Enabling Advanced Operational Analysis Through Multi-Subsystem Data Integration on Trinity

J. Brandt, D. DeBonis, A. Gentile, J. Lujan, C. Martin,
D. Martinez, S. Olivier, K. Pedretti, N. Taerat, R. Velarde

Cray Users Group
April 26-30, 2015

Motivation

New Mexico Alliance for Computing at Extreme Scale (ACES)

Large scale HPC platforms are continuing to push the limits of data center power and cooling infrastructure.

In order to maximize the efficiency of modern large scale platforms, a management approach that tightly integrates all information both internal and external to the platforms is essential.

Facilities for Trinity and Trinitite at LANL

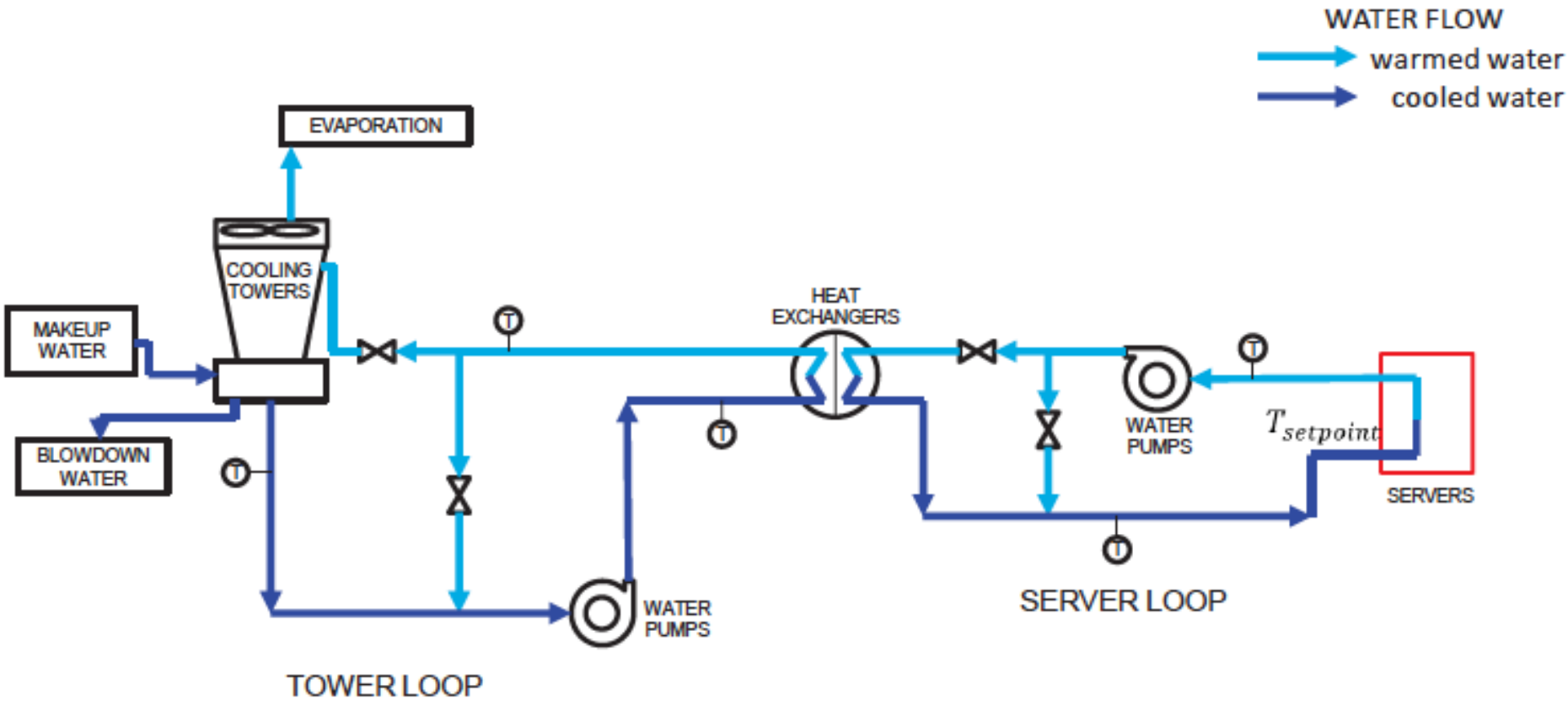
Nicholas C. Metropolis Center for Modeling and Simulation Strategic Computing Complex (SCC) – 2002

- 43,500 sq ft computer room floor
- Data center electrical capacity 19.2 MW
- Data center air cooling capacity 21 MW
- Data center water cooling capacity 15 MW

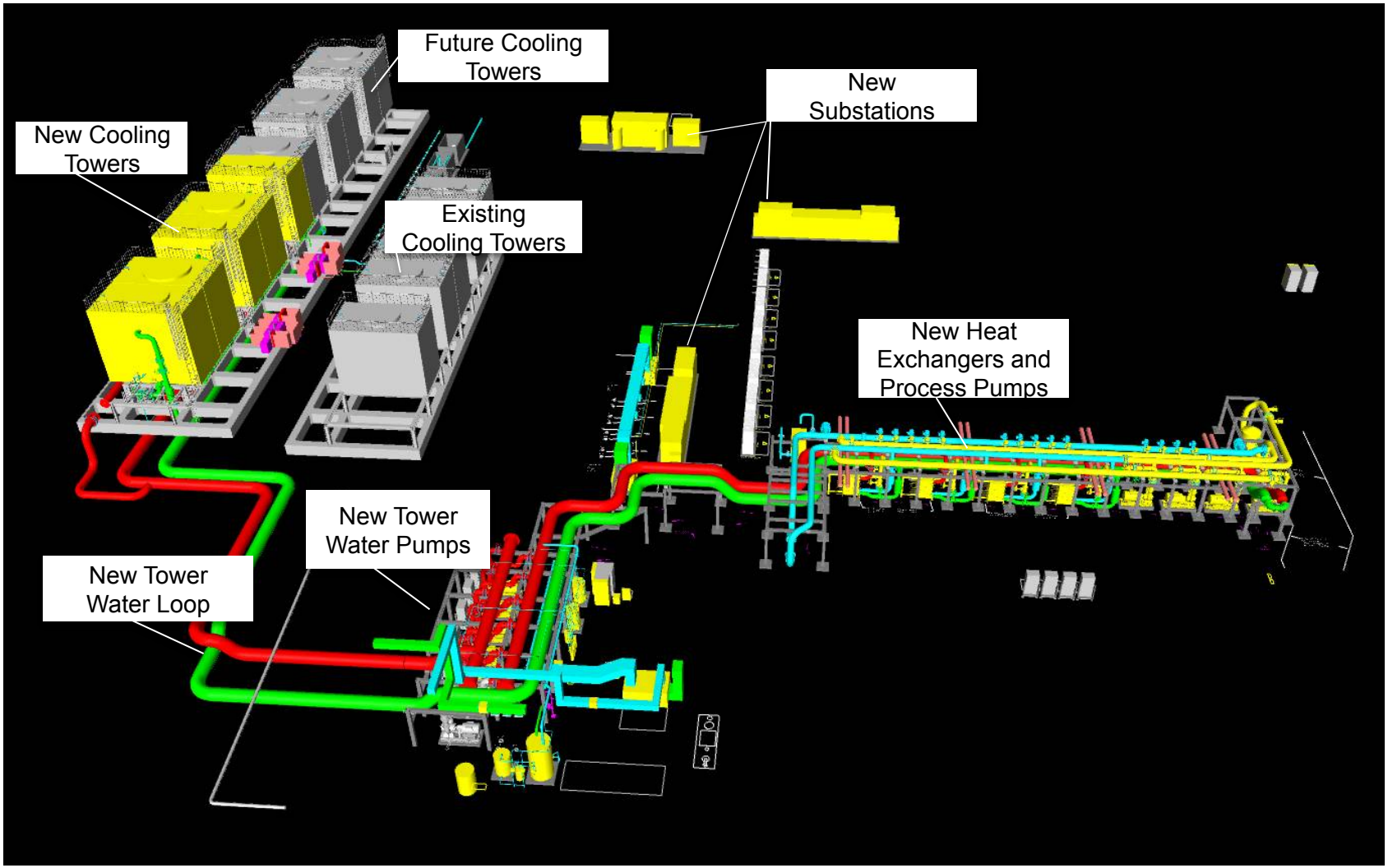
Laboratory Data Communications Complex (LDCC) – 1989

- Data center electrical capacity 8 MW
- Data center air cooling capacity 9 MW
- Data center water cooling capacity 2 MW

SCC Cooling Infrastructure



SCC Computer Cooling Equipment Project Overview



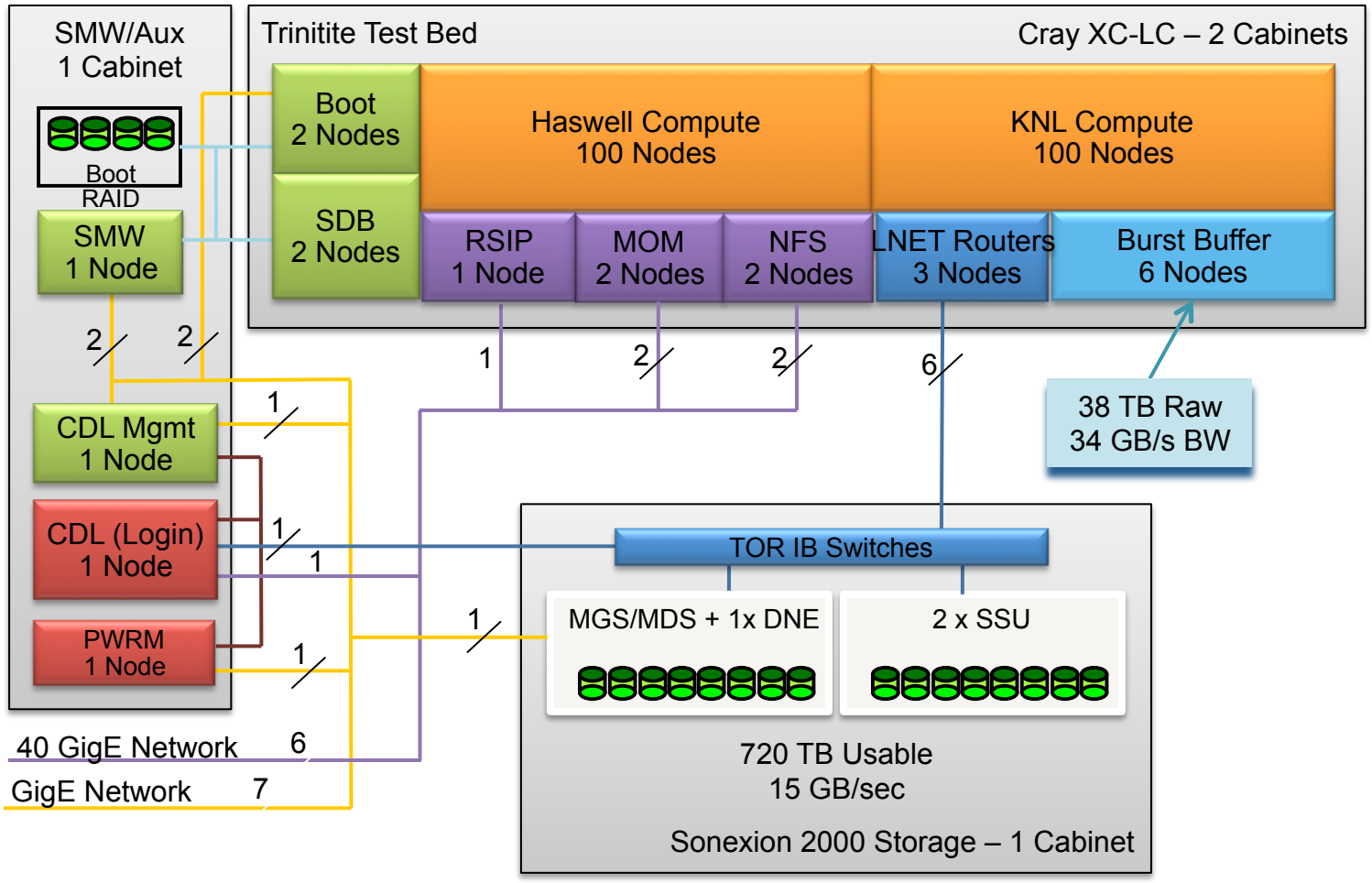
SCC Underfloor Infrastructure

New Mexico Alliance for Computing at Extreme Scale (ACES)
Cielo Configuration

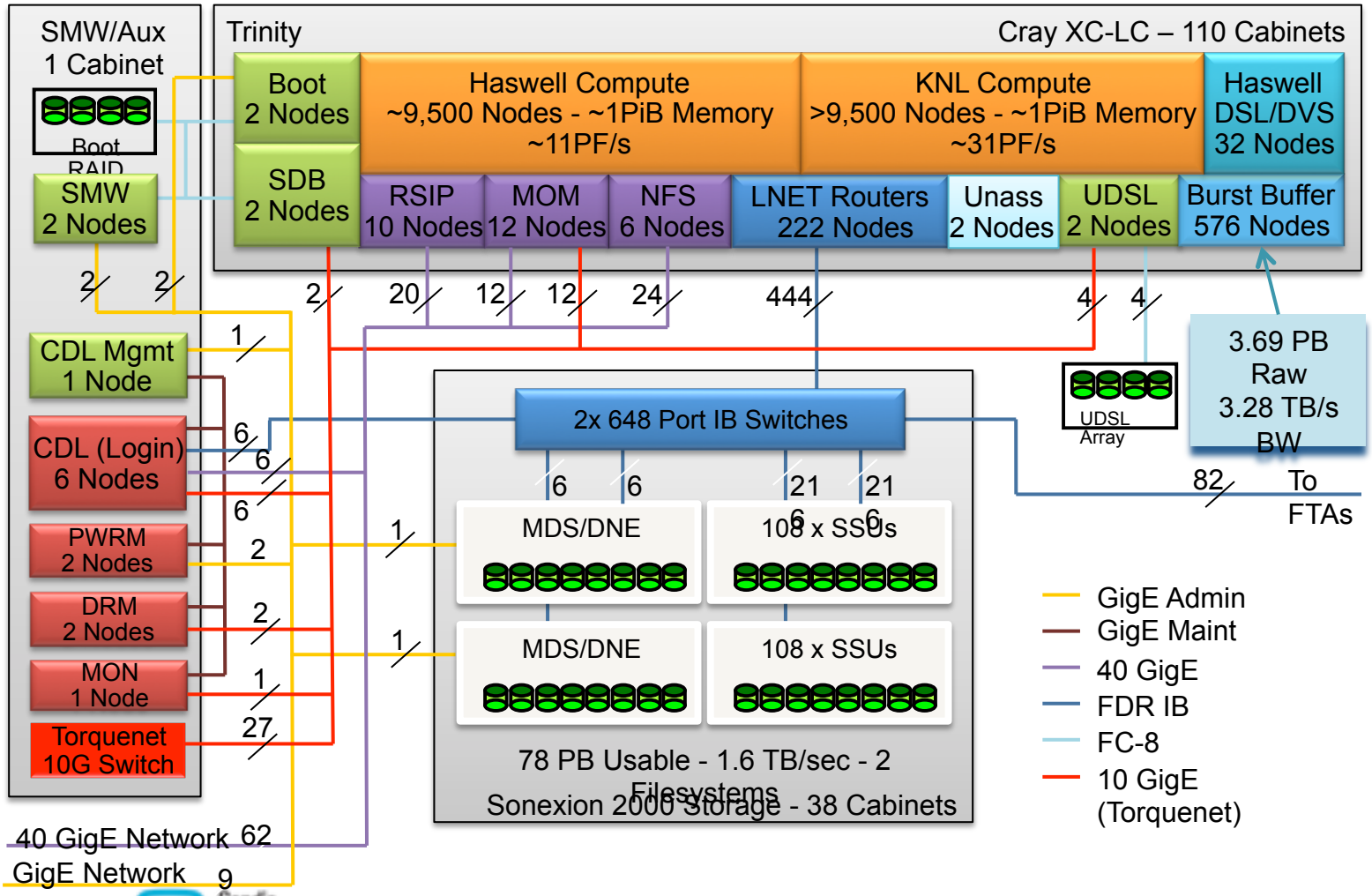
Cielo Specifications

Compute cabinets:	96	Peak:	**1.37** Pflops
Layout:	6 rows of 16	Memory:	291.5 TB
Compute Blades:	2236	Service Blades:	68
Compute Nodes:	8,944	Service Nodes:	272
Compute Cores:	143,104	Storage Bandwidth:	~271 GB/sec

ACES - Trinitite Configuration



ACES - Trinity Configuration



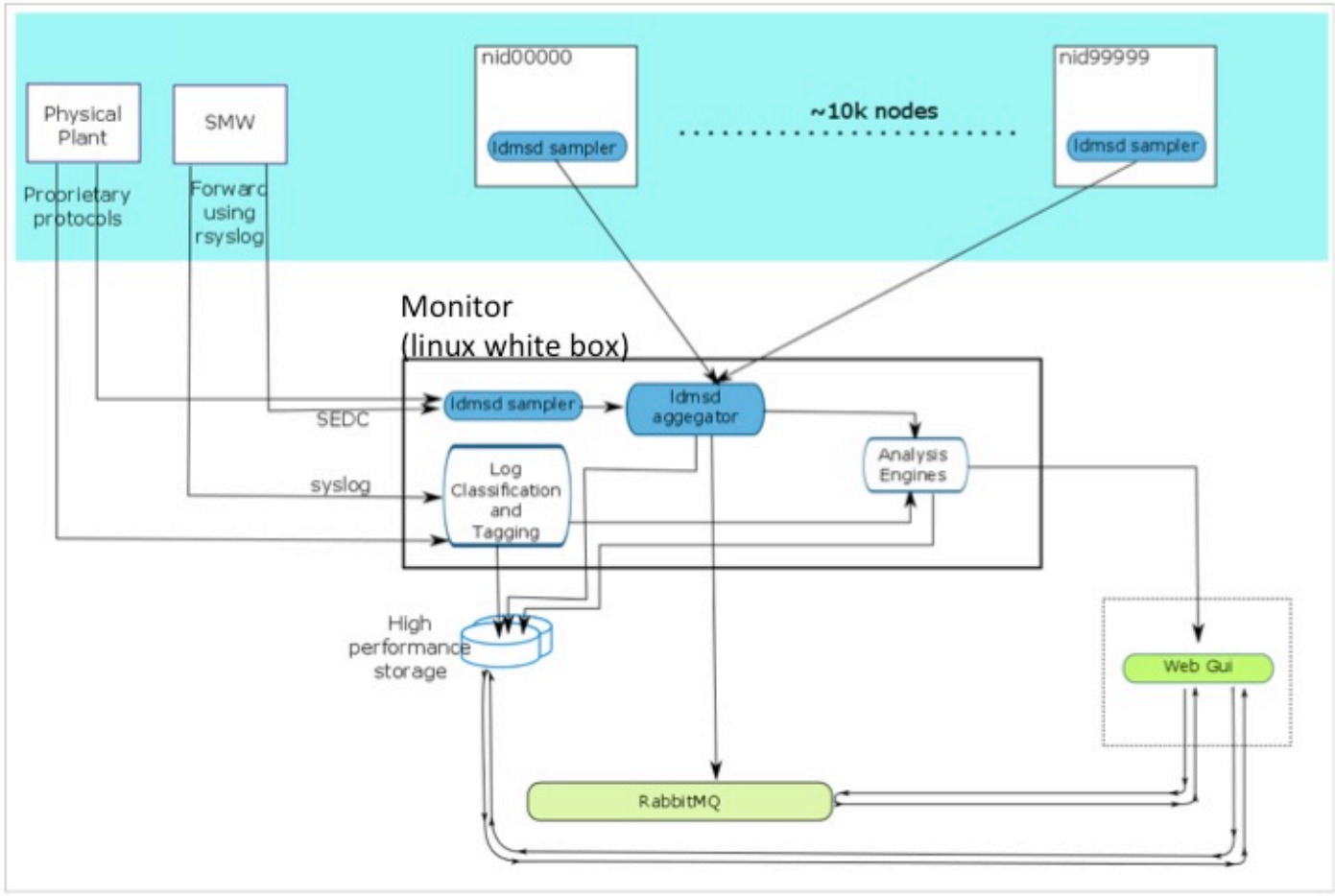
Workload

- Exercise various aspects of the machine:
  - Applications:
    - HPCG – memory bound
    - HPL - compute bound
    - Combustion Code – representative app
  - Single node and full machine
  - Power Capping: 0%, 50%, no capping (230, 322, 415 W)
  - Non-Turbo (base 2.3 GHz) and Turbo Modes (up to 3.6 Ghz)
- Particular focus on:
  - Power, cooling, temperature
  - Facilities, machine, component data
  - Behaviors, interactions, variations

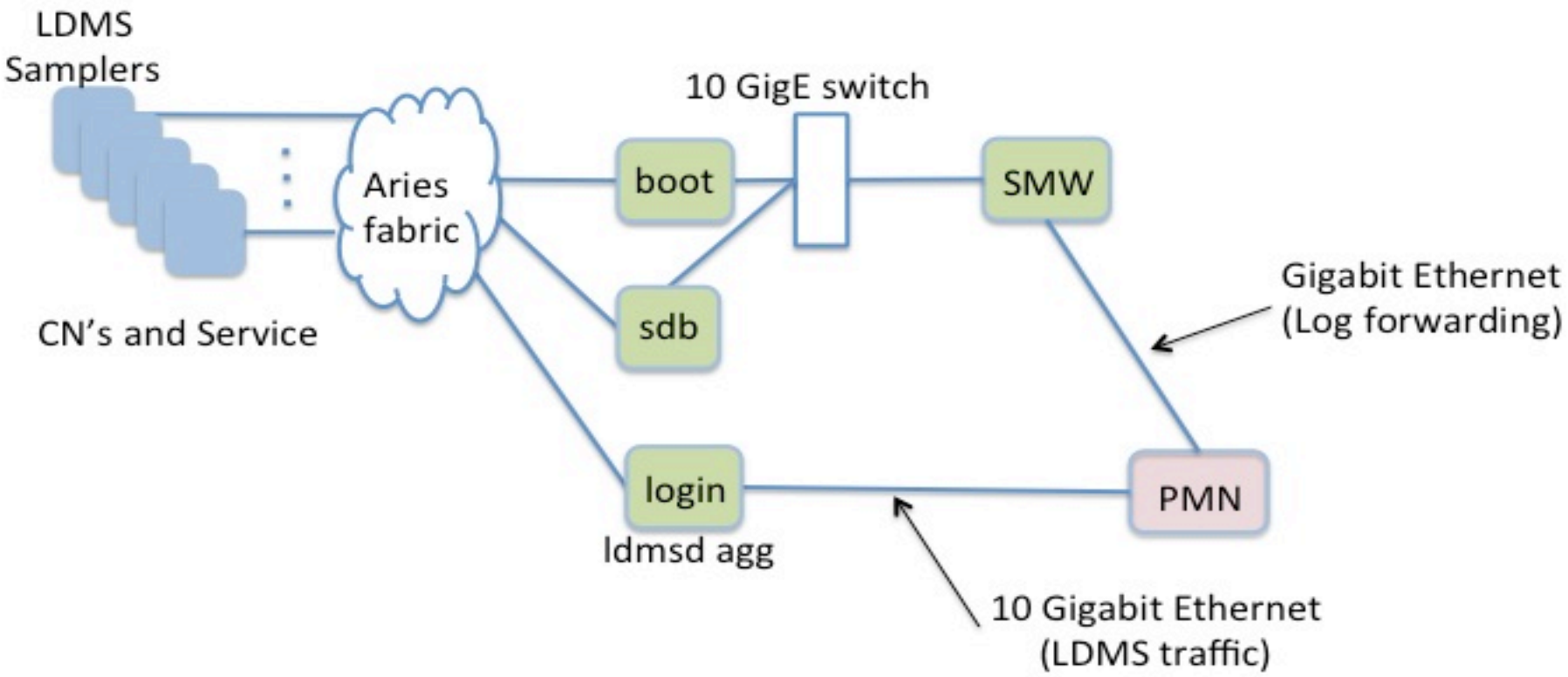
Data Sources

- SEDC
  - voltages, currents, and temperatures of a variety of components at the cabinet, blade, and node level
- Power API
  - prototype is a layered architecture of an implementation of the Power API specification version 1.0t o collect node level data at 10Hz
- LDMS
  - Lustre file system counters
  - CPU load averages
  - Current free memory
  - LNet traffic counters
  - ipogif counters
  - power and energy metrics via *sysfs*
- LLM
  - Using Cray's Lightweight Log Management to gather syslog, console, power management, smw, event, alps,etc.
- Envdata
  - water-related data directly from the cabinet (will soon be in SEDC data)

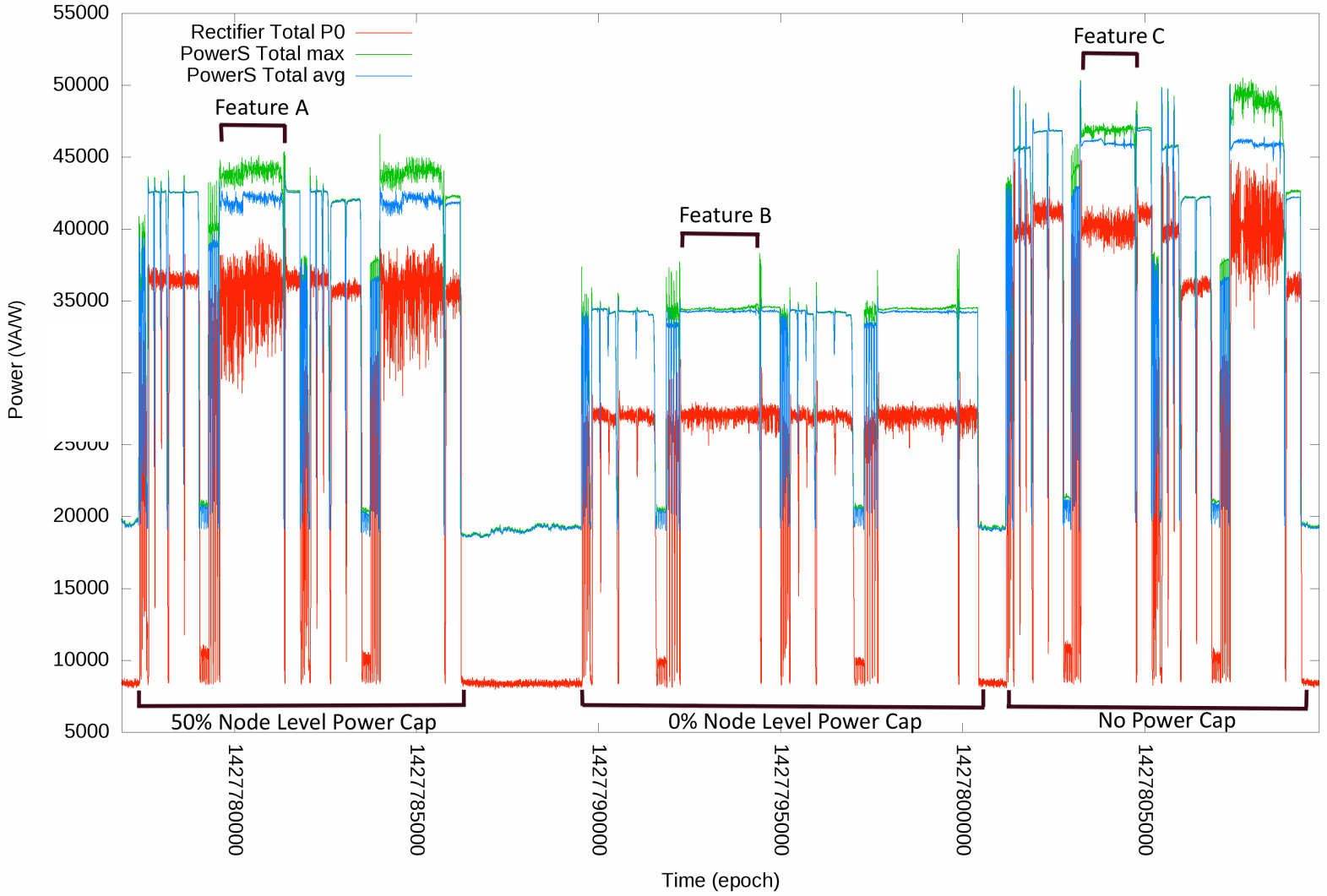
High-level Monitoring Diagram



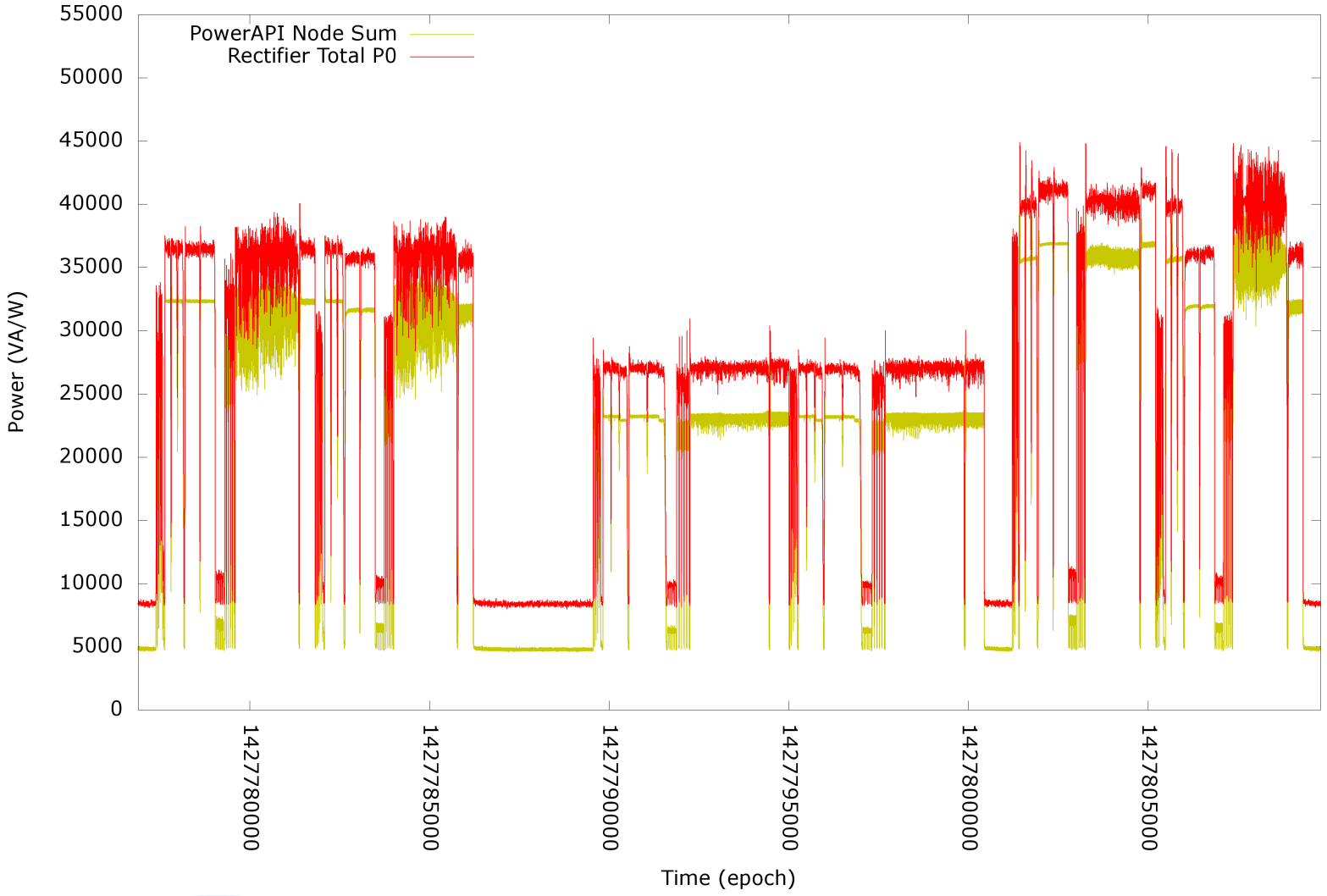
Art System Monitoring Configuration



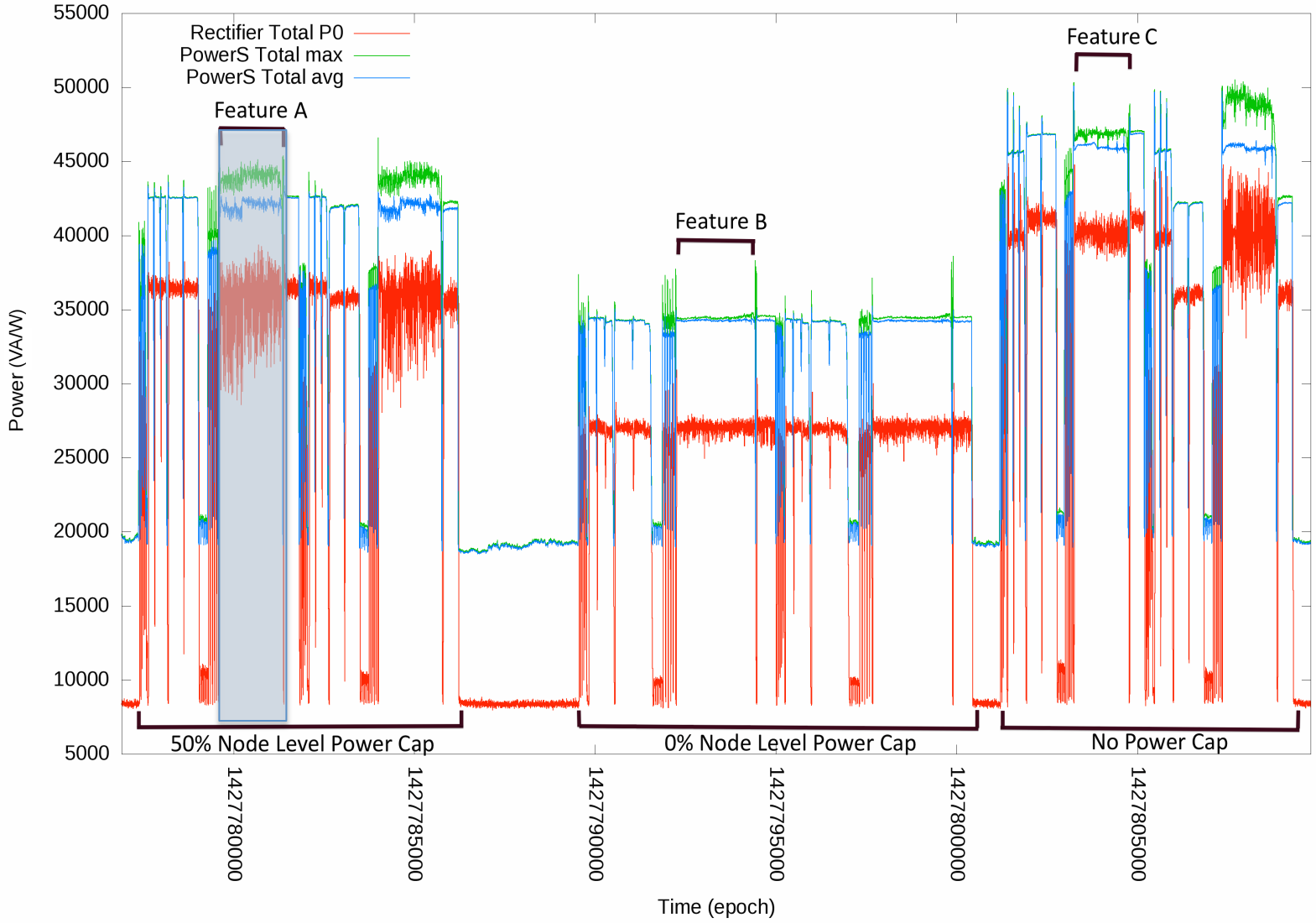
Facilities vs Machine



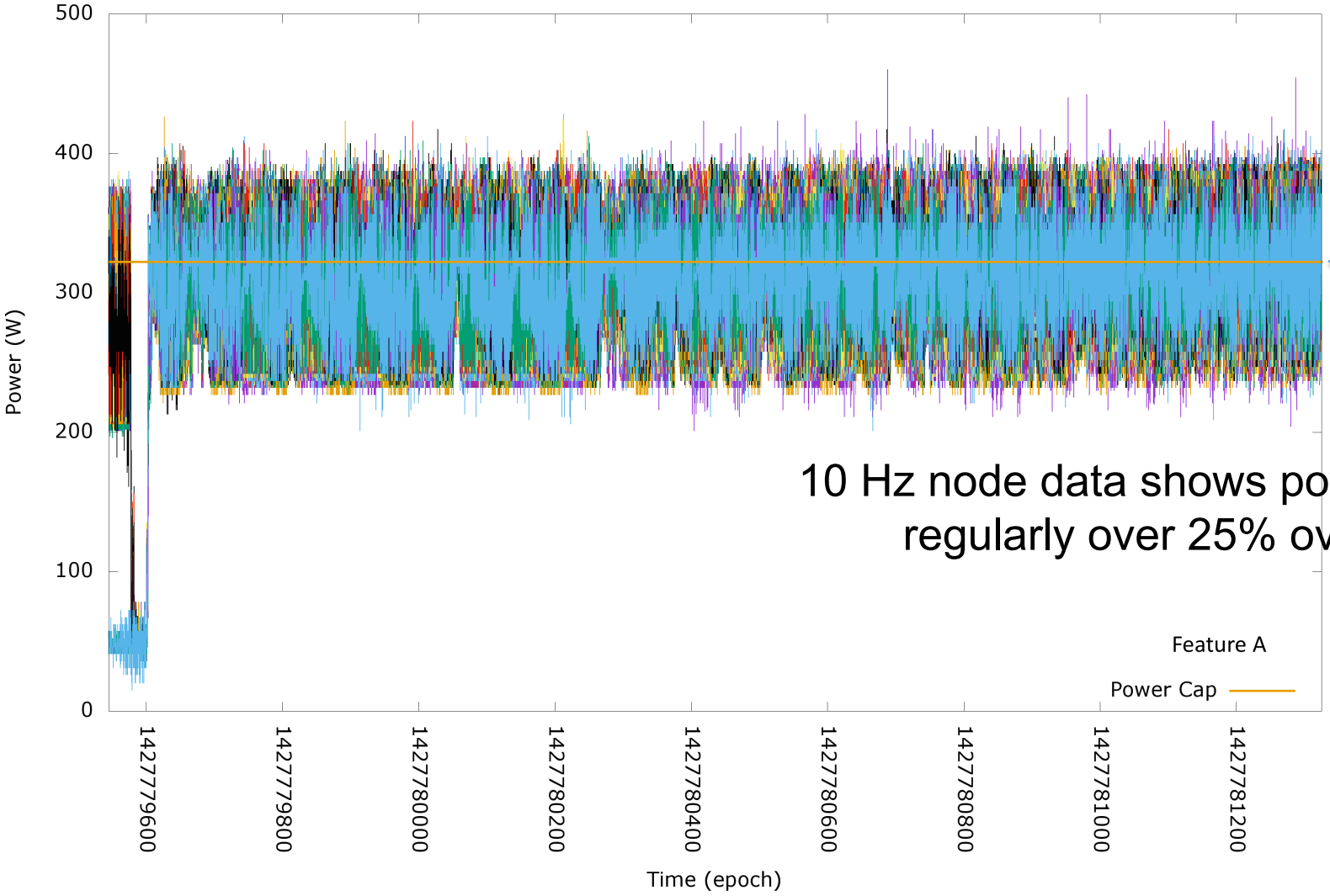
Machine vs 10Hz Node Data



Facilities vs Machine



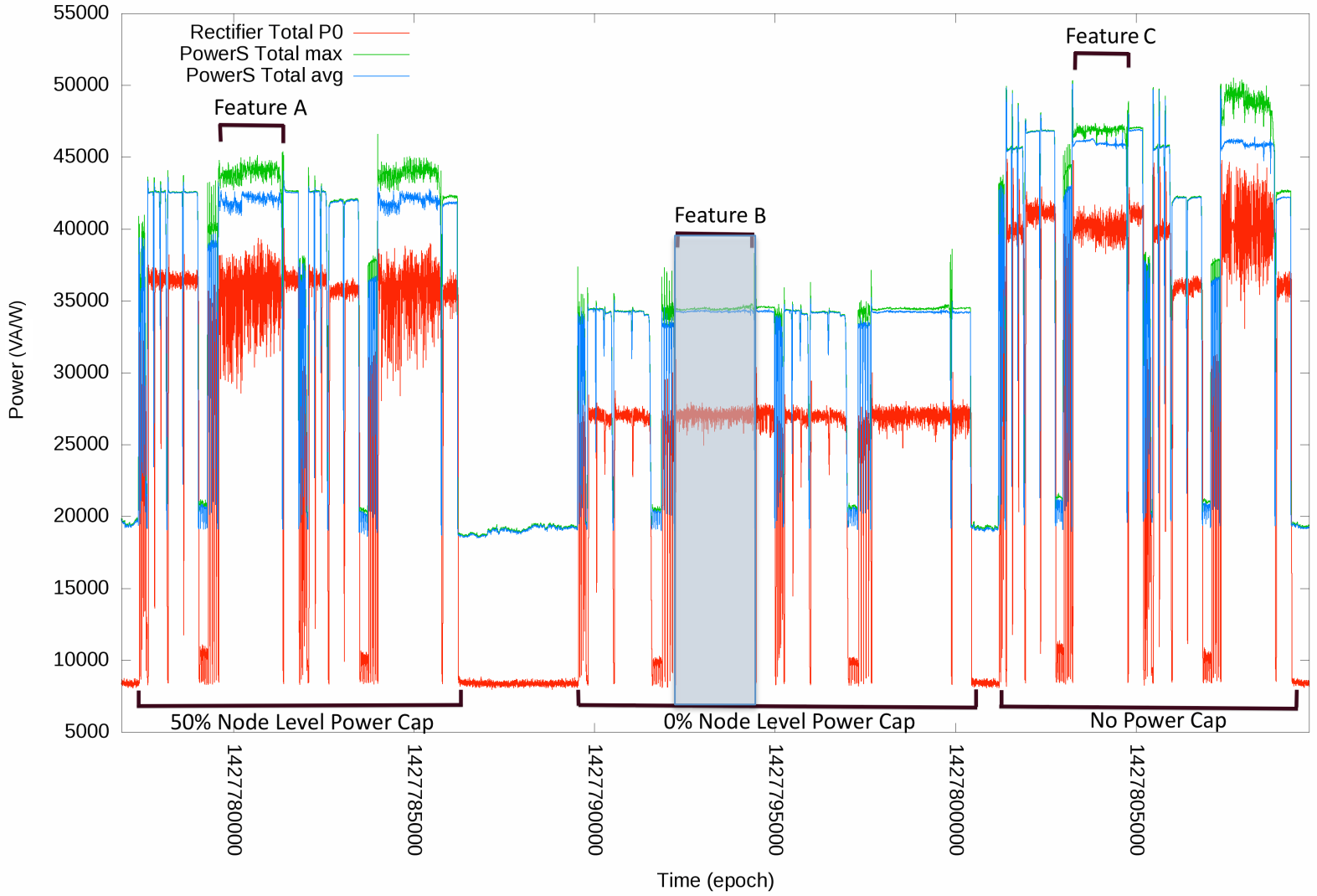
Power Capping (50% node level)



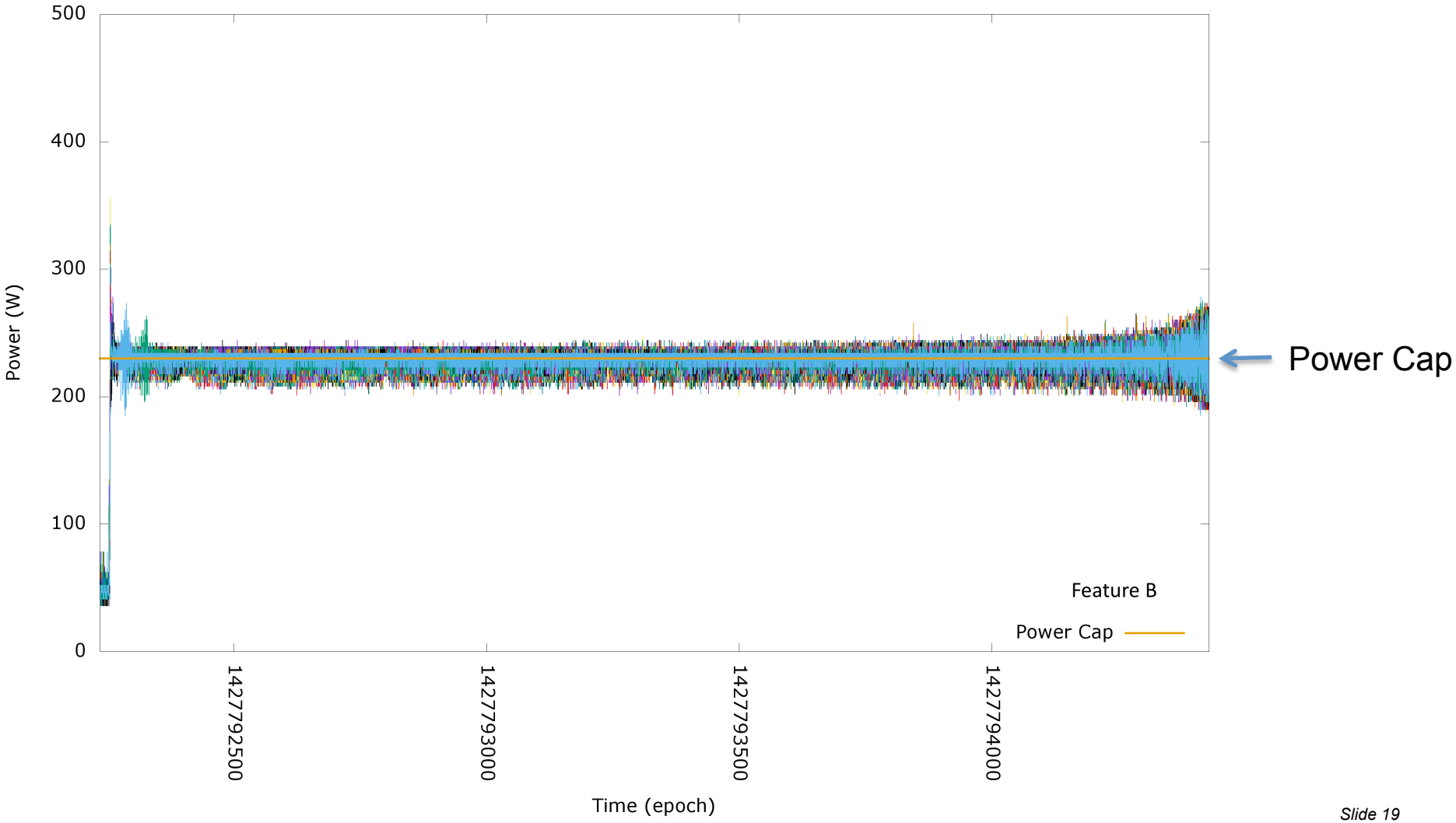
Power Cap

10 Hz node data shows power usage regularly over 25% over the cap

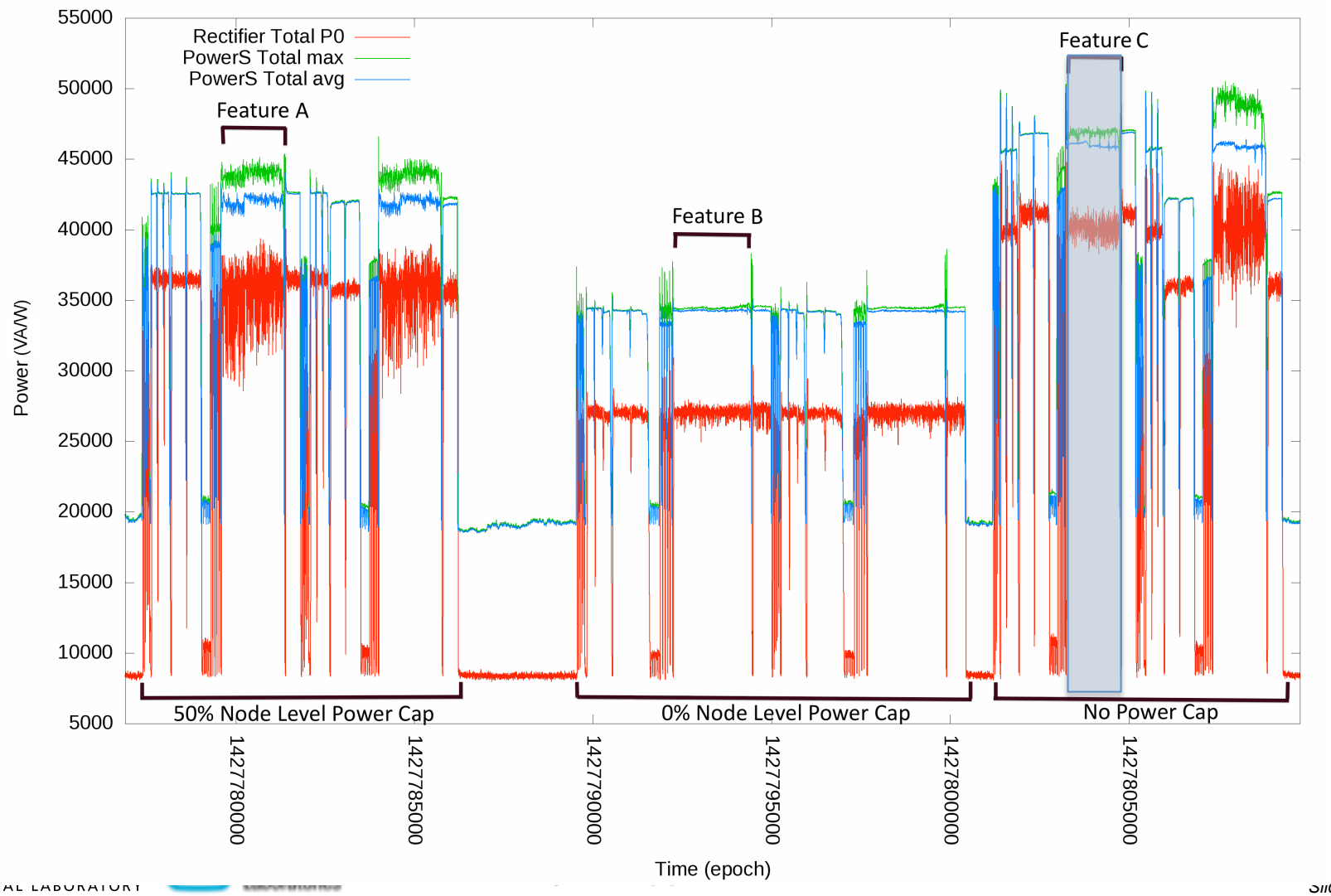
Facilities vs Machine



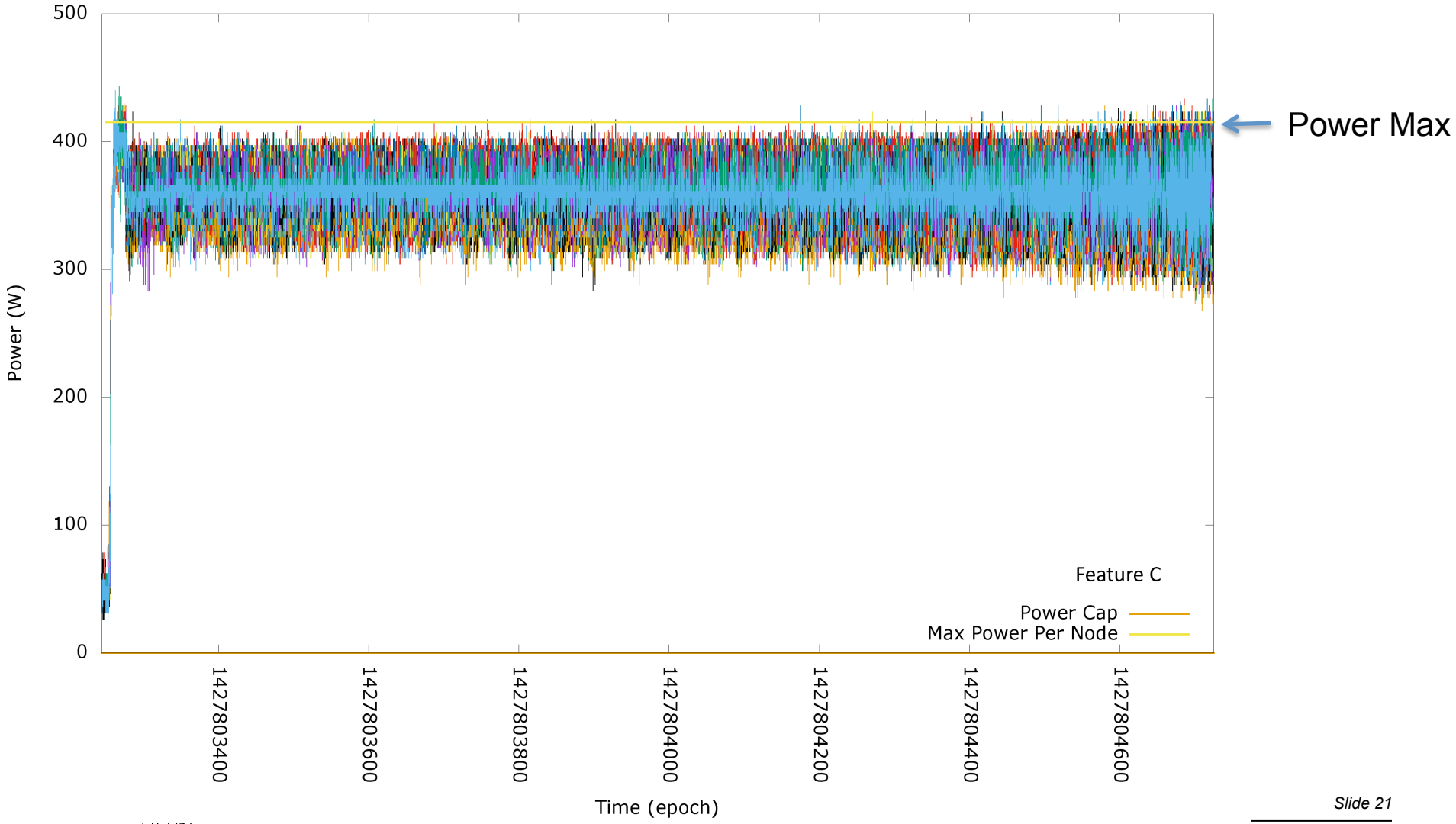
Power Capping (0% node level)



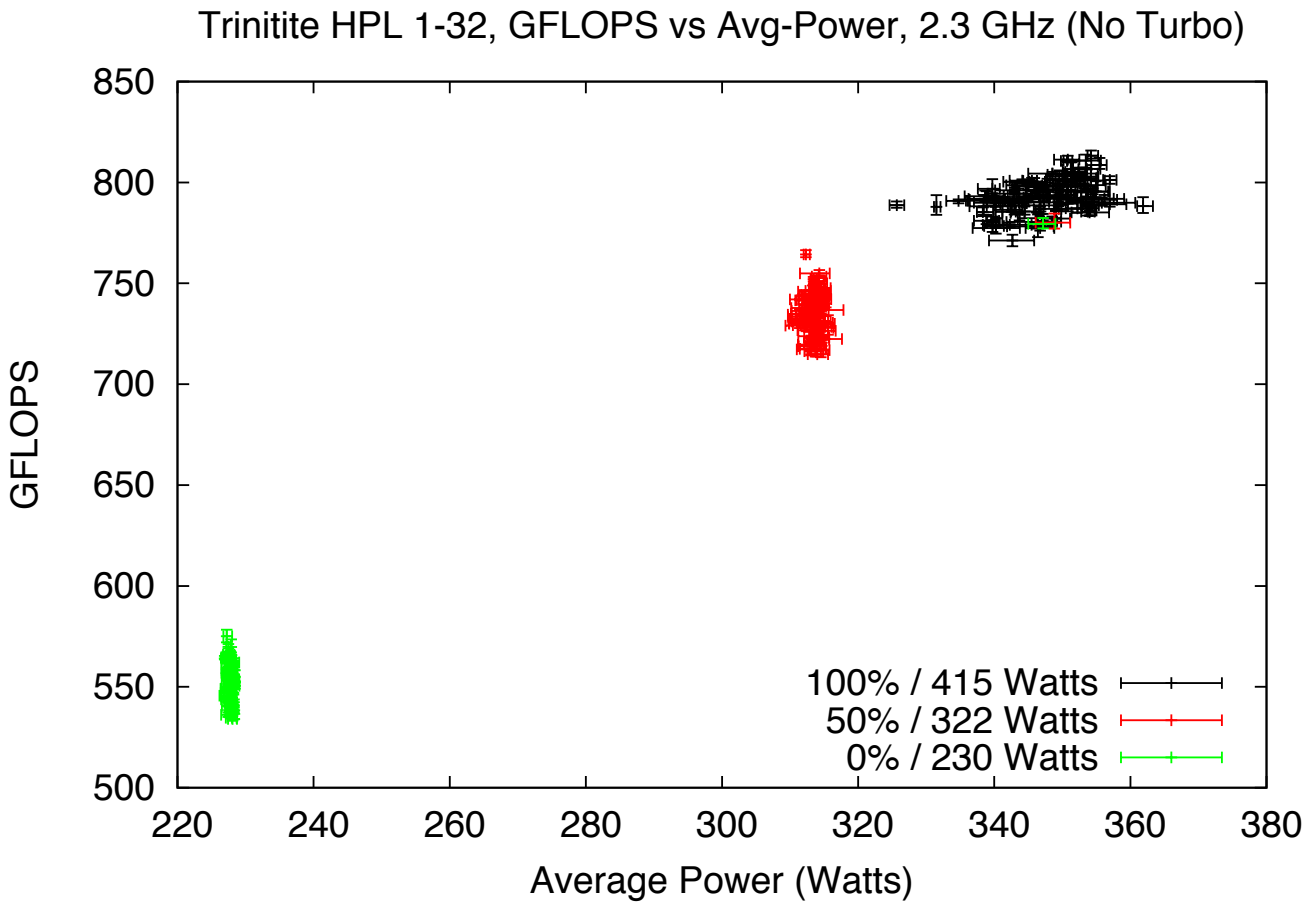
Facilities vs Machine



Power Capping (no cap)

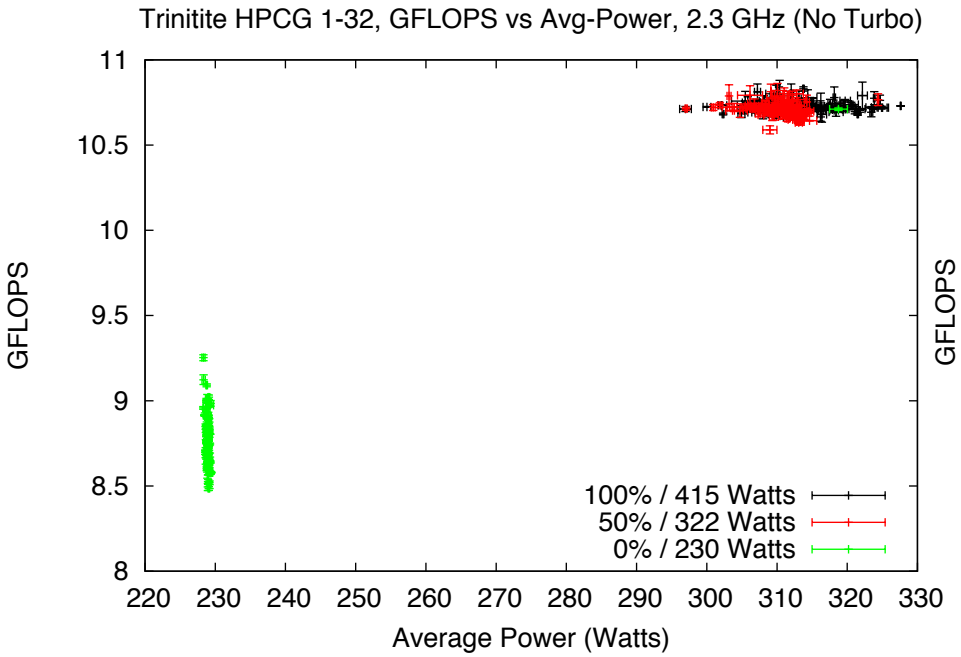


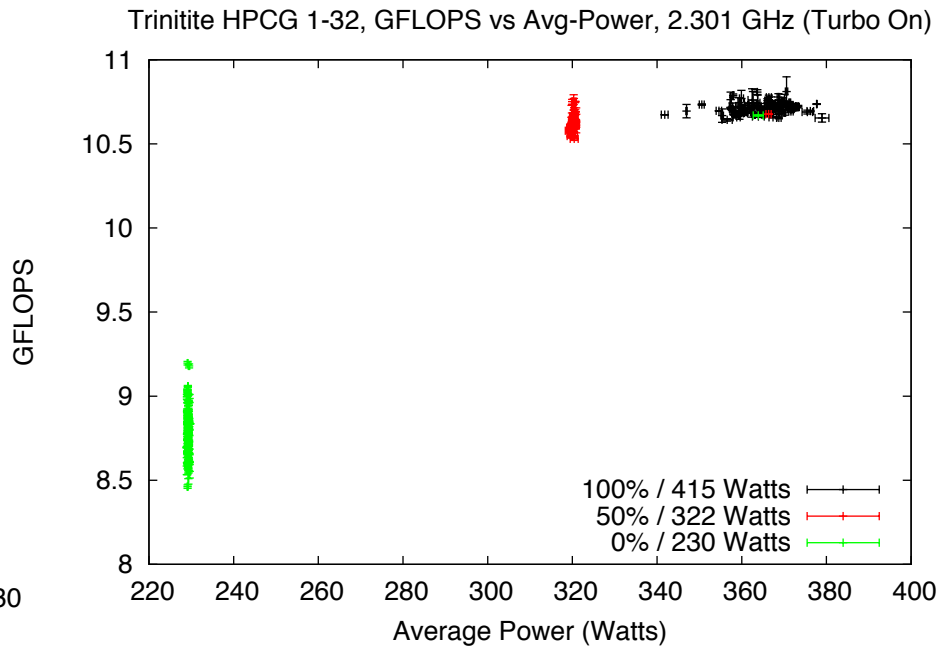
Performance vs Power Variations (HPL/No Turbo)



- Goal: Resource-aware scheduling matching component characteristics to application demands
- More vertical distributions = power cap being hit. Node performance depends on energy efficiency of the node's processors

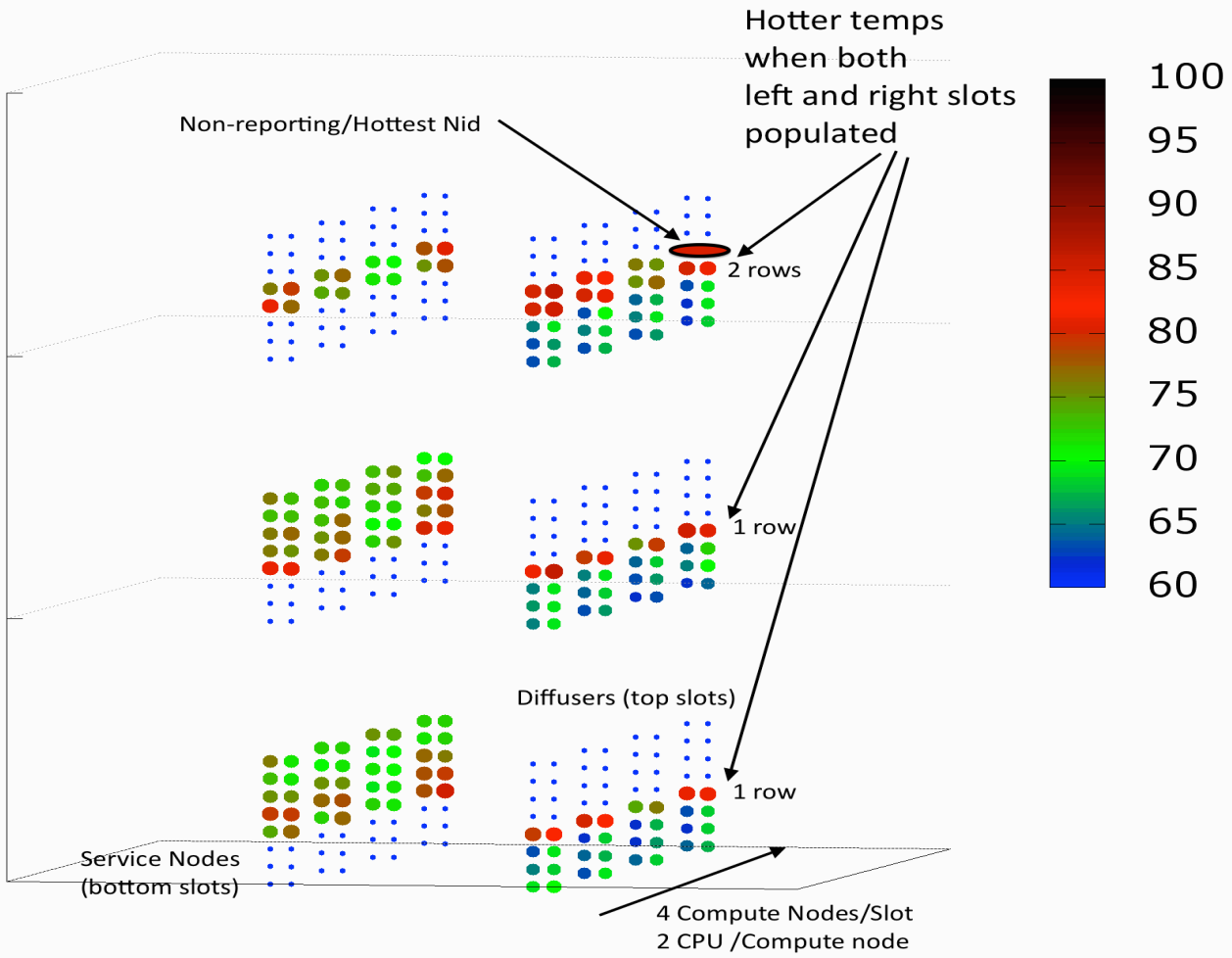
Performance vs Power Variations (HPCG)





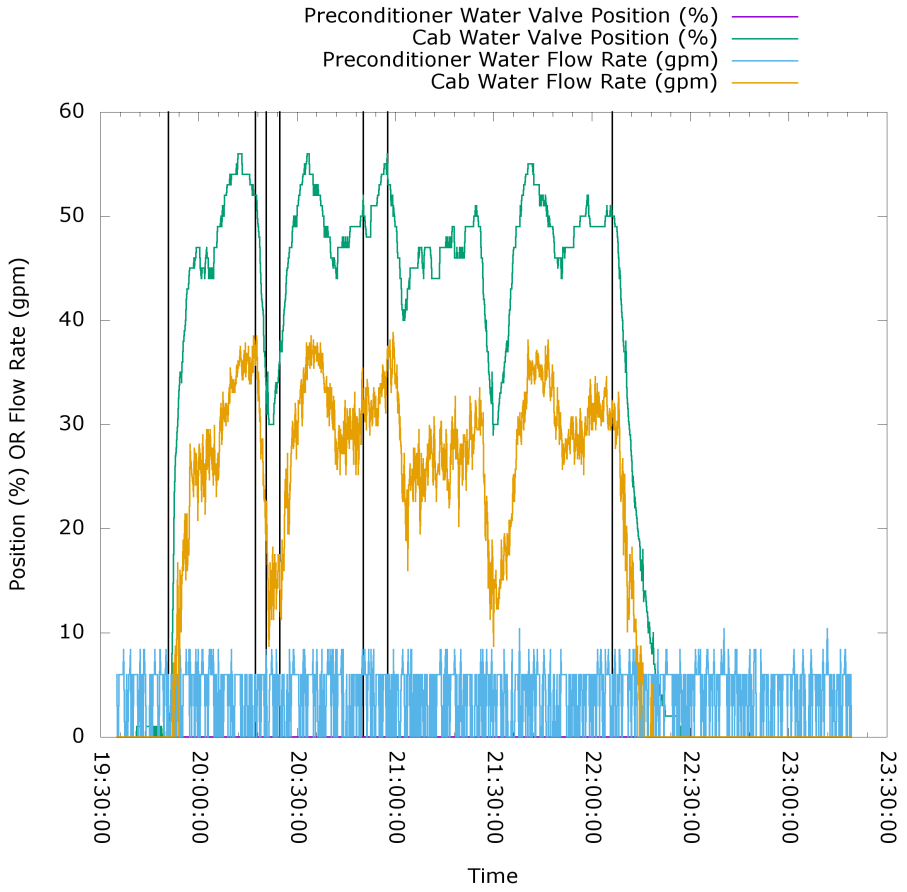
- No turbo: 50% and no capping results have similar power usage indicating that power cap is not being reached
- Turbo: more energy usage, but with no increase in performance

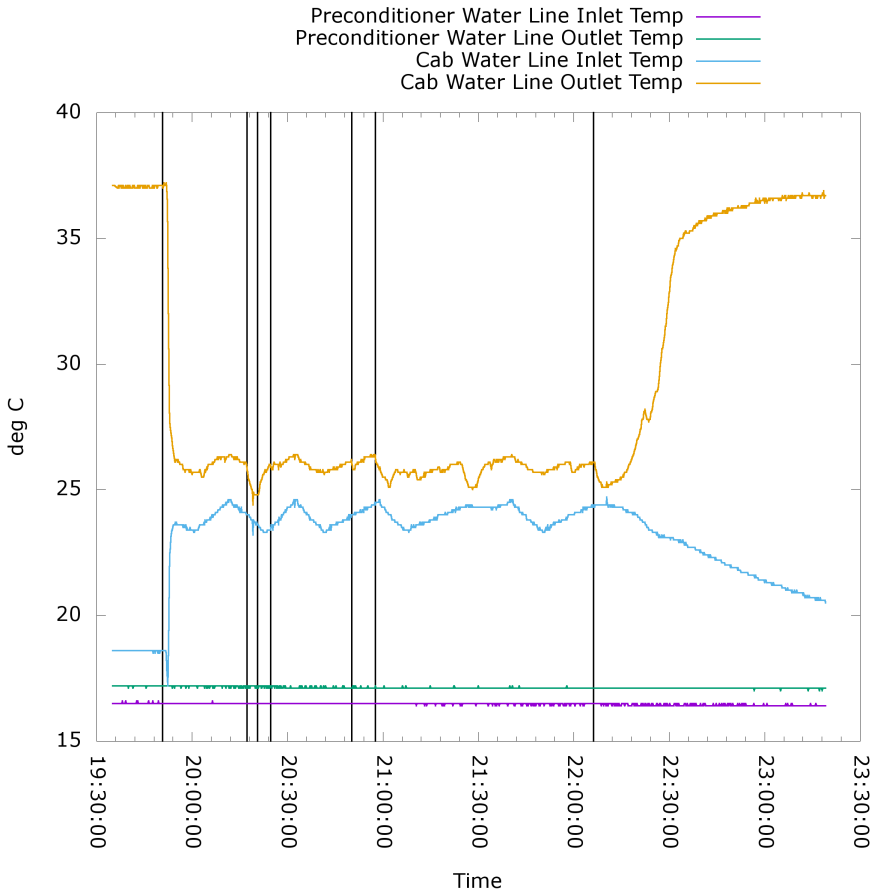
Thermal Mapping of Trinitite



- High temperatures result in thermal throttling and hence reduced performance
- High temperatures cause faster aging and higher failure rates
- Up to 30 degree variation across all CPU
- Up to 10+ degree variation for CPU in same slot

Cooling


Preconditioner Water Valve Position (%)
Cab Water Valve Position (%)
Preconditioner Water Flow Rate (gpm)
Cab Water Flow Rate (gpm)
Position (%) OR Flow Rate (gpm)
60
50
40
30
20
10
0
19:30:00
20:00:00
20:30:00
21:00:00
21:30:00
22:00:00
22:30:00
23:00:00
23:30:00
Time


Preconditioner Water Line Inlet Temp
Preconditioner Water Line Outlet Temp
Cab Water Line Inlet Temp
Cab Water Line Outlet Temp
deg C
40
35
30
25
20
15
19:30:00
20:00:00
20:30:00
21:00:00
21:30:00
22:00:00
22:30:00
23:00:00
23:30:00
Time

Log patterns from Baler

Baler Log Analysis tool discovers patterns from data with no user guidance

- New system analysis
- Deterministic patterns enable comparison cross-system, across time

Patterns 280 and 283 respectively:

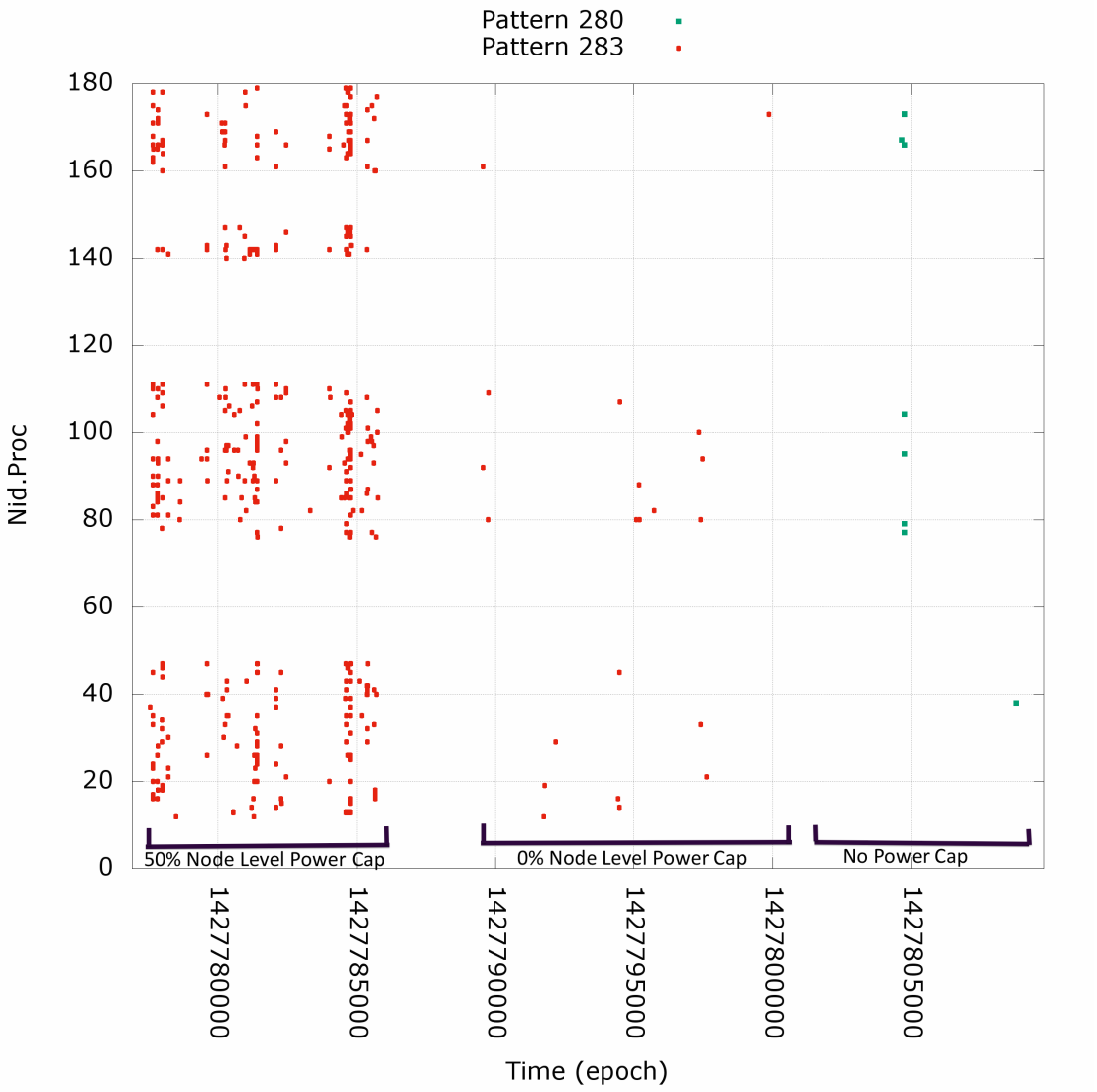
- 280: * * - - Node * interrupt *=*, *=*, *=* *[*]: * * * Processor Hot
- 283: * * - - Node * power budget exceeded! Power=*, Limit=*, * Correction Time=*

Example messages corresponding to patterns 280 and 283 respectively:

- 280: bcsysd 2080 - - Node 2 interrupt IREQ=0x20000, USRA=0x0, USRB=0x80 USRB[7]: C0_PROCHOT CPU 0 Processor Hot
- 283: bcpmd 2140 - - Node 2 power budget exceeded! Power=340, Limit=322, Max Correction Time=6

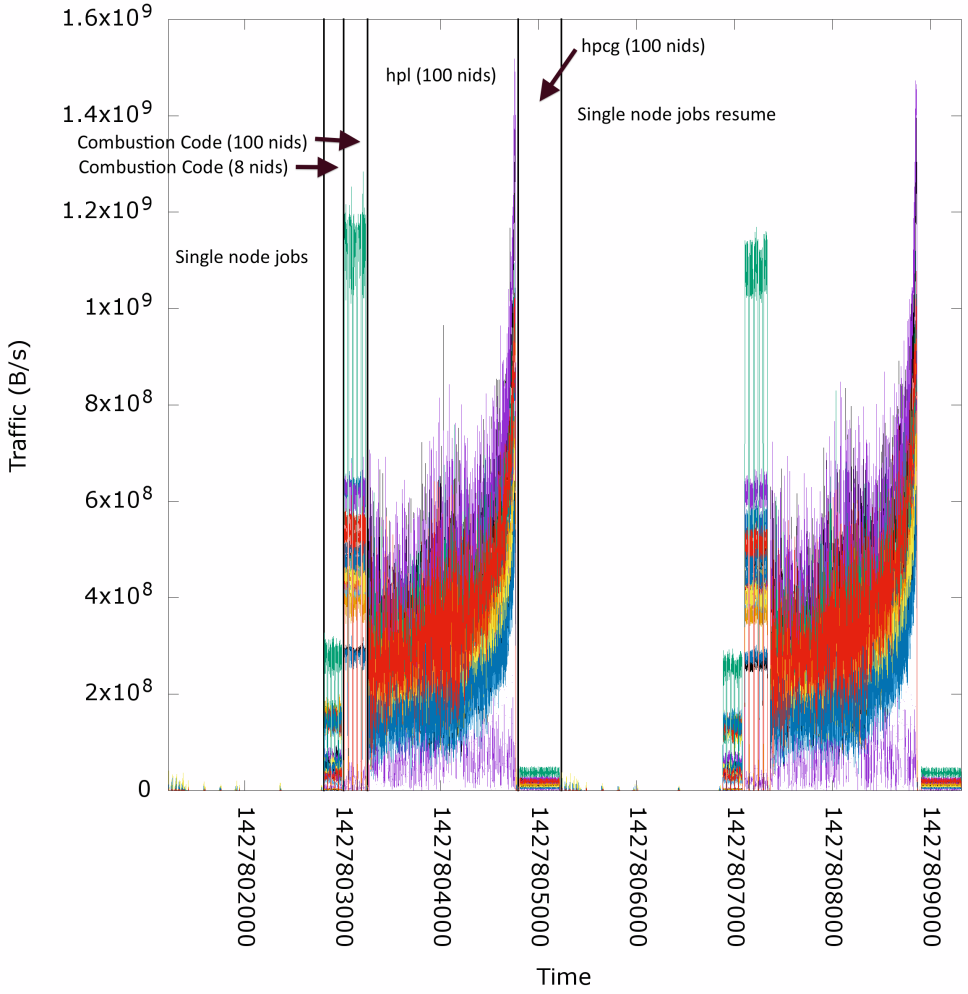
Integration of Baler Patterns with Workload Data

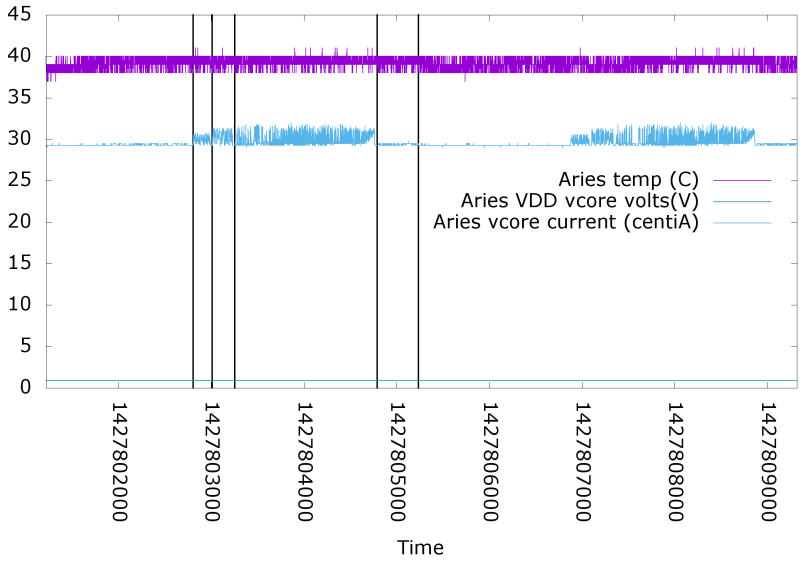
Power budget exceeded when under power cap (left and middle)



Processor Hot when no power cap (right)

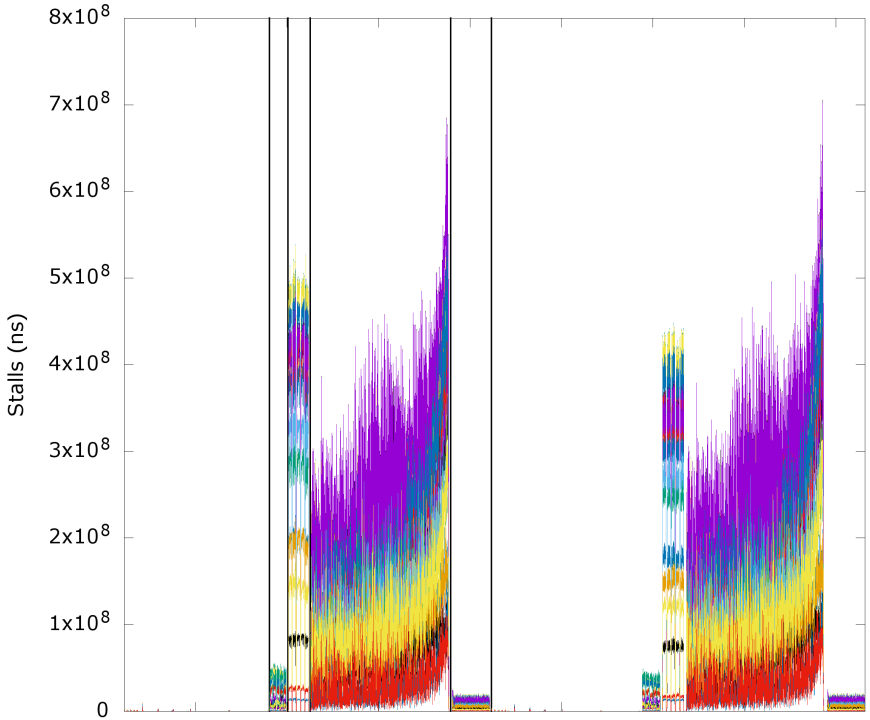
Aries Network Traffic and Machine Data





Future: Investigate dependencies of Aries temperature and power draw due to network utilization

Congestion Indicated by Stalls in the Aries Network



Time spent in network stalls due to contention for shared network

- Network congestion can significantly affect application run-time
- Monitor network to understand application performance variability
- Goal: include network data in resource-aware scheduling

Future Capabilities

- Ability to power cap and shed load from the platform to respond to facility power and cooling constraints.
- Maximize water temperatures to predict optimal curve ratios from a pump and tower perspective.
- Gain efficiencies in future procurements by comparing ratio of power usage within the computer (power supplies, DIMMS, and CPUs).
- Capture metrics from platform to drive infrastructure efficiencies by automating facility control schemes dynamically.