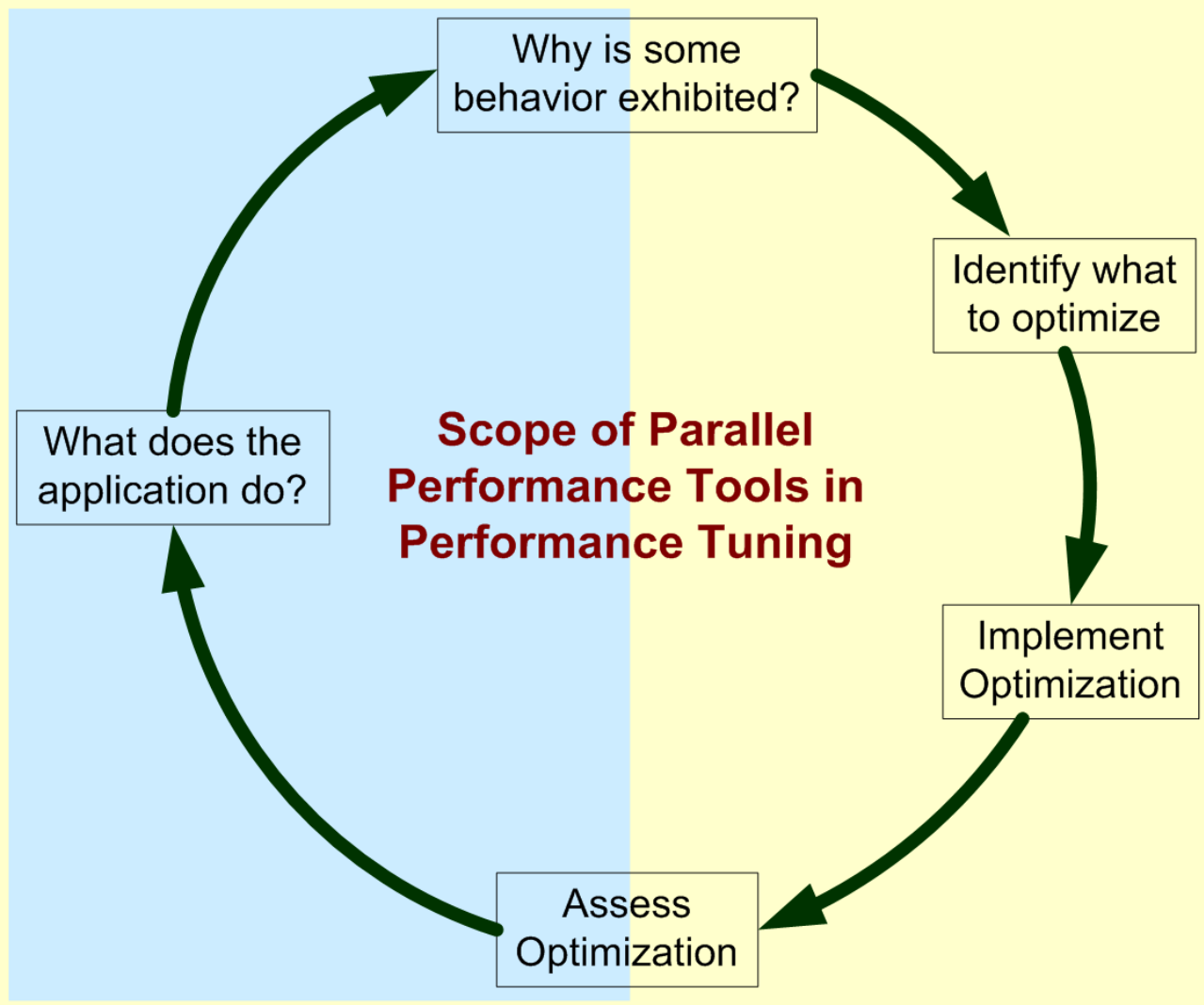
Power and Cooling Aware Parallel Performance Diagnosis

Karen L. Karavanic
Portland State University

Sriram Krishnamoorthy, Andrés Márquez
Pacific Northwest National Laboratory
Rashawn Knapp
Portland State University

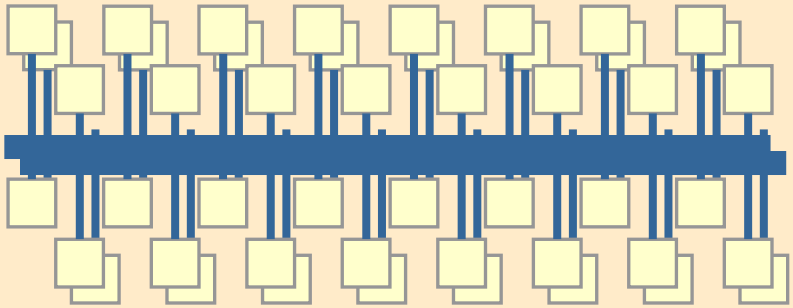
Portland State UNIVERSITY

HPC Performance Tuning and the Scope of Parallel Performance Tools

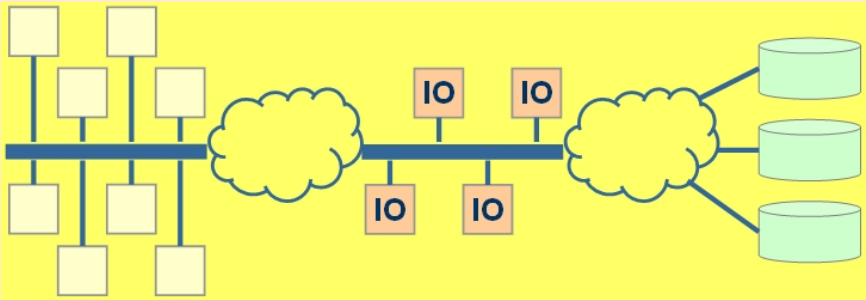


Limitations of Application Focused Performance Analysis

Faulty Hardware



OS Interference



File System

Traditional methods misdiagnose or do not detect an important class of performance problems

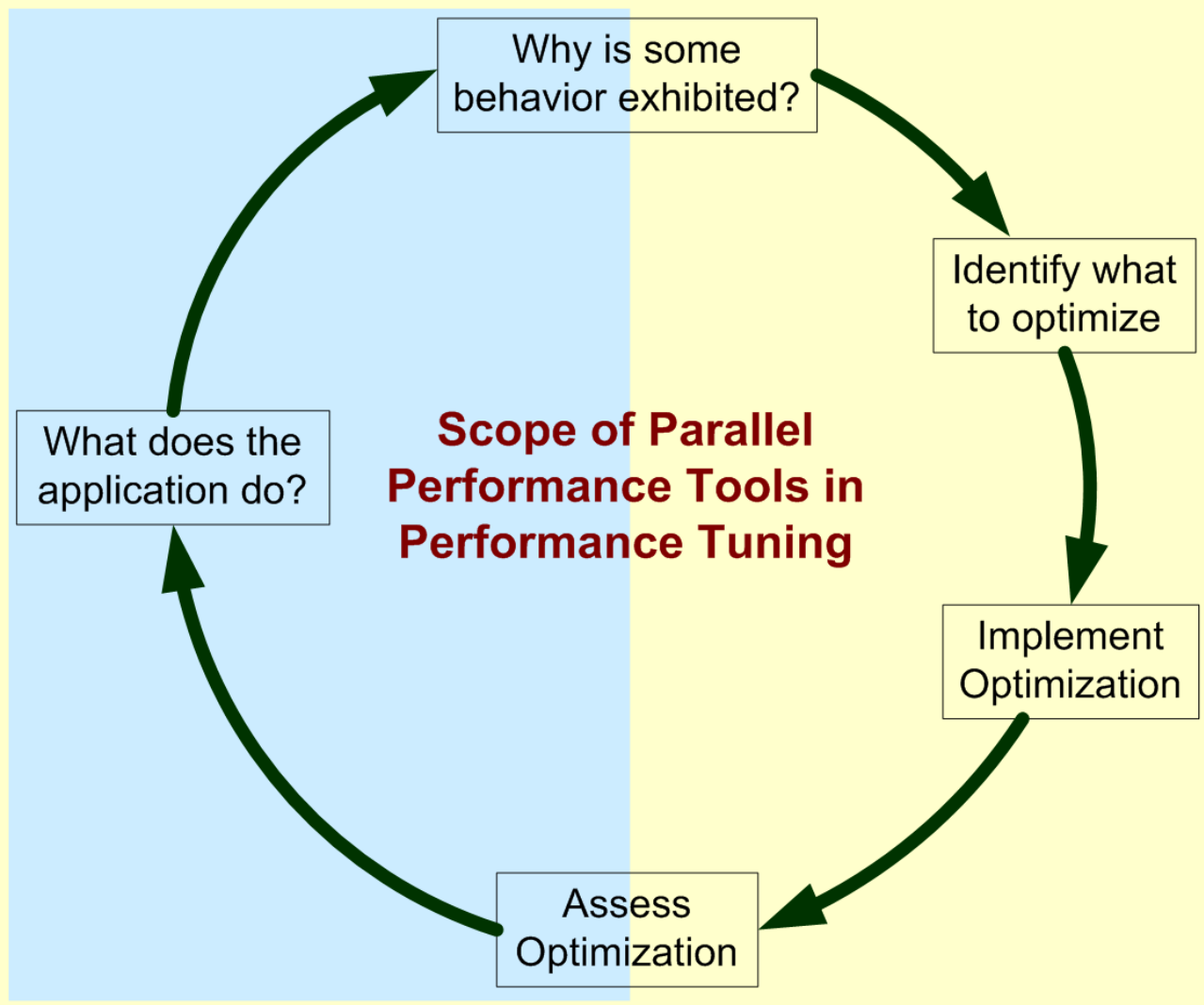
Examples:

- OS Noise
- File System Contention
- Processor throttling
- Inefficient Power / Cooling

Characterized by:

- Difficult diagnosis
- Application metrics (function profile, wait times) are not sufficient

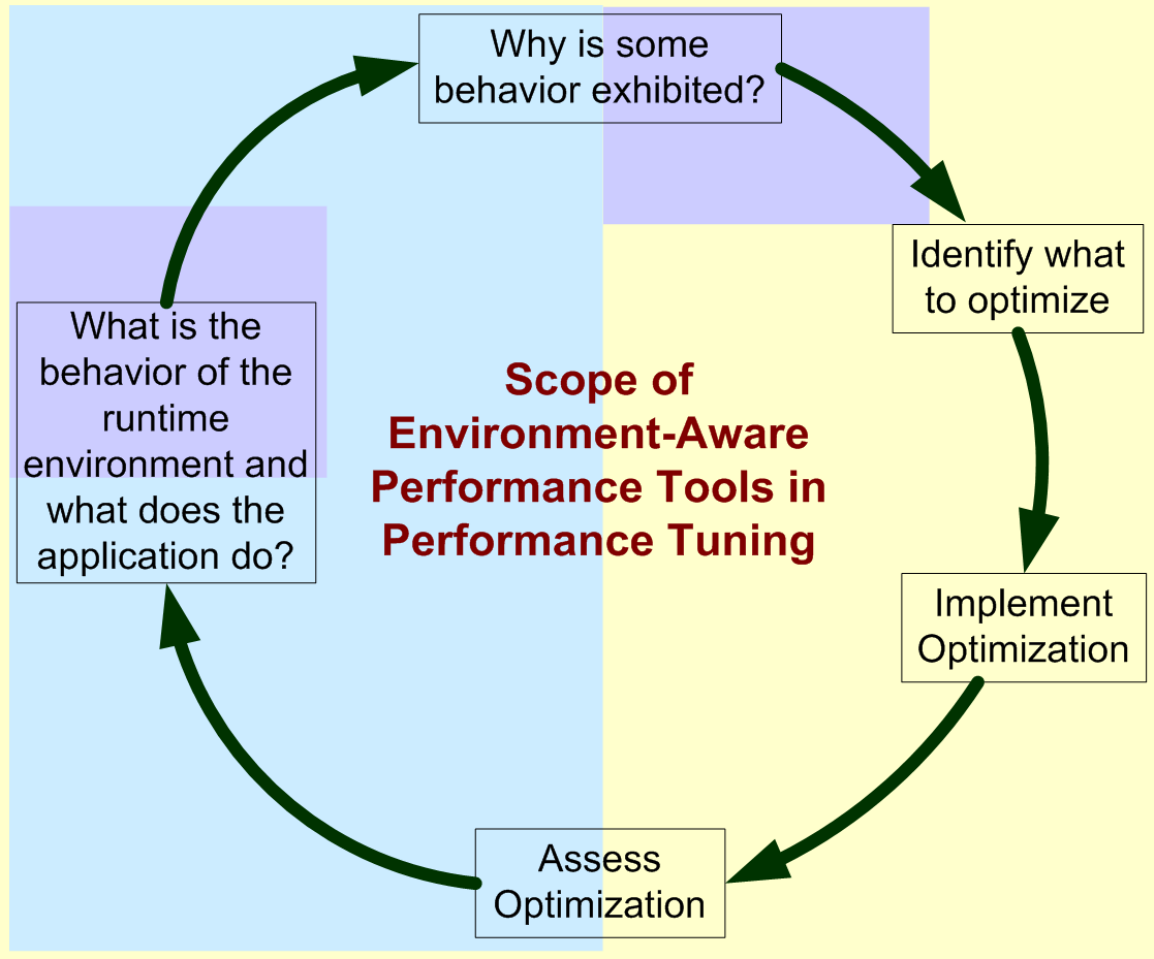
HPC Performance Tuning and the Scope of Parallel Performance Tools



Our Approach:

Environment Aware Performance Diagnosis (EAPD)

Online and *automated* methods



Application and *system* level and *room* level metrics

Project Overview

Goal: Develop performance diagnosis tool that integrates performance measurements from all levels:

- application: function profile, computation:communication
- system: file system contention, processor throttling,
- room: power consumption, cooling demand, node assignment

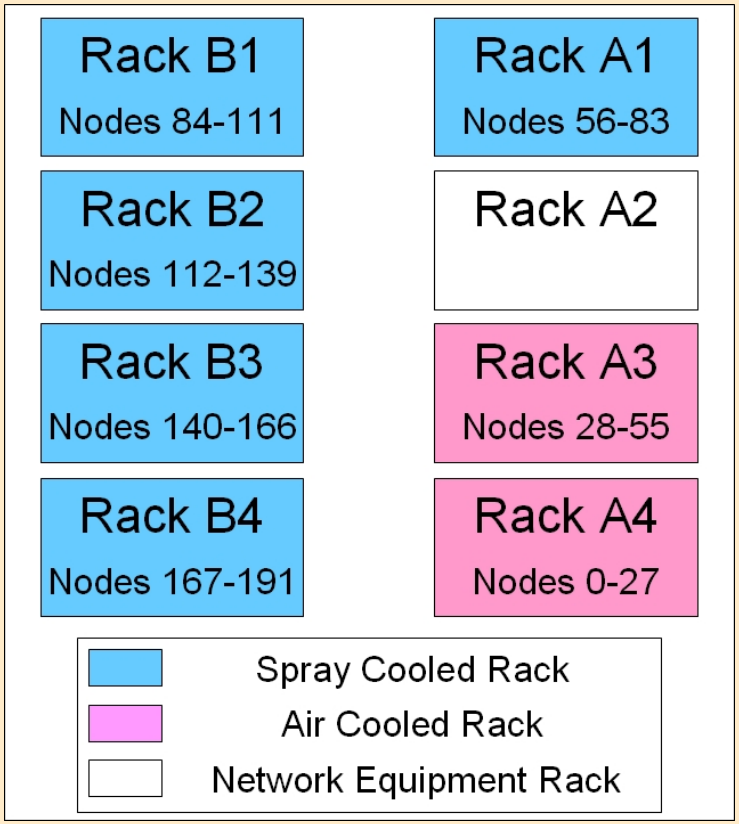
Approach:

Conduct case studies with full data collection to develop new post-mortem methods

Integrate system, power and cooling data into existing performance database, PerfTrack

Develop useful high level metrics

The Energy Smart Data Center (ESDC) at PNNL



Layout of NW-ICE's racks and nodes

NW-ICE:

- Linux cluster
- 192 nodes, each with dual Quad-Core Intel Xeon E5345 processors and 16 GB memory
- Spray-cooled and air-cooled racks

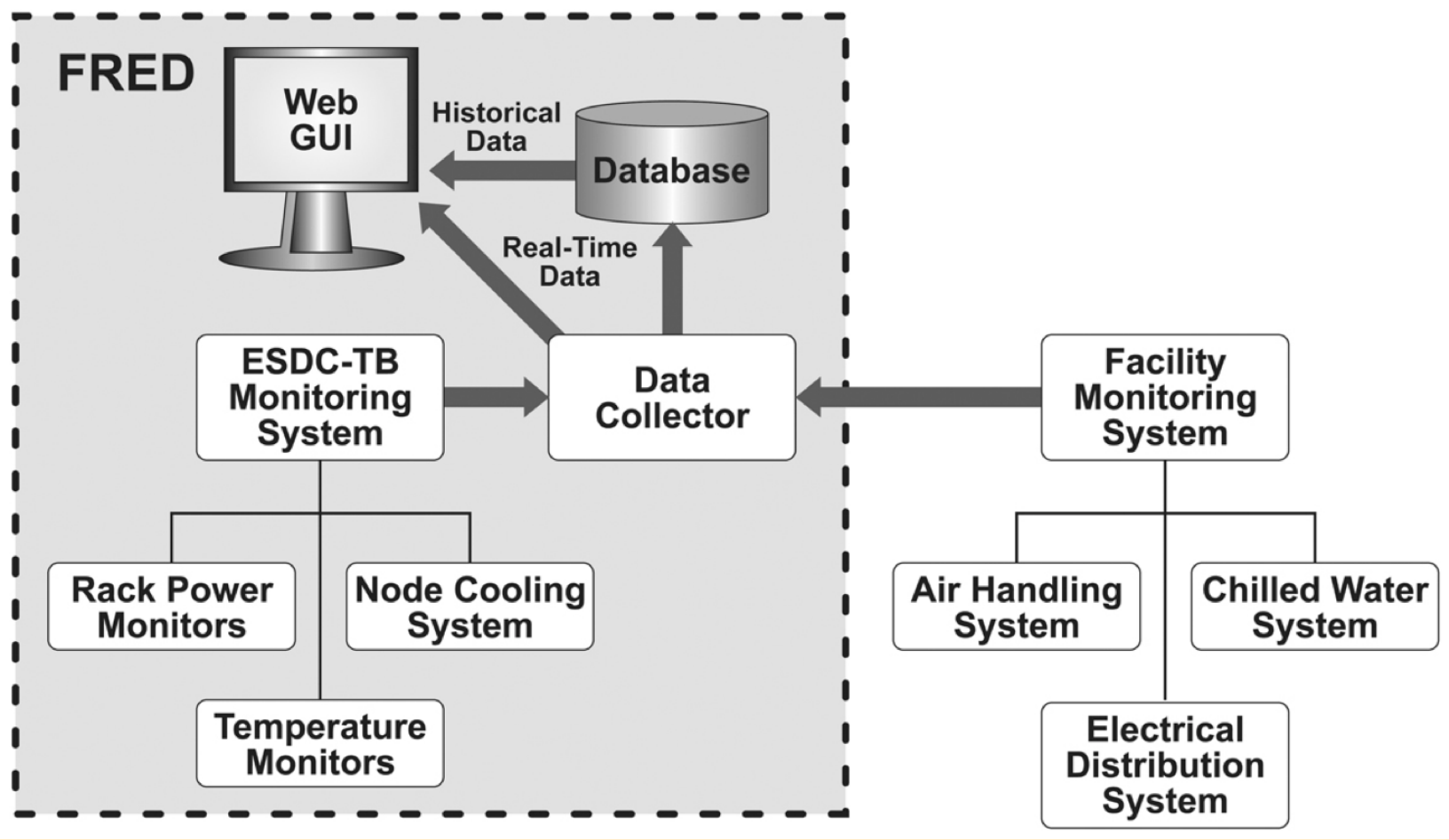
Room environment data:

- Temperature – CPUs, DIMMs, Racks, Room Air
- Power – Racks, Air conditioners, Chillers, Pumps

Fundamental Research in Energy-efficient Datacenters (FRED):

- Collects sensor data
- Stores data in a round-robin, recent history database and a permanent storage database
- Developed at PNNL

PNNL's Fundamental Research in Energy Efficient Data Centers (FRED)



Workload Placement Experiments

Application

- NWChem DFT calculation of $Si_{28}O_{67}H_{30}$
- Configure for large MPI process runs

High Density Experiments

- 8 MPI processes per node on everynode in a *single rack*

Low Density Experiments

- Spread the processes over *four racks*
- Several variations

High Density Experiments

Rack	Processes Per Node	Total Processes	NW-ICE Node IDs	Qty. Nodes	Avg. Runtime
A1	8	224	56-83	28	42.3 min.
B1	8	224	84-111	28	42.2 min.
B2	8	216	112-119,121-139*	27	38.2 min.
B3	8	216	140-166	27	40.2 min.
B4	8	184	167-187,189-190*	23	40.0 min.

**Nodes 120,188,191 unavailable*

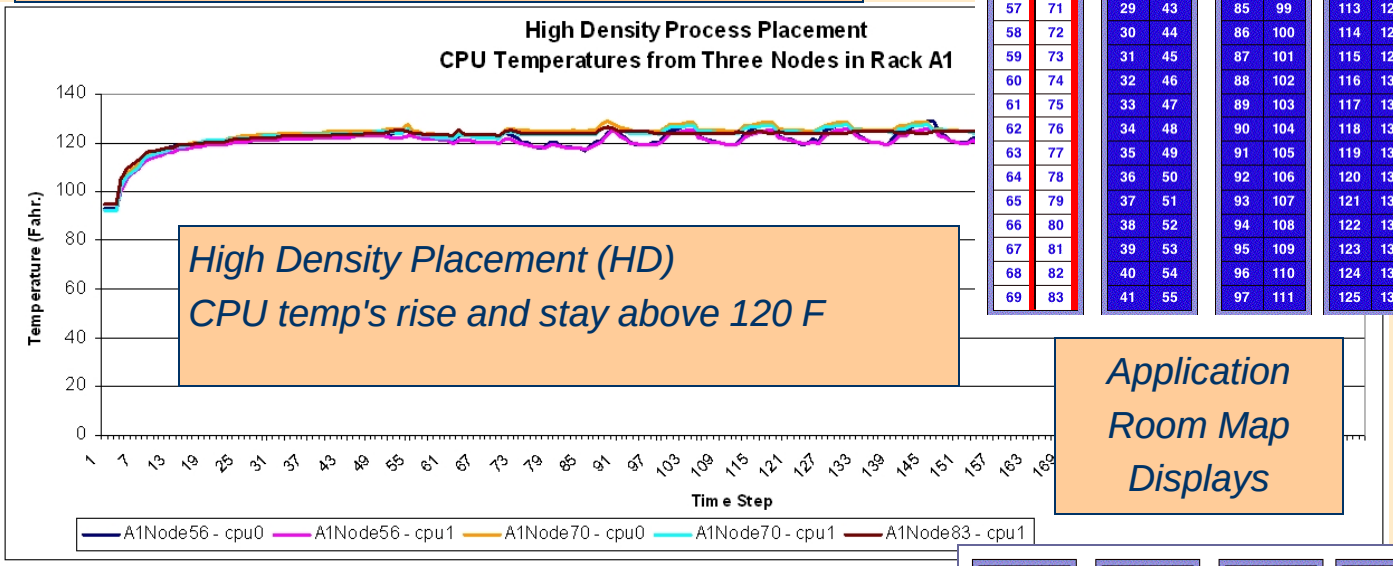
Low Density Experiments

Experiment	Racks	Processes Per Node	Total Processes	Qty. NW-ICE Nodes	Avg. Runtime
Even Distribution*	A1, B1, B2, B3	2	220	110	30.4 min.
Bottom of Rack	A1, B1, B2, B3	4	224	56	32.5 min.
Every Other	A1, B1, B2, B3	4	224	56	32.7 min.
Middle of Rack	A1, B1, B2, B3	4	224	56	32.4 min.
Top of Rack	A1, B1, B2, B3	4	224	56	32.6 min.

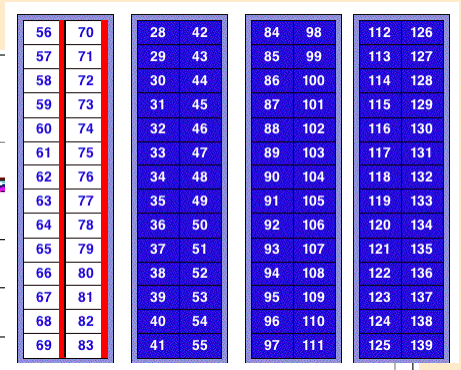
**Node 117 in rack B2 unavailable*

Integrating Power/Cooling Into Application Performance Analysis: Example

224 processes on 1 rack (8 proc.s per node)



High Density Placement (HD)
CPU temp's rise and stay above 120 F

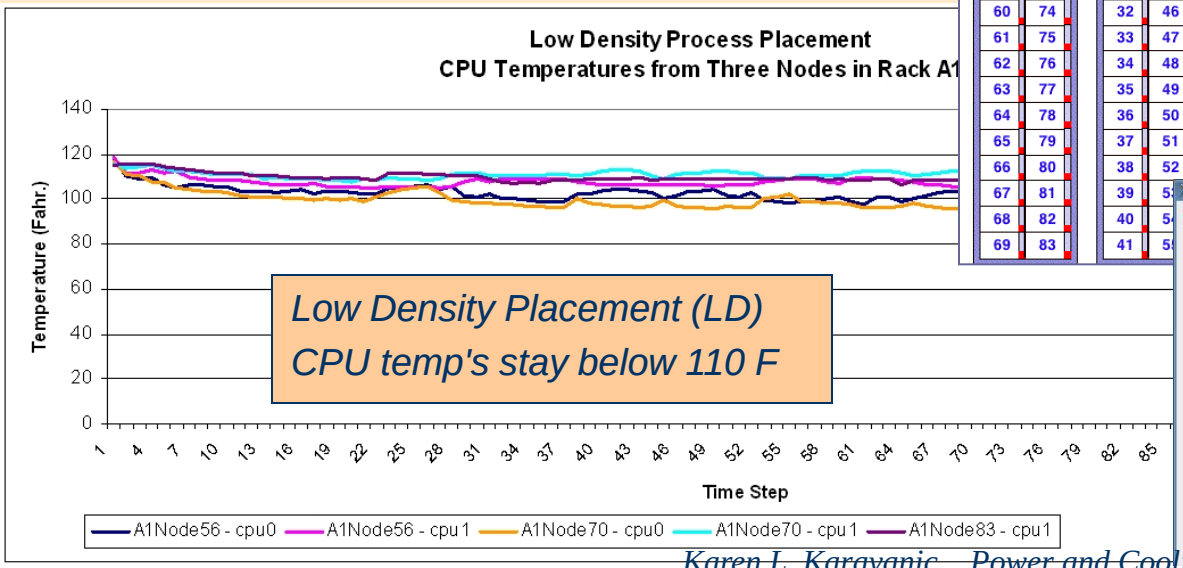


Application Room Map Displays

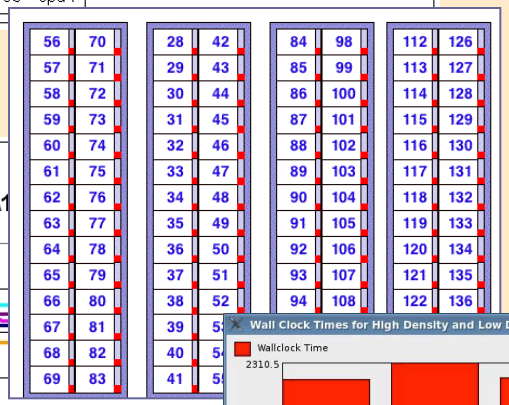
Total KWH

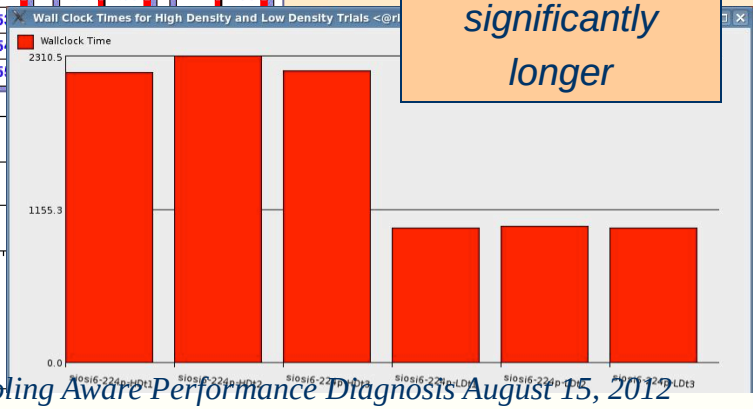
Rack	HD	LD
A1	6.397	2.448
A2	1.721	0.789
A3		2.299
B1		2.420
B2		2.443
ALL	8.119	10.401

224 processes on 4 racks (2 proc.s per node)



Low Density Placement (LD)
CPU temp's stay below 110 F

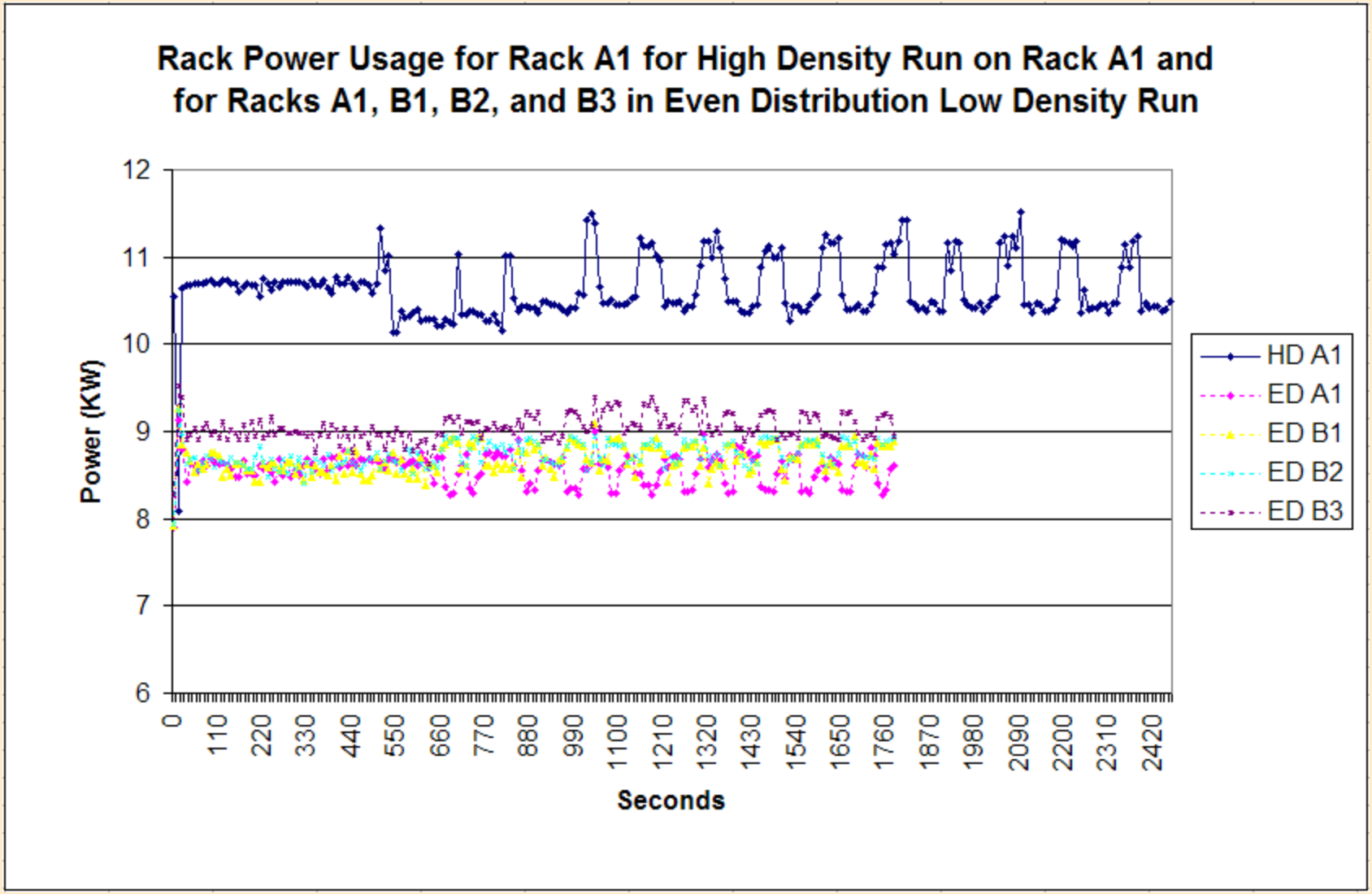




PerfTrack visi shows HD times significantly longer

Workload Placement Experiments – Performance and Room Environment

Rack Power Usage: High Density on A1 compared to *Even Distribution* Low Density



How to measure Energy Efficiency?

Consumption – Single Rack

$$I_i(RackEnergy) = 100 * ((R_i^{meas} - R_i^{\min}) / (R_i^{\max} - R_i^{\min})$$

Consumption - Total

$$I(TotalEnergy) = 100 * ((R^{meas} - R^{\min}) / (R^{\max} - R^{\min})$$

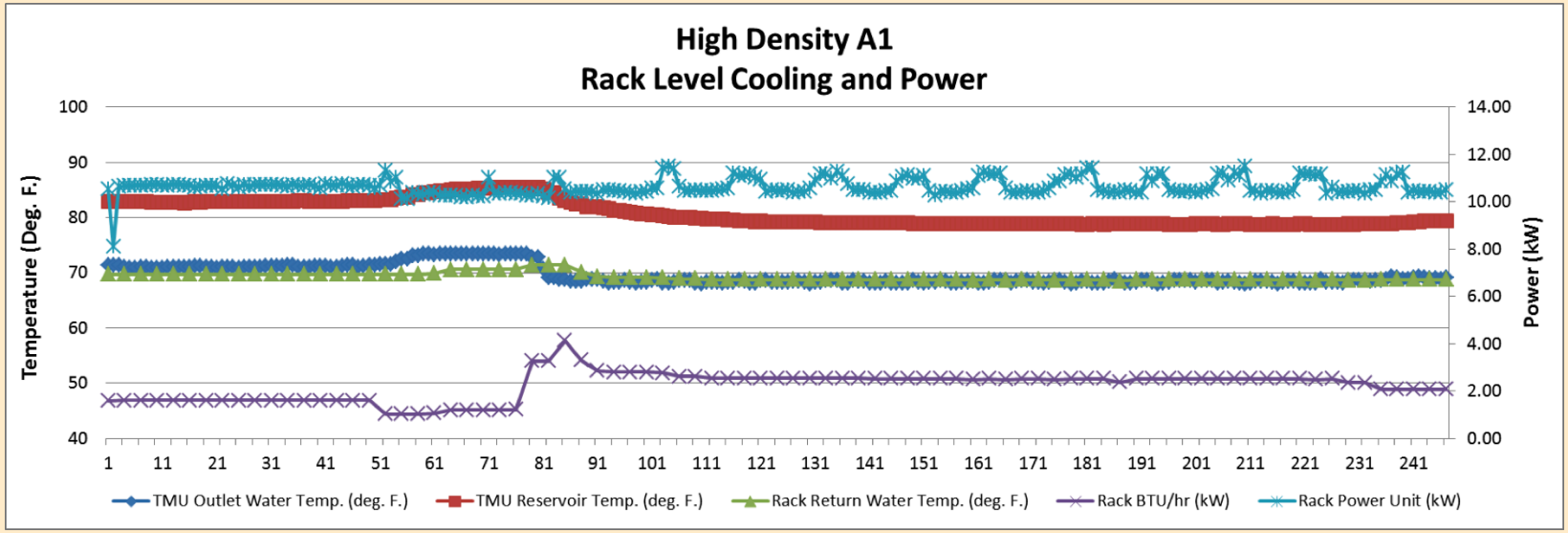
where

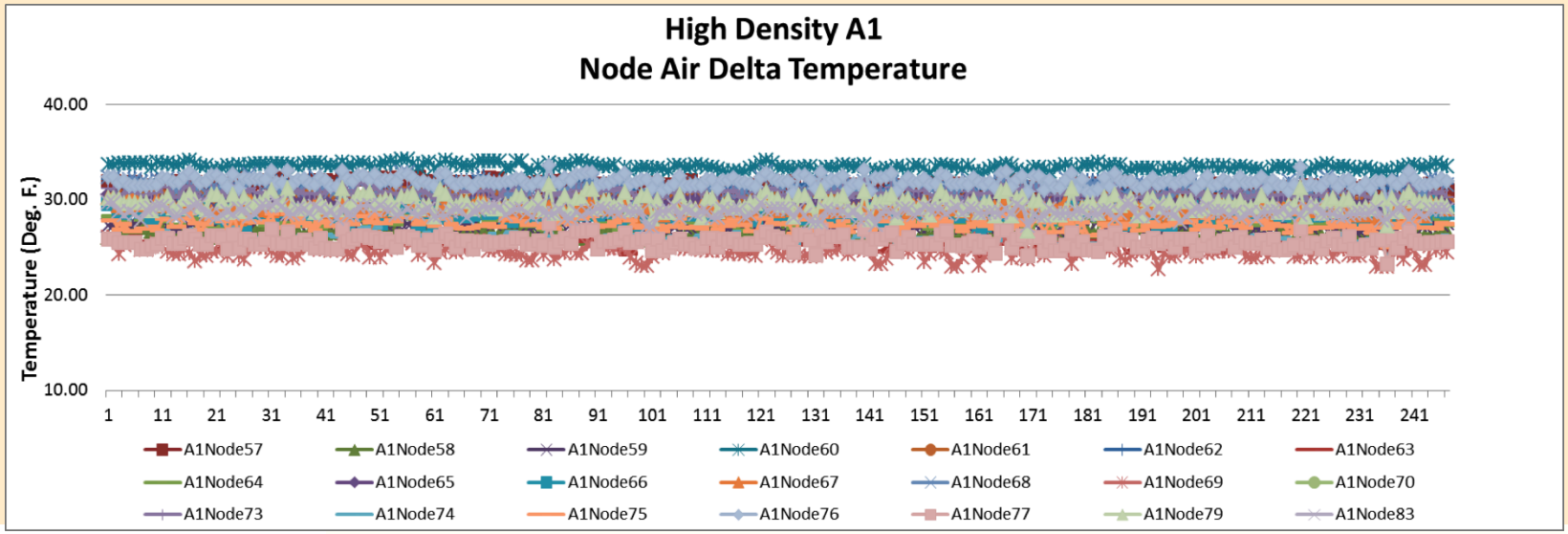
$$R^{meas} = \sum_{i=1}^{n} R_i^{meas}$$

$$R^{\min} = \sum_{i=1}^{n} R_i^{\min}, \quad R^{\max} = \sum_{i=1}^{n} R_i^{\max}$$

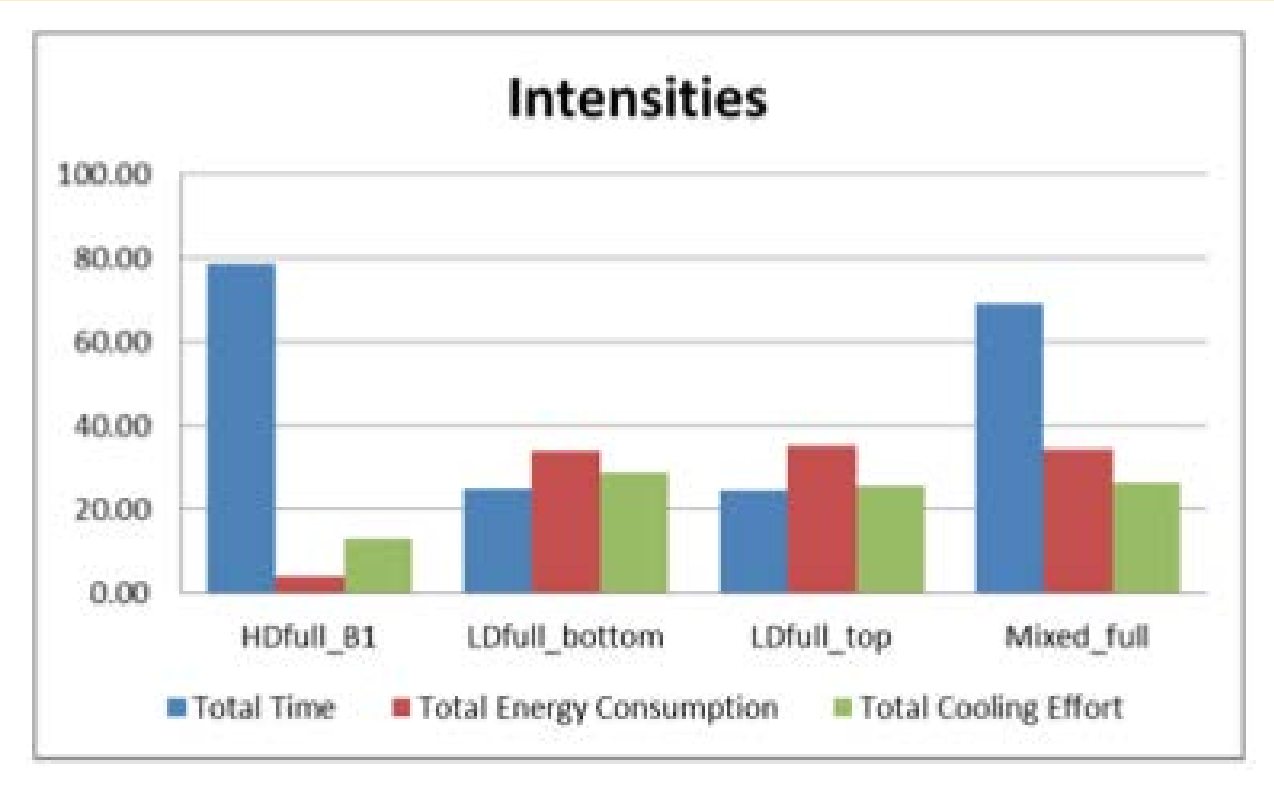
and n = *number of participating racks*

How to Measure Cooling Demand?





Intensities: High Level Metrics for Diagnosis



Goal: Normalize the values to rank the *severity* of the measurements
(Inspired by Kojak/Scalasca bottleneck severity approach)

The PerfTrack Performance Database

A tool for storing, exploring, and analyzing application performance data

- Collect and store description of each platform
- Collect and store description of each build and run of an application
- Integrate DBMS into a performance analysis tool
- Store a wide variety of performance data
  - Data from different measurement tools:

    TAU, Speedshop, HW counters, native application performance measurements, etc.
- GUI for data navigation and querying
- Shield tool user from DBMS internals

Originally developed in collaboration with LLNL (May)

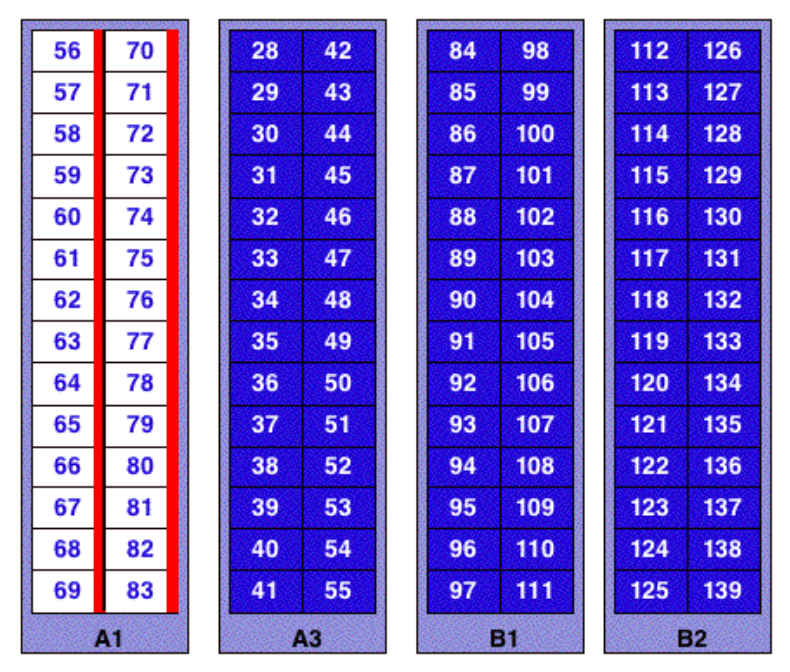
Scalability issues: Room Data File Sizes

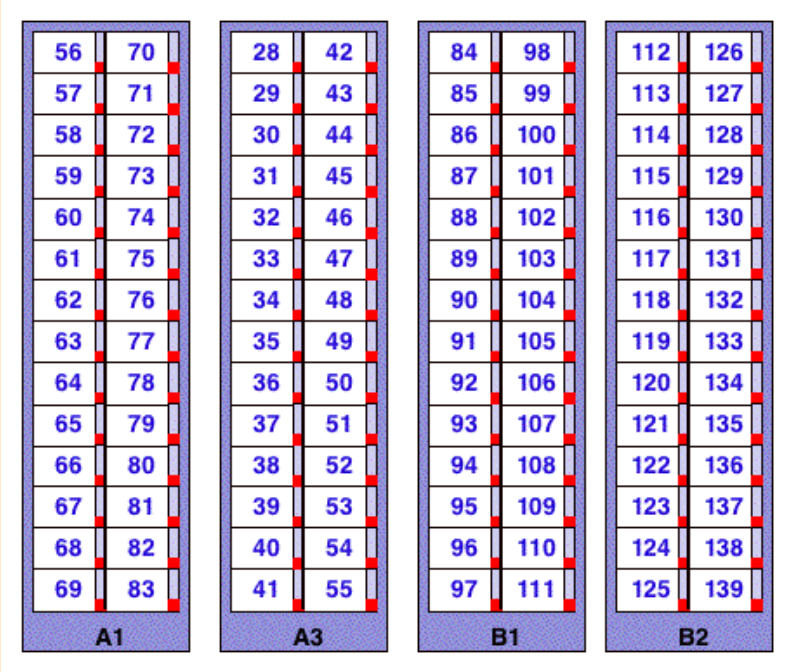
- HD experiments (approx. 60 minutes):
  - As csv (three time blocks) - about 48MB
  - As PTDF (three time blocks) -- about 748MB
- LD experiments (approx. 40 minutes):
  - As csv (two time blocks) - about 59 MB
  - As PTDF (two blocks) about 904.8 MB
- Selected measurement points only
- Temperature and Power only
- 24 hours LD: approx. 32,573 MB or 32.6 GB

Observations

- Continuous Measurement of room and system metrics is necessary for correct performance diagnosis in certain cases
- Data Volume must be addressed for scalability
- Room, System and Application data are linked by timestamp – request application run and get whole picture
- Our Current Focus:
  - Integration of system metrics (interrupts, cache misses) Adhikari, "Measuring system level performance to identify bottlenecks in high end systems," PSU CS Department Masters Thesis, 2010.
  - Implementing GUI enhancements and Clustering into PerfTrack
  - RackTrack !

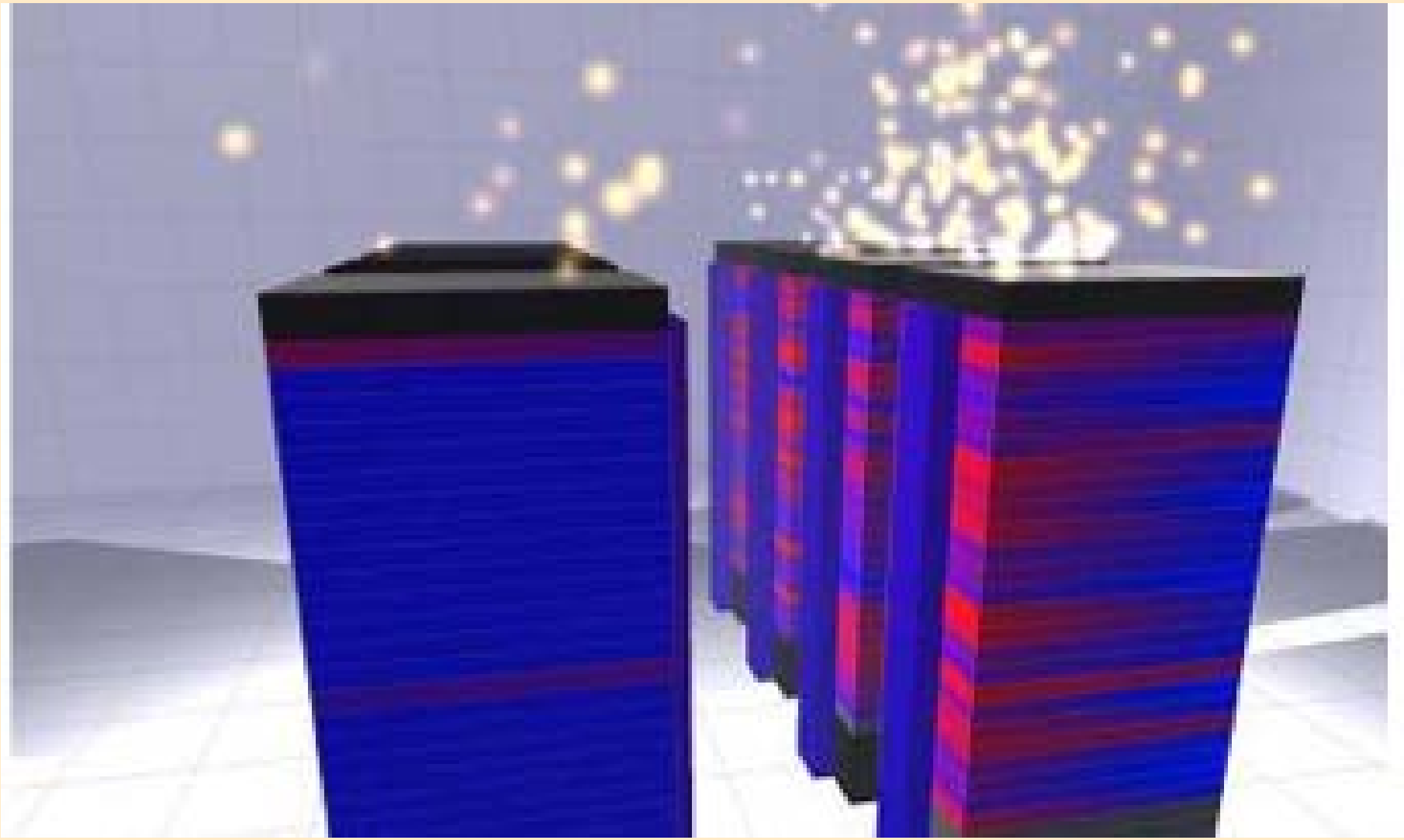
New Visualizations for Integrated Data





RackTrack

Michael Barger, Benjamin Brown, Charles Capps, Gaby Haddock, Tim Reilly



Karen L. Karavanic

karavan@cs.pdx.edu

http://www.cs.pdx.edu/~karavan

Acknowledgments

This work was supported in part by funding from PNNL/Battelle, the Portland State University Center for Sustainable Processes and Practices, the James F. and Marion L. Miller Foundation, UC/LLNL and the DOE Office of Science.

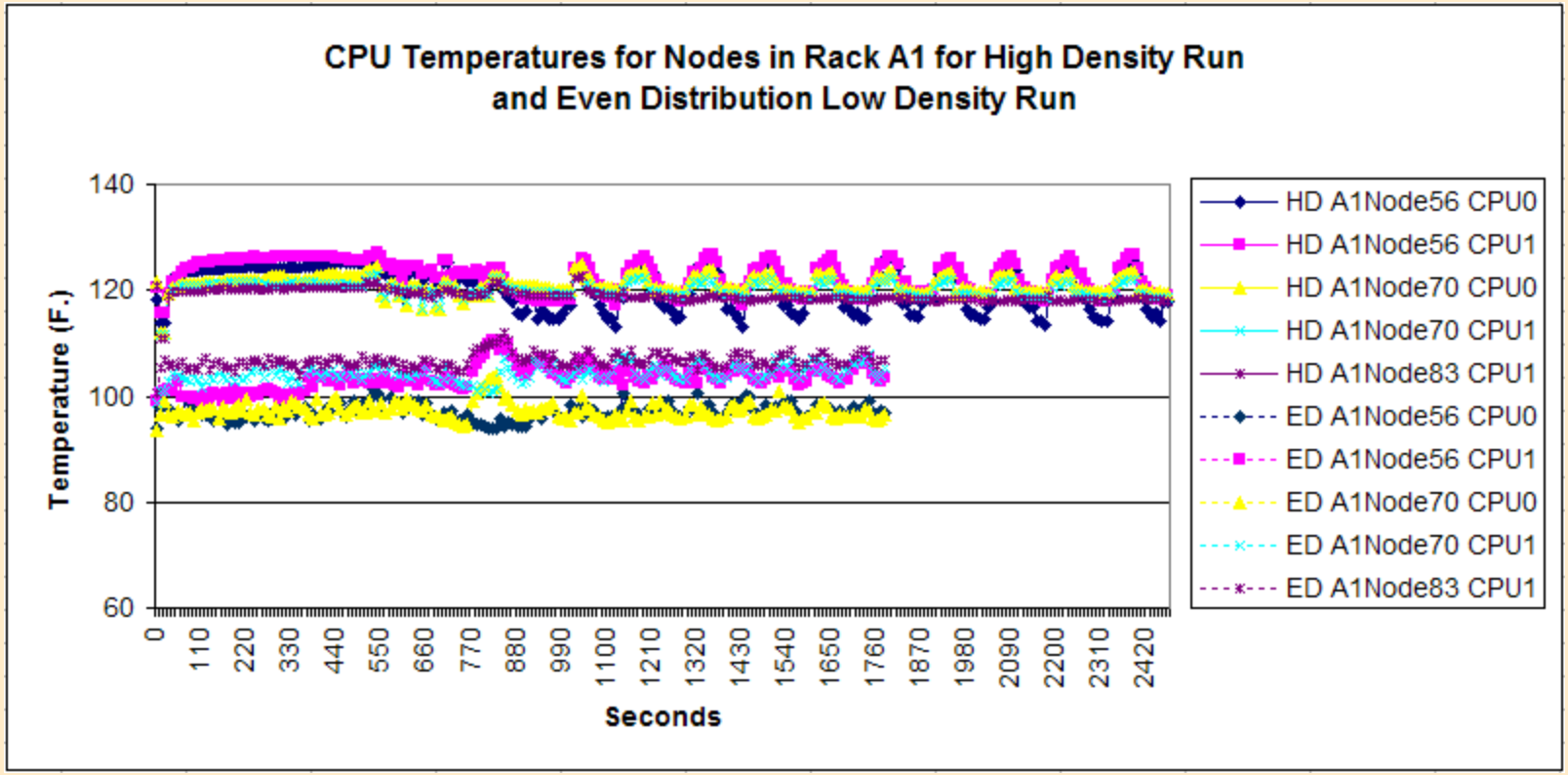
J. Evans, S. Gupta, K. Karavanic, A. Marquez, and G. Varsamopoulos, "Evaluating Performance, Power, and Cooling in High Performance Computing (HPC) Data Centers," Ahmad and Ranka, ed.s, Handbook of Energy-Aware and Green Computing, Chapman and Hall/CRC 2012.

Rashawn L. Knapp, Karen L. Karavanic, Sriram Krishnamoorthy, and Andres Marquez, "Power- and Cooling- Aware Parallel Performance Diagnosis," Parallel and Distributed Computing and Systems (PDCS 2011), Dallas, Texas, December 2011.

Rashawn L. Knapp, Karen L. Karavanic, and Andres Marquez, "Integrating Power and Cooling Data into Parallel Performance Analysis," The Second International Workshop on Green Computing (GreenCom 2010), San Diego, CA September 13, 2010.

Rashawn L. Knapp and Karen L. Karavanic, "Correct Diagnosis of Parallel Performance Problems Caused by the Runtime Environment," OSIHPA 2006, September 17, 2006, Seattle, WA.

Karen L. Karavanic, John May, Kathryn Mohror, Brian Miller, Kevin Huck, Rashawn Knapp, Brian Pugh, "Integrating Database Technology with Comparison-based Parallel Performance Diagnosis: The PerfTrack Performance Experiment Management Tool," SC2005, November 2005, Seattle, WA.

Workload Placement Experiments – Performance and Room Environment

CPU Temperatures: High Density on A1 compared to Even Distribution Low Density