Energy Smart High Performance Computing

April 27-30, 2009

The Salishan Conference on High-Speed Computing

Moe A. Khaleel[1], Andrés Márquez[1], Landon Sego[1], Steve Elbert[1], Tahir Cader[2], Rashawn Knapp[3], Karen Karavanic[3]

1. Pacific Northwest National Laboratory
2. Hewlett-Packard
3. Portland State University

Office of Science
U.S. DEPARTMENT OF ENERGY

Pacific Northwest
NATIONAL LABORATORY

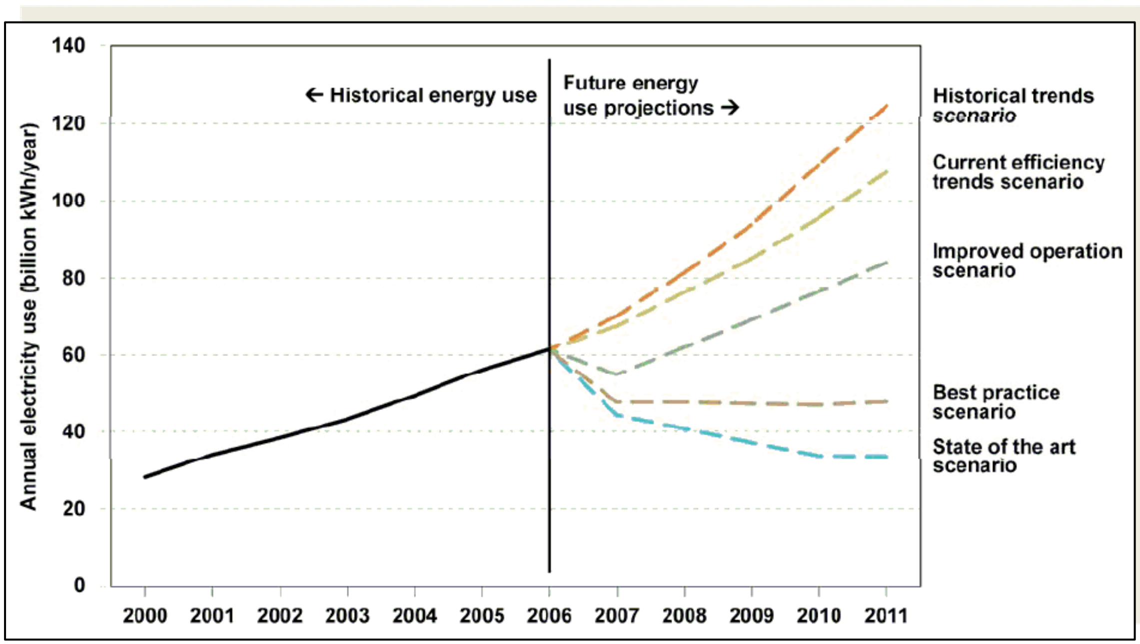
Proudly Operated by Battelle Since 1965

Overview

- **Power Consumption Trends for Data Centers and HPC: The white elephant in the room**
- The Energy Smart Data Center (ESDC) at PNNL
- Selected Research Topics at ESDC
  - Advanced Cooling Solutions
  - Metrics
  - Power Aware Computing

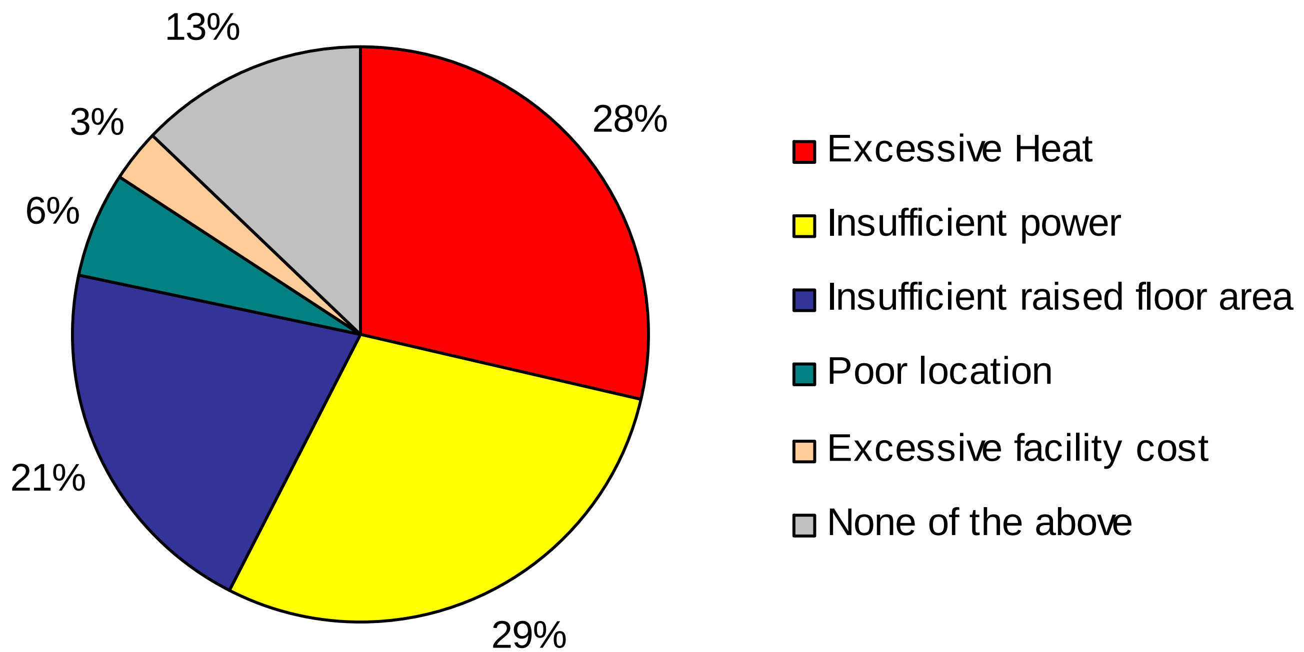
National Challenge

- Current efficiency trends estimate energy use in data centers could double by 2011 from a 2006 baseline
- A combination of improved operations, best practices and state of the art technologies could reduce electricity use by up to 55% compared to 2006 efficiency trends



EPA Report to Congress on Server and Data Center Energy Efficiency
Released On August 2, 2007 and in response to [Public Law 109-431]

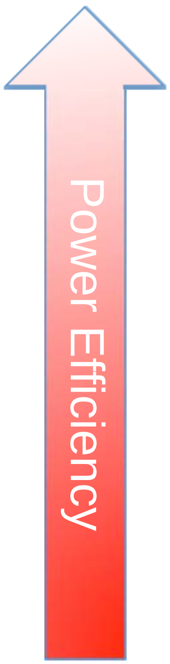
Excessive heat and insufficient power: Biggest concerns for data center managers



Source: AFCOM 2006. Five Bold Predictions For The Data Center Industry That Will Change Your Future [Keynote Slides]. AFCOM Data Center Institute

Top500 Power Consumption

newer systems



older systems

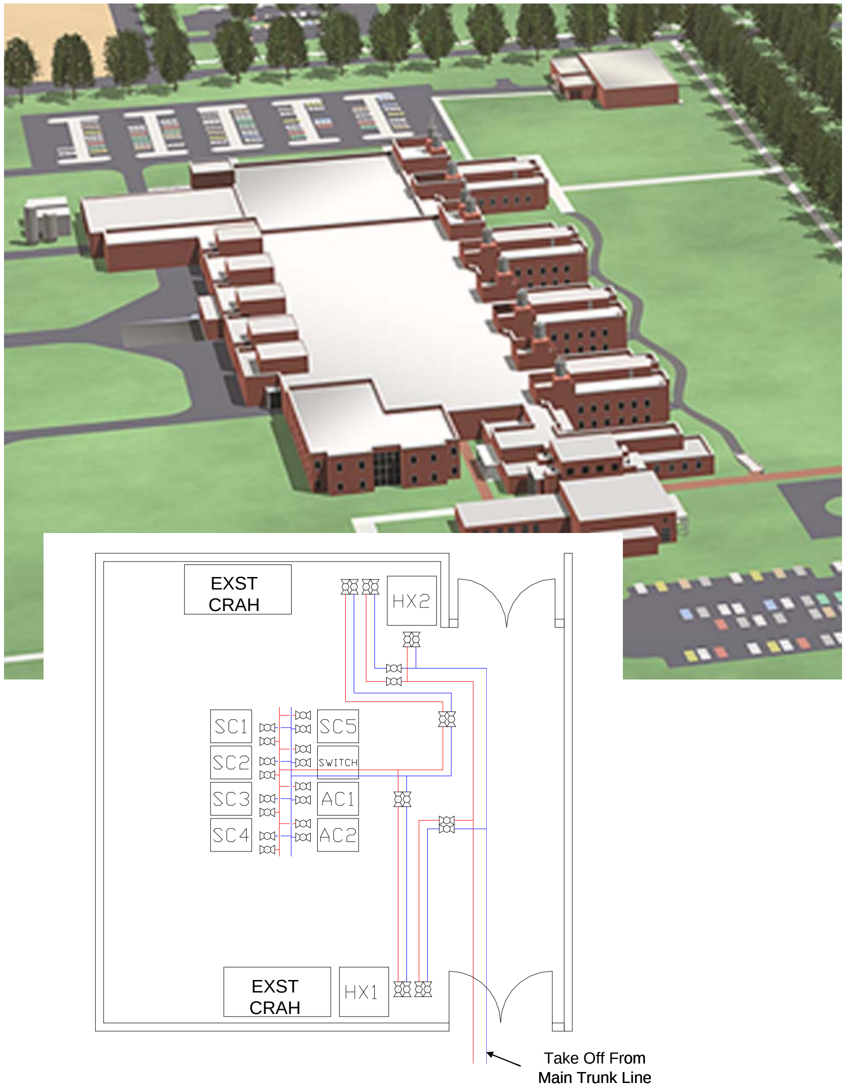
- TOP10 System
  - average power draw: 1.32 MW
  - average power efficiency: 248 Mflop/s/W
- TOP50 System
  - average power draw: 908 kW
  - average power efficiency: 193 Mflop/s/W
- TOP500 System
  - average power draw: 257 kW
  - average power efficiency: 122 Mflop/s/W

Source: http://www.top500.org/lists/2008/06/highlights/power

Overview

- Power Consumption Trends for Data Centers and HPC: The white elephant in the room
- **The Energy Smart Data Center (ESDC) at PNNL**
- Selected Research Topics at ESDC
  - Advanced Cooling Solutions
  - Metrics
  - Power Aware Computing

Energy Smart Data Center

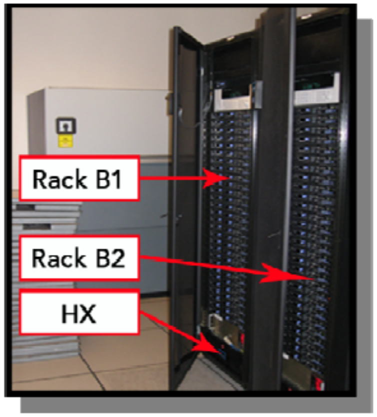


- Fully observable, almost fully controllable 700 sf Data Center
- Data Center integrated in a mixed used facility, sharing power distribution and cooling provisioning
- Over 1000 sensors providing data at the chip, server, rack, room and facility level.

http://esdc.pnl.gov

Device Under Test (Hardware): NW-ICE

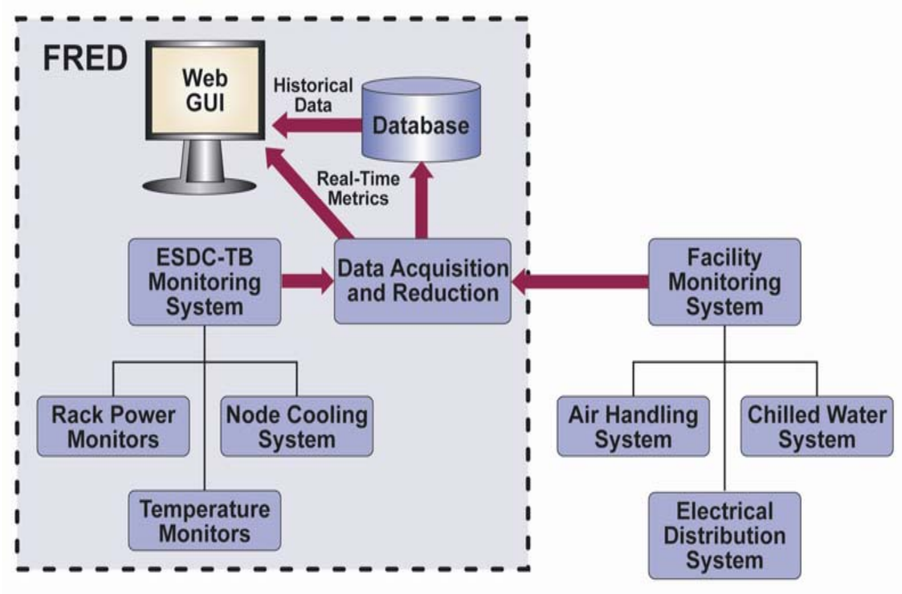
- 192 servers, each with two 2.3 GHz Intel (quad-core) Clovertown, 16 GB DDR2 FBDIMM memory, 160 GB SATA local scratch, DDR2 Infiniband NIC
- Five racks with evaporative cooling at processors
- Two racks completely air cooled
- Lustre Global File System
  - 34TB mounted
  - 49TB provisioned



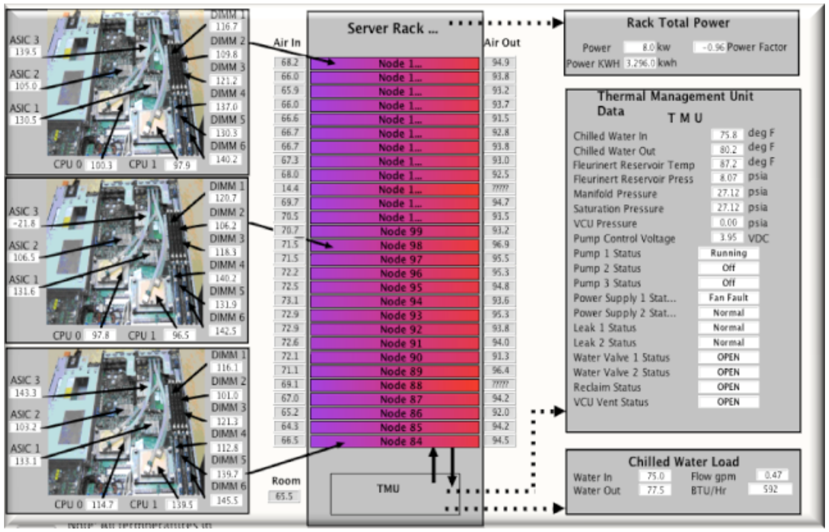
Measurement Harness

- Over 1000 sensors at the chip, server, rack, room and facility-level measuring air/liquid temperatures, humidity ratios flows, pressure differences and electric currents

- FRED software to monitor environmental data; based on in-house developed industrial strength supervision and diagnostic tool DSOM



Contributors to Power Consumption: Power Distribution

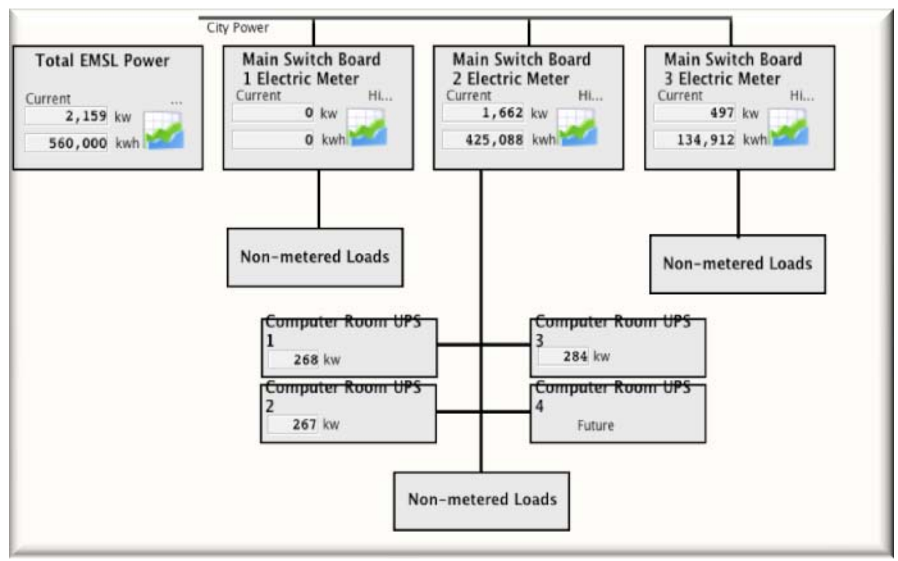


Facility:

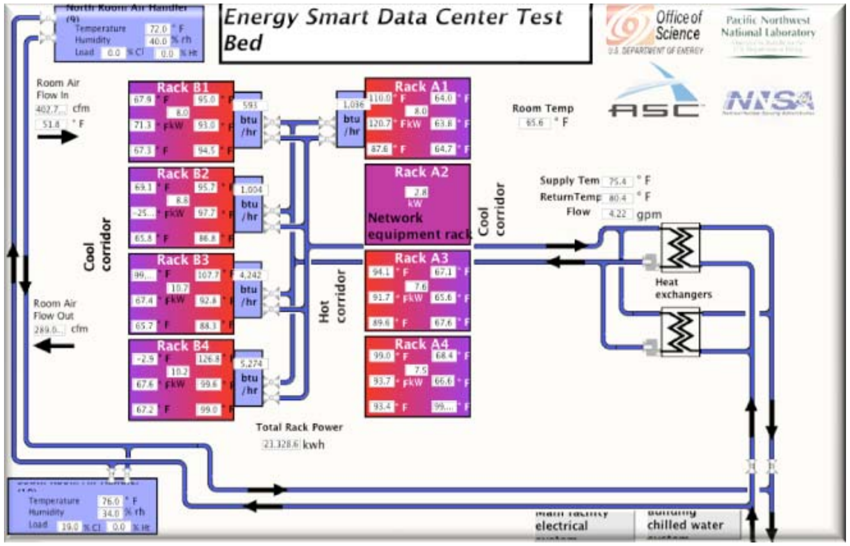
- Transformers
- Rectifiers
- UPS
- Inverters

Data Center:

- Power Management Modules
- Power Supply Units
- Voltage Regulators



Contributors to Power Consumption: Cooling Chain

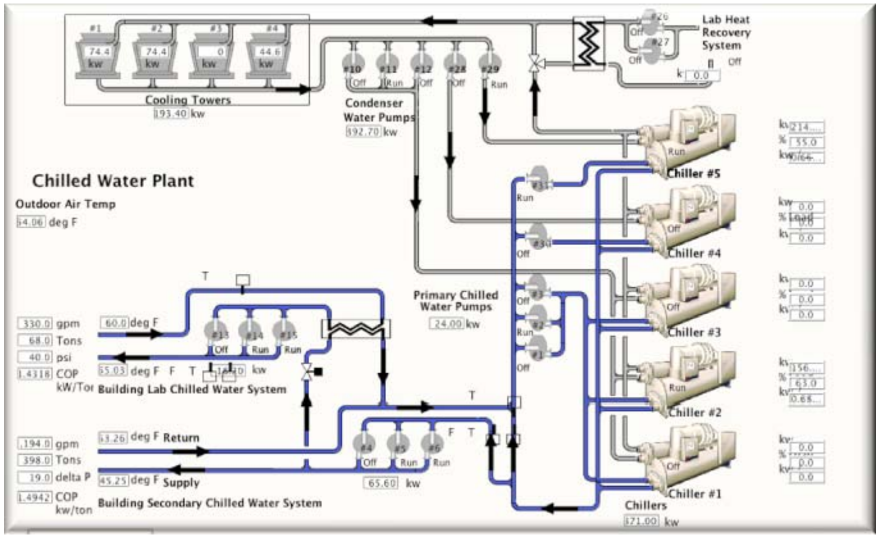


Machine Plant:

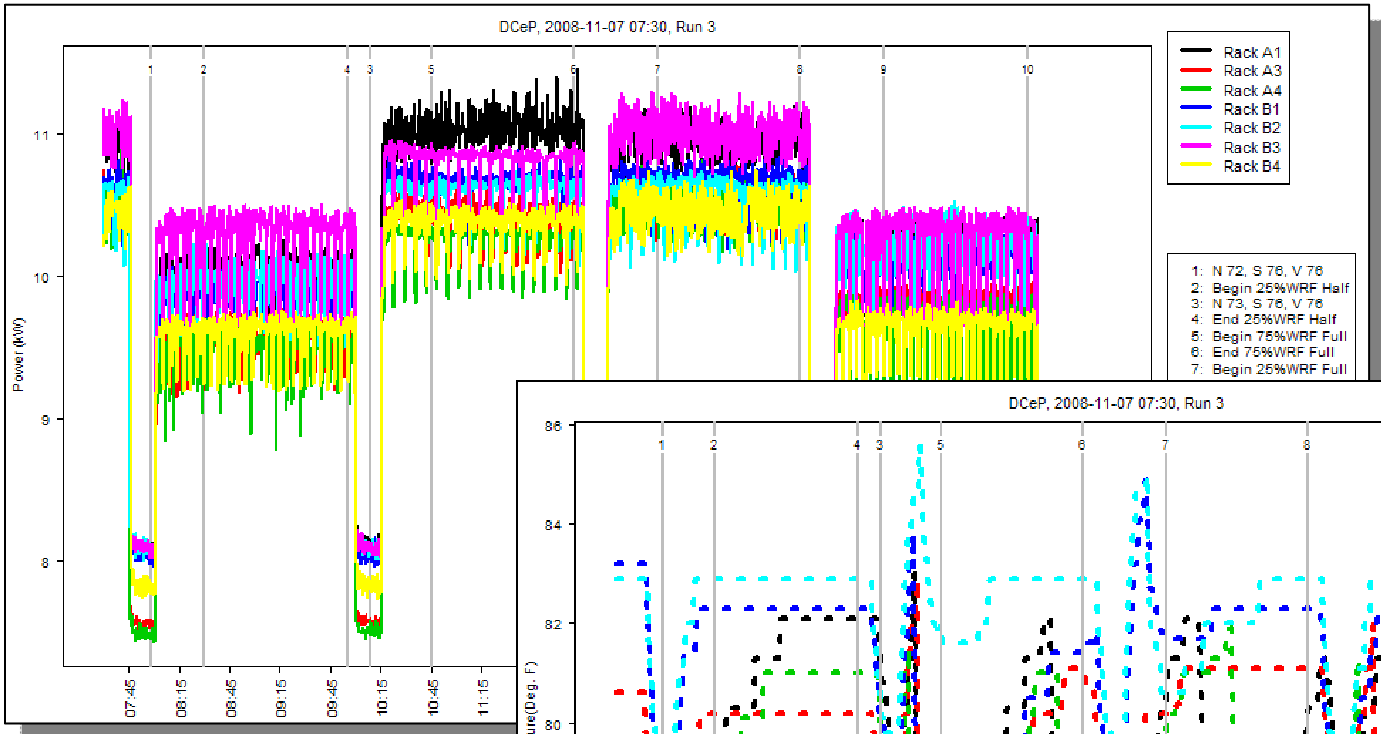
- Pumps
- Chillers
- Cooling Towers
- Economizers

Data Center:

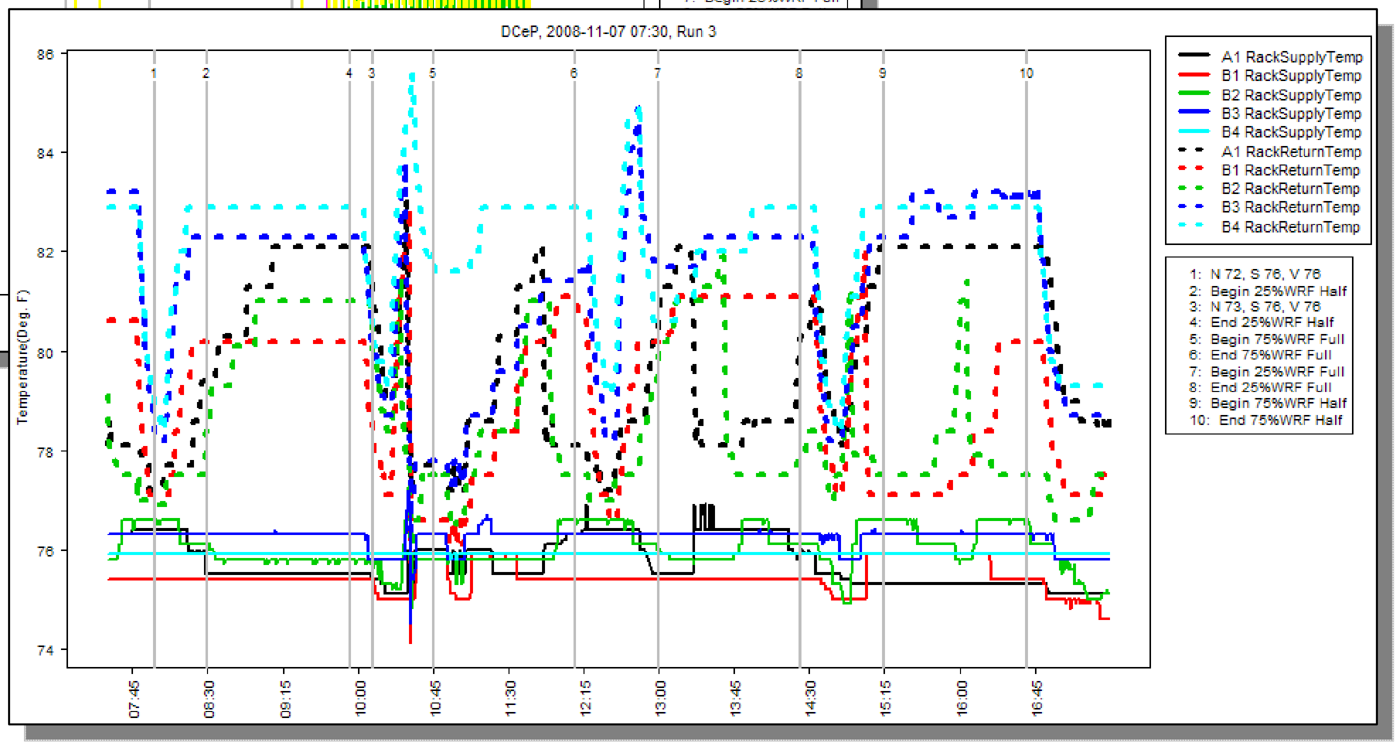
- Air Handlers
- Closely Coupled Cooling Systems
- HVAC



Power and Water Temperature Signatures



Rack Power



Water Temperature

Overview

- Power Consumption Trends for Data Centers and HPC: The white elephant in the room
- The Energy Smart Data Center (ESDC) at PNNL
- **Selected Research Topics at ESDC**
  - **Advanced Cooling Solutions**
  - Metrics
  - Power Aware Computing

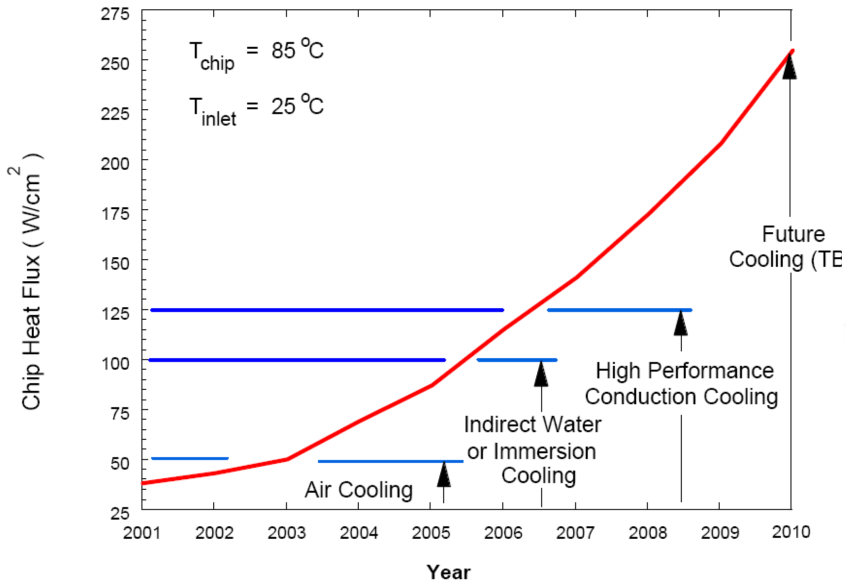
Advanced Cooling Solutions

- Challenges: evaluating existing cooling solutions for HPC
  - Are existing cooling solutions energy efficient?
  - Are best practices applied?
  - Do existing cooling solutions scale with high density racks?

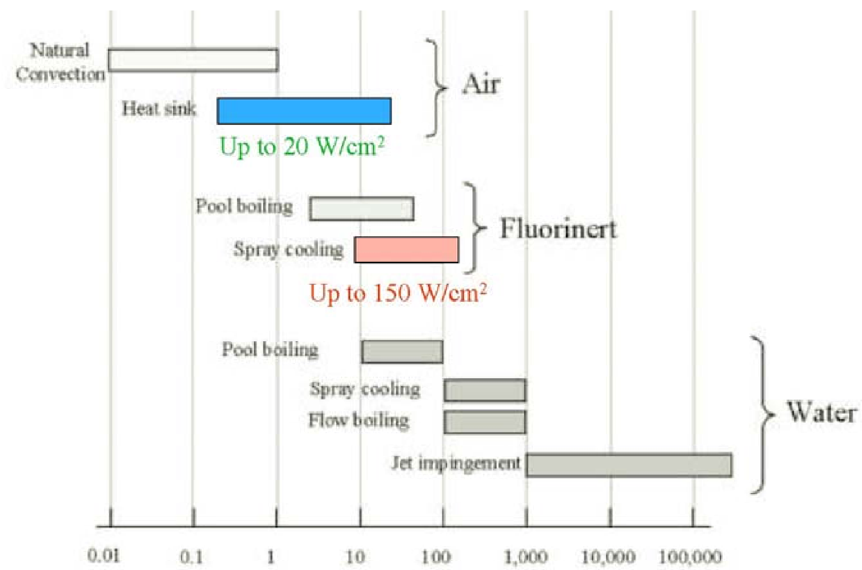
- Our answers:
  - Evaluate advanced cooling solutions that act close to heat sources
  - Explore hybrid cooling solutions, e.g., air and spray

Krell Institute Study: Energy Efficient HPC Data Center Infrastructure Issues

- John Ziebarth, Gary Johnson



Projected Heat-Flux W/cm²



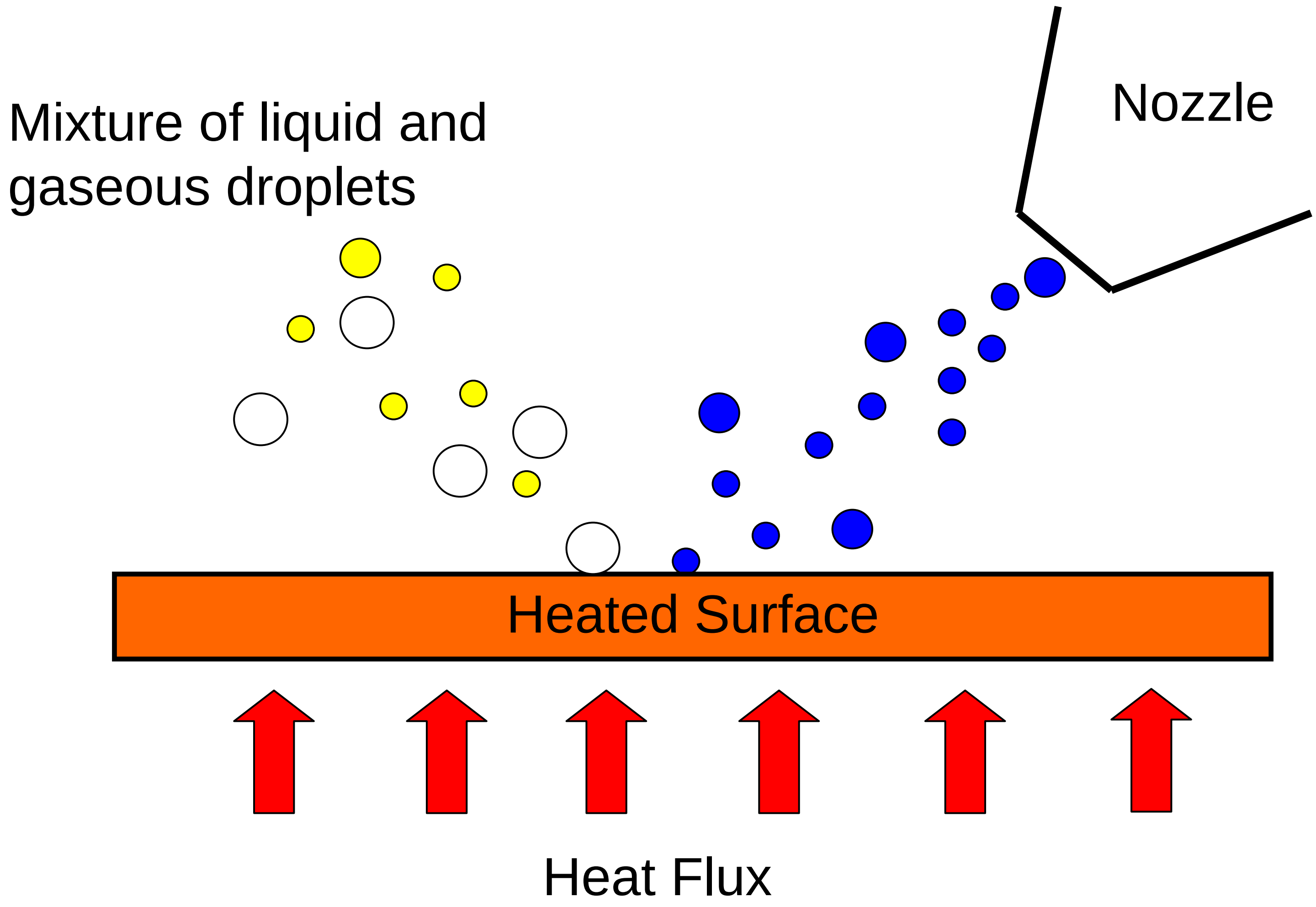
Critical Heat-Flux W/cm²

Office of Science
U.S. DEPARTMENT OF ENERGY

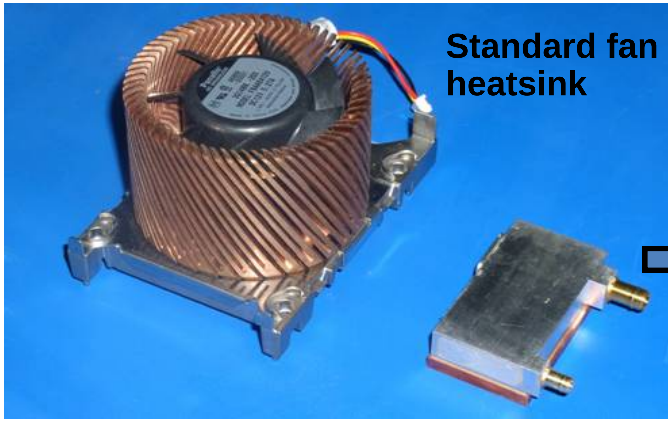
Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Two-Phase Cooling Regime

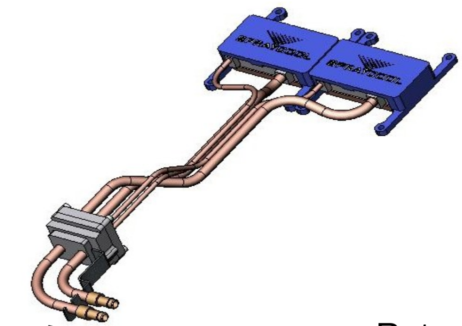


Spray Spot Cooling: Server Conversion

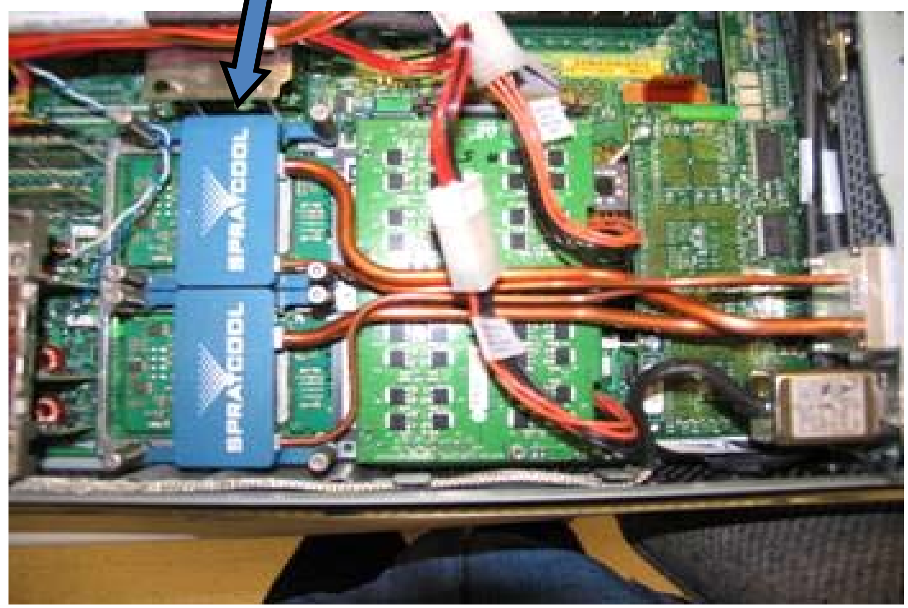


SprayModule™

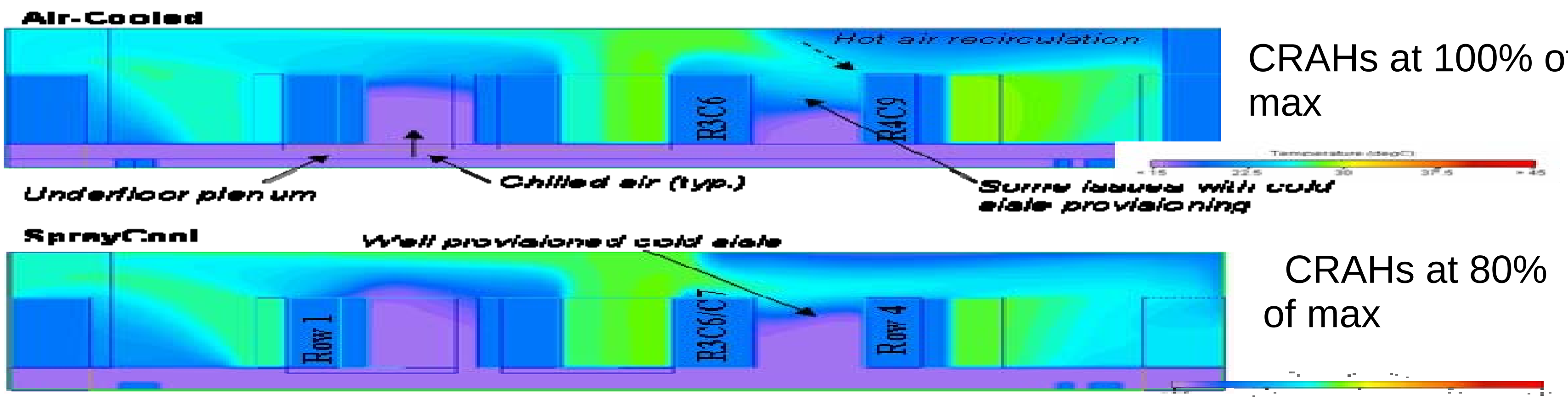
SprayModule Kit



Retrofitted Server



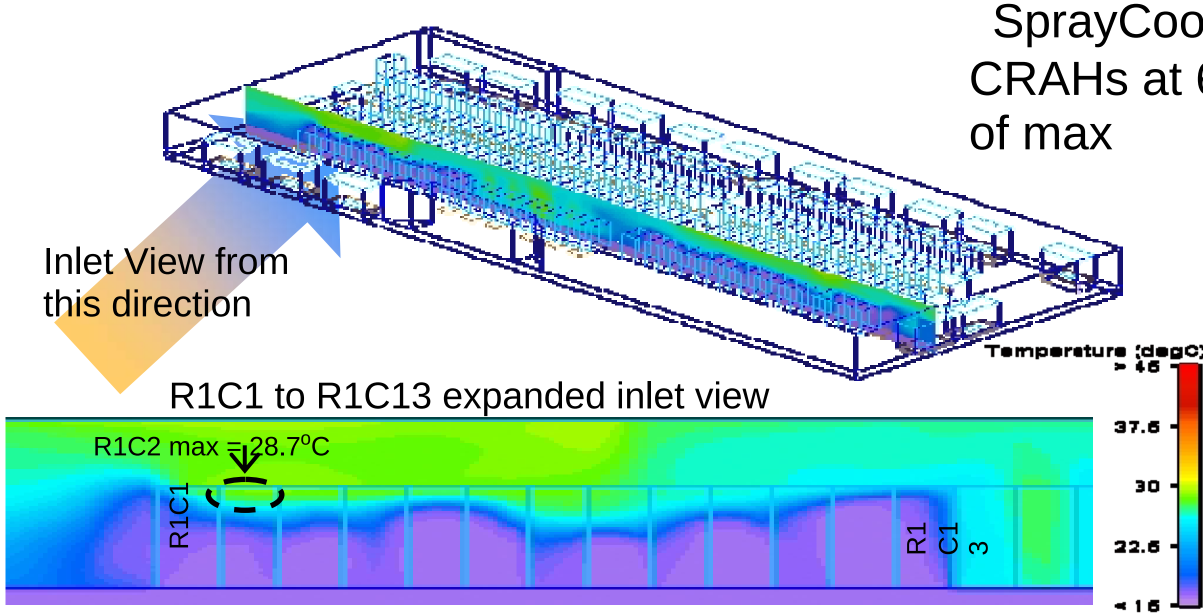
Replacing Processor Air-Cooling with Spraycooling (study conducted with HP)



CRAHs at 100% of max

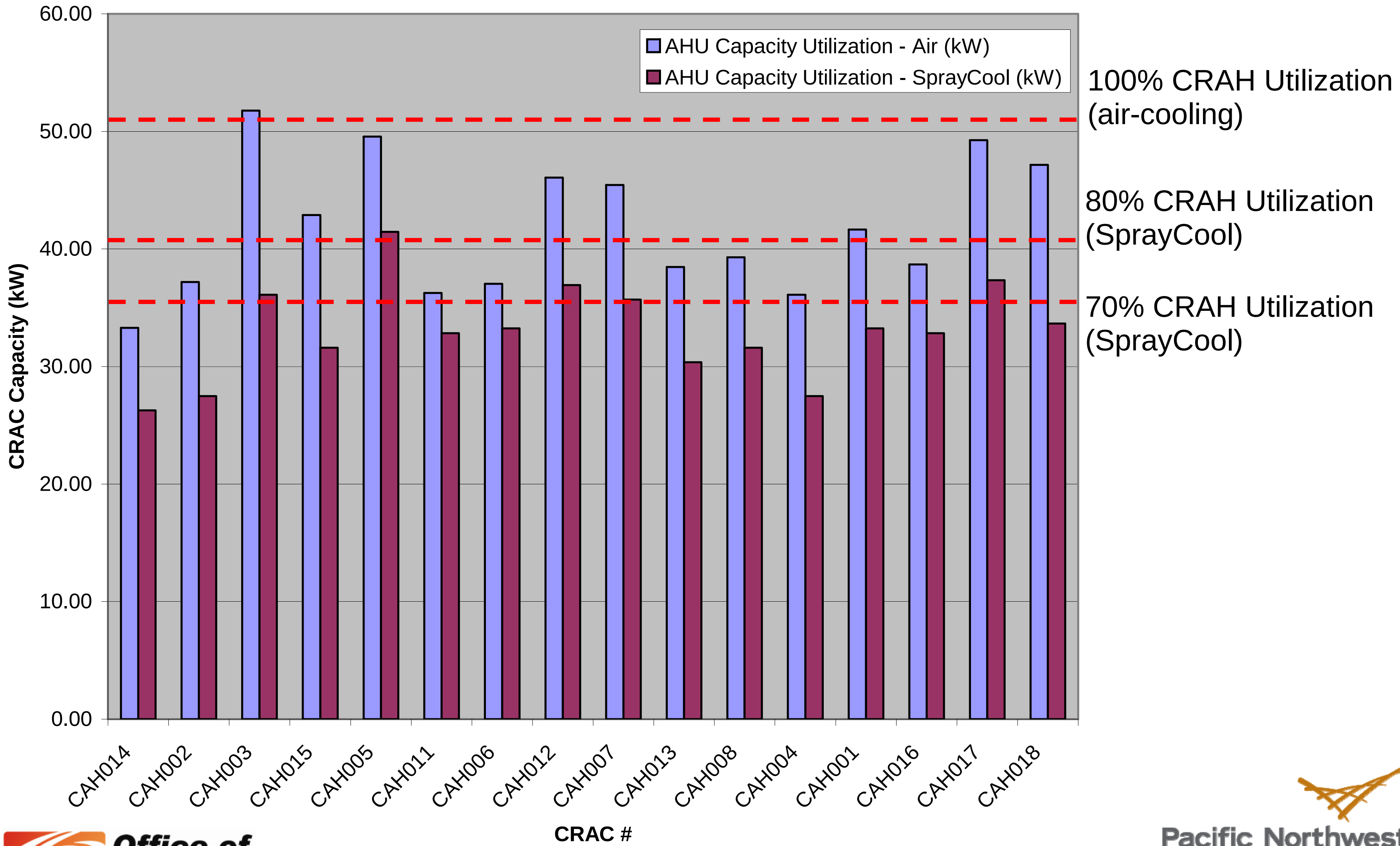
CRAHs at 80% of max

SprayCool - CRAHs at 60% of max

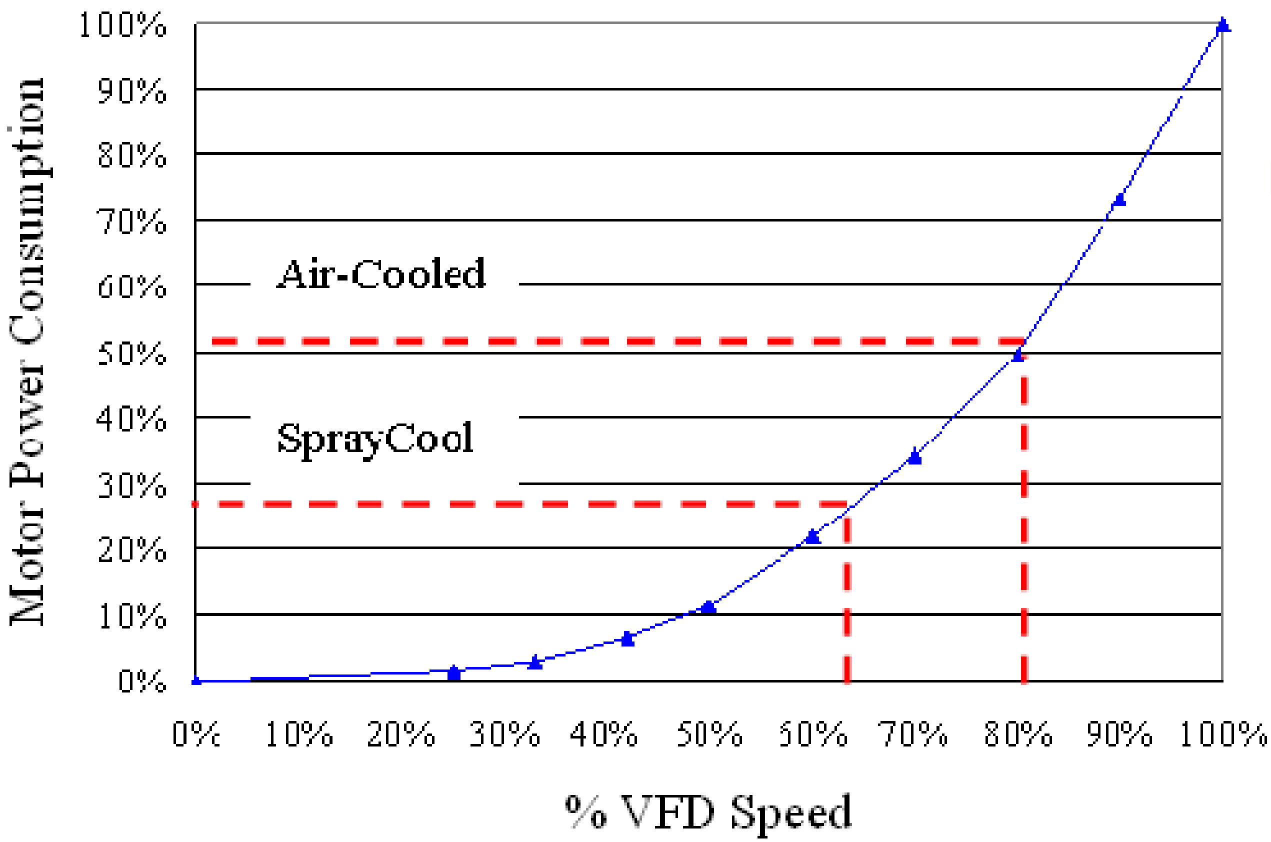


- Poor cold aisle provisioning at this CRAH setting
- Ideal CRAH setting for SprayCool is approx 70% of max

How Much Airflow is Required?



Energy Savings VFD



- Air-cooled facility's blowers use 82% more power

Source: Cullen Bash, HP Labs.

Office of Science
U.S. DEPARTMENT OF ENERGY

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Overview

- Power Consumption Trends for Data Centers and HPC: The white elephant in the room
- The Energy Smart Data Center (ESDC) at PNNL
- **Selected Research Topics at ESDC**
  - Advanced Cooling Solutions
  - **Metrics**
  - Power Aware Computing

Metrics

Challenges: evaluating usability of existing metrics for HPC

- Do existing metrics penalize HPC?
  - Not considering space/density
  - Not considering output: product (as in time-to-solution).

Our answers:

- Introduce productivity metrics in conjunction with The Green Grid
- Establish realistic test cases.

Metrics: Why important

We need metrics to measure power efficiency

Why should we care about existing metrics?

- Recognized and accepted by a large community
- Use to drive
  - Next-generation of HW/SW and infrastructure development
  - Regulation and mandates in energy efficiency.

Popular Data Center Metrics: Infrastructure Efficiency

PUE (Power Usage Efficiency)

$$\frac{\text{Total Facility Power}}{\text{Computer Power}}$$

Range: 1 - ∞

DCiE (Data Center Infrastructure Efficiency)

$$\frac{\text{Computer Power}}{\text{Total Facility Power}}$$

Range: 0 - 1

- No productivity measured
  - Computer could be idling
- No space considered
  - Computer could be a distributed web server farm
- Ratio can be misleading
  - Computer could be drawing large power
- Scope of "Total Facility Power", "Computer Power" may not be consistent

Motivation: DCP and DCeP

Existing scientific computing metrics do not adequately address the total energy cost of producing a computational result:

- Metrics such as MegaFLOPs/W ignore the power delivery and cooling energy costs
- Data center energy-efficiency metrics such as PUE/DCiE focus only on the efficiency of equipment in the data center used to deliver conditioned power and eliminate heat generated by the computing equipment
- These metrics are not designed to quantify the useful work output of a data center in relationship to the total energy cost of the facility.

Source: The Green Grid

Motivation: DCP and DCeP (contd)

- DCeP looks at energy consumption of the whole facility, not just the computing equipment
- Provides a means to benchmark computational energy productivity
  - Specific changes to workload mix or facility configuration can be assessed in terms of their overall effect on the energy productivity of the facility
- PNNL is an early adopter of this new productivity metric

The Green Grid's DCP and DCeP

A family of metrics:

Data Center Productivity (DCP)

$$\frac{\text{Useful Work Produced}}{\text{Total Quantity of Resource Consumed Producing this Work}}$$

Range: 0 - ∞

A particular metric that fits well our model A:

Data Center Energy Productivity (DCeP)

$$\frac{\text{Useful Work Produced}}{\text{Total Data Center Energy Consumed Producing this Work}}$$

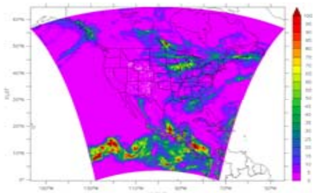
We interpret Data Center to include the computer and the supporting power/cooling equipment in the facility

Range: 0 - ∞

Source: The Green Grid

Device Under Test (Software): Typical Production PNNL HPC Workload Mix

WRF

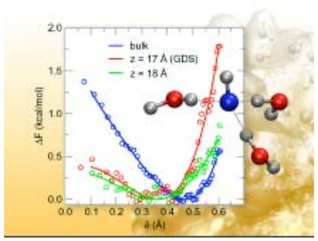


Multiple concurrent basic 4.5 days weather forecasts for North & Central America

- Initialization: 1° Global Forecast System analysis from National Weather Service
- Decomposition: 480x480 Cartesian grid (15km) with 45 levels
- Solver: Horizontal: Explicit High-Order Runge-Kutta; Vertical: Implicit
- Output: Asynchronous 2.3GB netCDF every 3 model-hours per forecast

Multiple concurrent liquid-vapor interface model simulations

CP2K



- Initialization: Standard slab geometry (15x15x71Å^3)
- Decomposition: 215 H_2O with single hydroxide ion
- Solver: Density Functional Theory with dual basis set (Gaussian & Plane-Wave) in conjunction with molecular dynamics and umbrella sampling
- Output: Synchronous 75MB per 20k 0.5fs model-steps (MD time step)

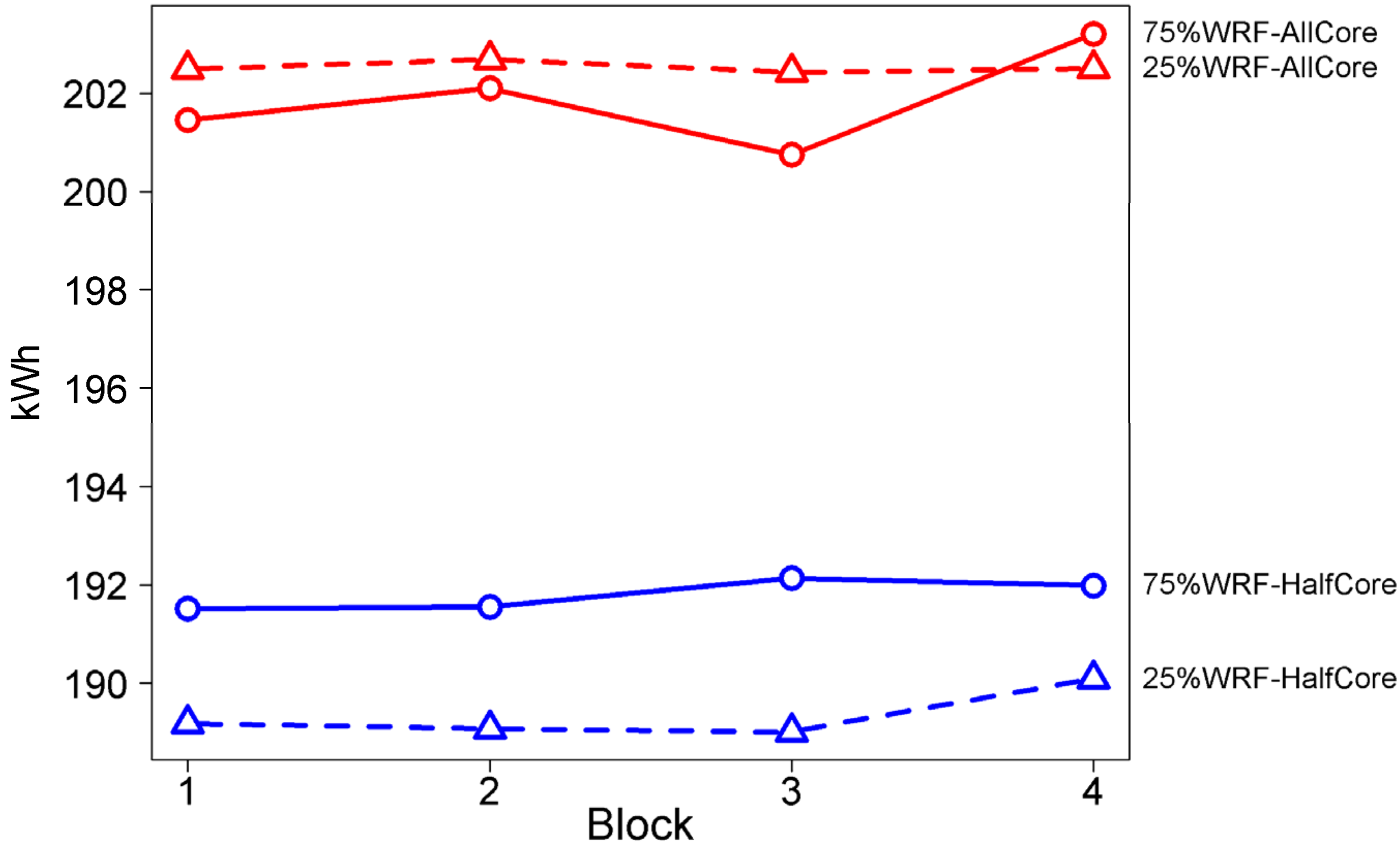
Experimental Plan: A PNNL HPC Workload

Completely randomized block design with a 2^2 factorial treatment structure:

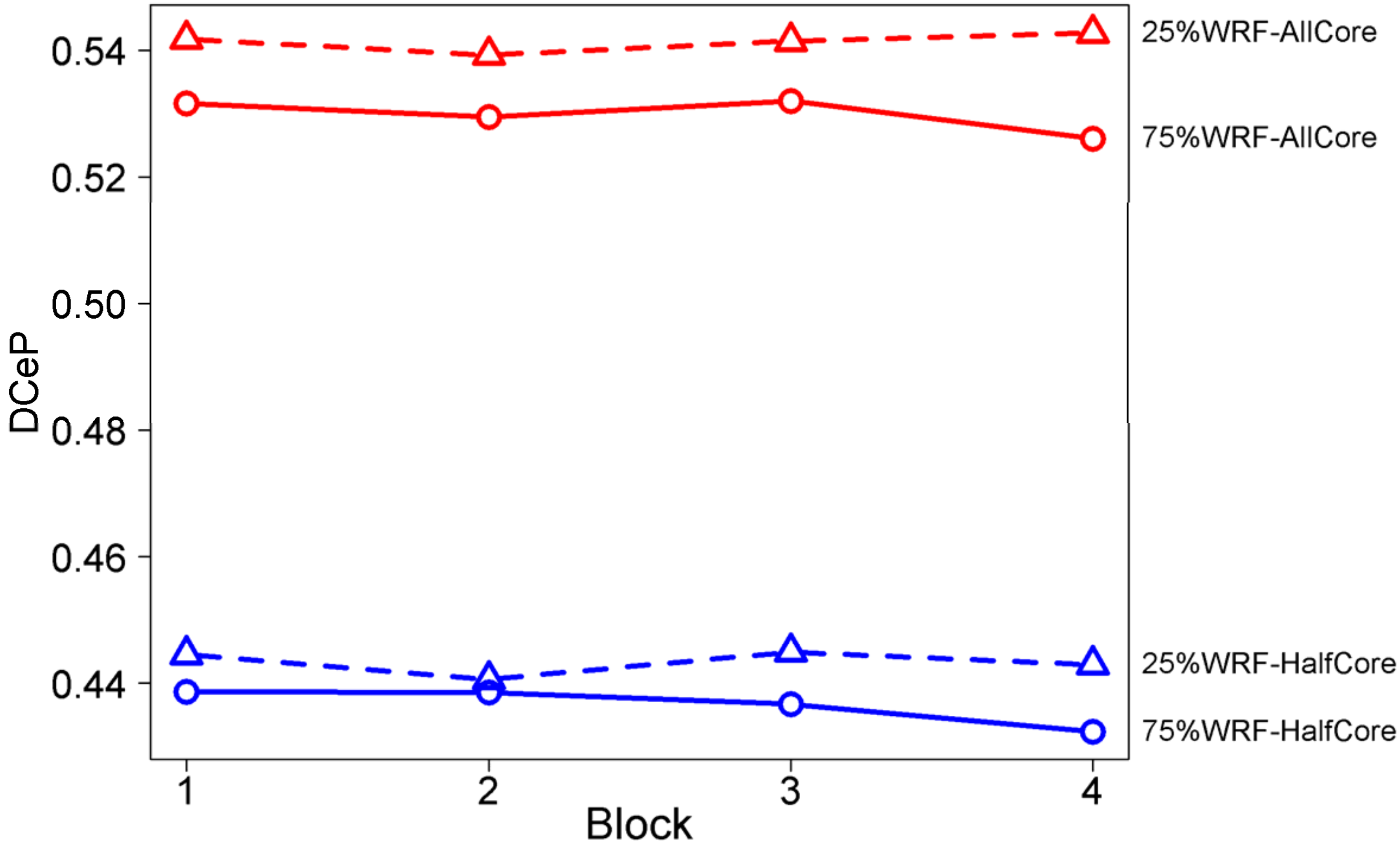
- **Treatment 1**: application's machine load:
  - 75% WRF, 25% WRF
- **Treatment 2**: number of cores per server:
  - full-core, half-core
- **Block**: day of the week and time of run:
  - weekday, weekend, day, night

Each treatment produces Useful Computational Units (UCU) extracted from a stable, ~1.5 hour long assessment window

Energy Use in kWh



DCeP



Useful Work = $\Sigma\ V_i T_i$

T_i = 1 if task completed in assessment window, 0 otherwise

V_i = 0.10 for CP2k, V_i = 1 for WRF

(normalized to same sampling rate, same weight)

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

Summary of Experimental Results

- DCeP ***can*** be used to distinguish between different operational states in a data center and guide load balancing
- Full core implementations use more energy than half core, but are also more efficient (regardless of weighting scheme)
- Treatments with 25% WRF load are more efficient than 75% (given weighting scheme where each CP2K unit is with 10% of a WRF unit)

Overview

- Power Consumption Trends for Data Centers and HPC: The white elephant in the room
- The Energy Smart Data Center (ESDC) at PNNL
- **Selected Research Topics at ESDC**
  - Advanced Cooling Solutions
  - Metrics
  - **Power Aware Computing**

Integrating Power/Cooling Into Application Performance Analysis

Challenges: combining two disparate worlds of tools and data

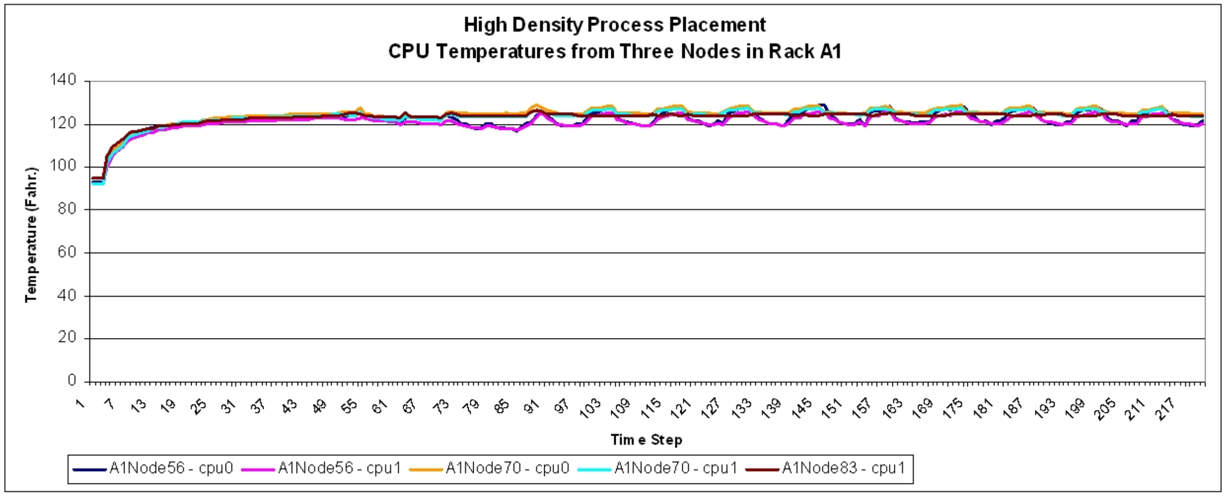
- Infrastructure for collecting, storing and analyzing data
  - combine data from the application and room environment perspectives
  - Ex. "What was the rack temperature during this run?"

Our answers:

- PerfTrack performance database extended to hold room data
- Job placement study currently being conducted at ESDC facility
  - "Can we save $ on cooling by changing job placement within the cluster?"
  - "Is it more efficient to use one rack, or use them all?"

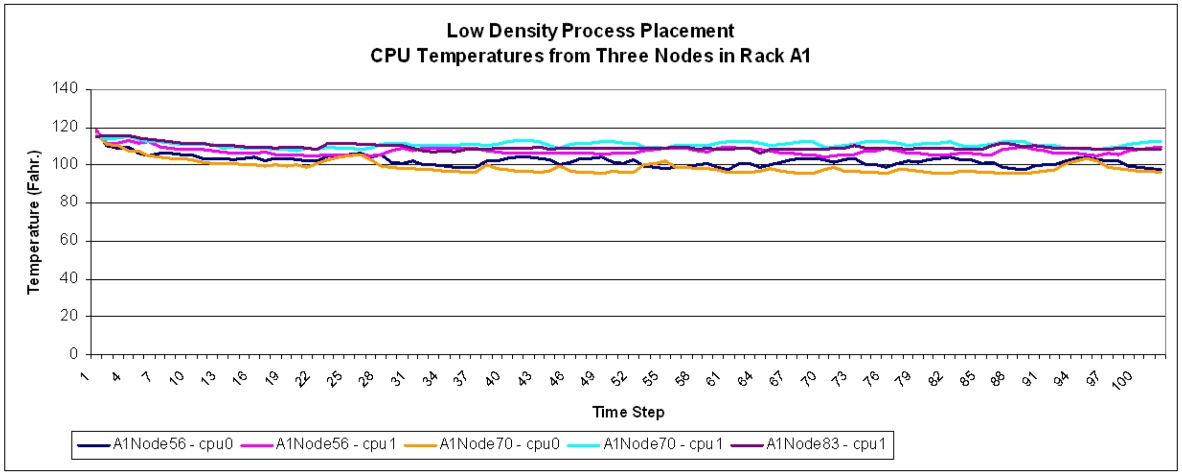
Thermal Profiles of Running Applications

Goal: use integrated data to inform job scheduler



- “High Density” Placement
- 224 processes on 1 rack
- (8 processes per node)
- CPU temperatures stay above 120ºF

- “Low Density” Placement
- 224 processes on 4 racks
- (2 processes per node)
- CPU temperatures stay below 110ºF



Summary

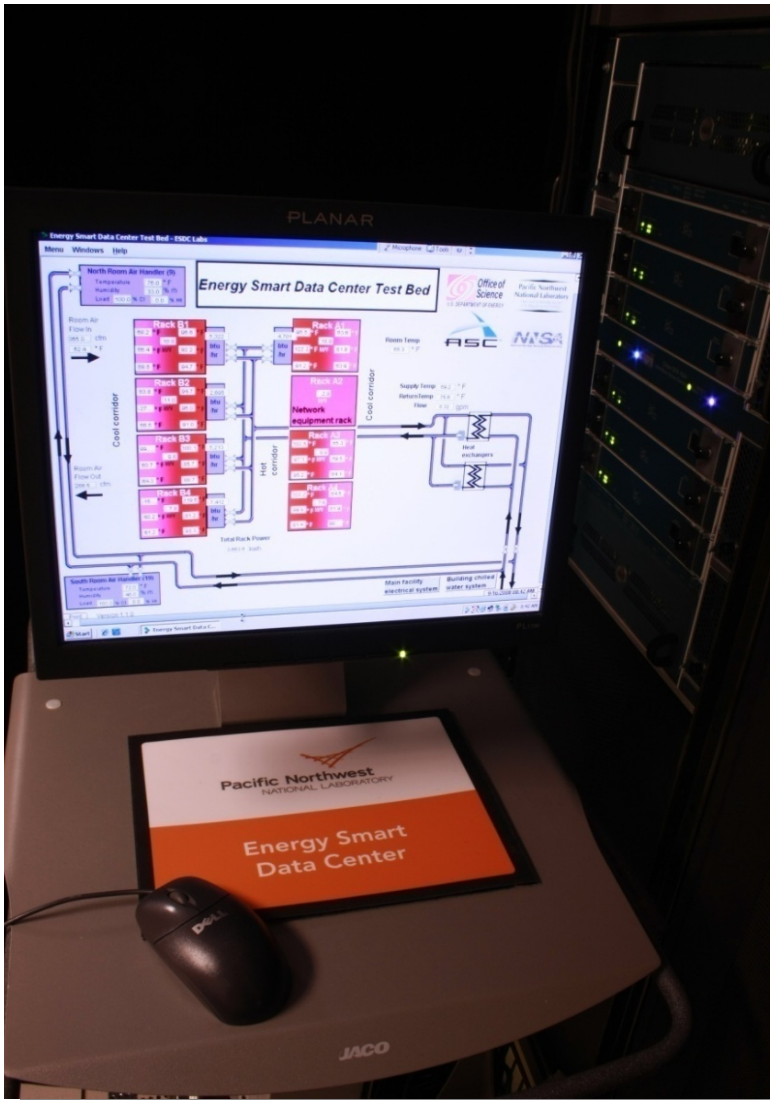
- **Power Consumption trends for Data Centers and HPC: The white elephant in the room**

  Energy use is an increasingly acute problem
  - At the national level
  - At the data center level
  - As HPC user

- **The Energy Smart Data Center (ESDC) at PNNL**

  Provides the building blocks to conduct energy efficiency studies by providing monitoring and control tools
  - At the mechanical side
    - Advanced Cooling solutions
    - Advanced Power Distribution
  - At the software side
    - Sensible HPC Metrics
    - Power Aware Computing

Questions?



Energy Smart Data Center Research

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*