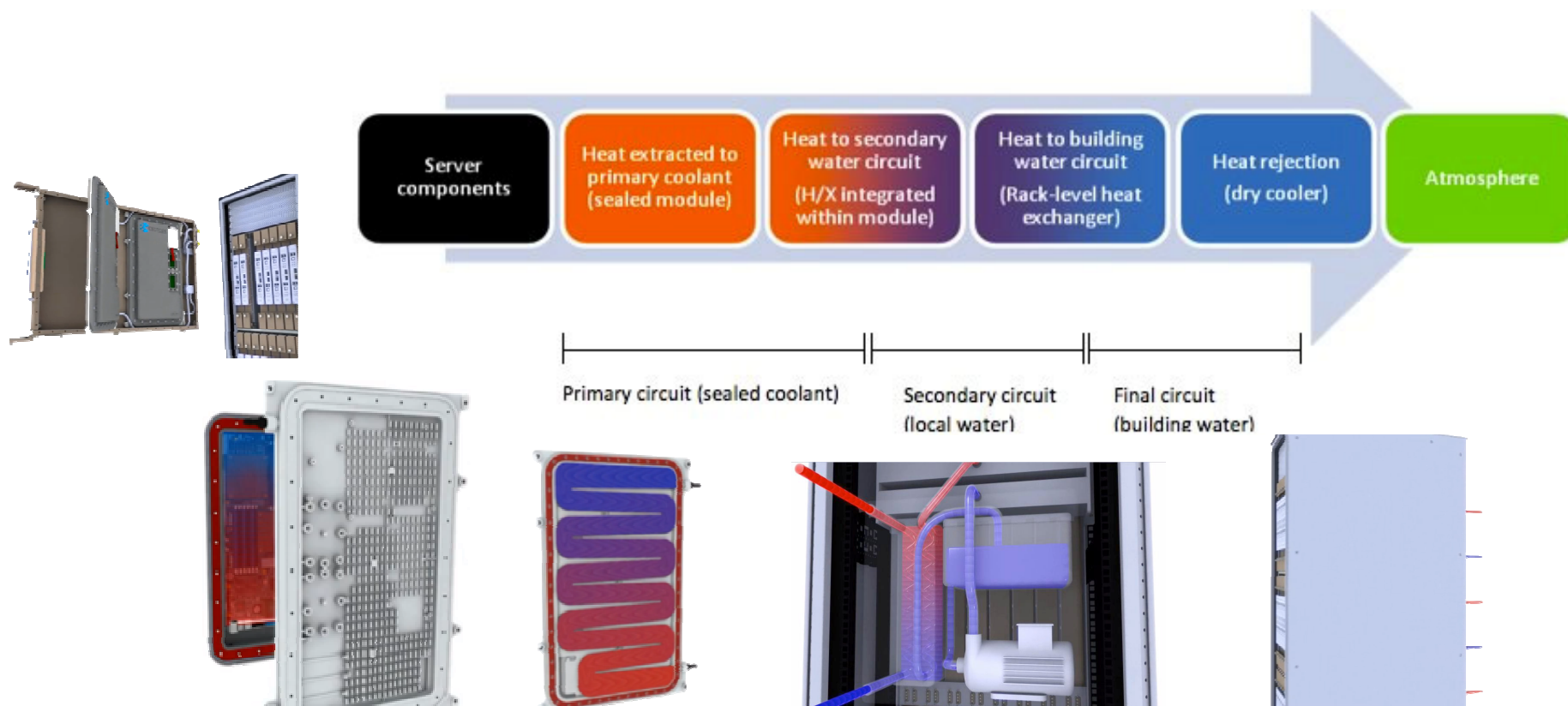
- SuperMUC is based on **hot-water-cooling technology** invented for QPACE, Aquasar and iDataCool. (prototype at University of Regensburg)
- In operation at the Leibniz Supercomputing Centre (LRZ) in Munich, Germany, by 2012.
- Energy-efficiency:
 - PUE 1.1 – Green IT
 - 40% less energy consumption compared to air-cooled systems
 - 90% of waste heat will be reused
- Based on an IBM System X iDataPlex ®:
 - Peak performance of 3 PF/s
 - 9531 Nodes with total 19476 Intel Xeon CPUs / 157464 Cores, 324 TB Memory
 - InfiniBand FDR10 Interconnect with ~ 11900 (optical) IB cables
 - 10 PetaByte File Space based on IBM GPFS and 2 PetaByte NAS Storage

Source: Bruno Michel, IBM

Hot-water-cooled datacenters – towards zero emission



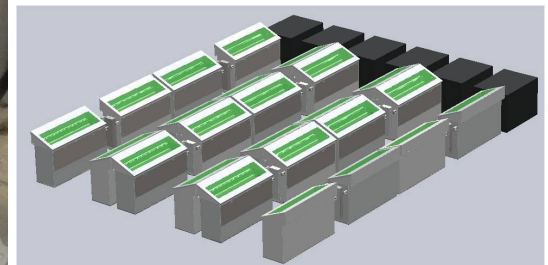
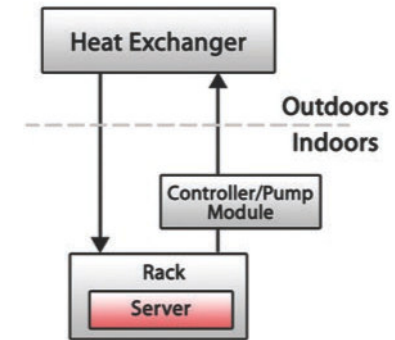
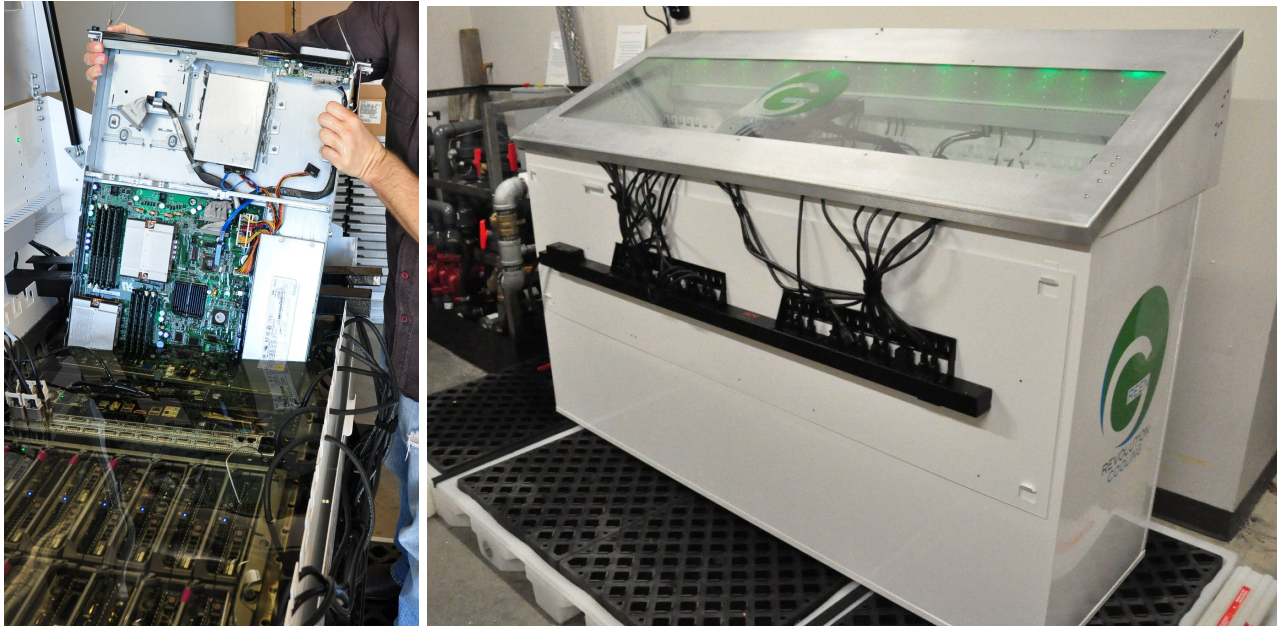
Iceotope: Two-stage modular Liquid Cooling



Images from web-site of Iceotope Limited, Sheffield, UK

- Heat captured by individually sealed liquid coolant in a primary circuit.
- Transferred to water through a secondary circuit, which is in turn cooled by the building water in a final circuit
- End to end liquid cooling without messing with coolant during maintenance

Total Immersion-in-Oil Cooling



Photos courtesy Green Revolution Cooling, Austin Texas

- High heat capacity coolant (1,300x by volume than air)
 - Direct contact to CPU reduces its temperature (10-15 deg C reduction reported)
 - Lower power for cooling
 - less coolant volume to circulate (95% cooling power reduction claimed)
 - 10-20% less server power due to elimination of internal server fans
 - Improved reliability
 - Fan failures are eliminated by removing fans
 - Disk drive reliability improved, with temperatures at coolant temperature level, and reduced vibrations associated with fans and pressurized air.
- Advertise upto 100kW per rack power density.

Intelligent Management of Power Distribution in a Data Center

▪ Problem

- Overprovisioning of power distribution components in data centers for availability and to handle workload spikes

▪ Solutions

- Provision for average load => reducing *stranded power*, use *power capping*.
- Oversubscribe with redundancy and power cap upon failure of one of the supplies/PDUs.
- Employ power distribution topologies with overhead power busses to spread secondary power feeds over larger number of PDUs, reducing the reserve PDU capacity at each PDU.
- Use power-distribution-aware workload scheduling strategies to match load more evenly with power availability.

▪ Challenges

- Separated IT and facilities operations, not enough instrumentation – no integrated, complete view of power consumption versus availability for optimizations.
- Existing methods for increased availability of the power delivery infrastructure have high energy/power costs.

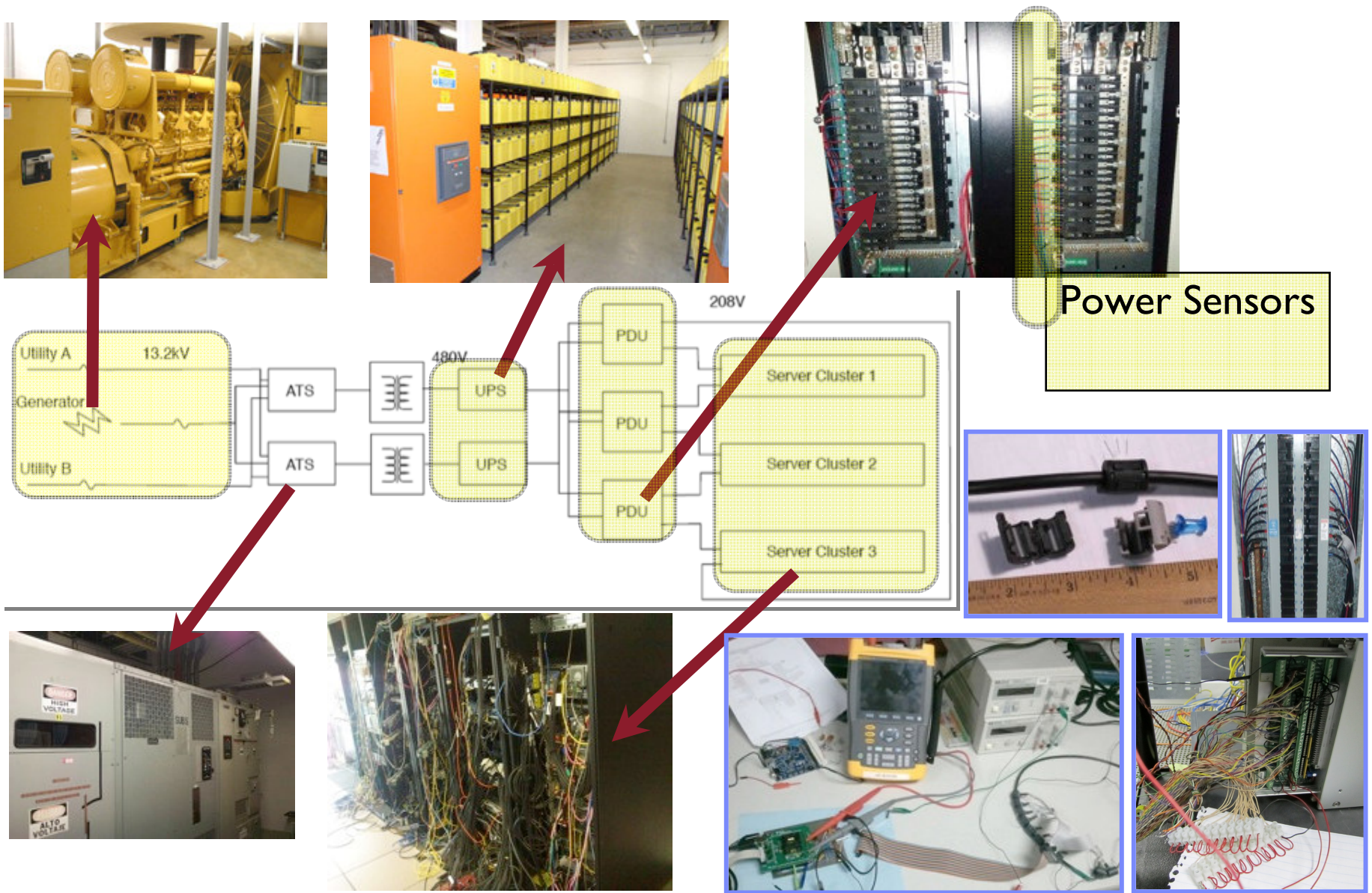
*Power Routing: Dynamic Power Provisioning in the Data Center, Steven Pelley, David Meisner, Pooya Zandevakili, Thomas F Wenisch, Jack Underwood, ASPLOS 2010

New Technologies for Datacenter Power distribution Management

- Datacenter Power Management – Vision for Projects at IBM Research Austin
 - Develop new technologies which enable impact of integrated management without actual merger of IT and Facilities operations, and
 - Develop optimization and management techniques which demonstrate enhanced benefits where integrated control of IT and facilities infrastructure is possible

- Intelligent Power Distribution and Control
 - Low-cost power sensors
 - Power monitoring and management infrastructure prototype
 - Branch Circuit Identification

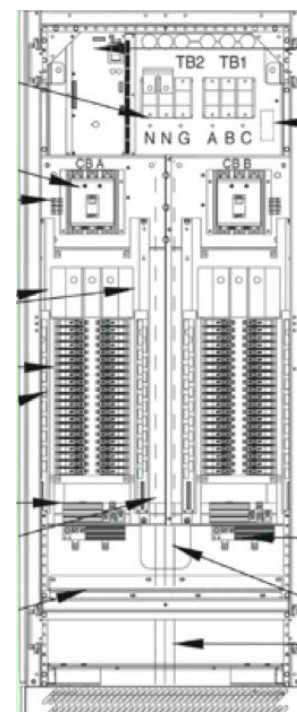
Branch Circuit Power Sensors in the Data Center



Power Monitoring Infrastructure (1/3)

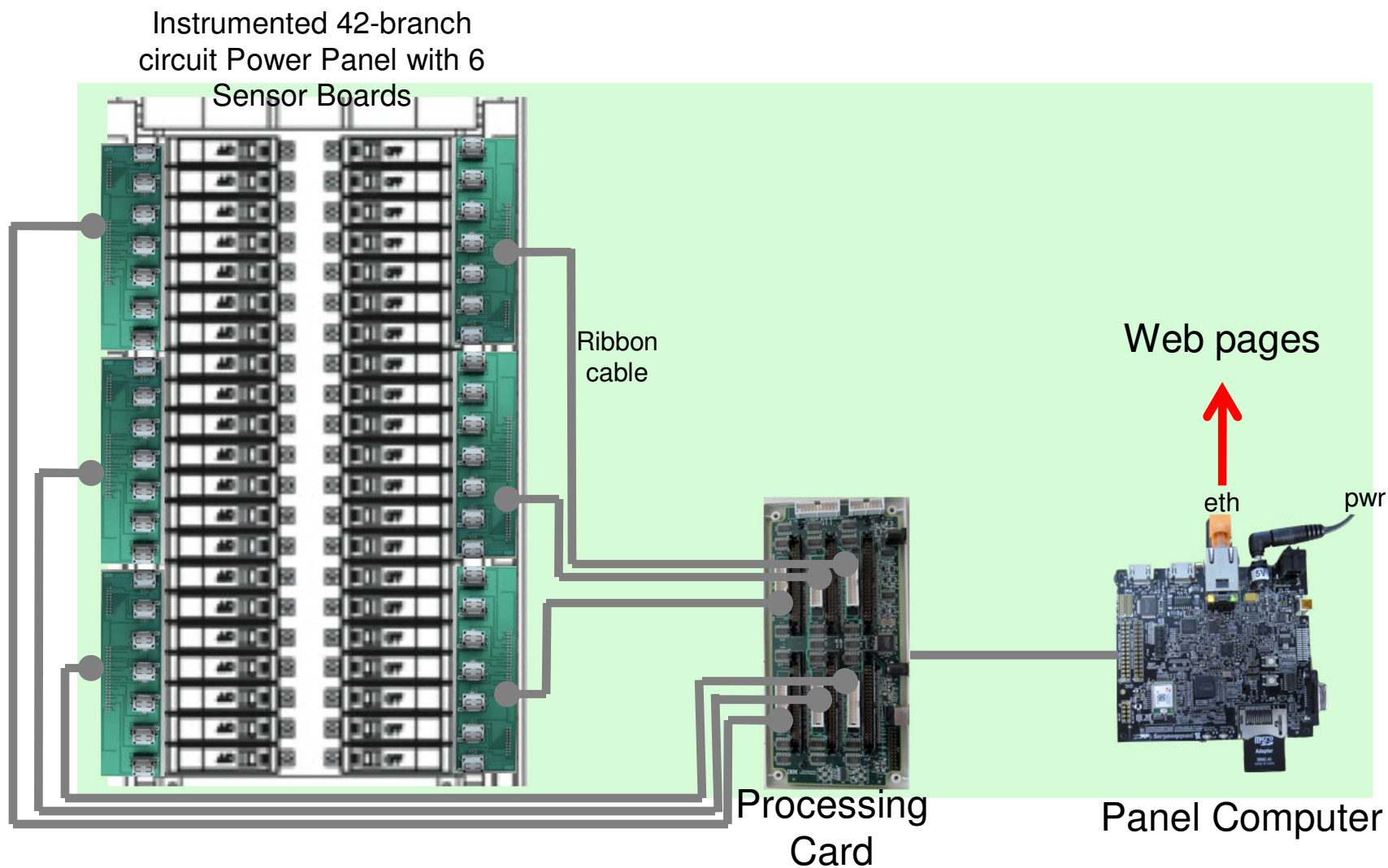


Power Distribution Units in a Data Center



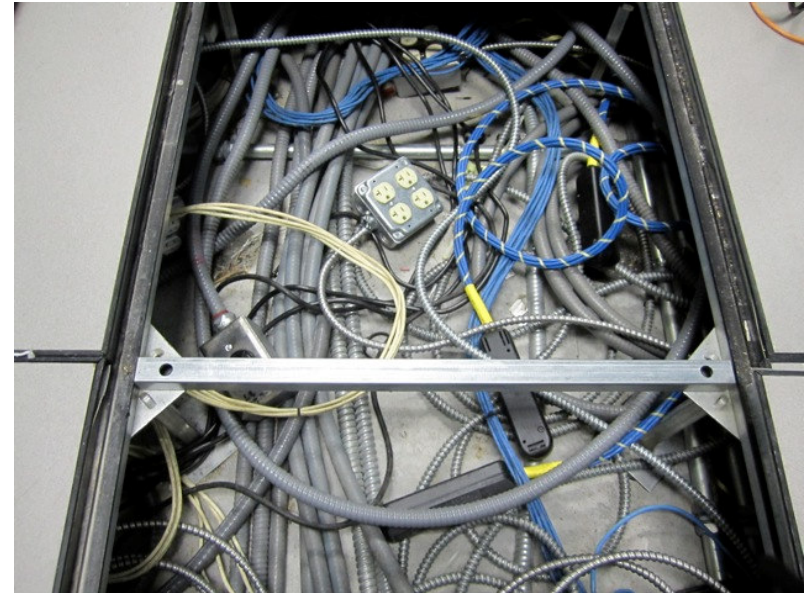
Power Distribution Unit showing 2 panels of circuit breakers

Power Monitoring Infrastructure (3/3)



Branch Circuit Identification (BCID)

- BCID – determining connectivity of power panel's branch circuit by **novel power signal generation & detection technology**
- Applications of BCID information
 - Installation of equipment on desired branch circuits
 - Load balancing among phases/circuits and consumption-aware placement
 - Power distribution based on load
 - Load-aware failure mitigation



► Three methods for generating signals for BCID

- *IBM systems with EnergyScale – custom power-signaling technology.*
- Systems with USB interface – USB-clicker
- Available power outlet in a rack PDU – AC clicker

BCID Demo

- ▶ Displays branch circuit identification in real time using power measurement infrastructure

Power Monitoring Demo

- ▶ Displays live monitoring of currents on branch circuits in a power-panel, showing load over time.

Power capping demo

- ▶ A real-time demonstration of gracefully managing overcurrent in a branch circuit via power capping of connected servers.

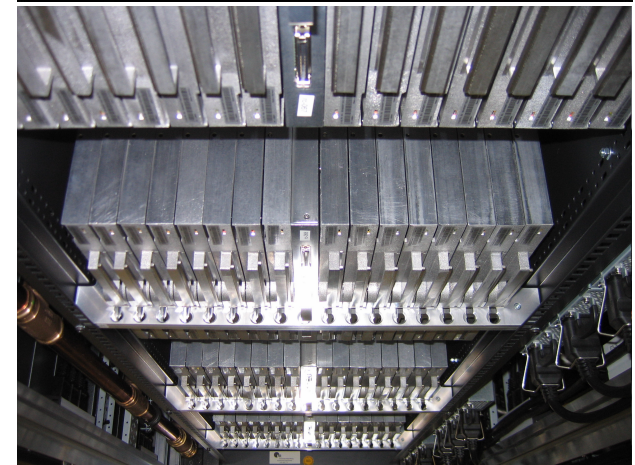
Workload-optimized Systems

Workload-optimized Systems for Energy-efficiency

- Custom designs/accelerators for specific functions at greater efficiency.
- Lower-power processors driven by focus on throughput versus costlier single-thread performance.
 - Fewer active components (functions) not pertinent to current execution.
- Application-specific integrated systems (ASIS \leq ASIC)
 - Reduced components to improve cooling, lower power.

Quantum Chromodynamics Parallel Computing on the Cell Broadband Engine

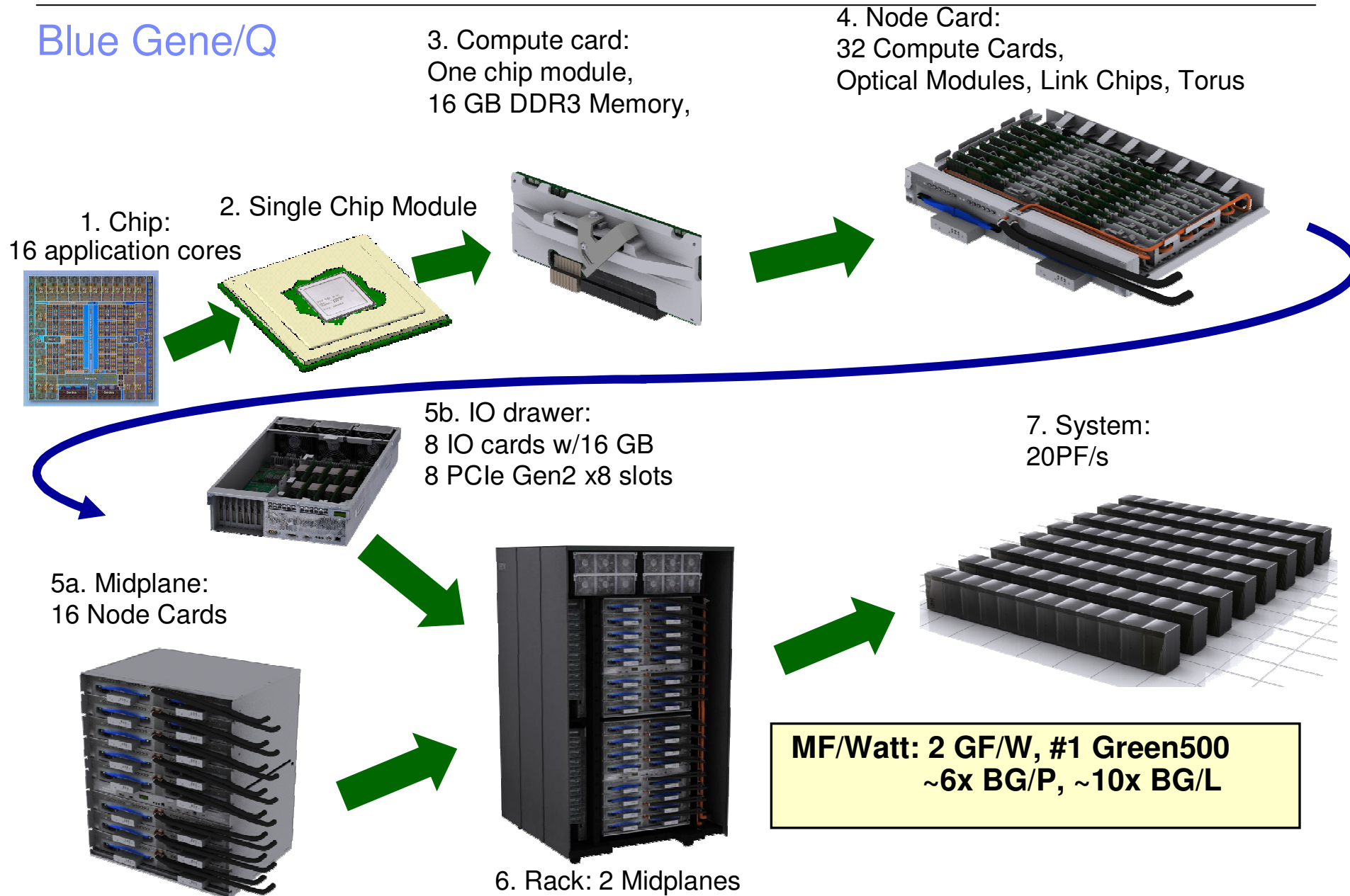
- Computer optimized for lattice quantum chromodynamics
 - QPACE System @ Forschungszentrum Jülich 2010
 - Commodity PowerXCell 8i processor
 - Custom FPGA-based network chip
 - Custom communication protocol for LQCD torus network
 - Custom voltage tuning
 - Custom liquid cooling
 - LQCD performance
 - 544-681 MFLOPS/W (QPACE)
 - 492 MFLOPS/W (Intel + Nvidia GPU-based Dawning Nebulae @ National Supercomputing Centre in Shenzhen)
 - #7 on Green500 June 2011 with 773.4 MFLOPS/W



H. Baier et al., “QPACE: Power-efficient parallel architecture based on IBM PowerXCell 8i”, First Intl. Conf. on Energy-Aware High-Performance Computing, 2010.

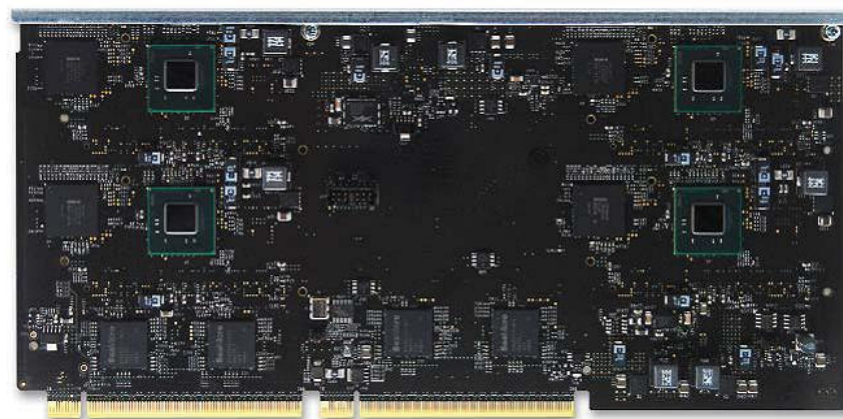
http://www.ena-hpc.org/2010/talks/EnA-HPC2010-Pleiter-QPACE_Power-efficient_parallel_architecture_based_on_IBM_PowerXCell_8i.pdf

Blue Gene/Q



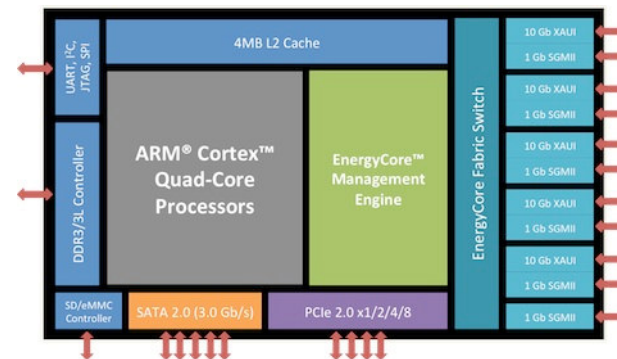
SeaMicro SM10000-64

- Goal is improve compute/power and compute/space metrics
- 10U server
 - 512 ATOM 64 bit cores (256 sockets)
 - 4GB per socket
 - 2.5KW
 - Operates as 256 node cluster
 - High-speed internal network – 1.2Tbit/s
 - External 64 x 1Gb ethernet ports
 - Virtualized I/O
 - All I/O is shared between sockets
 - Improves efficiency – very low overhead per socket.
- Designed for high volume of modest computational workloads
 - Web servers
 - Hadoop



ARM servers on the horizon

- ARM still 32bits (limited to 4GB per socket)
 - 64bits version announced - A15
- Marvell, Calxeda, STM announcements
 - Dual (STM) or quad core (Marvell, Calxeda) A9 from 1 to 2GHz
 - Calxeda: 5W for CPU+4GB DRAM DDR3
- ZT systems R1801e (product)
 - 8 modules with each module:
 - dual core STM processor, 1GB DRAM 1333MHz DDR3, 1GB Flash, 1Gbit ethernet, USB and SATA.
 - 80GB SSD
 - 80W in a 1U system.



Calxeda ECX-1000



ZT systems R1801e

Maybe the future would look like this.....

- Cluster of 'low-power' nodes
- Each node has a processor-memory socket
 - 3D stack with processor and DRAM
 - Stacks of PCM-Memory and PCM-Storage, connected to the processor-memory stack via Silicon Interposer.
 - Integrated switch routers for inter-node connectivity
- Inter-cluster connectivity with optics.
- Powered from renewable resources
- Cooled for free.

But surely we cannot predict the future !