Towards real-world HPC energy efficiency and productivity metrics in a fully instrumented Datacenter

Andrés Márquez

L. Sego, K. Fox, D. Sisk, D. Hatley, D. Johnson, S. Elbert, M. Khaleel

http://esdc.pnl.gov/

Pacific Northwest
NATIONAL LABORATORY

EPA reports energy used in U.S. for servers and data centers is significant.

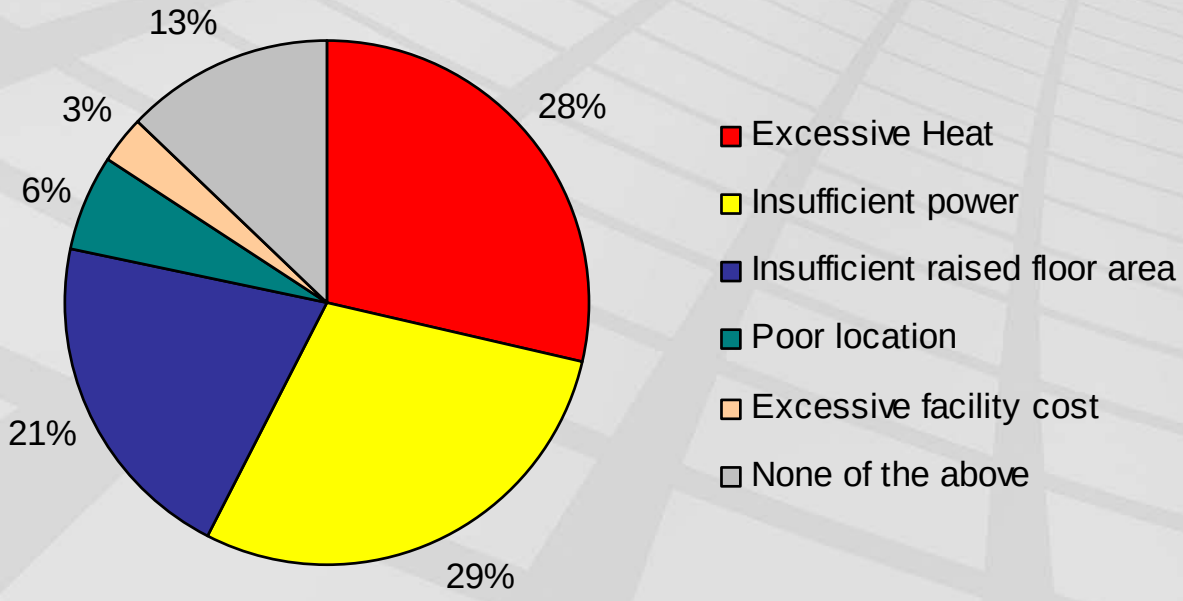
- ~ 61 billion kilowatt-hours (kWh) in 2006
- 1.5% of total electricity consumption
- Total electricity cost of about $4.5 billion.
- Similar to the amount of electricity consumed by approximately 5.8 million average U.S. households (or about five percent of the total housing stock).
- Federal servers and data centers alone
  - ~ 6 billion kWh
  - 10% of electricity used for servers and data centers
  - Total electricity cost of about $450 million annually.

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

Pacific Northwest
NATIONAL LABORATORY

Two biggest concerns for data center managers?

Excessive heat and insufficient power



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

Pacific Northwest
NATIONAL LABORATORY

Top500 statistics

newer systems

older systems

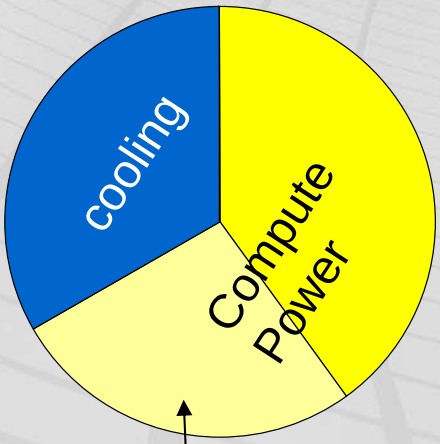
- **TOP10 System**
  - **average power draw: 1.32 MW**
  - **average power efficiency: 248 Mflop/s/W**
- **TOP50 System**
  - **average power draw: 908 kW**
  - **average power efficiency: 193 Mflop/s/W**
- **TOP500 System**
  - **average power draw: 257 kW**
  - **average power efficiency: 122 Mflop/s/W**

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

Current Power Usage by Chinook, Molecular Science Computing Flagship System at PNNL

- Chinook (160TF peak), has 2310 dual socket quad-core AMD Opteron (2.2GHz) based servers from HP each with 16 GB memory, 365 GB local disk, a DDR Infiniband interconnect, and 297 TB global disk
- Consumes nearly 1.9 MW
  - ~ 1/3 for cooling
  - ~ 2/3 compute power (1.25 MW)
    - 40% of compute power is lost to power delivery (rectifier, UPS, Feed, PDU, power supply, voltage regulator)
- Average power efficiency for HPL
  - no losses: 133MFlop/s/W
  - with power delivery losses: 80MFlop/s/W
  - with power- and cooling delivery losses: 52MFlop/s/W

Top500 measures here



40% of compute power lost in power delivery

Pacific Northwest NATIONAL LABORATORY

We need metrics to measure power efficiency

Why should we care about existing metrics?

- Recognized and accepted by a large community
- Used to drive
  - next generation of HW/SW and infrastructure development
  - regulation and mandates in energy efficiency

A Bunch of Popular Metrics
Infrastructure Efficiency

PUE (Power Usage Efficiency)

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

Range: 1 - ∞

DCiE (Datacenter Infrastructure Efficiency)

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

Range: 0 - 1

- No productivity measured
  - Computer could be idling
- No space considered
  - Computer could be distributed web server farm
- Range is restricted
  - Computer could be drawing large power

Pacific Northwest
NATIONAL LABORATORY

A Bunch of Popular Metrics
Computing Efficiency

SWaP (Space Watts Performance)

$$\frac{\text{Performance}}{\text{Power} * \text{Space}}$$

default: Flop/s

Range: 0 - ∞

Power Efficiency

$$\frac{\text{Performance}}{\text{Power}}$$

default: Flop/s

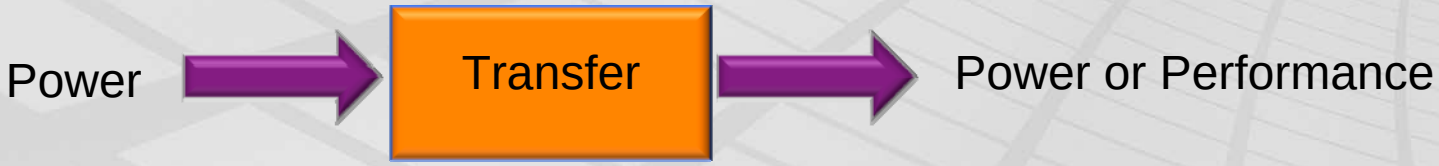
Range: 0 - ∞

- Both
  - Performance: default: Flop Rate
    - Low level. Might not be what the user requires
- SWaP
  - Space considered
    - Rewards density without considering impact on infrastructure
- Power Efficiency
  - No space considered
    - Density not explicitly rewarded

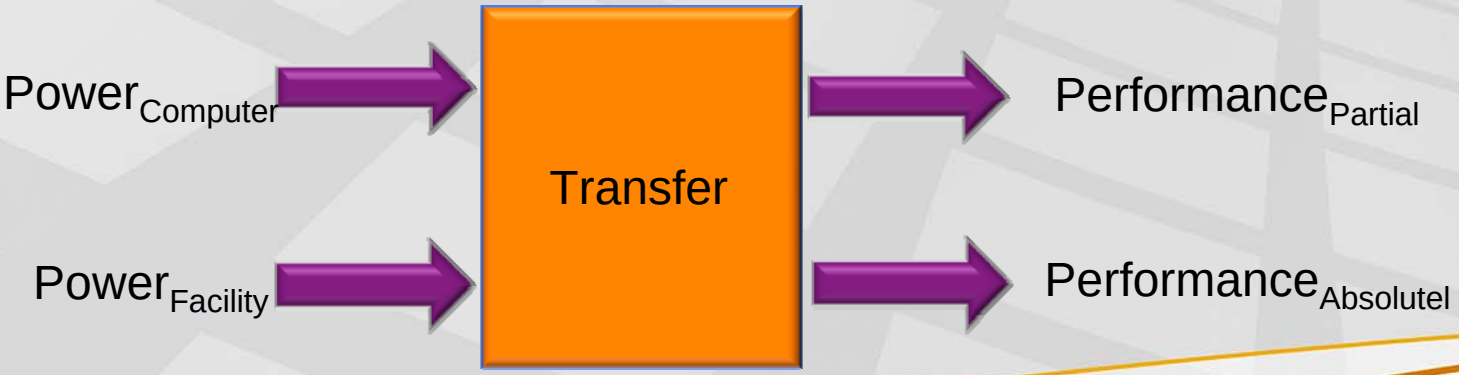
Pacific Northwest
NATIONAL LABORATORY

Chasing the Holistic Metric

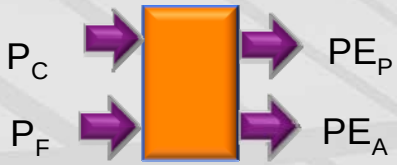
Current popular power/energy efficiency metrics have a "2-poles" view that limits a holistic facility/datacenter/computer/component view



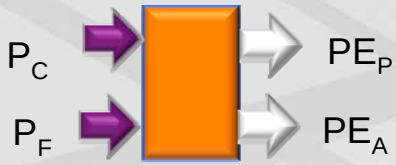
A "4-poles" view provides a comprehensive and partial view at the same time.



Metrics under “4-poles” view

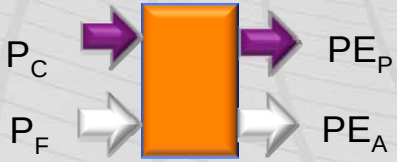


- P_F : Power drawn by facility related to computer
- P_C : Power drawn by computer
- PE_A : Absolute performance (e.g. time-to-solution rate)
- PE_P : Partial performance (e.g. Flop/s , Packets/s rates)

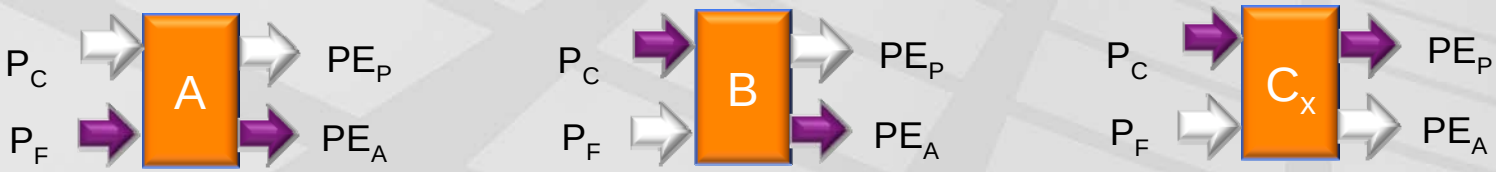


Infrastructure metrics only capture left side

Productivity metrics tend to capture top side



We need A, B and C_x metrics under a unified model



A: true power efficiency (as used in TCO)
B: computer power efficiency (as used in “green” resource allocation)
C_x: figure of merit (as used in unit to unit comparisons)

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 (substitute Facility for Data Center):

Data Center Energy Productivity (DCeP)

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

Range: 0 - ∞

Source: The Green Grid

Pacific Northwest
NATIONAL LABORATORY

Useful Work for DCP and DCeP

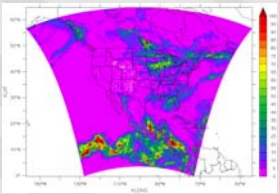
Useful Work = $\sum V_i U_i(t,T) T_i$

- M is the number of tasks initiated during the assessment window
- V_i is a normalization factor that allows the tasks to be summed numerically
- T_i = 1 if task i completes during the assessment window, and 0 otherwise.
- $U_i(t,T)$ is a time-based utility function for each task, where the parameter t is elapsed time from initiation to completion of the task, and T is the absolute time of completion of the task.

Source: The Green Grid

Experimental Plan:
A PNNL HPC workload

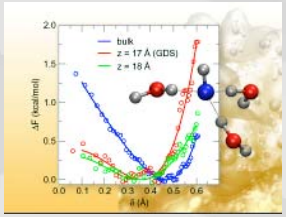
WRF



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

CP2K



Multiple concurrent liquid-vapor interface model simulations

- 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)

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

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

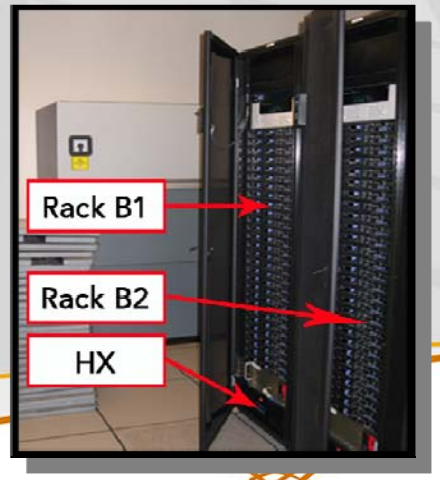
Each treatment produces Useful Computational Units (UCU) extracted from a stable, ~1.5 hour long assessment window.

The aggregation of the UCU constitutes Useful Work Produced

Pacific Northwest
NATIONAL LABORATORY

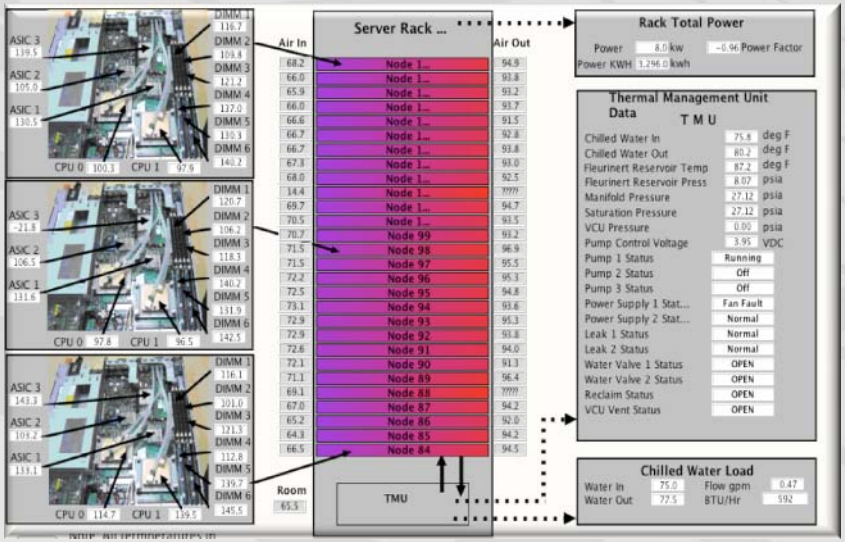
Device Under Test: NW-Ice

- 192 servers, 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



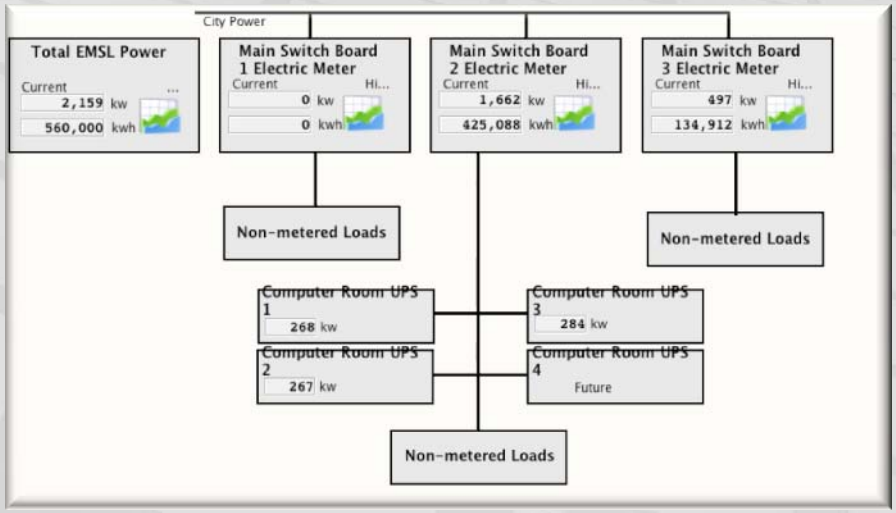
Pacific Northwest
NATIONAL LABORATORY

Contributors to Power Consumption: Power Distribution



Facility:

- Transformers
- Rectifiers
- UPS
- Inverters

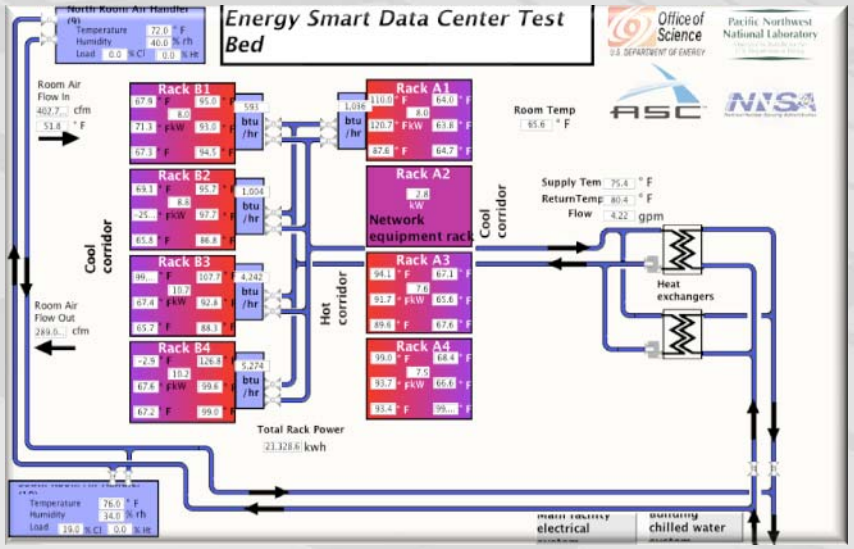


Data Center:

- Power Management Modules
- Power Supply Units
- Voltage Regulators

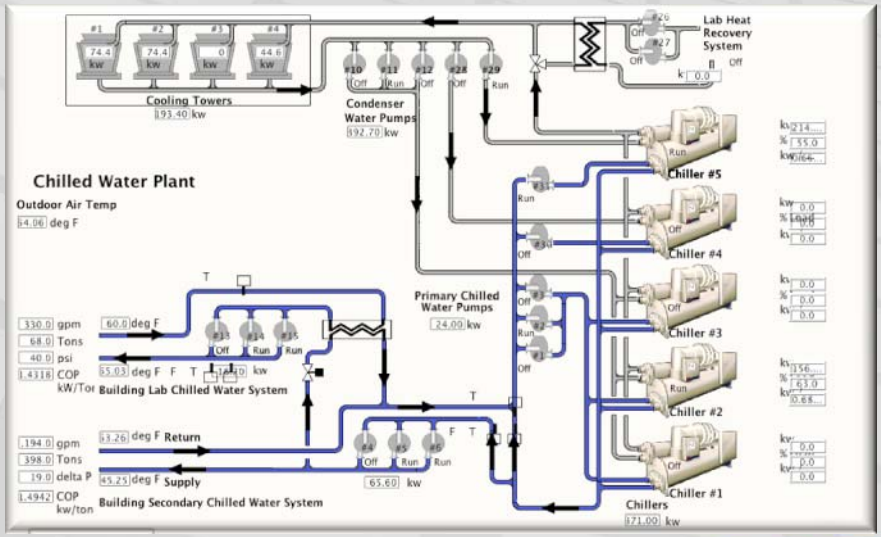
Pacific Northwest
NATIONAL LABORATORY

Contributors to Power Consumption: Cooling Chain



Machine Plant:

- Pumps
- Chillers
- Cooling Towers
- Economizers



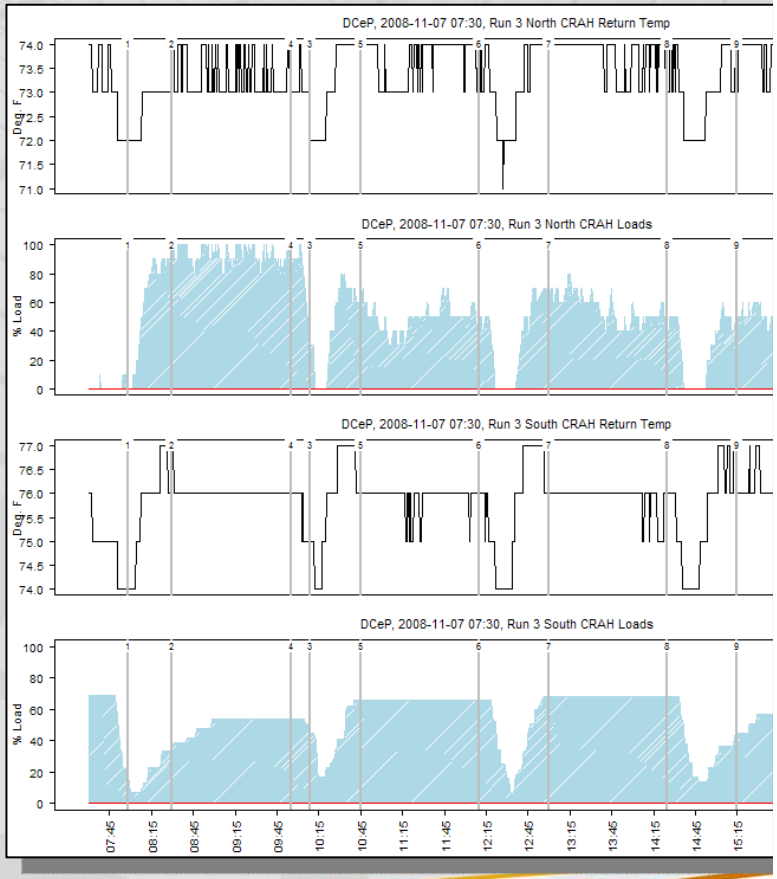
Data Center:

- Air Handlers
- Closely Coupled Cooling Systems
- HVAC

Pacific Northwest
NATIONAL LABORATORY

Stability criteria for captured data in assessment window

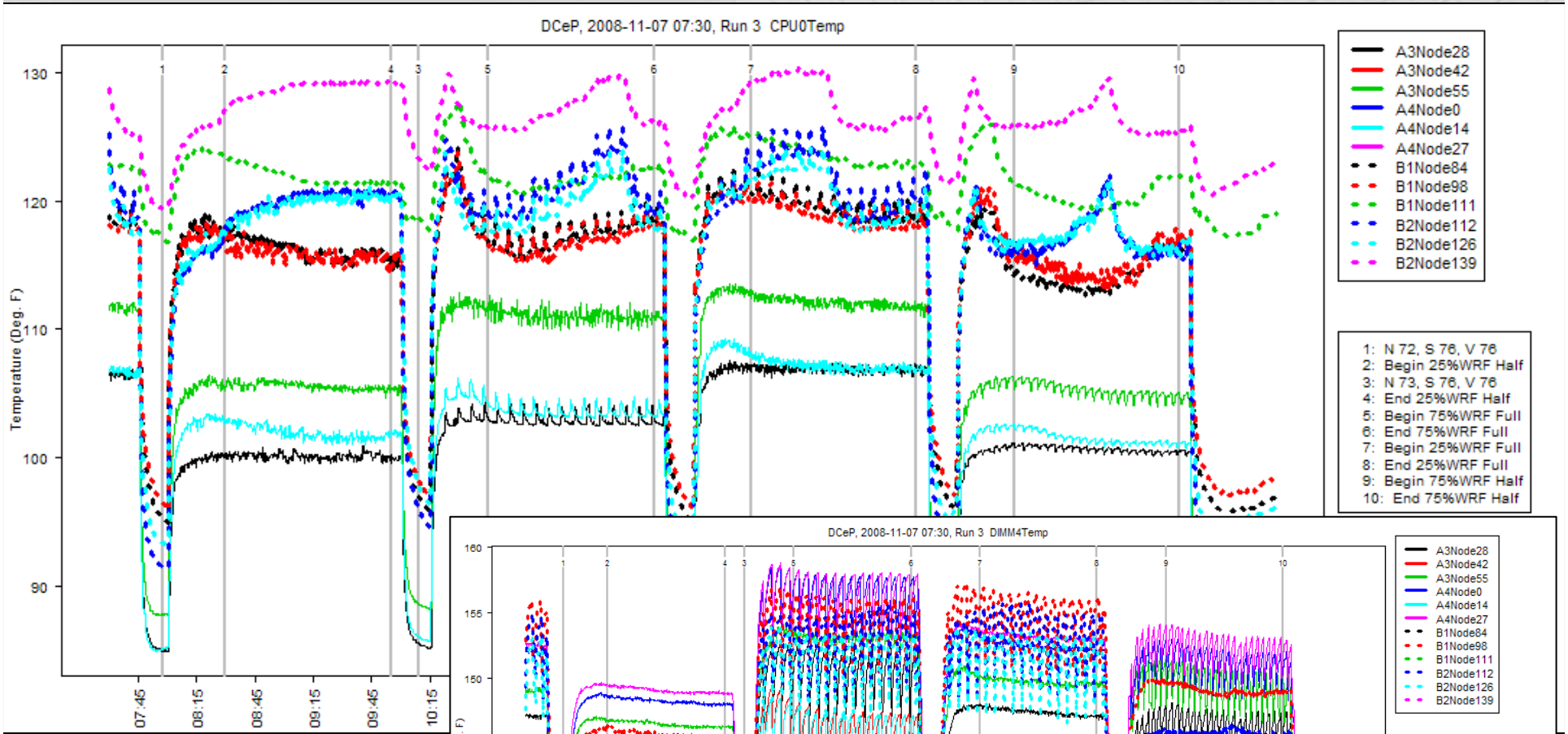
- No saturation of air handlers
  - Never both at 100%
  - Never both at 0%
- Temperatures and humidity within ASHRAE standard recommendations
- Time-intervals excluded from assessment window (larger than system time constants):
  - ramp-up phase
  - clean-up phase
  - ramp-down phase



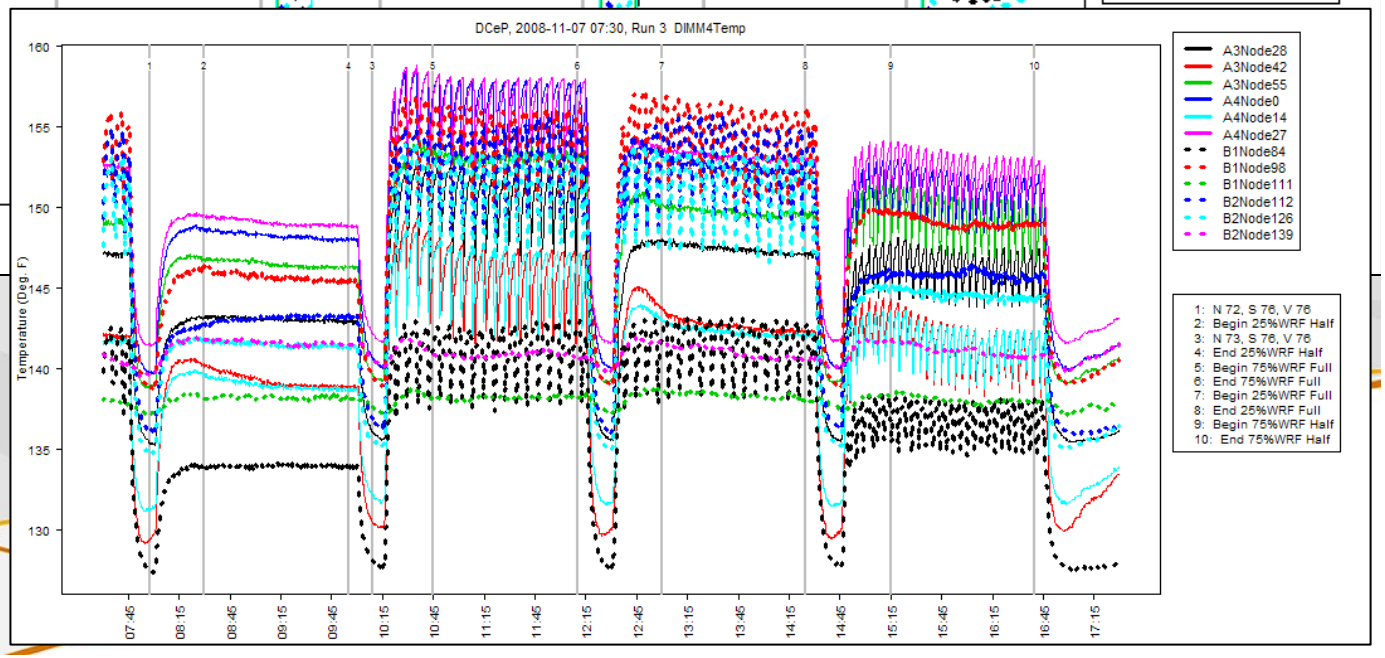
Pacific Northwest
NATIONAL LABORATORY

Chips Temps within Specs

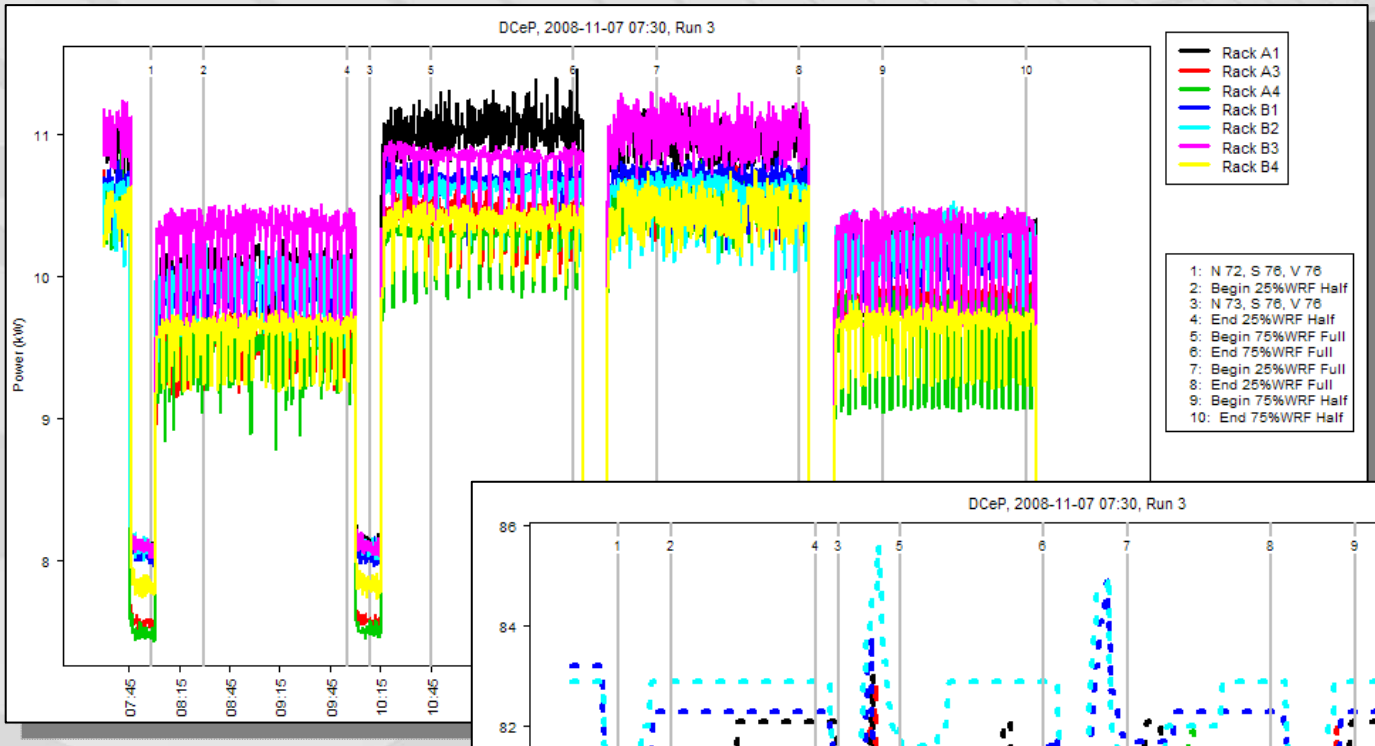
CPU Temps



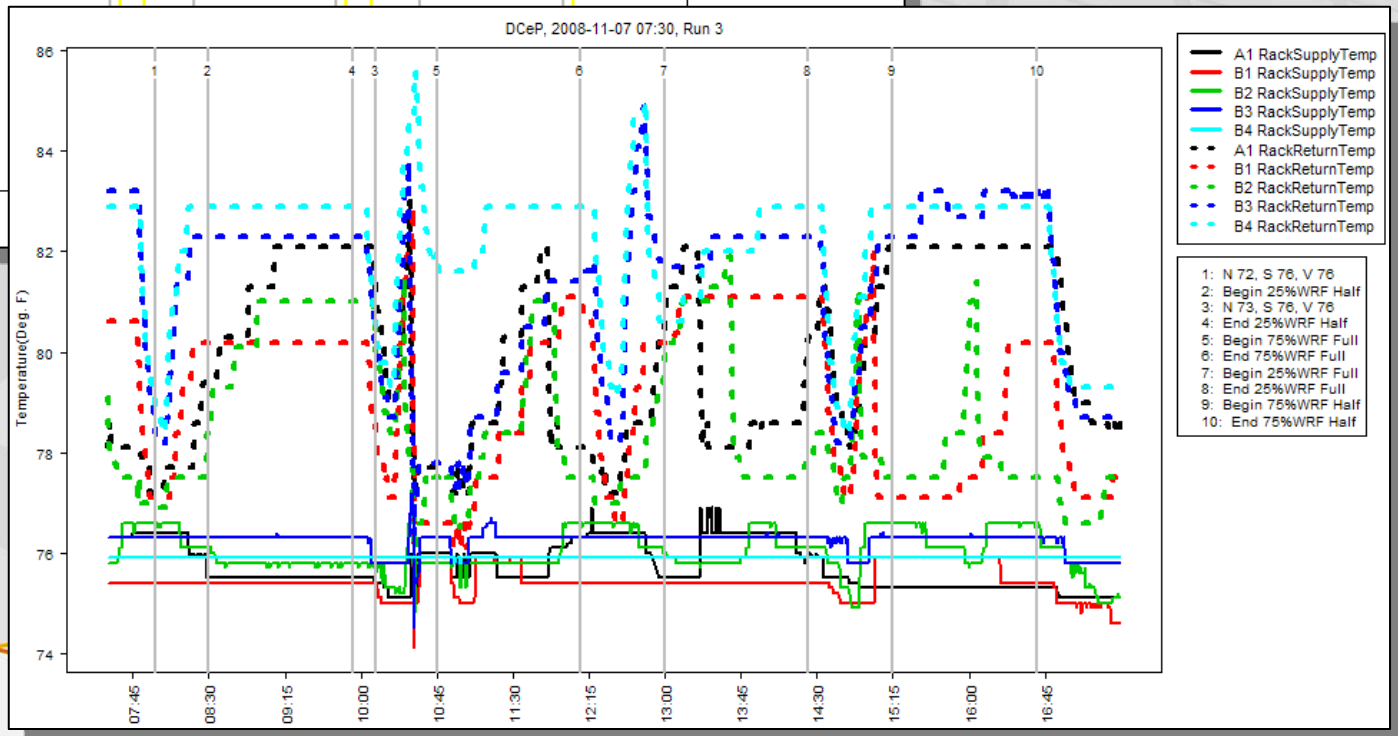
DIMM Temps



Power and Water Temperature Signatures



Rack Power



Water Temperature

Energy Use in kWh

Core WRF/CP2k	Block 1	Block 2	Block 3	Block 4	Treatment Means
Full 75/25	201.5	202.1	200.7	203.2	201.9
Full 25/75	202.5	202.7	202.4	202.5	202.5
Half 75/25	191.5	191.6	192.1	192.0	191.8
Half 25/75	189.2	189.1	189.0	190.1	189.3
Block Means	196.2	196.4	196.1	197.0	

Pacific Northwest
NATIONAL LABORATORY

DCeP

Core WRF/CP2k	Block 1	Block 2	Block 3	Block 4	Treatment Means
Full 75/25	0.532	0.529	0.532	0.526	0.530
Full 25/75	0.542	0.539	0.541	0.543	0.541
Half 75/25	0.439	0.439	0.437	0.432	0.437
Half 25/75	0.445	0.441	0.445	0.443	0.443
Block Means	0.489	0.487	0.489	0.486	

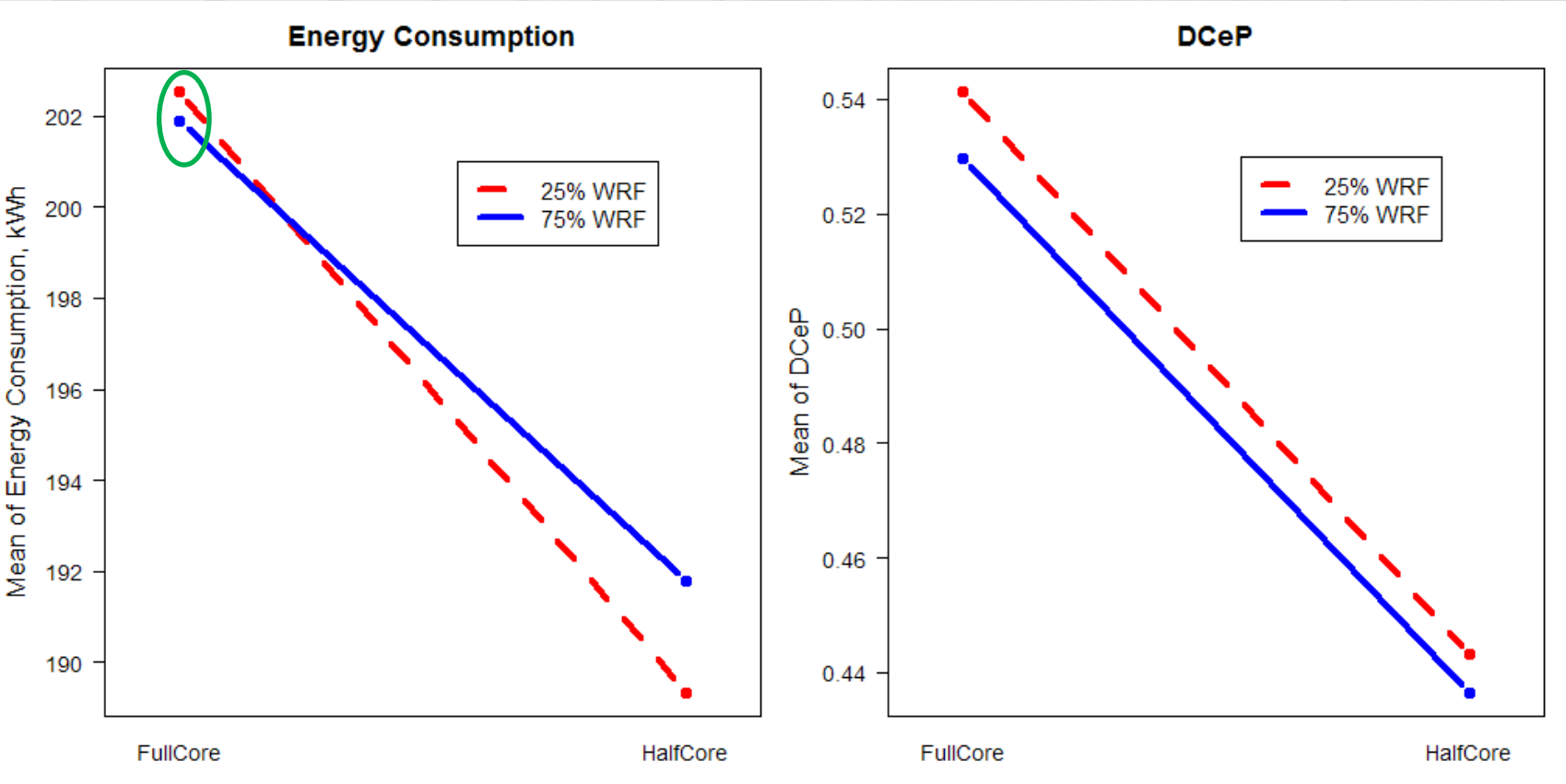
$$\text{Useful Work} = \sum V_i T_i$$

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)

Interaction-Plots
Mean Energy Use, DCeP



All pairs of treatment means differ significantly at the 5% level, except the pair circled in green

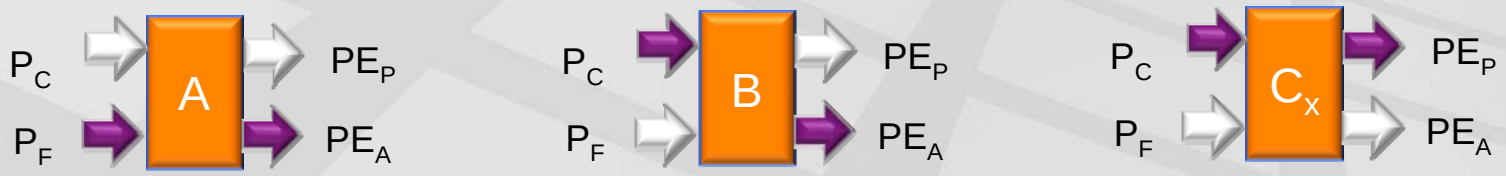
Pacific Northwest
NATIONAL LABORATORY

Summary of Experimental Results

- Full Core implementations use more energy than half core, but are also more efficient
- Given our weighting scheme: Treatments with 25% WRF load are more efficient than 75% (given our chosen weighting scheme)
- DCeP can be used to distinguish between different operational states in a data center and guide load balancing
- Energy Use, DCeP consistent from one experimental period to another (days, nights, and weekends show similar results)
- Results are consistent, reproducible, and show very little variability

Conclusions and Future Work

- Just looking at one aspect, the “two-pole” view, can be misleading
- DCP family of metrics fits nicely with 4-pole model “A” aspect
  - In our case we “expanded” the data center to the whole mixed use facility
- One of the first exercises of real-world DCeP figures with HPC workloads
- We can “contract” DCP to its 4-pole “B” aspect by considering only computer power drawn
- In collaboration with our partners, we are in the process of integrating performance analysis tools into our environmental monitoring and analysis tools to enable the “C” aspect as well
- Power Aware Computing over whole Facility



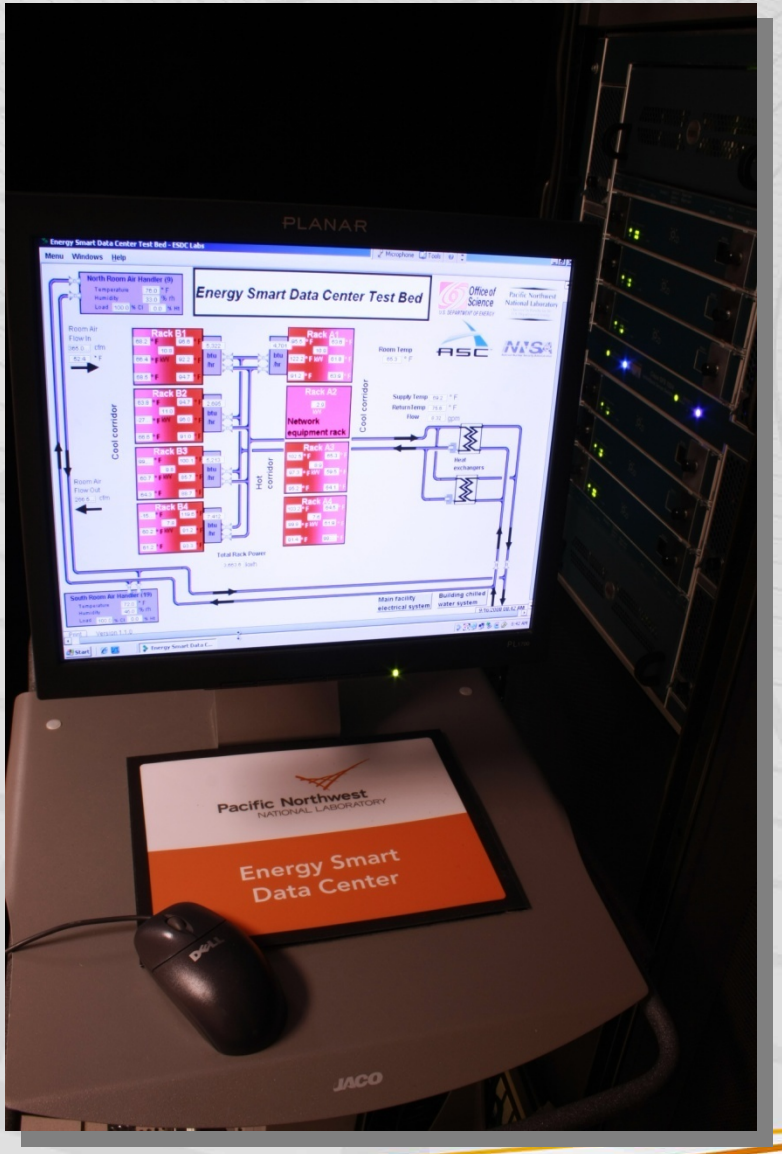
A: true power efficiency (as used in TCO)
B: computer power efficiency (as used in “green” resource allocation)
C_x: figure of merit (as used in unit to unit comparisons)

Pacific Northwest
NATIONAL LABORATORY

Questions?

Fundamental Research of Efficient Datacenters (FRED)

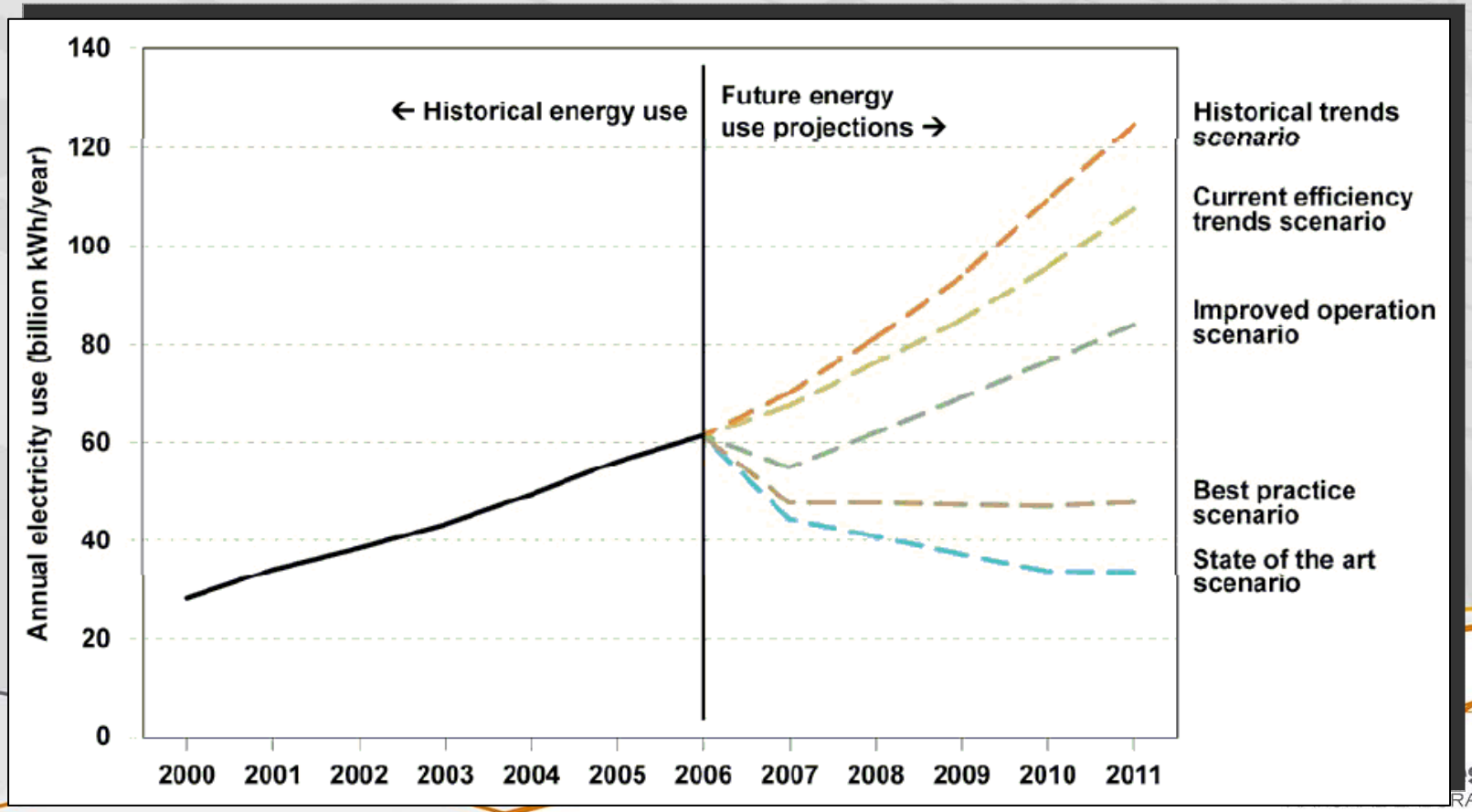
A Graphical Interface to Real Time and Historical testbed data.



Pacific Northwest
NATIONAL LABORATORY

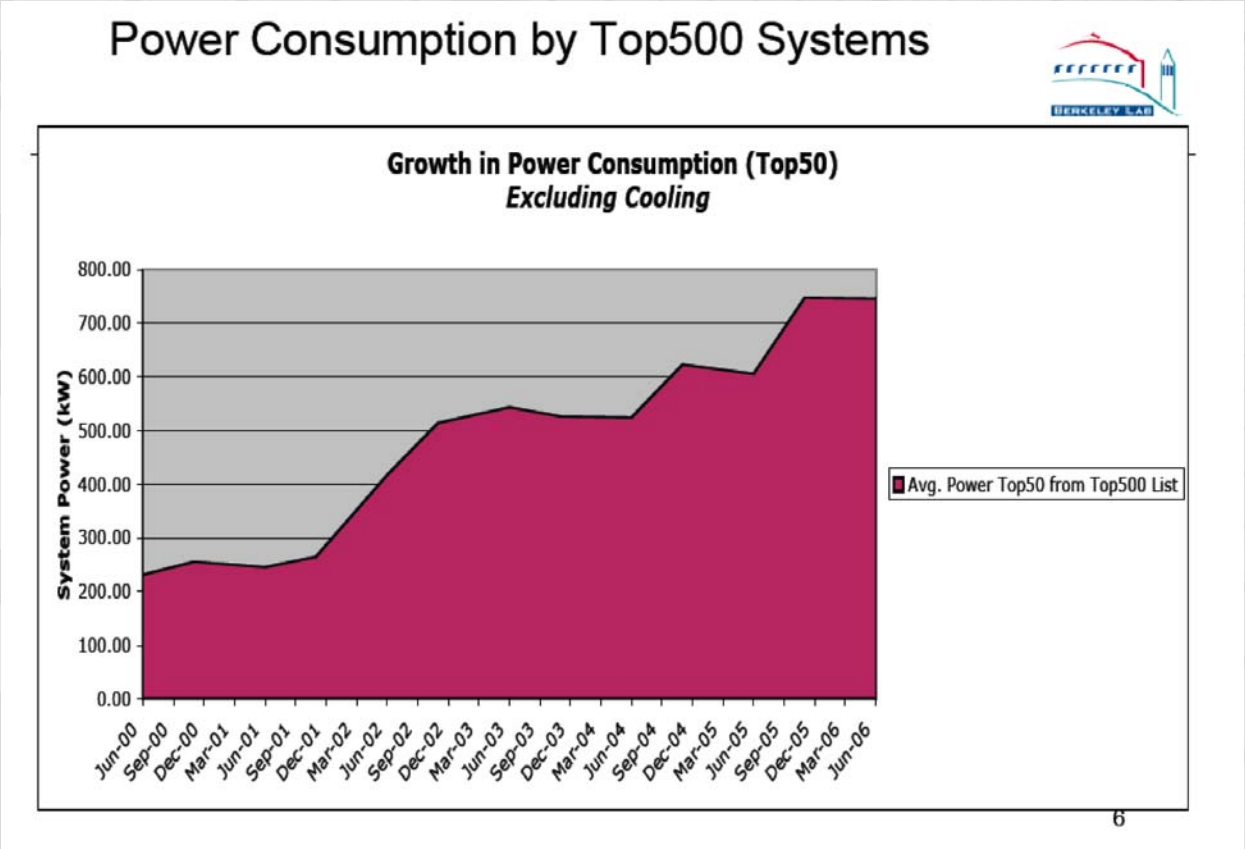
National Challenge

- Current efficiency trends estimate energy use in datacenters 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 current 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

Top500 Power consumption



Source: Berkely Labs

Pacific Northwest
NATIONAL LABORATORY