SAND2014-15556PE

Exceptional service in the national interest

Sandia National Laboratories

Lightweight Distributed Metric Service (LDMS)

Capacity and Capability Application Impact Testing and Deployment

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

LDMS Overview

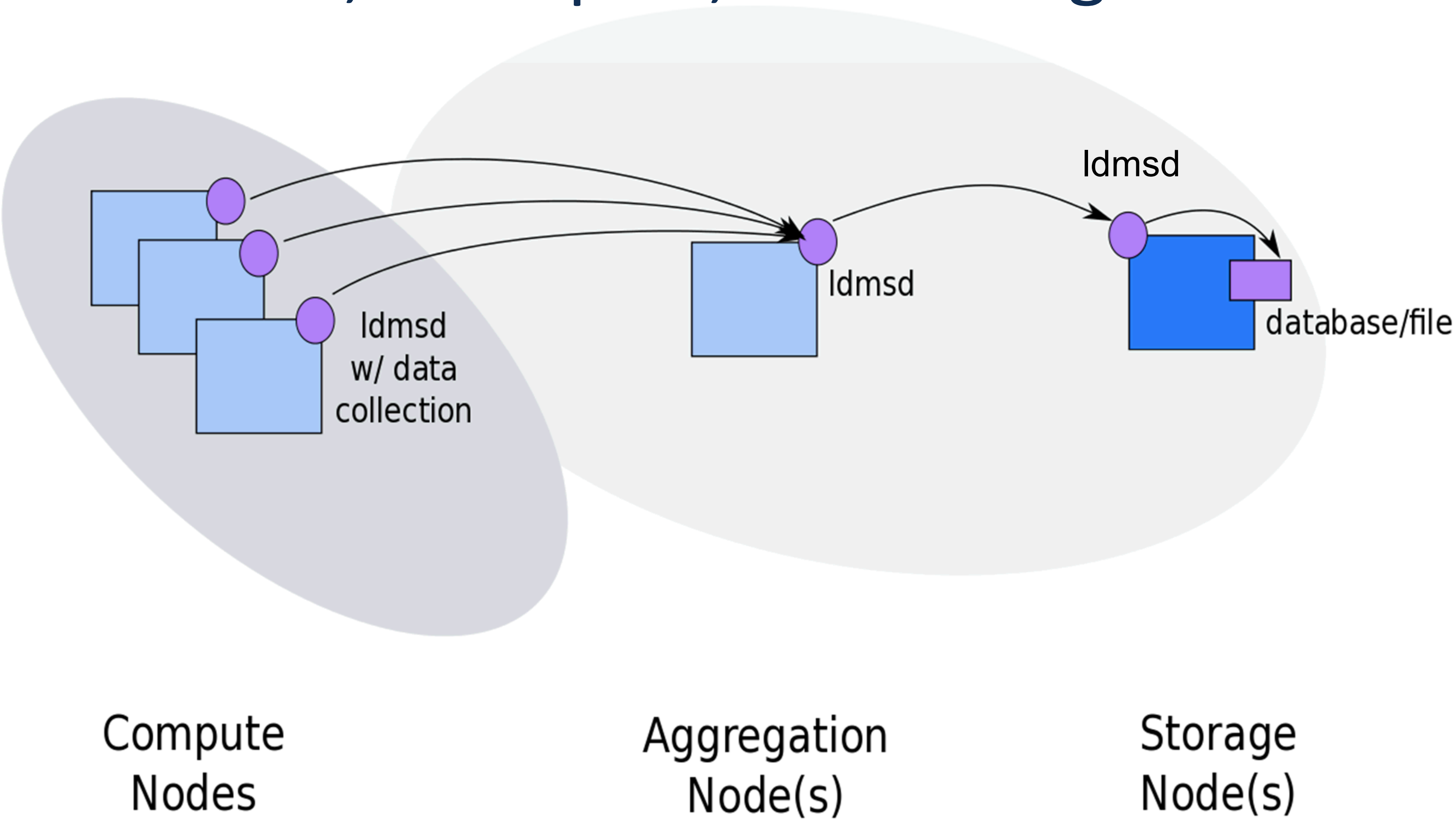
- Project Goals
- Deployment configuration and architecture
- Major functional components
- Impact testing
- Deployments and data (system and application)
- Related Software In Progress

Project Goals

Monitor system resource state and utilization as a system service, running asynchronous to applications, for both administrators and users of High Performance Computing (HPC) systems

- Provide:
  - Run-time collection of metrics of interest on time scales of interest
  - Per-node resource utilization understanding
  - Application profiles
- Lightweight (negligible application impact)
  - ~1.5MB memory footprint
  - CPU overhead is dependent on number of metrics, frequency of collection, and communication technology
- High fidelity (~< 100Hz)
  - Typically run at intervals of seconds
- Platform independent (i.e. portable across Linux OSs)
  - Fedora, RHEL, SUSE, CentOS, Mint, Ubuntu
- Support for Socket and RDMA on major network technologies
  - Ethernet, Infiniband, Cray Gemini/Aries
- Simple configuration

Deployment Configuration: Data Collection, Transport, and Storage



LDMS Metric Set Examples

shuttle-cray.ran.sandia.gov_1/meminfo

U64	160032	MemFree
U64	181728	Buffers
U64	3443332	Cached
U64	33076	SwapCached
U64	2987544	Active

shuttle-cray.ran.sandia.gov_1/procstatutil

U64	1826564	cpu0_user_raw
U64	699631	cpu0_sys_raw
U64	663843760	cpu0_idle_raw
U64	201018	cpu0_iowait_raw

shuttle-cray.ran.sandia.gov_1/vmstat

U64	40008	nr_free_pages
U64	122286	nr_interactive_anon
U64	321902	nr_active_anon
U64	465532	nr_inactive_file
U64	424986	nr_active_file

Metric sets:

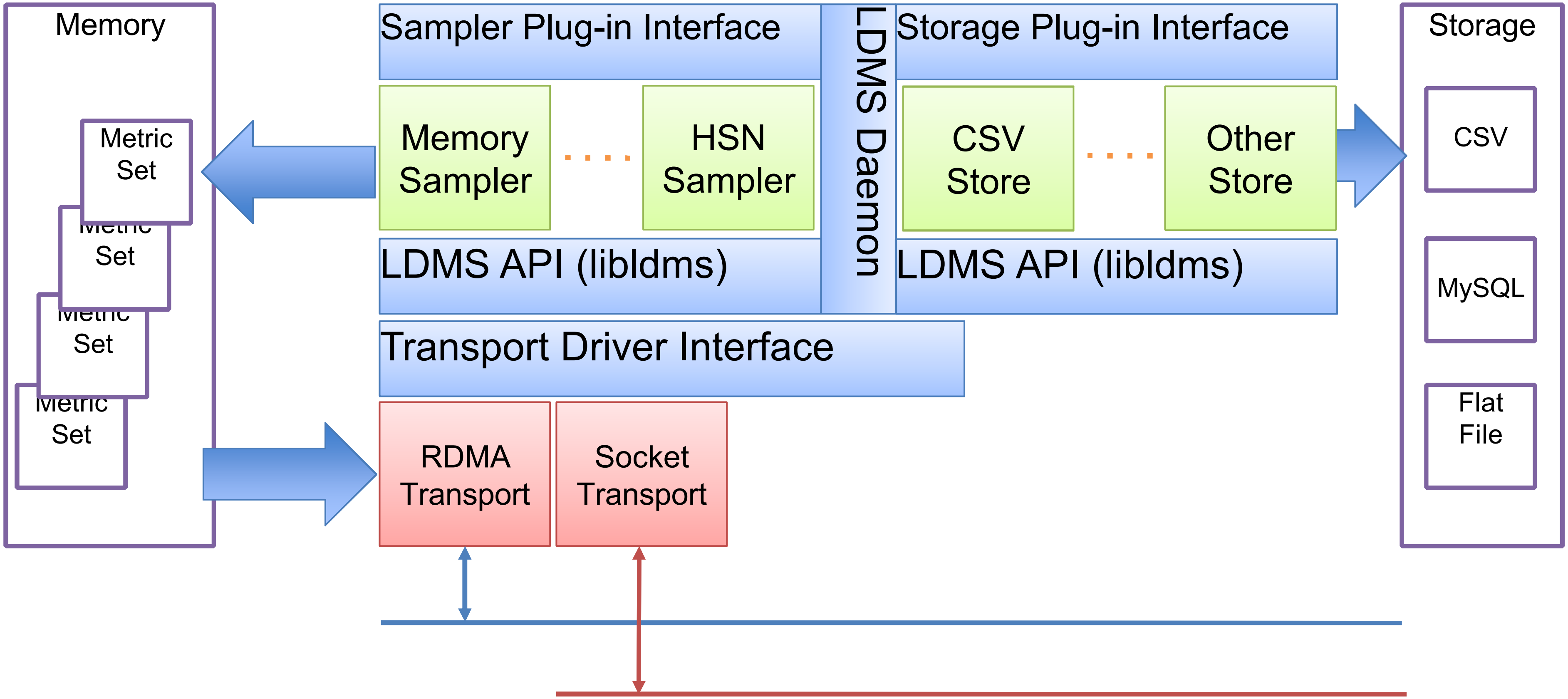
- (datatype, value, metricname) tuples
- optional per metric user metadata e.g., component id

API:

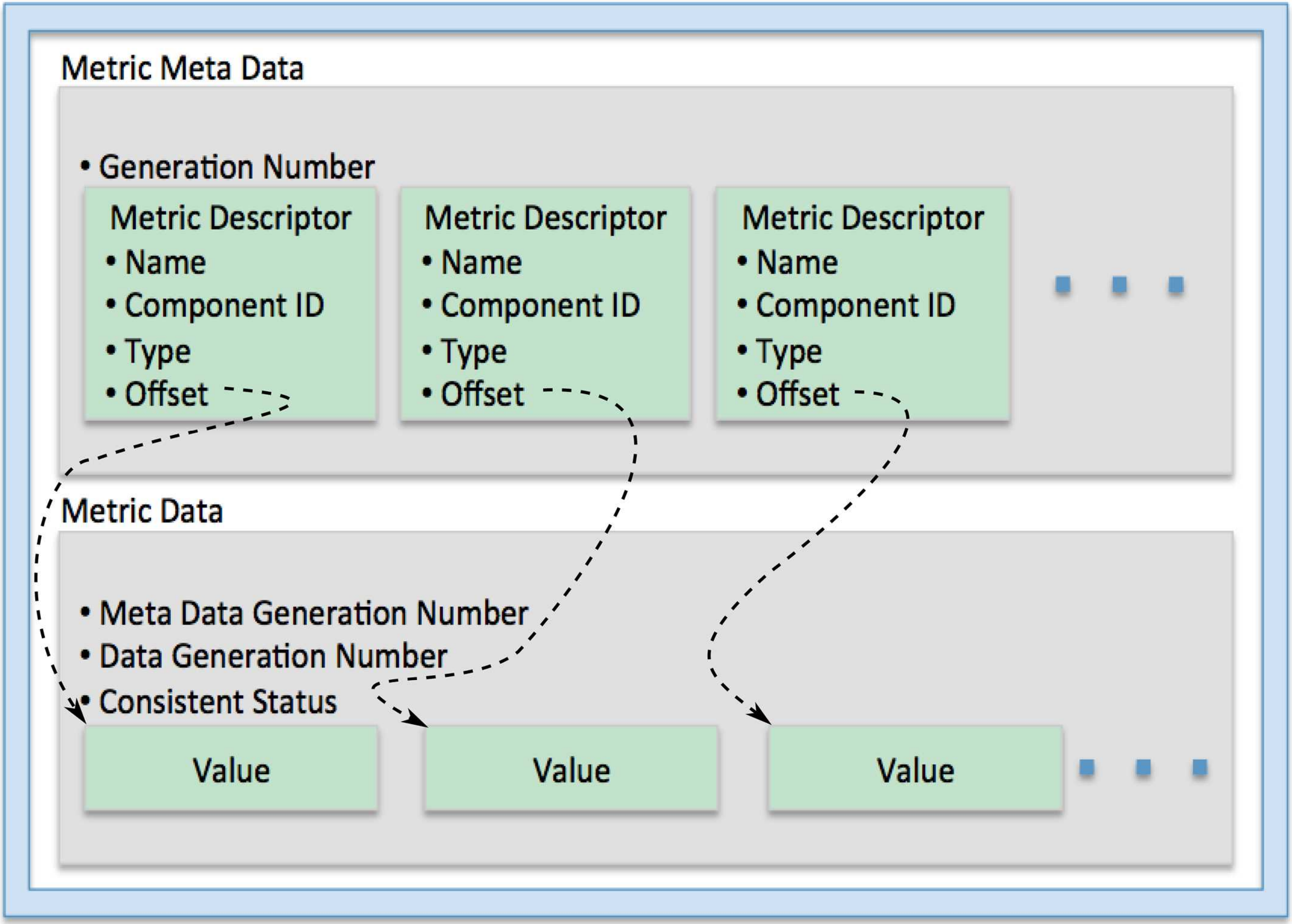
- *ldms_get_set*
- *ldms_get_metric*
- *ldms_get_u64*

- Same API for on-node and off-node transport

LDMS Architecture



Metric Set Memory



Metric Meta Data

- Generation Number

Metric Descriptor
- Name
- Component ID
- Type
- Offset

Metric Descriptor
- Name
- Component ID
- Type
- Offset

Metric Descriptor
- Name
- Component ID
- Type
- Offset

Metric Data

- Meta Data Generation Number
- Data Generation Number
- Consistent Status

Value

Value

Value

Major Functional Components and Features

- Collection
  - Run-time loadable monitor plugins (collect data into a “metric set”)
  - Run-time configurable sampling period from ~100Hz to days
  - Variety of collectors draw from /proc, /sys, lm-sensors, perf-event
  - Control is at the granularity of a “metric set”
  - Synchronous option enables “system view” to within clock skew
- Aggregation
  - Fan-in of thousands to 1
  - Support for failover configuration
  - Supports daisy-chaining
    - Aggregate from collectors and/or aggregators
- Storage
  - Support for: CSV, flatfile, MySQL, custom
  - CSV header can be first line of output or separate file
  - Support for derived metrics in “store_derived_csv” plugin

Additional Recent Feature Adds

- Separate connection thread pool
- Authentication
- IB sampler supports 64 bit counters
- Template for multi-source samplers (e.g. cray_system_sampler)
- Support Cray's gpcdr kernel module for both Gemini and Aries network performance counters

Current Monitor Plugins

- /proc
  - meminfo, vmstat, net/dev/stat, interrupts, nfs
  - kgnilnd (Cray specific)
  - Lustre
- cray_system_sampler (Cray specific supports XE, XK, XC)
  - Gemini/Aries Tile and NIC counters w/ link aggregation
  - Lustre llite counters
  - A variety of metrics from other sources
- perf_event
  - Generic interface for acquisition of hardware counters e.g., data cache misses, instruction cache misses, hyper-transport bandwidth
- rsyslog (Cray specific)
  - SEDC (RAS) and ALPSdata
- lmsensors (/sys)
  - Temperatures, fan speeds, voltages
- IB traffic counters (/sys)

IMPACT TESTING

Synchronous Collection

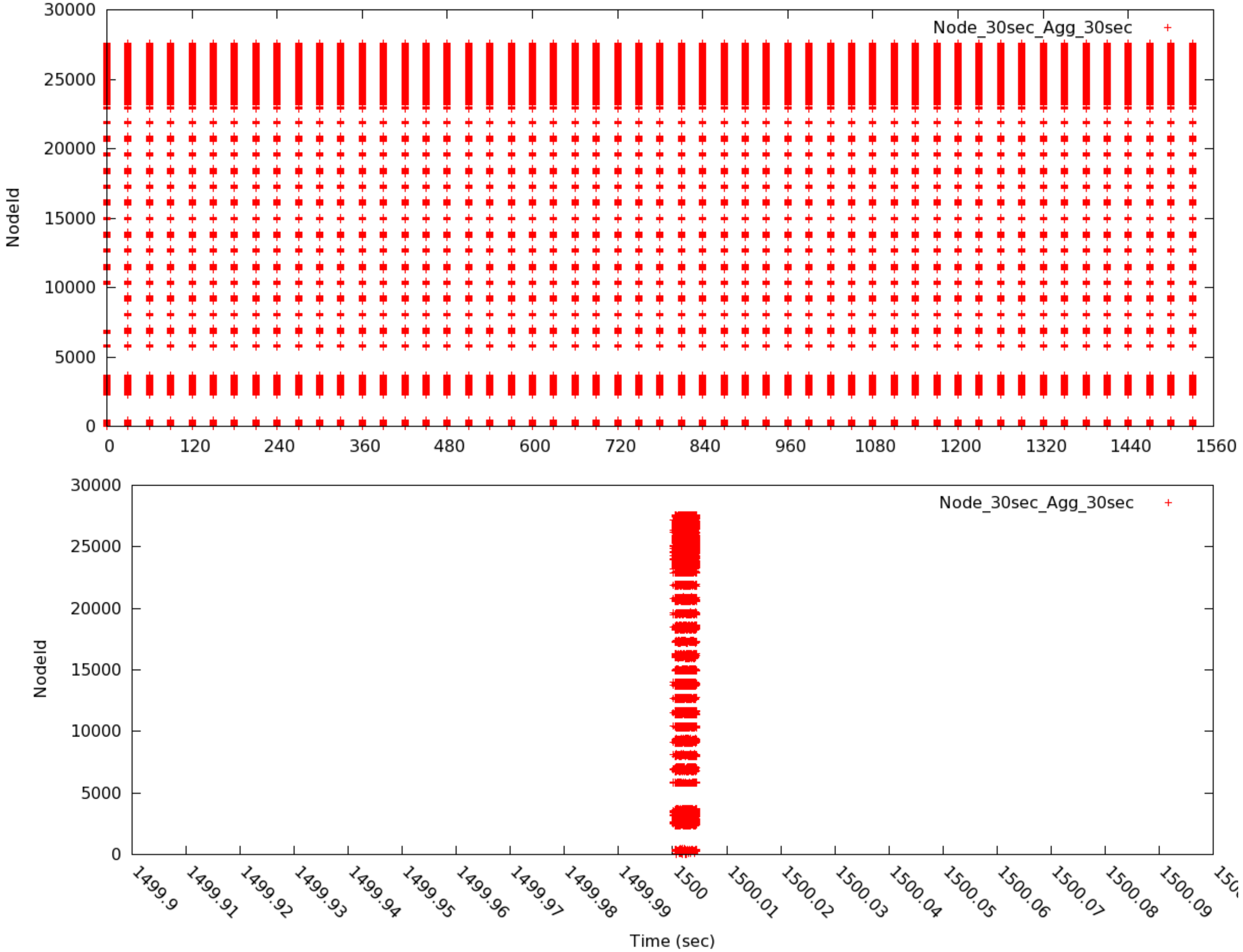
Synchronized sampling across all nodes:

- Enables a coherent system snapshot

Synchronous:

- Variance in collection timestamps ~ 4ms

Note: Clock skew not accounted for



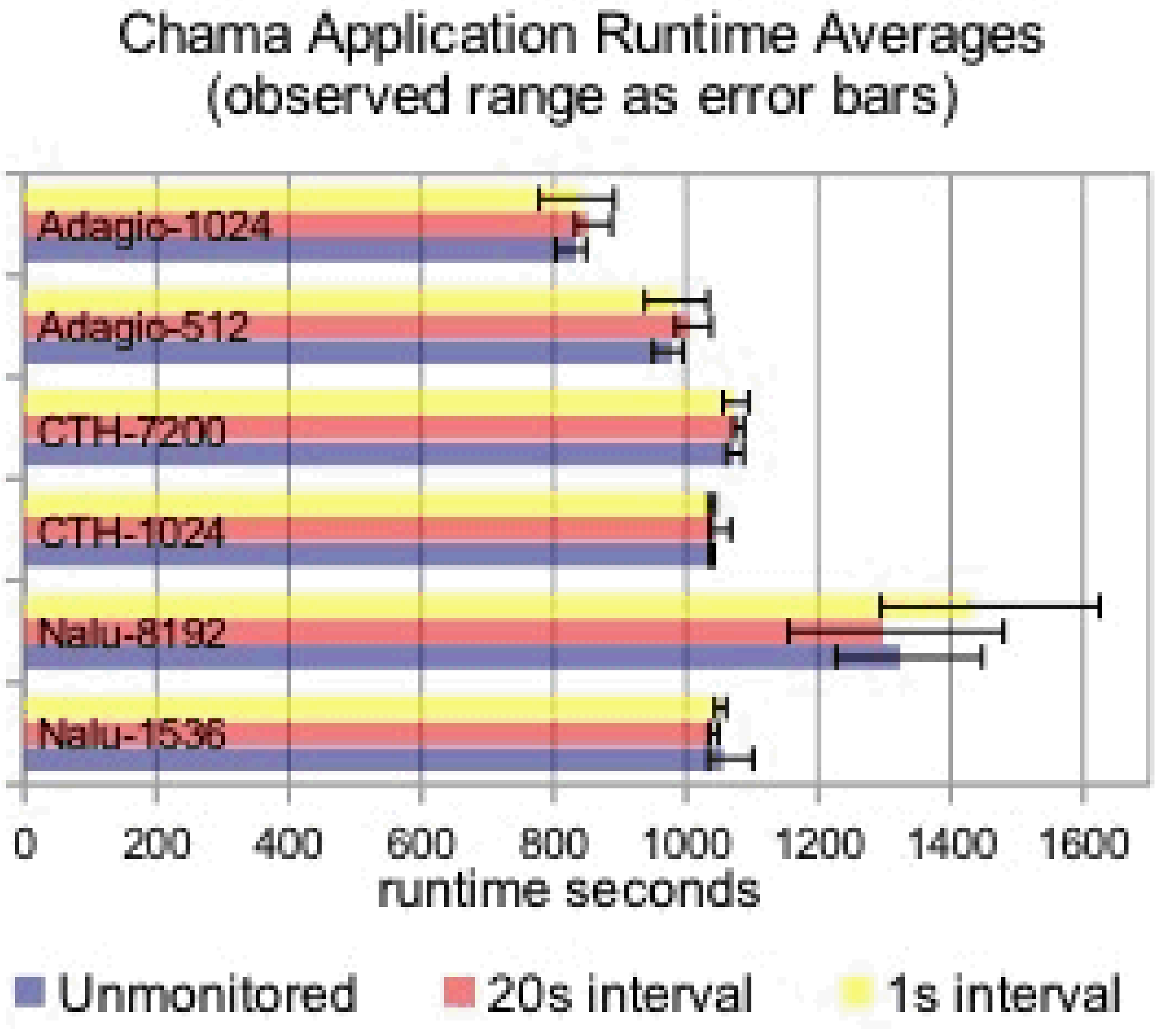
Collection occurrences over 10000 nodes on Blue Waters

CHAMA DAT

Goal: Quantify LDMS impact on application performance on TLCC2 System

- 7 metric sets: lustre_llite, procstatutil, procmeminfo, procnetdev, procnfs, vmstat, sysclassib
- 3 major SNL apps sensitive to network and node noise, some include I/O
- PSNAP

No Discernable Application Impact

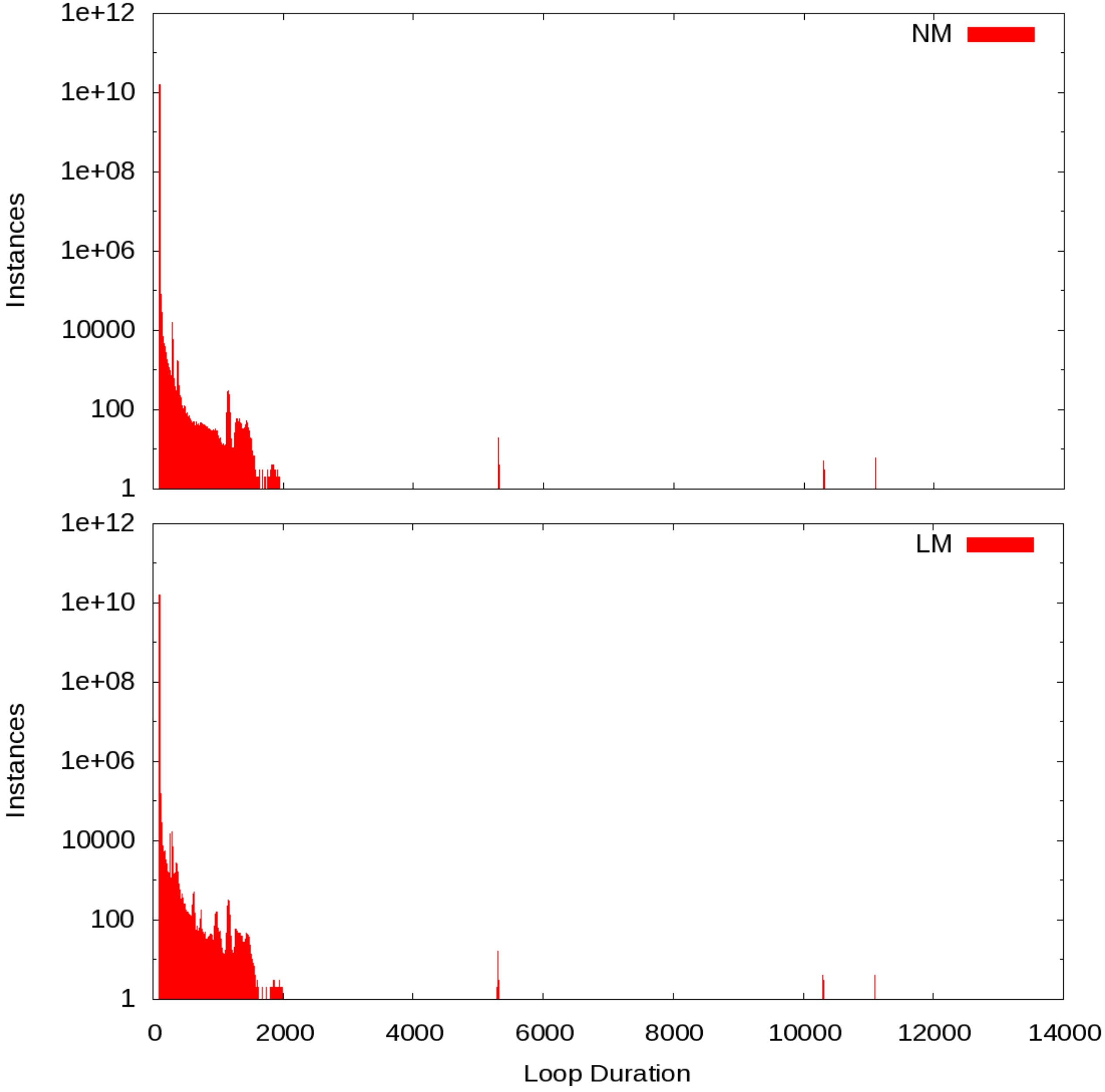


PSNAP:

NM: No Monitoring

LM: Low Monitoring

- 20 sec intervals



NCSA’s Blue Waters DST

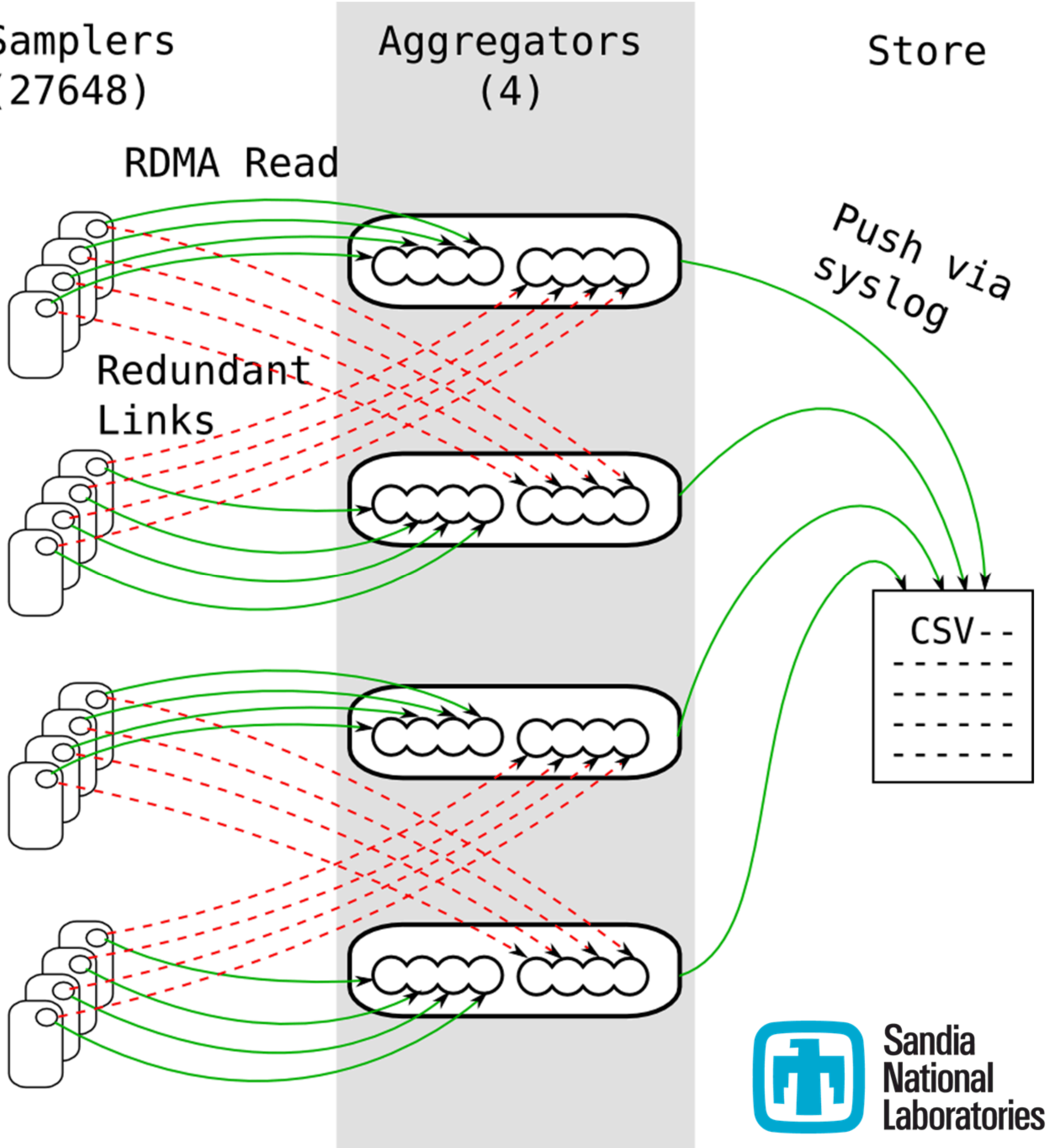
(Large Scale Capability)

Goal: Quantify LDMS impact on application performance

- 1 metric set: cray_system_sampler
- 3 apps sensitive to network and node noise
- PSNAP

Blue Waters Configuration

- All metric sets identical independent of node
  - 194 metrics
- Sample period
  - 60 seconds (normal)
  - 1 second (high)
- Each aggregator primary for 6912 nodes
  - Pull model using RDMA read
- Each aggregator secondary for 6912 nodes
  - RDMA connection established
- In event of failover aggregator collects from 13824 nodes
- Data is pushed to store (MySQL database) using syslog-ng
- One day data set for 60 second collection period contains ~40 million data points per metric and 7.7 billion data points overall

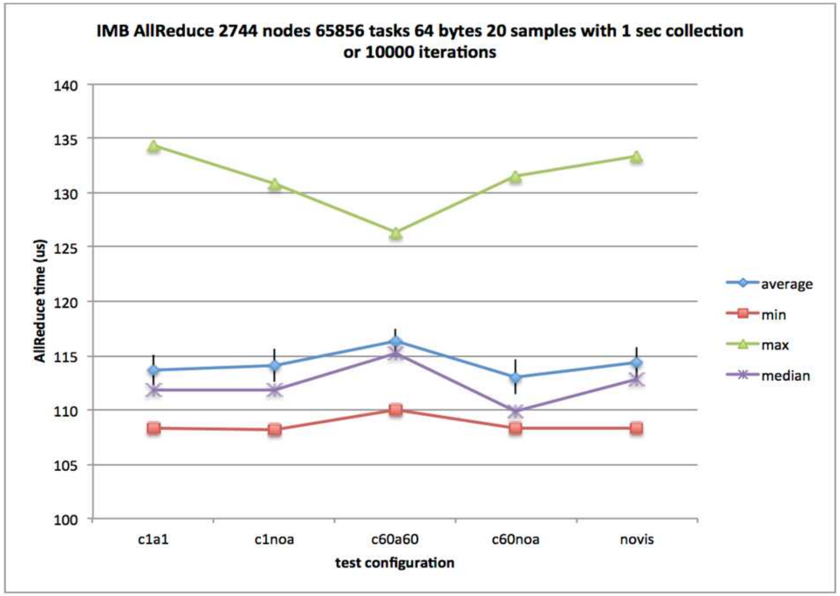


BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Sandia National Laboratories

Impact Testing: Applications

- Intel MPI Benchmark
  - *No correlation of performance with sampling*
- MILC
  - 2774 node run 50 steps
  - 5 phases + Step time
  - *No statistically significant impact*



MILC/CG	novis	c60noa	c60a60	c1noa	c1a1
Ave	5.20e-3	5.21e-3	5.20e-3	5.20e-3	5.19e-3
Min	5.00e-3	5.20e-3	5.00e-3	5.01e-3	5.00e-3
Max	5.43e-3	5.44e-3	5.44e-3	5.45e-3	5.41e-3

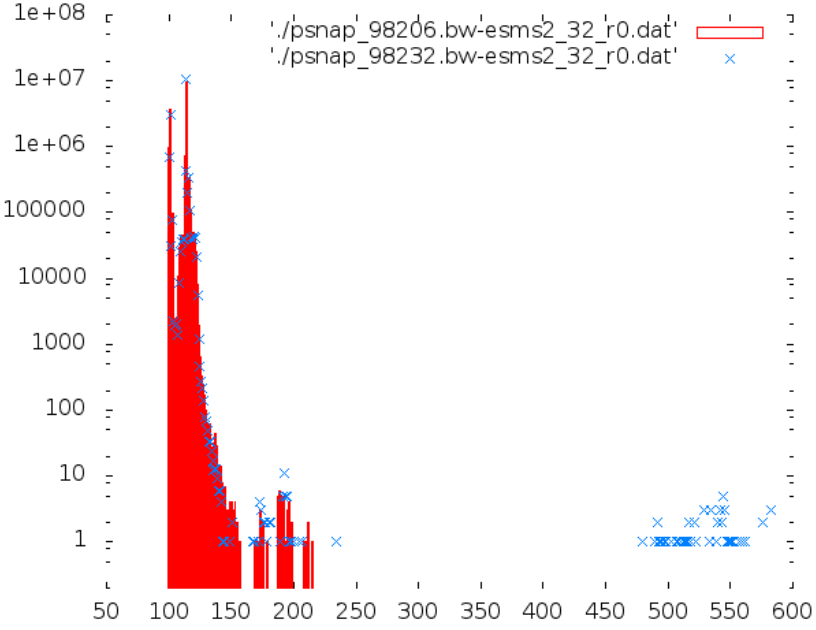
Impact Testing: Applications

- SNL MiniGhost
  - Instrumented for runtime, communication time, time which includes the barrier
  - 8192 nodes, 3 reps
  - *No statistically significant impact*

Total Runtime	**novis**	**c1a1**
Rep1	98.5	92.3
Rep2	95.3	90.2
Rep3	91.8	90.8

Impact Testing: Benchmarks

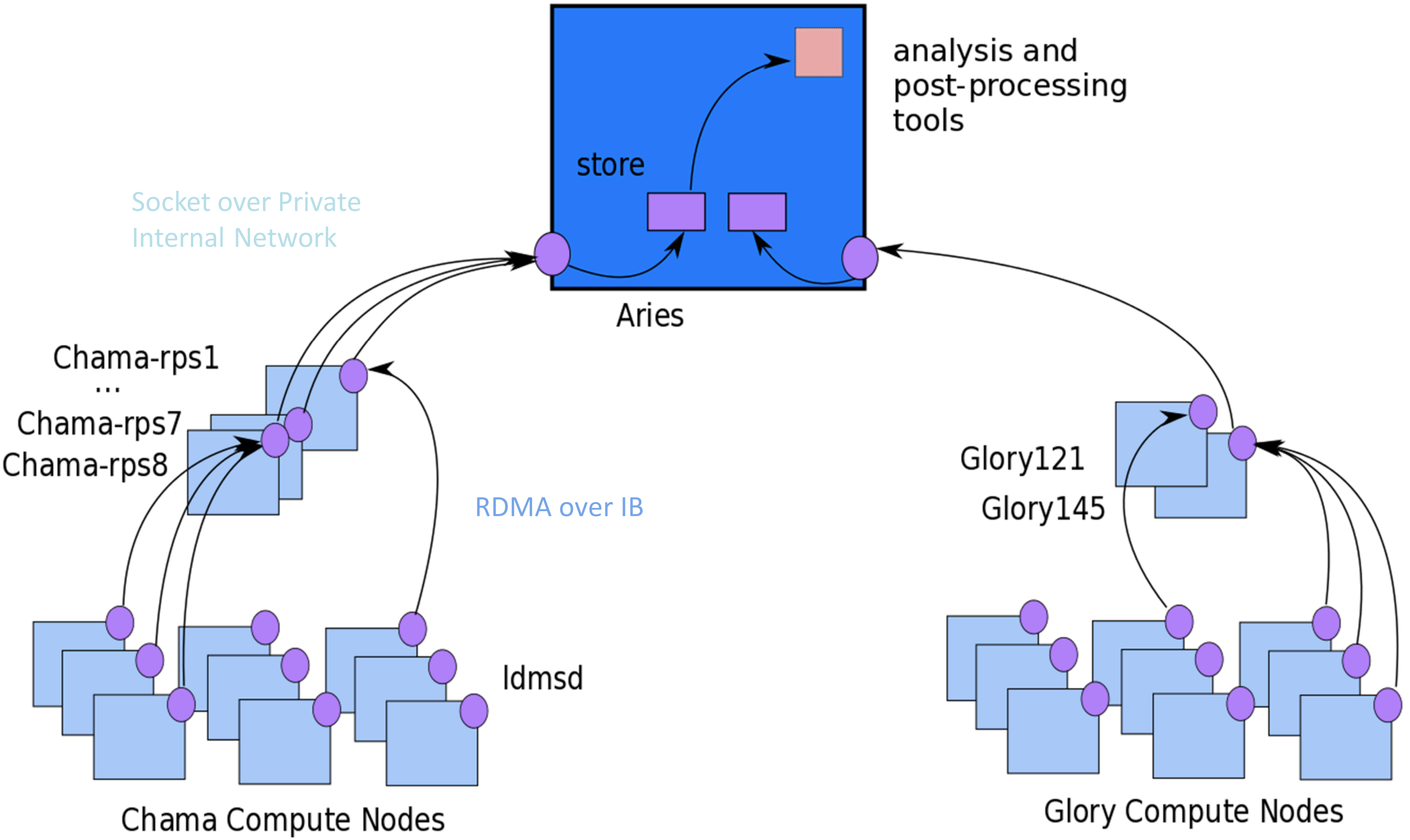
- PSNAP
  - No sampling (red)
  - 1 sec sampling (blue)
  - 60/16M points shifted by sampling time of ~450 usec
  - *Effect on application bounded by synchronized sampling*
- Cray's LinkTest
  - 10,000 iterations of 8kB messages.
  - The no sampling result is 1.74278 ms/packet
  - Sampling result is 20 ns shorter
  - *No significant impact*



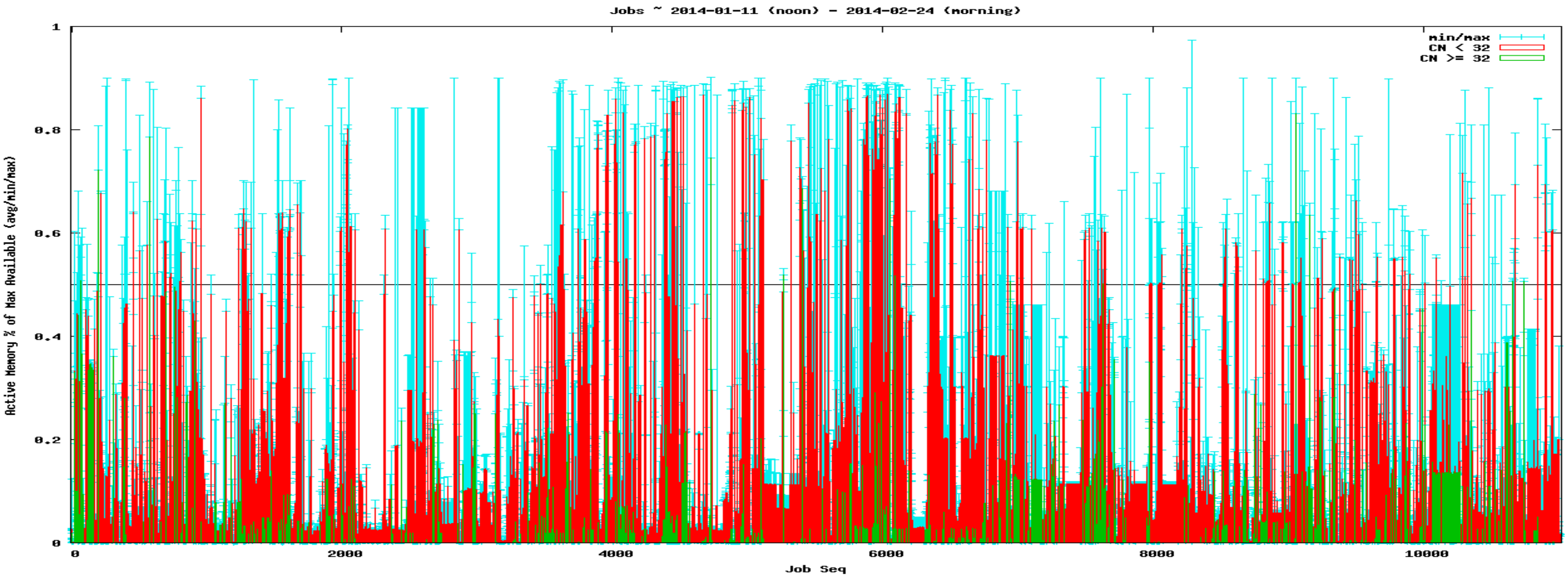
Loop duration (usec)

SNL CAPACITY DEPLOYMENT

Chama: 1232 node TLCC2 cluster (SNL)
Glory: 288 node TLCC1 cluster (SNL)

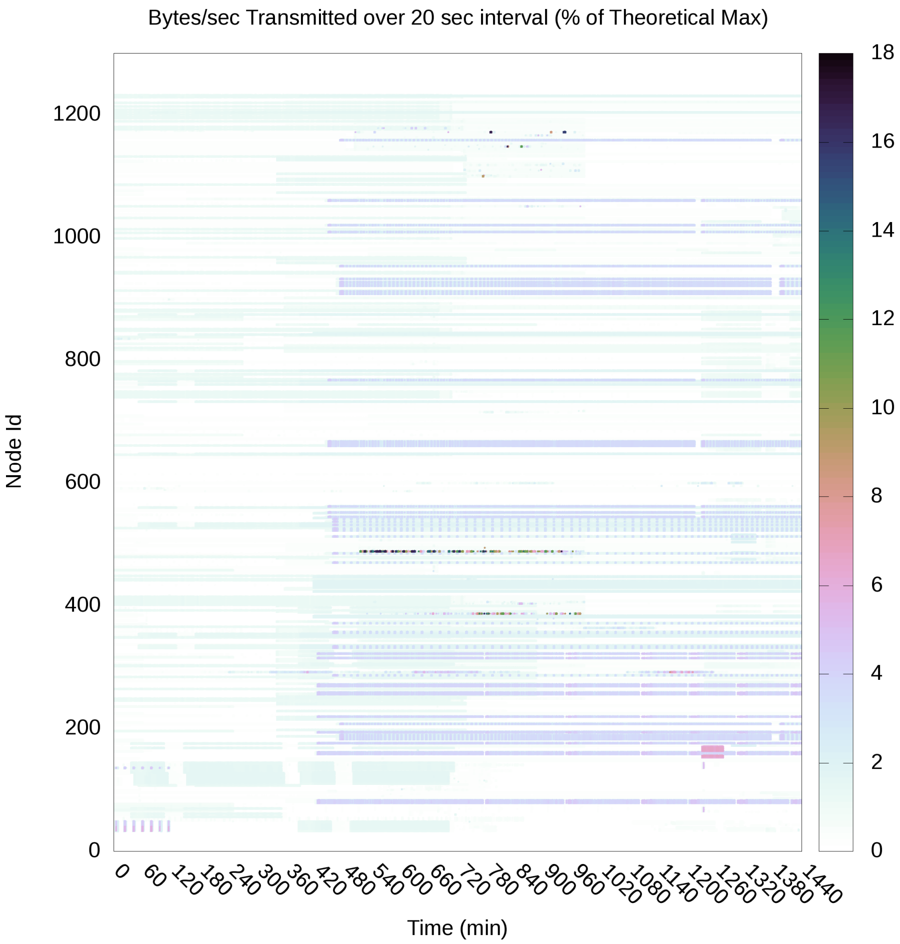


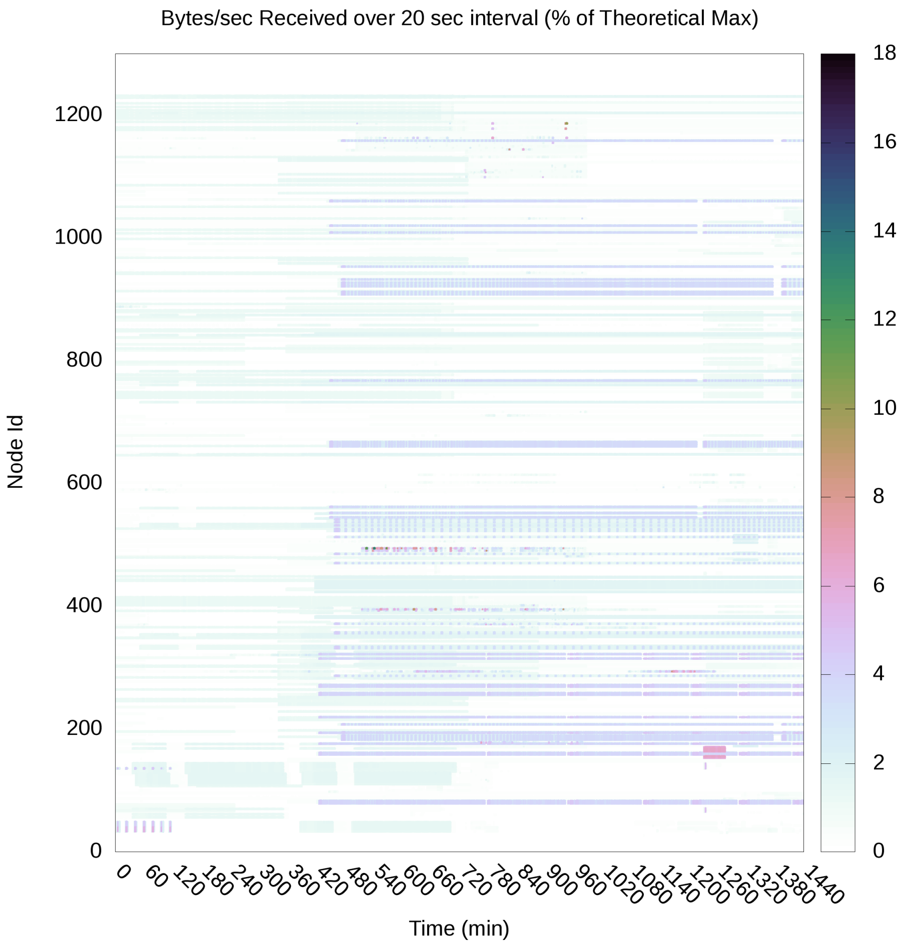
Chama: Memory Use Across All Jobs (30 Days)



- Green represents jobs with greater than or equal to 32 nodes
- Blue (error bars) shows high water mark while green and red are average over job

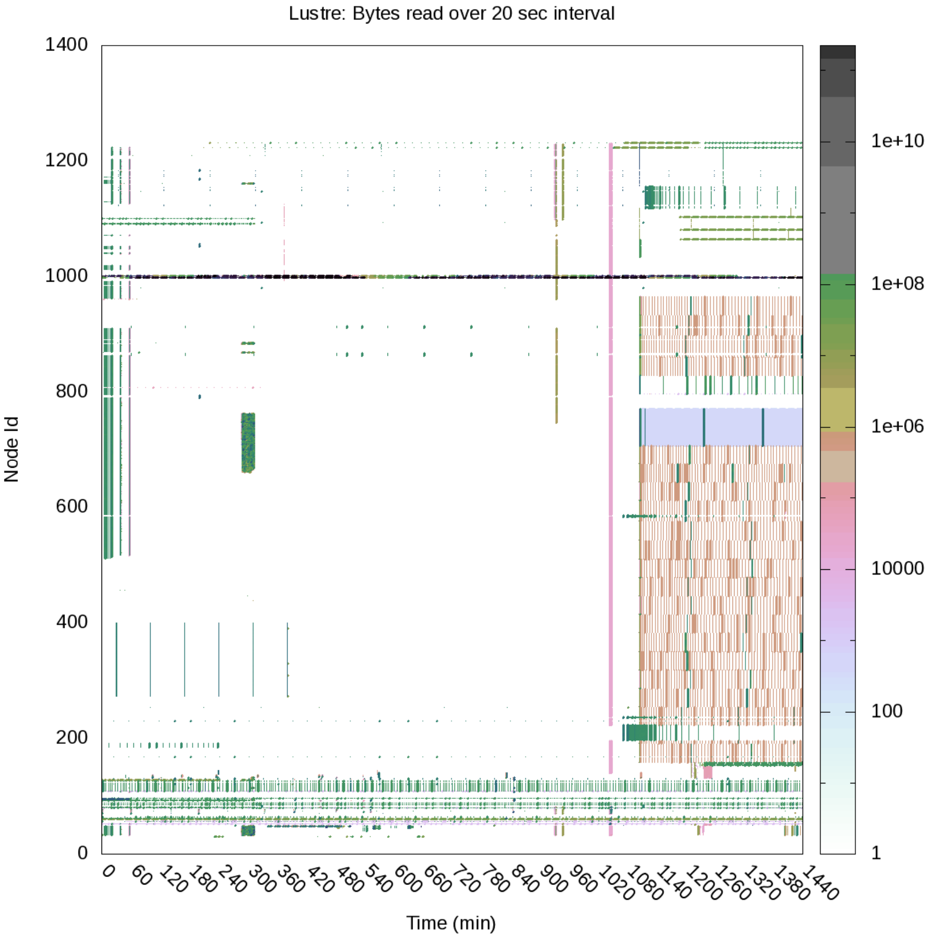
Chama System Maps: IB Bandwidth

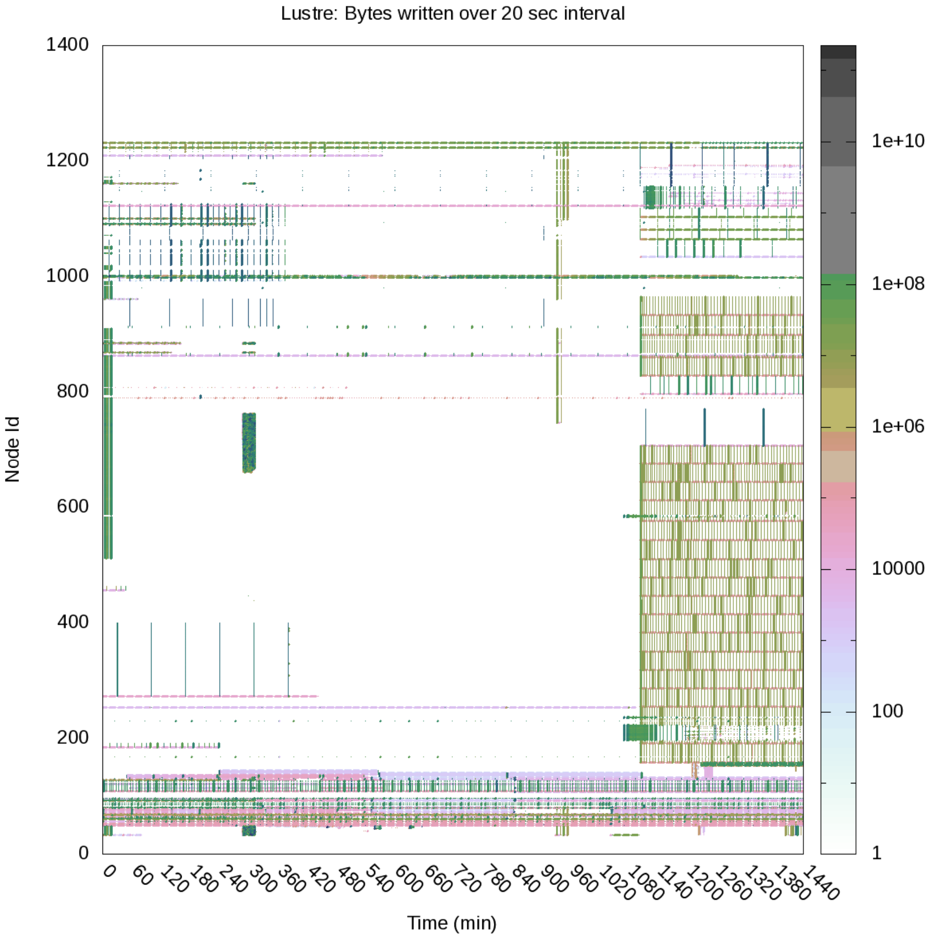

Bytes/sec Transmitted over 20 sec interval (% of Theoretical Max)


Bytes/sec Received over 20 sec interval (% of Theoretical Max)

- IB Transmits and Receives over 24 hours

Chama System Maps: Lustre

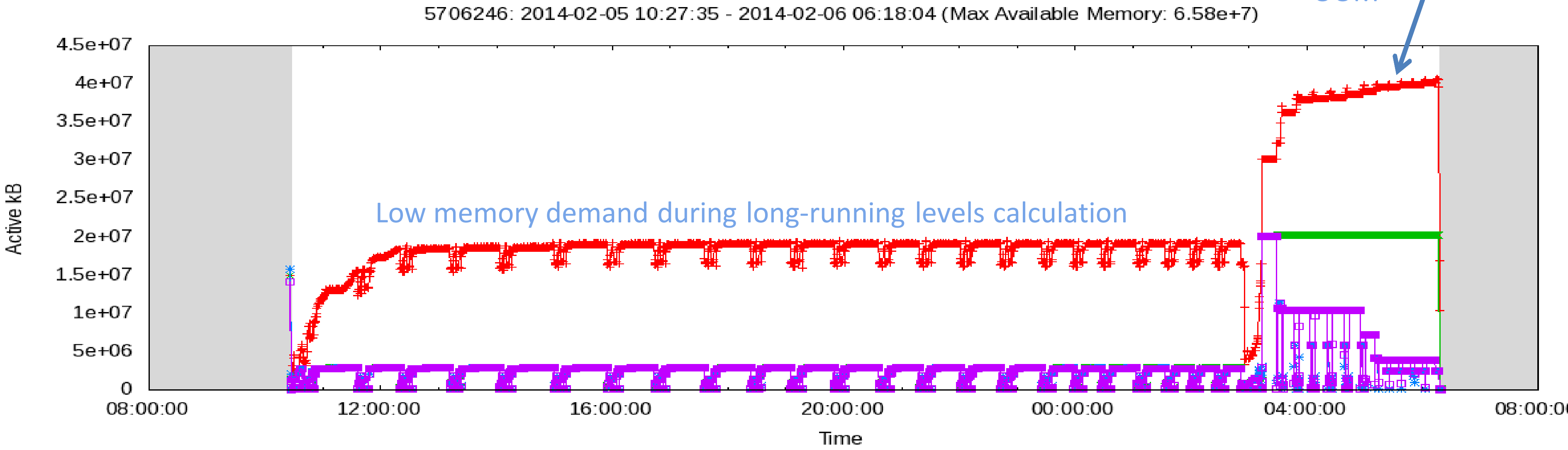




Application Insights

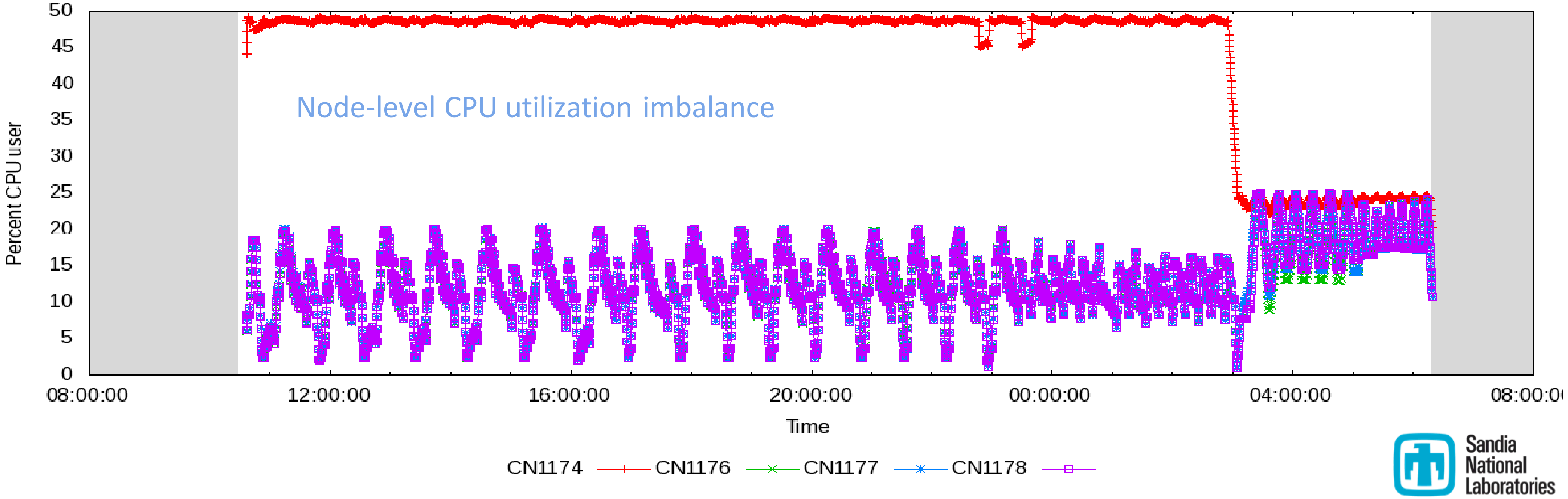
Application Profile: Gaussian

High memory demand during DFT may result in OOM



Low memory demand during long-running levels calculation

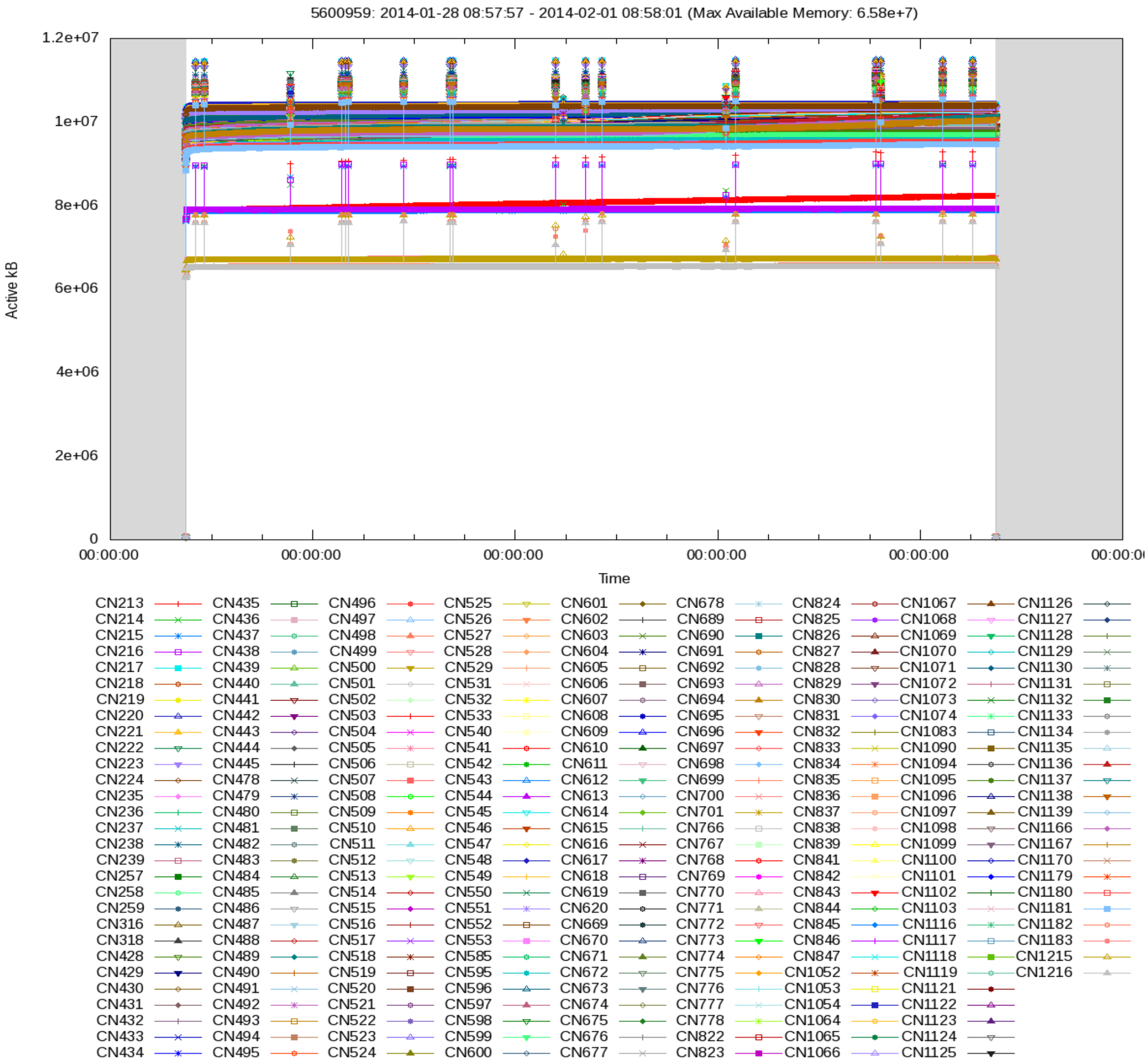
Two-execution phases are potentially separable enabling better application-to-resource mapping



Node-level CPU utilization imbalance

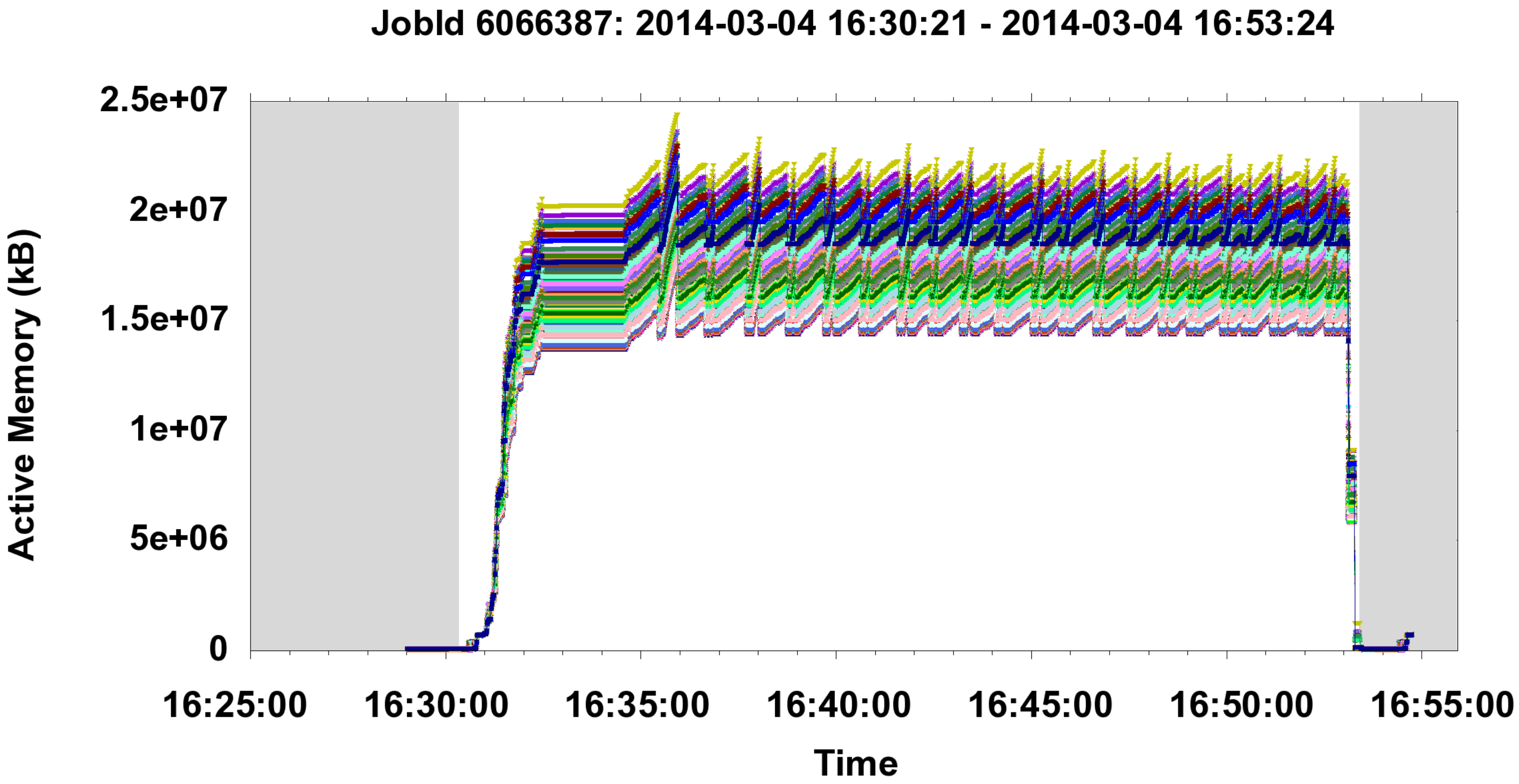
Sandia National Laboratories

Application Profile: LAMMPS



- Generally well balanced in memory usage
- Running on fewer nodes can increase memory usage (17%->25%)
- However application is CPU bound, running nearly 100% user time on all cores.

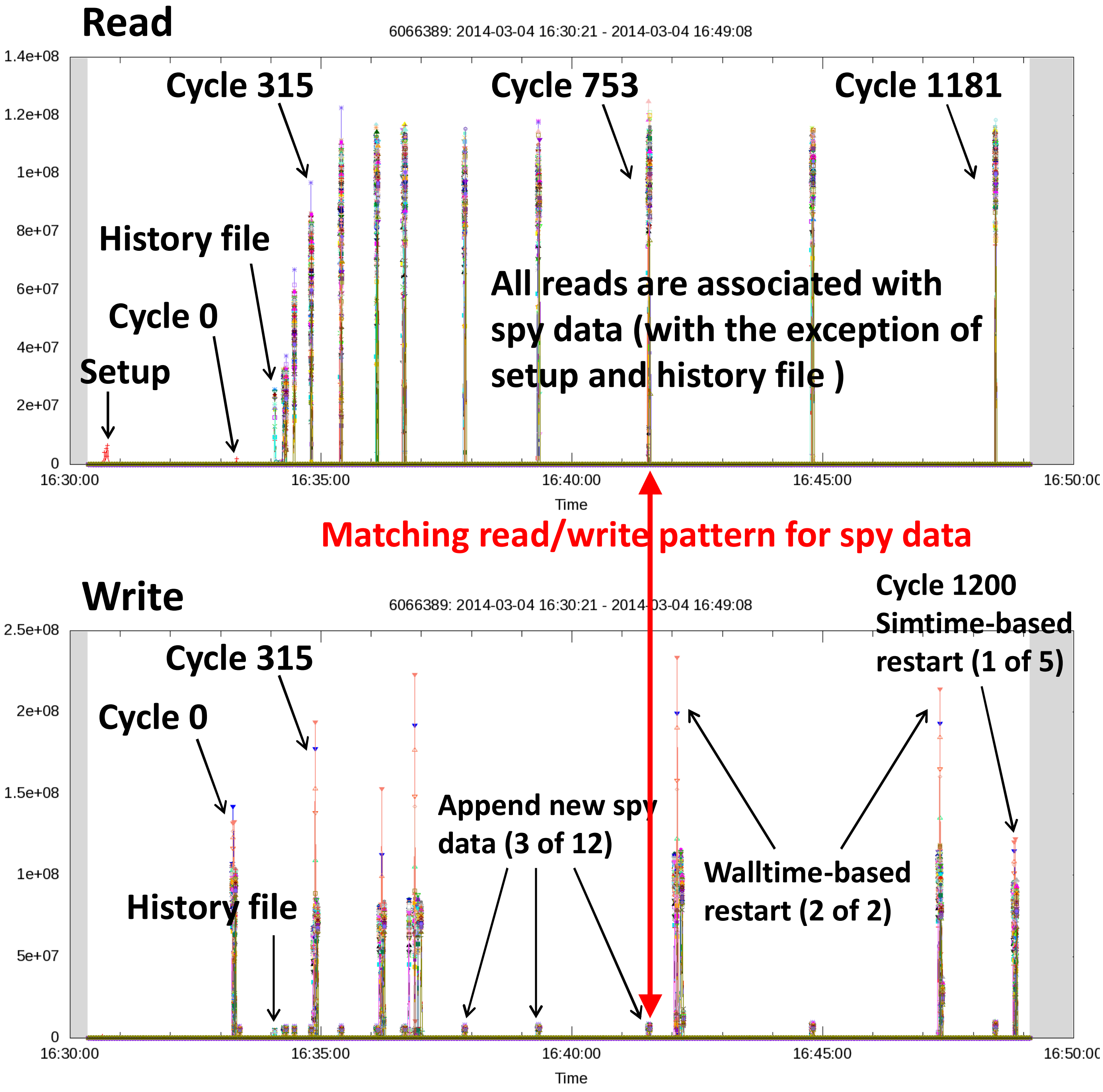
NALU



Memory Imbalance: Variation is 10% of the node’s total memory.

CTH

- Writes are preceded by significant Reads
- FS performance can affect application performance



Read

Cycle 315

Cycle 753

Cycle 1181

History file

Cycle 0

Setup

All reads are associated with spy data (with the exception of setup and history file)

Matching read/write pattern for spy data

Write

Cycle 1200 Simtime-based restart (1 of 5)

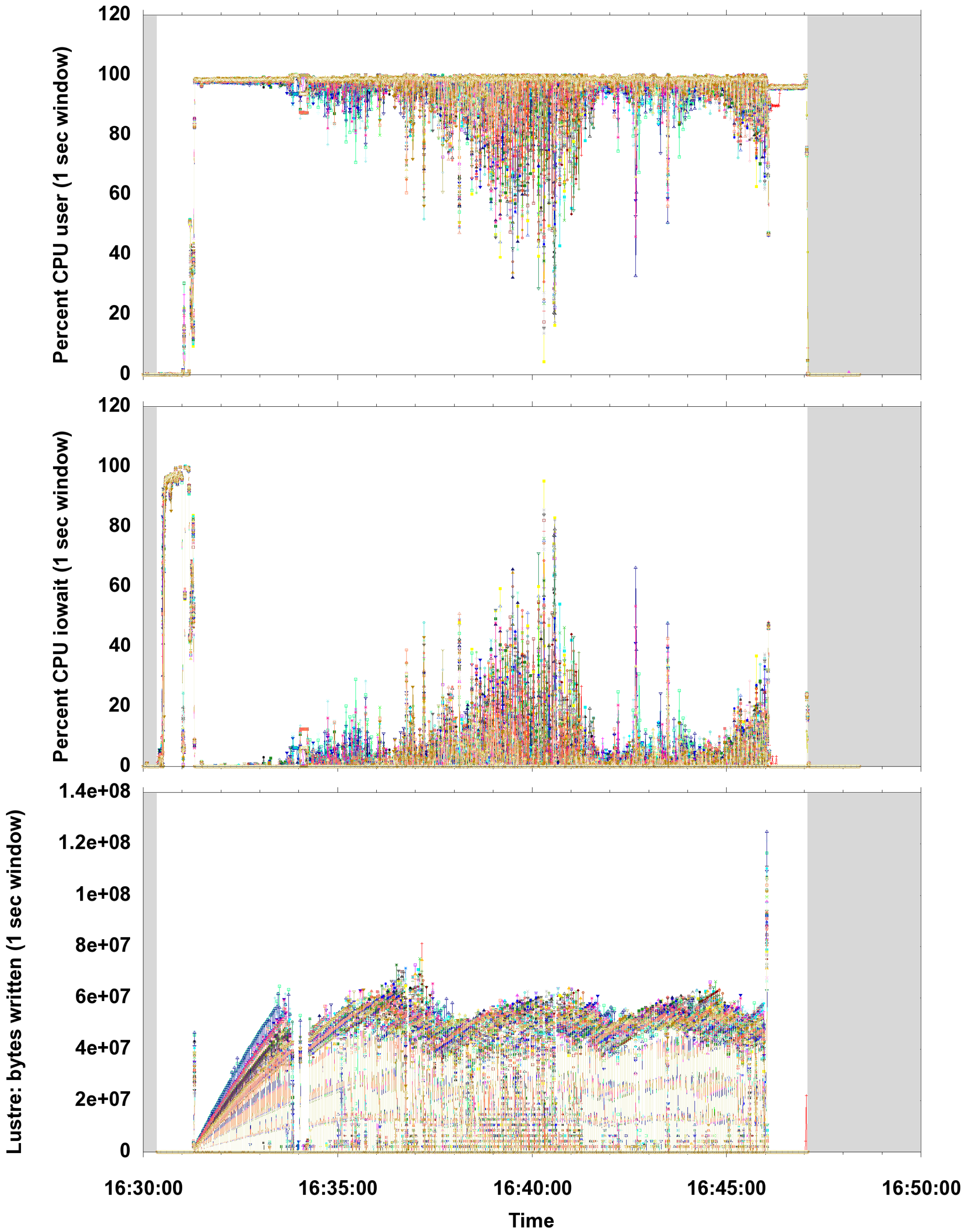
Cycle 315

Cycle 0

Append new spy data (3 of 12)

Walltime-based restart (2 of 2)

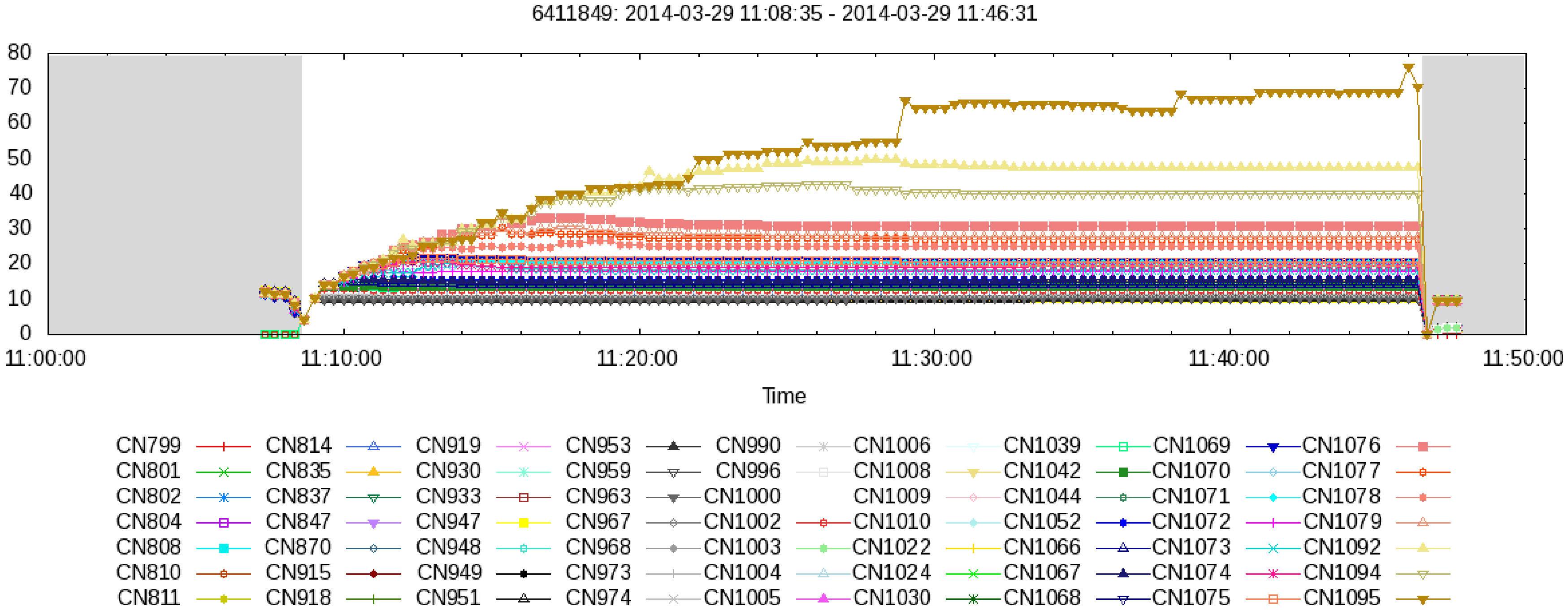
History file

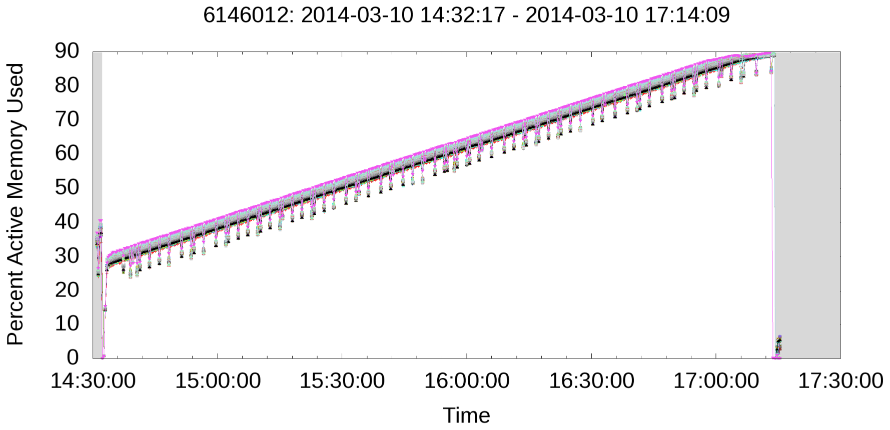


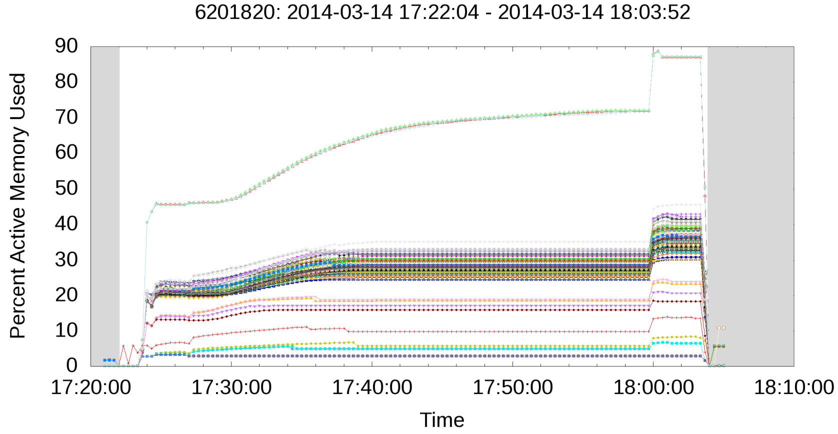
Adagio

- Significant time spent in I/O wait, rather than computation, due to constant writes

OOM Profiles

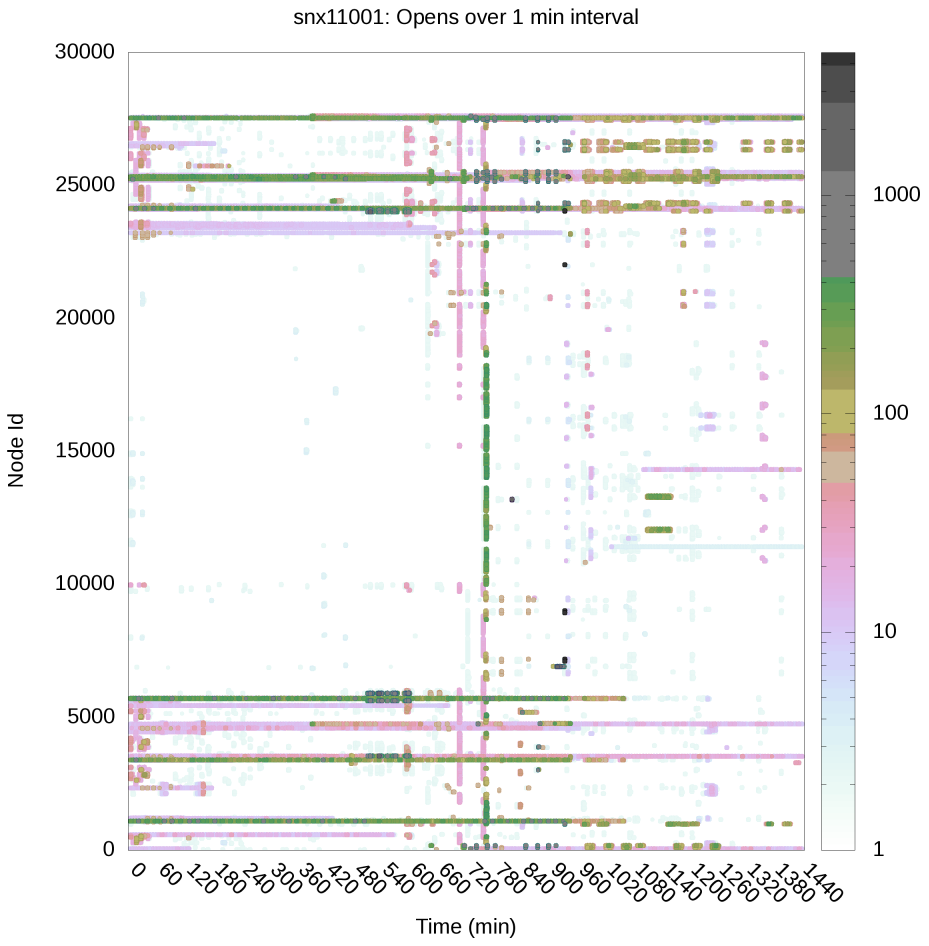

6411849: 2014-03-29 11:08:35 - 2014-03-29 11:46:31

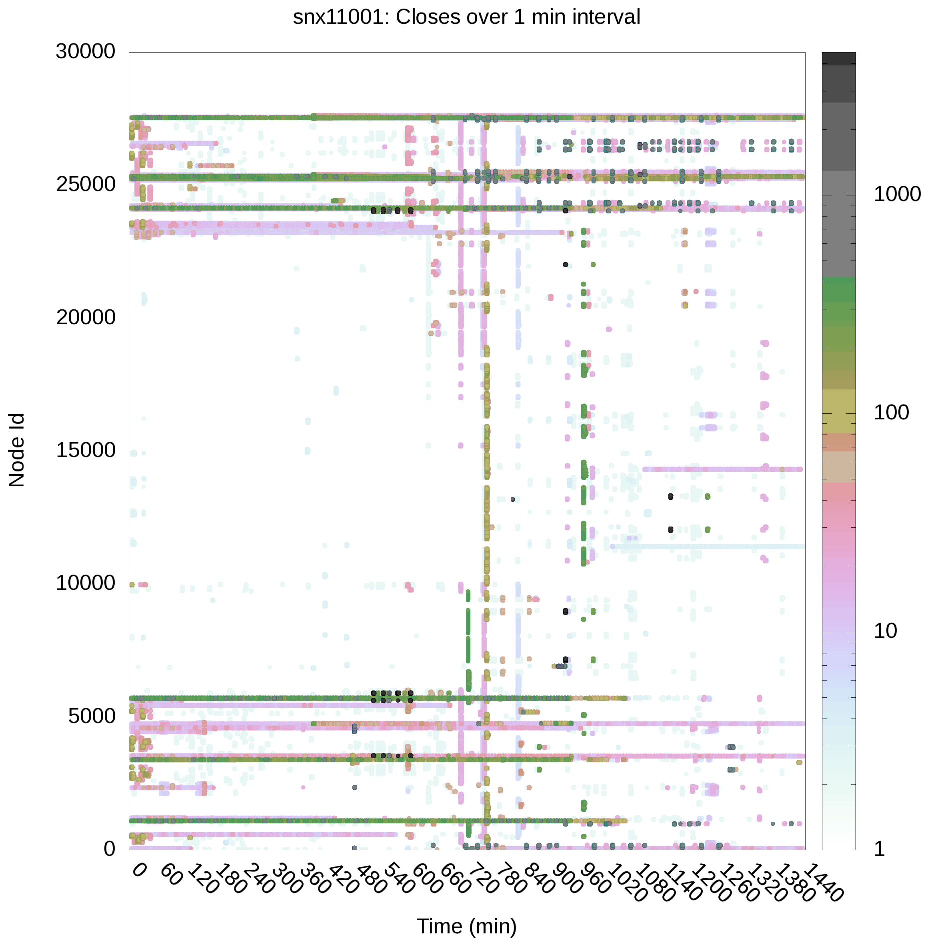

6146012: 2014-03-10 14:32:17 - 2014-03-10 17:14:09


6201820: 2014-03-14 17:22:04 - 2014-03-14 18:03:52

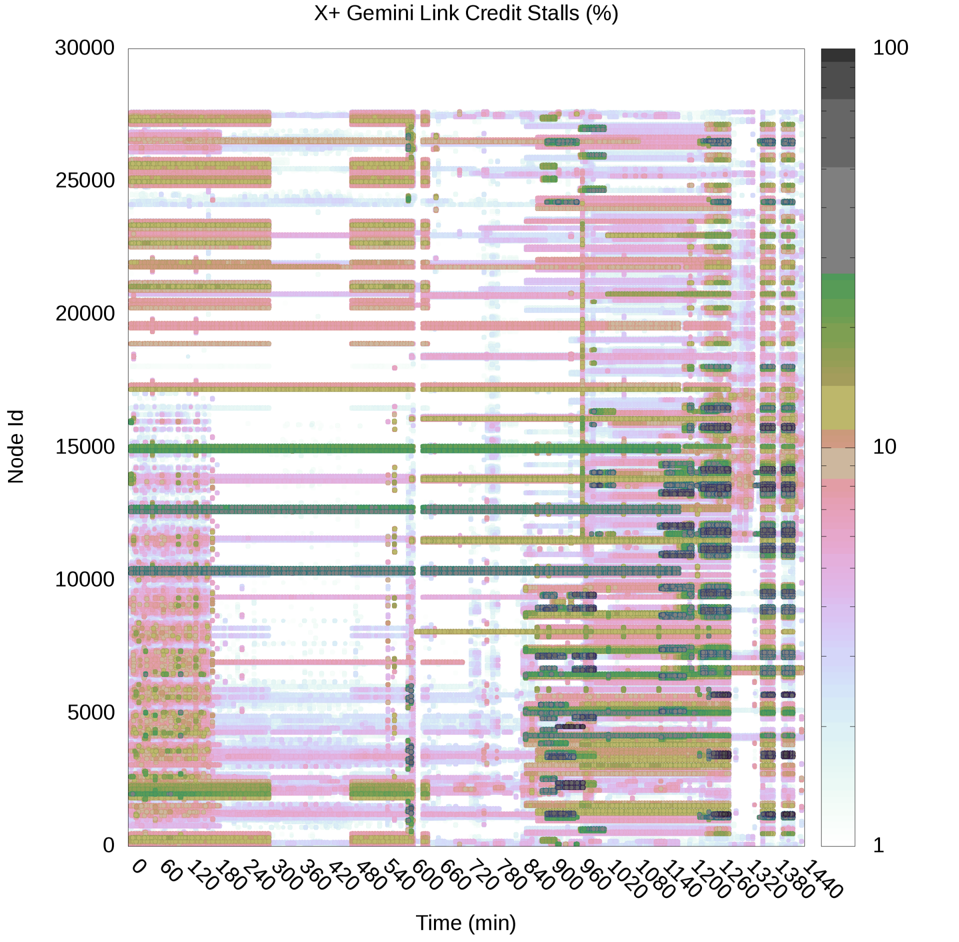
NCSA BLUE WATERS DEPLOYMENT

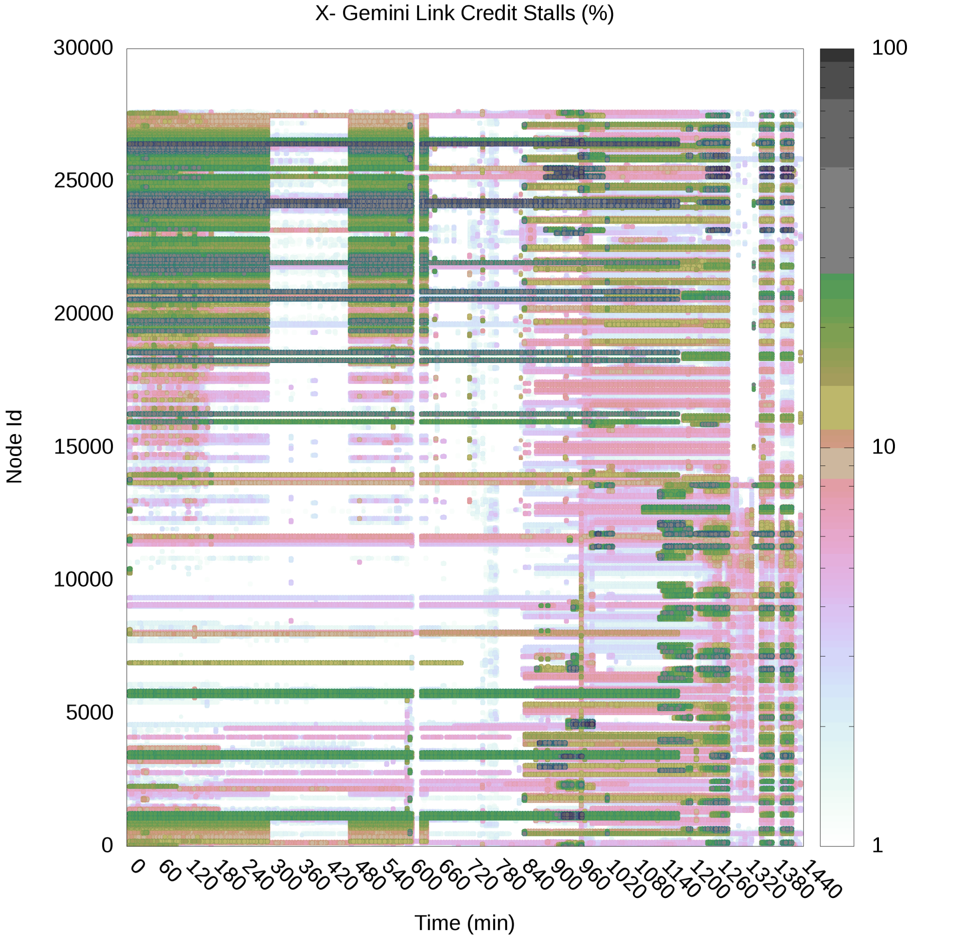
Lustre Opens/Closes



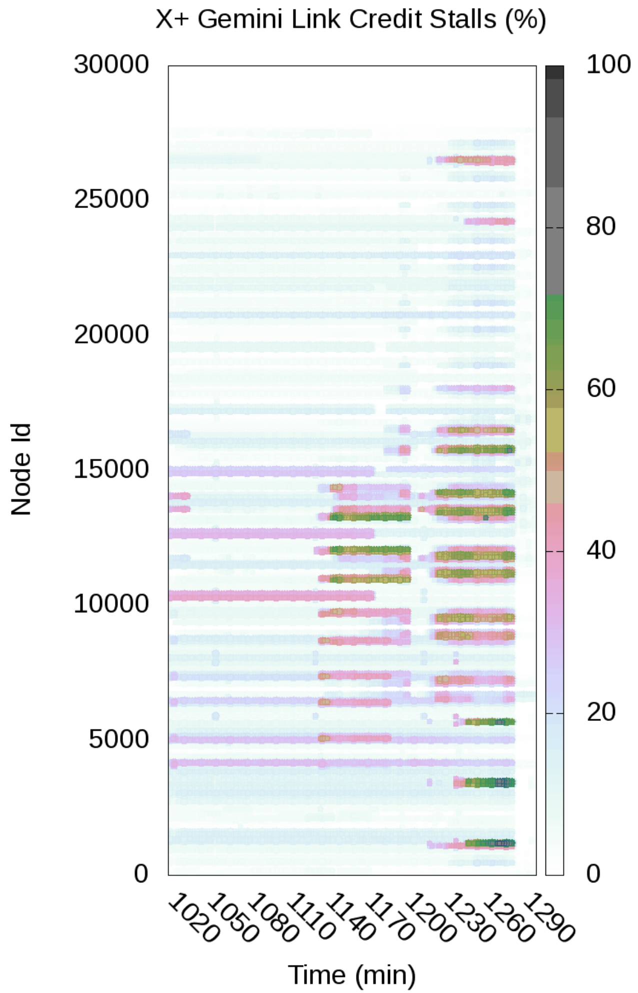


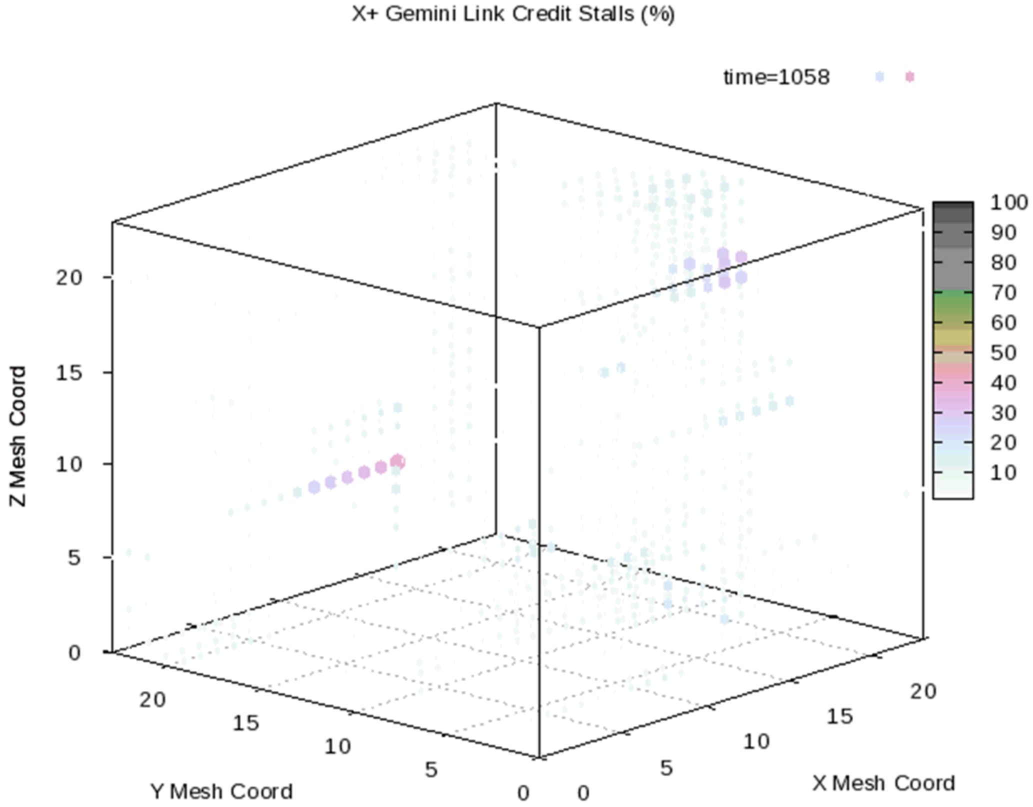
HSN Output Stalls (X)





Mesh Topology Representation
X+ Animation: 4 hrs @ 1059





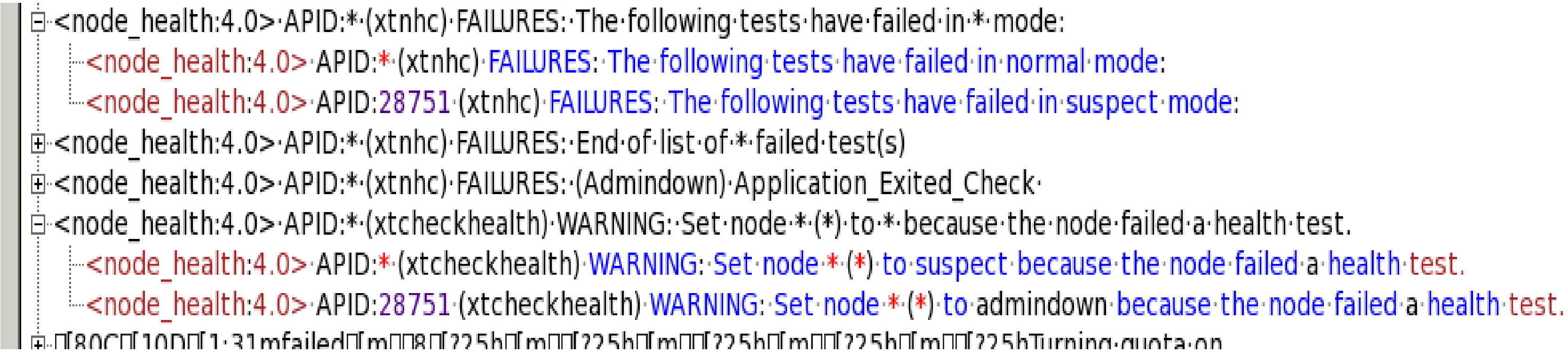
Related Software In Progress

Baler

- Parallel log file analysis
  - Automated pattern discovery and grouping
    - Discovers similarities and bundles log messages for human interaction and automated correlation
- Text search enables discovery of text strings of interest
- Visualization enables discovery of patterns of interest
  - Global pattern_id
  - Meta-clustering
  - Web GUI
  - Association rule mining for combined log and binned metric data

Baler

- Converts log message to patterns for data reduction and event analysis
- Discovers patterns in messages (* = variable data)



```
⊟<node_health:4.0>·APID:*·(xtnhc)·FAILURES:·The·following·tests·have·failed·in·*·mode:
   <node_health:4.0>·APID:*·(xtnhc)·FAILURES:·The·following·tests·have·failed·in·normal·mode:
   <node_health:4.0>·APID:28751·(xtnhc)·FAILURES:·The·following·tests·have·failed·in·suspect·mode:
⊞<node_health:4.0>·APID:*·(xtnhc)·FAILURES:·End·of·list·of·*·failed·test(s)
⊞<node_health:4.0>·APID:*·(xtnhc)·FAILURES:·(Admindown)·Application_Exited_Check·
⊟<node_health:4.0>·APID:*·(xtcheckhealth)·WARNING:·Set·node·*·(*)·to·*·because·the·node·failed·a·health·test.
   <node_health:4.0>·APID:*·(xtcheckhealth)·WARNING:·Set·node·*·(*)·to·suspect·because·the·node·failed·a·health·test.
   <node_health:4.0>·APID:28751·(xtcheckhealth)·WARNING:·Set·node·*·(*)·to·admindown·because·the·node·failed·a·health·test.
⊞□[80C□[10D□[1;31mfailed□[m□□8□[?25h□[m□□[?25h□[m□□[?25h□[m□□[?25h□[m□□[?25hTurning·quota·on
```

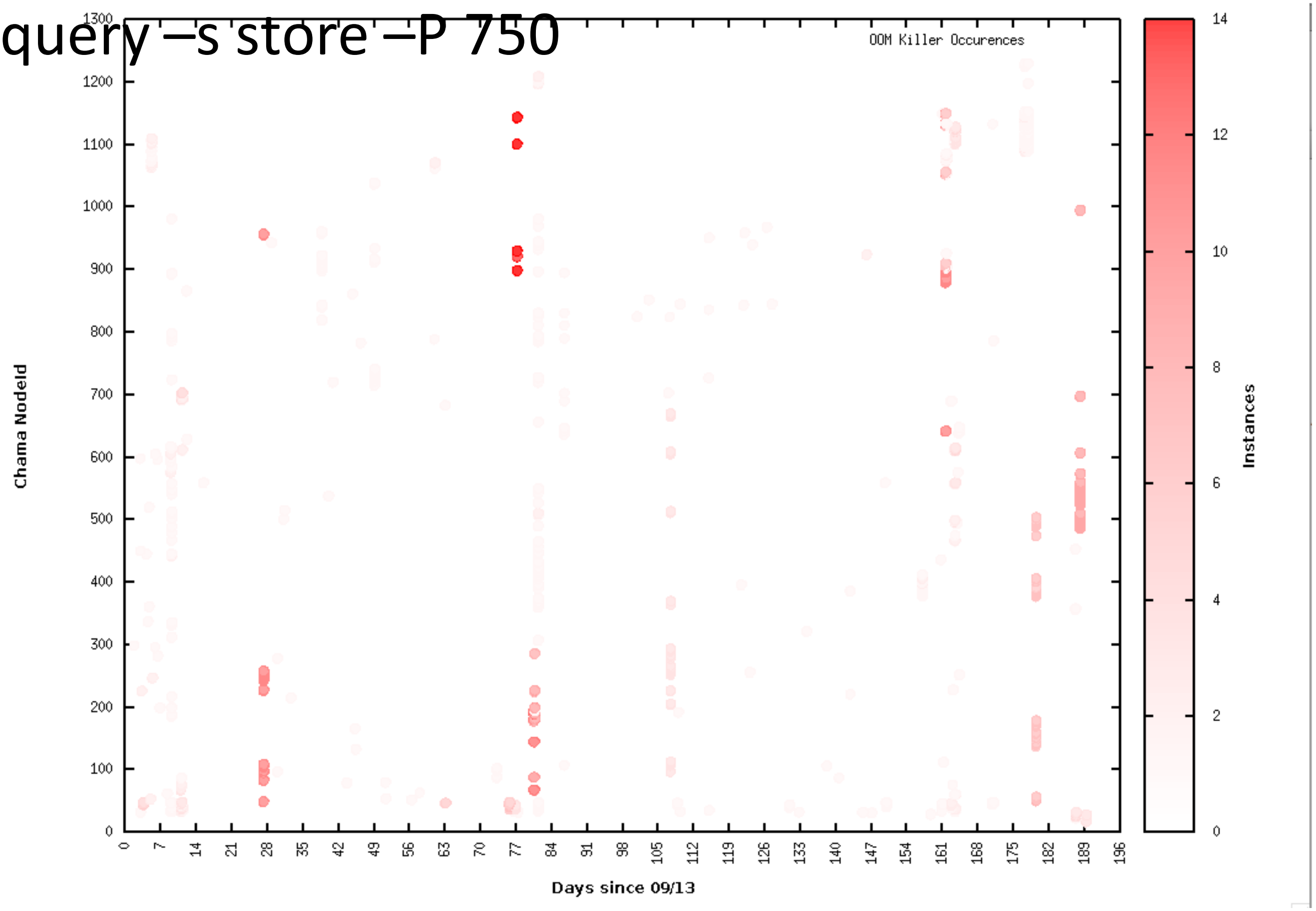
- Chama:
  - 1200 Nodes
  - 3 months of logs = 81 million lines
  - 35 min processing time
  - 60K unique 1st level patterns; few thousand higher level patterns

Baler (cont’d): Query tools and Plots

- bquery –s store –t PTN

750 kernel: [*.*] Memory * out of memory: Kill process * (*) score * or sacrifice child

- bimg_query –s store –P 750

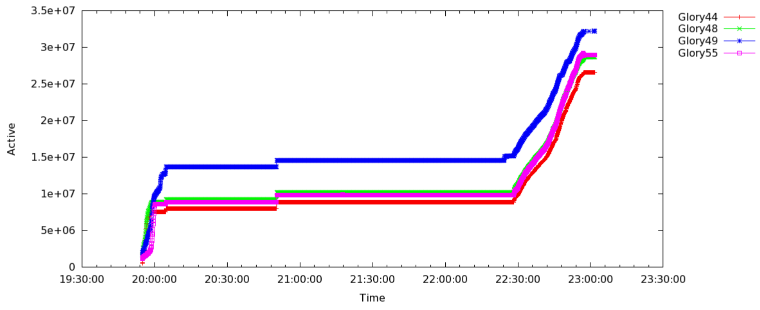


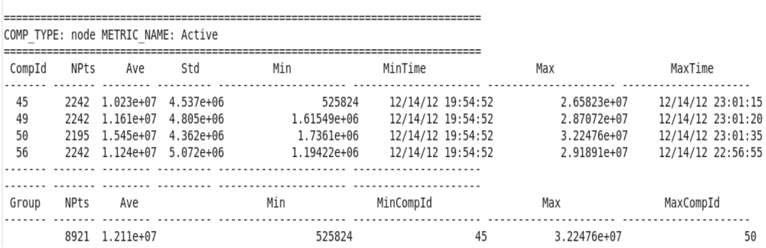
Web GUI Support

- Web GUI support for data browsing by both admins and users
  - Raw data with appropriate processing (e.g. differences for counter data)
  - Utilize Slurm/Moab logs to present job centric views
  - Time series graphs
  - Component layout
  - Statistics (e.g. min, max, mean, std dev)
    - Job, node list, cluster

Questions?

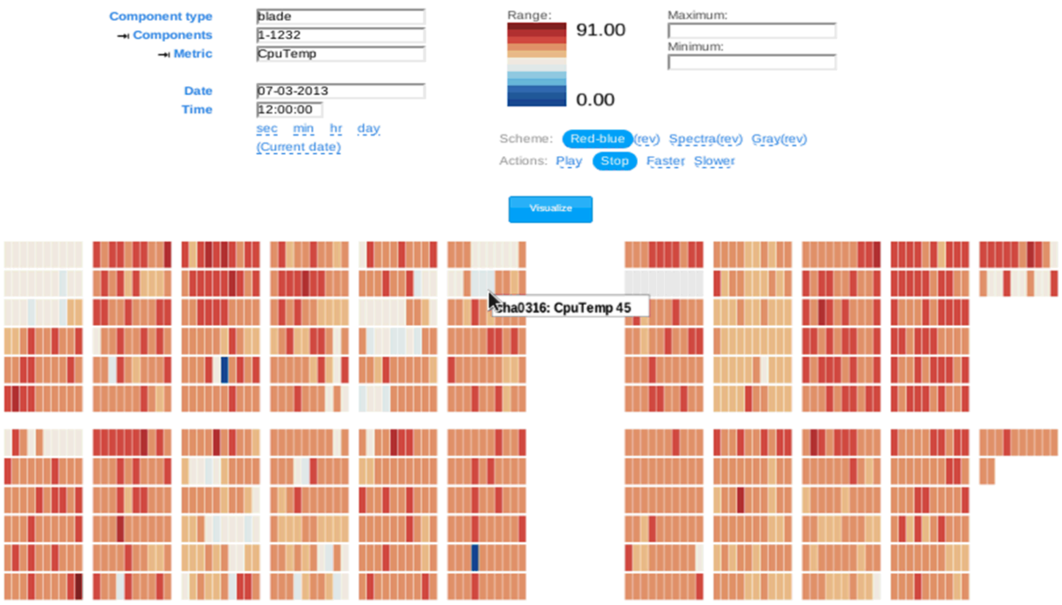
Current Analysis and Post Processing





```
============================================================================
COMP_TYPE: node METRIC_NAME: Active
============================================================================
 CompId   NPts    Ave      Std            Min              MinTime              Max            MaxTime
------- ------- -------- -------- -------------------- -------------------- -------------------- --------------------
  45     2242  1.023e+07  4.537e+06           525824     12/14/12 19:54:52       2.65823e+07    12/14/12 23:01:15
  49     2242  1.161e+07  4.805e+06      1.61549e+06     12/14/12 19:54:52       2.87072e+07    12/14/12 23:01:20
  50     2195  1.545e+07  4.362e+06       1.7361e+06     12/14/12 19:54:52       3.22476e+07    12/14/12 23:01:35
  56     2242  1.124e+07  5.072e+06      1.19422e+06     12/14/12 19:54:52       2.91891e+07    12/14/12 22:56:55
------- ------- -------- -------- -------------------- --------------------
------- ------- -------- -------- -------------------- --------------------
 Group   NPts    Ave                  Min              MinCompId             Max              MaxCompId
------- ------- -------- -------- -------------------- -------------------- -------------------- --------------------
         8921  1.211e+07                 525824                   45       3.22476e+07                    50
```

Post-Job Stats and Plots

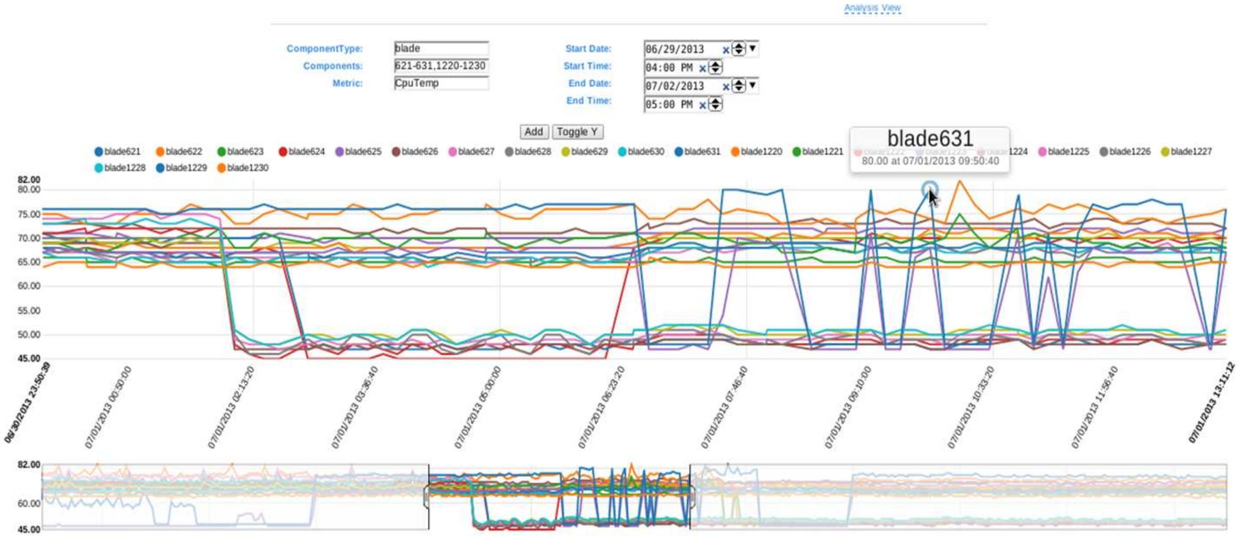


Component type: blade
Components: 1-1232
Metric: CpuTemp

Date: 07-03-2013
Time: 12:00:00

Range: 91.00 – 0.00
Maximum:
Minimum:

Scheme: Red-blue (rev) Spectra(rev) Gray(rev)
Actions: Play Stop Faster Slower



Analysis View

ComponentType: blade
Components: 621-631,1220-1230
Metric: CpuTemp

Start Date: 06/29/2013
Start Time: 04:00 PM
End Date: 07/02/2013
End Time: 05:00 PM

blade631
80.00 at 07/01/2013 09:50:40

Interactive Web Interface:
System Layout and Time Series