SAND2016-5127C

Lightweight Distributed Metric Service

Production overview and development status

June, 2016

Presented by Benjamin Allan

Sandia National Laboratories

Exceptional service in the national interest

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. 2014-3659C

Acknowledgements (developers and leading users in 2016)

Nichamon Naksinehaboon, Narate Taerat, Tom Tucker (OGC)

Jim Brandt, Ann Gentile, Ben Allan, Sophia Lefantzi, Ed Walsh (SNL)

Open Grid Computing, Austin, TX

Sandia National Laboratories, Albuquerque, NM

NCSA, Urbana-Champaign, IL

Los Alamos National Laboratories, Los Alamos, NM

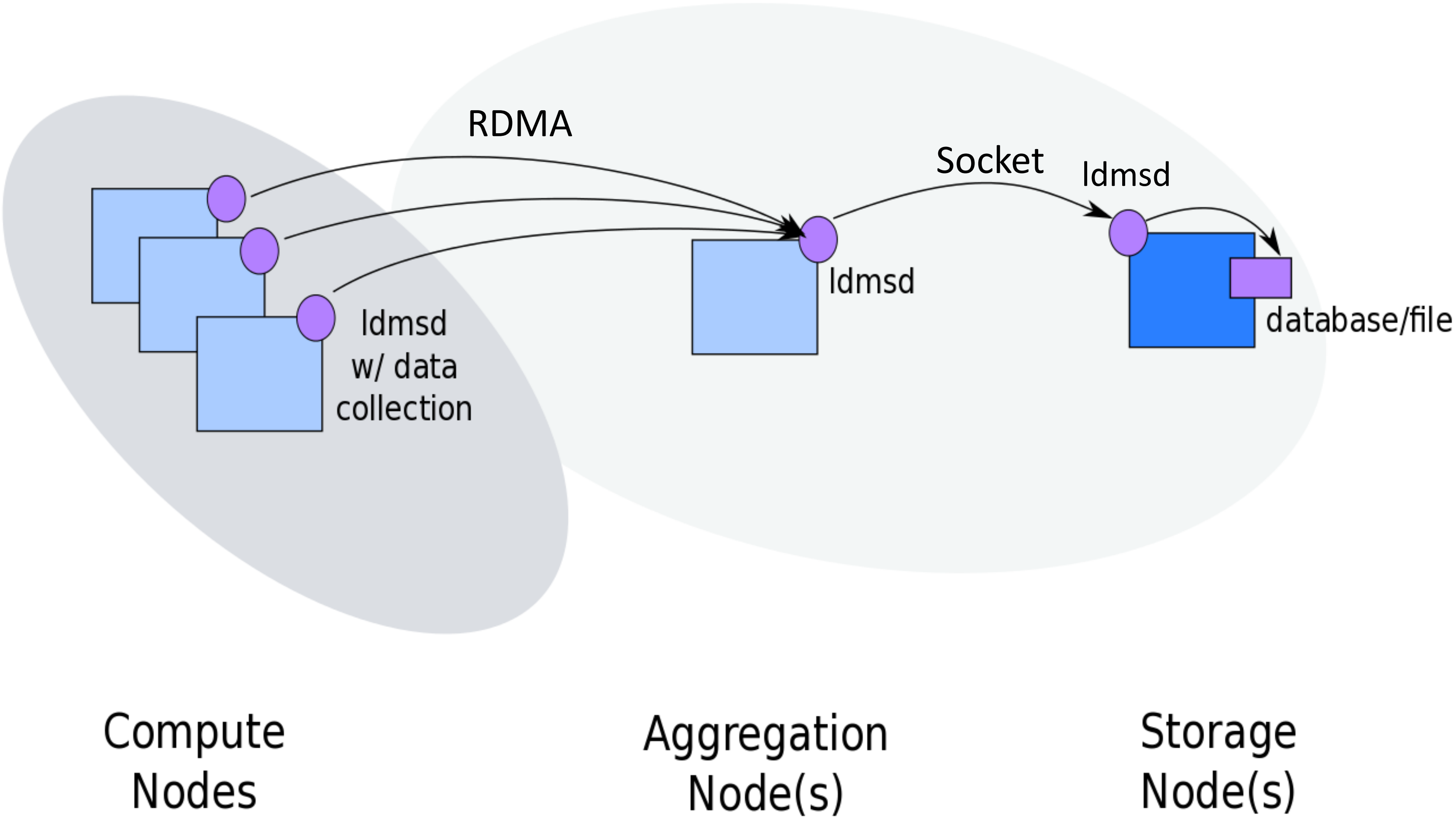
Lawrence Livermore National Laboratory, Livermore, CA

Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

Outline

- What is LDMS
- New live analysis capabilities
- New job scoring capabilities
- Development status

In-band Data Collection, Transport, and Storage



Sandia
National
Laboratories

What is LDMS?

An open framework for scalable lightweight data collection from High Performance Computer (HPC) systems

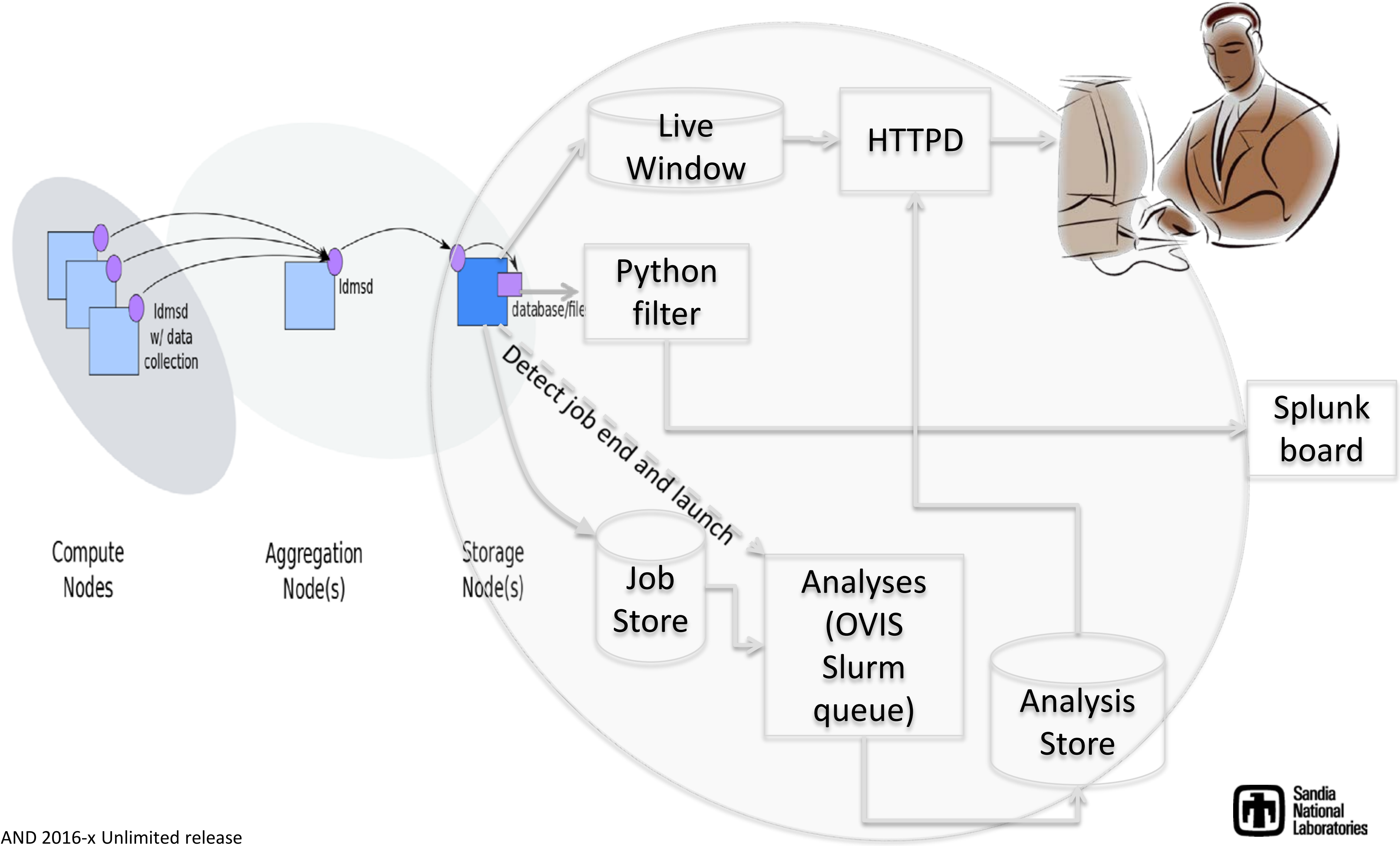
- Since J34-2015:
  - LDMS v2 released to TOSS approximately quarterly with feature additions.
  - LDMS v3 development version tracking changes in Cray's Trinity (XC30).
- Provide system performance "snapshots" including hundreds of metrics at frequencies ***near 1Hz***. (1kHz capability planned in v3 for some metrics).
- Fraction of a percent of a core and memory
- Scale to tens or hundreds of thousands of compute nodes
  - Currently deployed on NCSA's 27,648 node Blue Waters system with as few as 2 aggregators (4 for redundant failover) and data routed to syslog.

- Uses asynchronous transport of data to storage
  - Captures metric sets across nodes at selected time intervals.
- Spans multiple networks with asymmetric access policies
- Robust to failing data collector, node, network, & aggregator.

Functional Overview

- Synchronously collected data is transported asynchronously as fixed “sets” defined by plugins, e.g.
  - /proc/meminfo metrics plugin
- Data is pulled over the network by aggregators
- Data is consumed by store plugins.
  - CSV archiving
  - Web store for live analyses
  - Job store for batch analyses

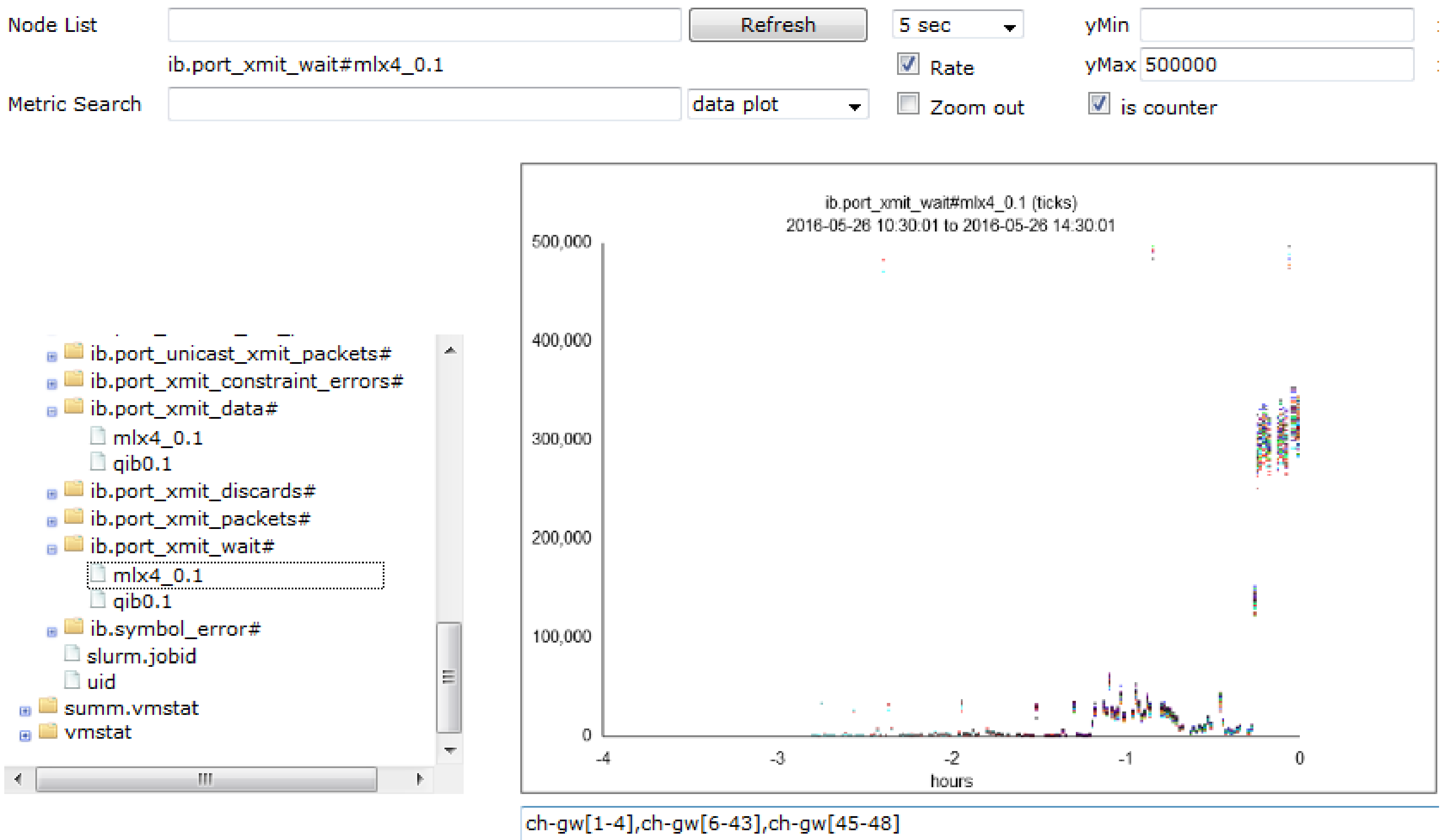
Commodity data analysis



SAND 2016-x Unlimited release

Sandia National Laboratories

Live Window (cluster view)



Node List

Refresh

5 sec

yMin

ib.port_xmit_wait#mlx4_0.1

☑ Rate

yMax 500000

Metric Search

data plot

☐ Zoom out

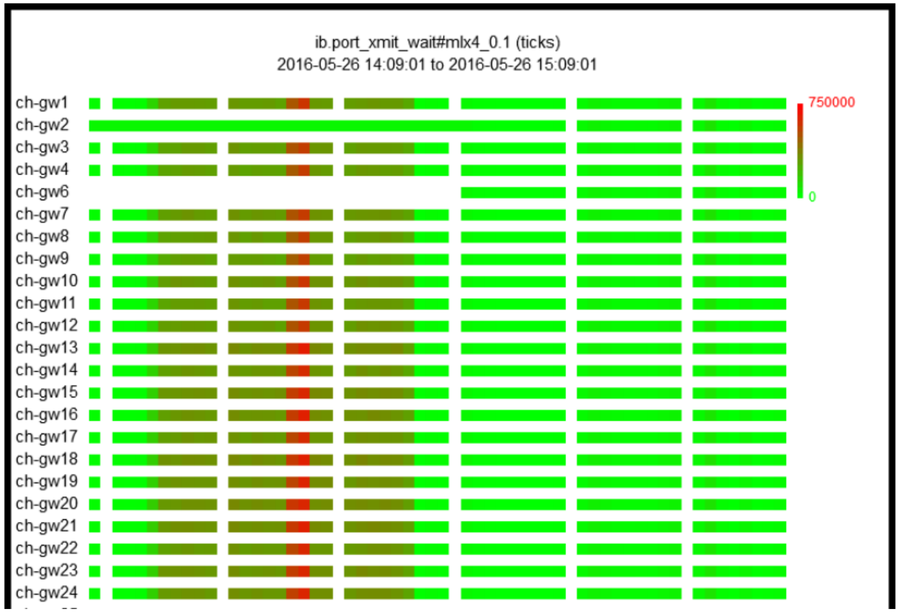
☑ is counter

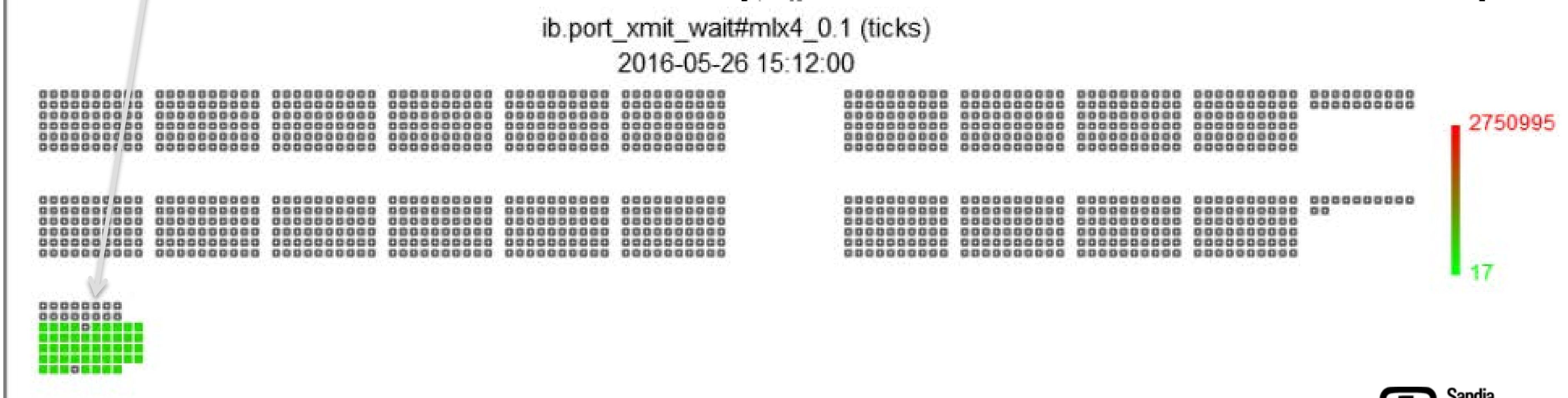
ch-gw[1-4],ch-gw[6-43],ch-gw[45-48]

- Node local and collective values for all metrics.
- Stream results to CSV archive and splunk, E.g. CPU, Memory, IB traffic, Storage IO
- Live since Dec. 2015.

Live window: other renderings

- Xmit wait heat map
  - (node vs time)
- Machine room now
  - (note 2 Lustre OOS)





Job Analyses: Scores and Plots

For *all* metrics:

- Collective value line plot
- Per node line plot
- Per node heat plot

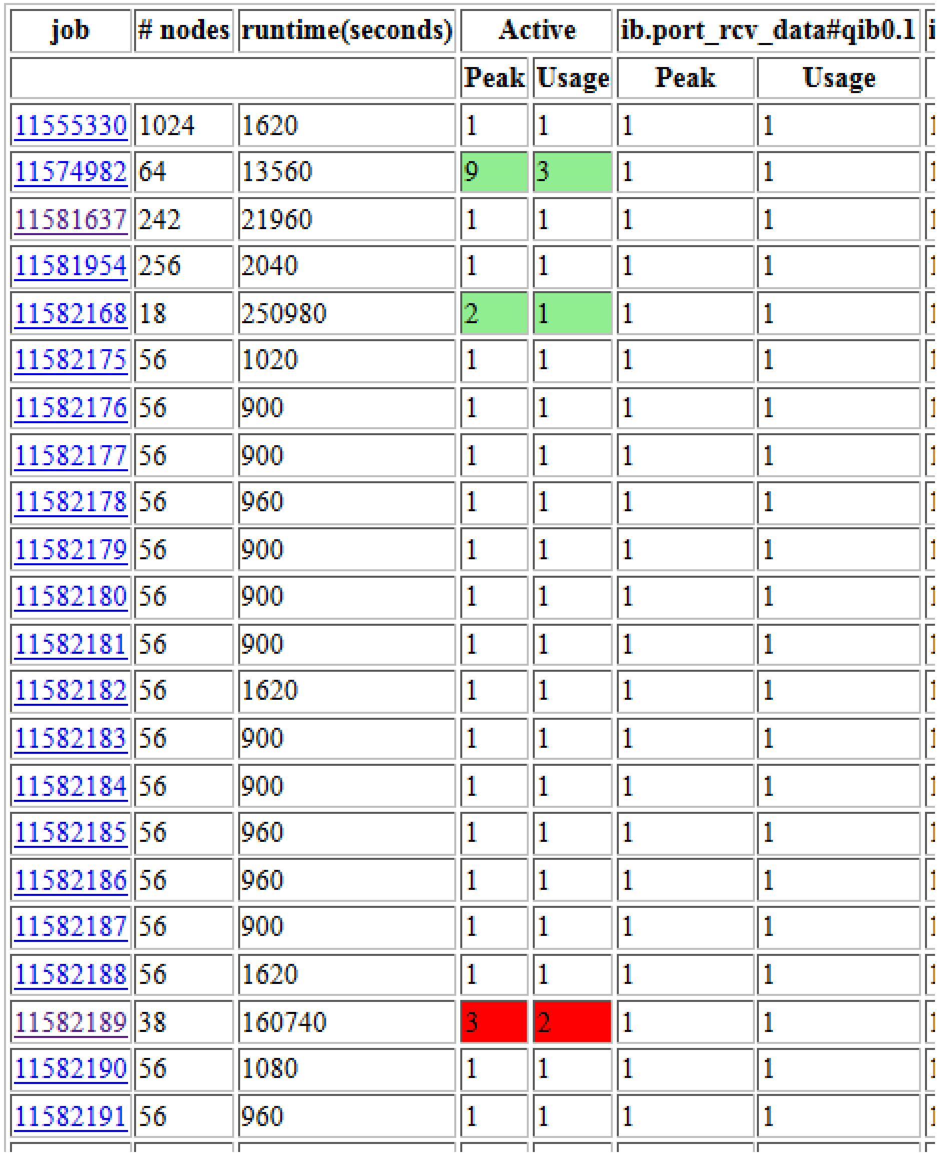
For *constrained* resource metrics:

- Applicable collective limit is determined.
- Plots are given as the percentage of a resource limit.
- Scores (1-10) generated for:
  - Peak local usage on any node.
  - Time average usage across nodes.
  - Balance (coefficient of variance) in usage across nodes.
- A metric metadata inventory is needed (and has been made) to perform these computations.

Job score list

Highlight the atypical

Summary of key metrics

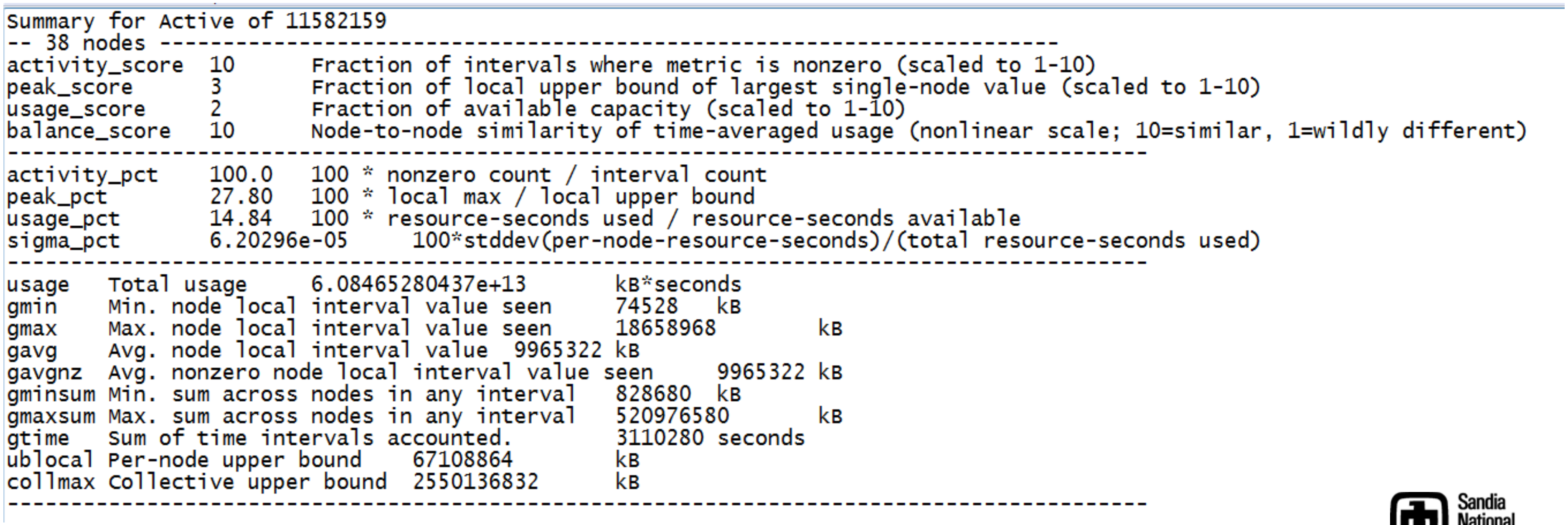


job	# nodes	runtime(seconds)	Active		ib.port_rcv_data#qib0.1		i
			Peak	Usage	Peak	Usage	
11555330	1024	1620	1	1	1	1	1
11574982	64	13560	9	3	1	1	1
11581637	242	21960	1	1	1	1	1
11581954	256	2040	1	1	1	1	1
11582168	18	250980	2	1	1	1	1
11582175	56	1020	1	1	1	1	1
11582176	56	900	1	1	1	1	1
11582177	56	900	1	1	1	1	1
11582178	56	960	1	1	1	1	1
11582179	56	900	1	1	1	1	1
11582180	56	900	1	1	1	1	1
11582181	56	900	1	1	1	1	1
11582182	56	1620	1	1	1	1	1
11582183	56	900	1	1	1	1	1
11582184	56	900	1	1	1	1	1
11582185	56	960	1	1	1	1	1
11582186	56	960	1	1	1	1	1
11582187	56	900	1	1	1	1	1
11582188	56	1620	1	1	1	1	1
11582189	38	160740	3	2	1	1	1
11582190	56	1080	1	1	1	1	1
11582191	56	960	1	1	1	1	1

Job statistics

In addition to scores

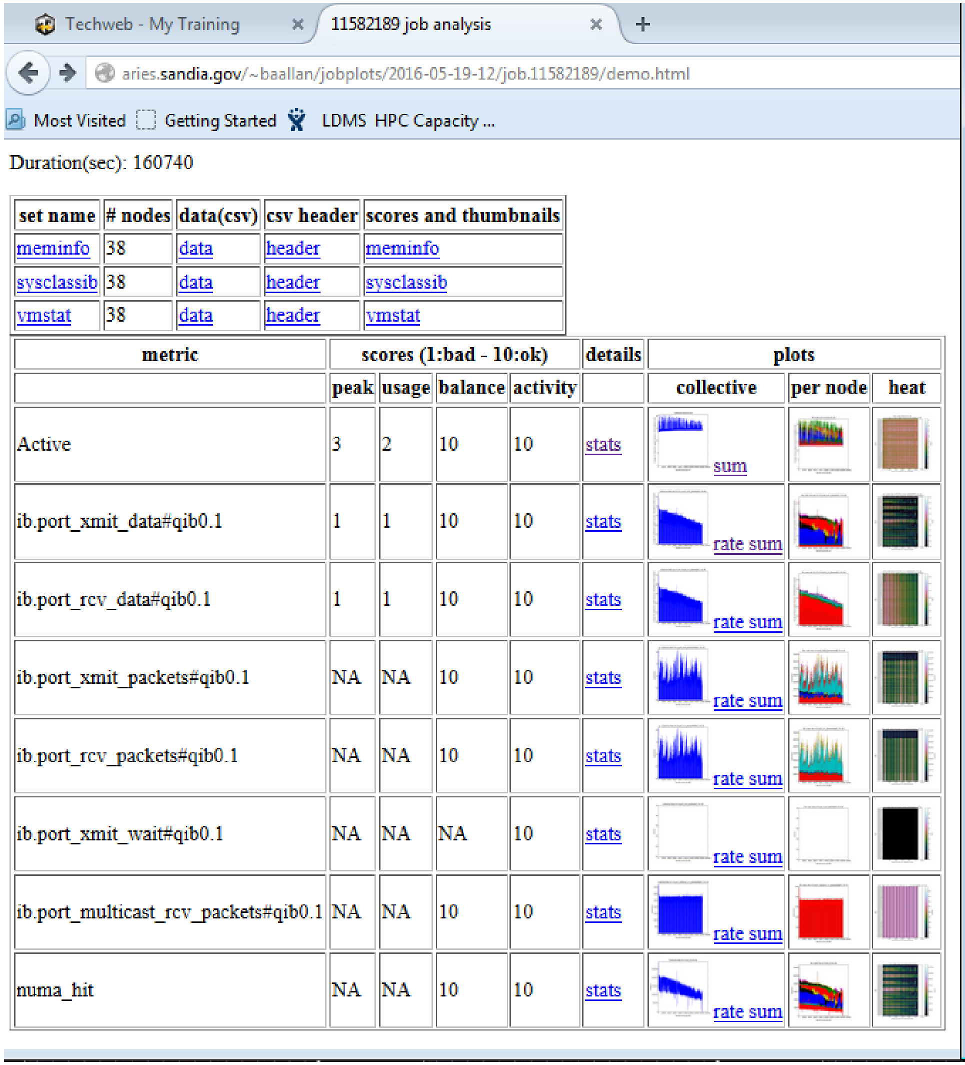
- Fractional minimum, maximum, average
- Minimum, maximum, average, standard deviation
- Minimum, maximum, average, std. dev. excluding zero-value readings



```
Summary for Active of 11582159
-- 38 nodes -----------------------------------------------------------------------
activity_score  10      Fraction of intervals where metric is nonzero (scaled to 1-10)
peak_score      3       Fraction of local upper bound of largest single-node value (scaled to 1-10)
usage_score     2       Fraction of available capacity (scaled to 1-10)
balance_score   10      Node-to-node similarity of time-averaged usage (nonlinear scale; 10=similar, 1=wildly different)
------------------------------------------------------------------------------------------
activity_pct    100.0   100 * nonzero count / interval count
peak_pct        27.80   100 * local max / local upper bound
usage_pct       14.84   100 * resource-seconds used / resource-seconds available
sigma_pct       6.20296e-05     100*stddev(per-node-resource-seconds)/(total resource-seconds used)
------------------------------------------------------------------------------------------
usage   Total usage     6.08465280437e+13       kB*seconds
gmin    Min. node local interval value seen     74528   kB
gmax    Max. node local interval value seen     18658968        kB
gavg    Avg. node local interval value  9965322 kB
gavgnz  Avg. nonzero node local interval value seen     9965322 kB
gminsum Min. sum across nodes in any interval   828680  kB
gmaxsum Max. sum across nodes in any interval   520976580       kB
gtime   Sum of time intervals accounted.        3110280 seconds
ublocal Per-node upper bound    67108864        kB
collmax Collective upper bound  2550136832      kB
------------------------------------------------------------------------------------------
```

Job plots (static analyses)

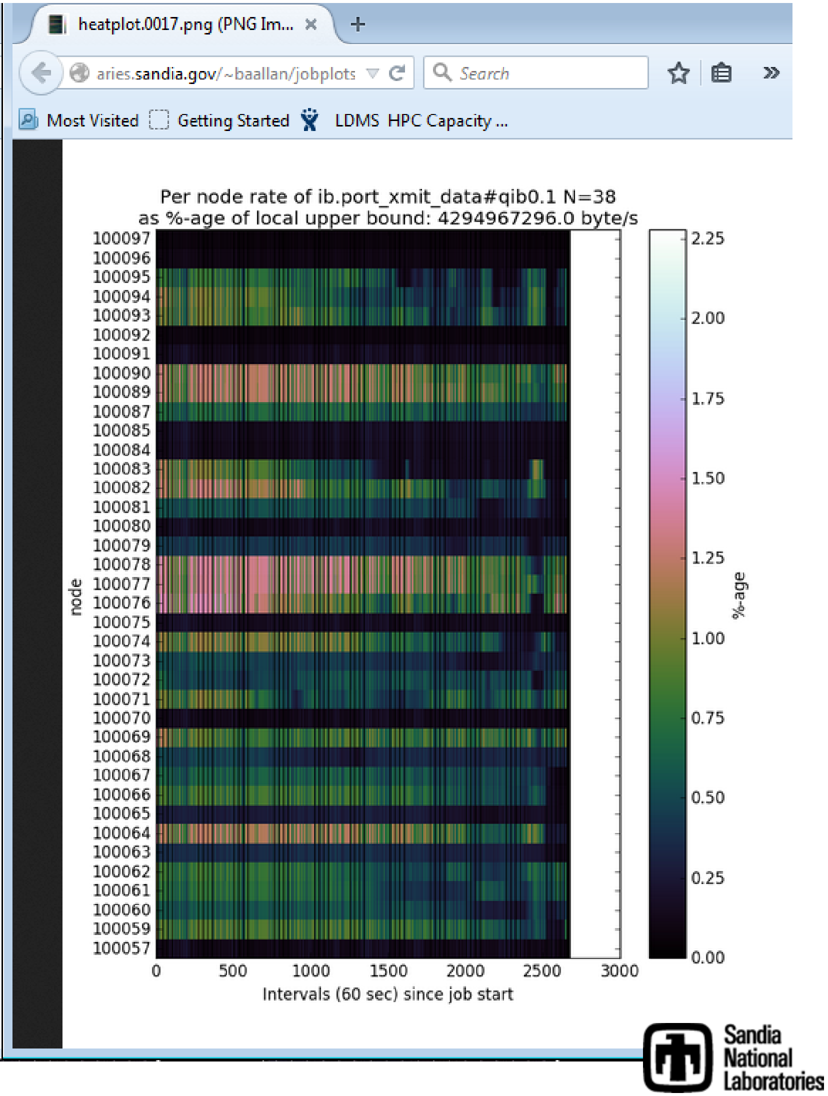
- Hundreds of metrics, several plots each…



Duration(sec): 160740

set name	# nodes	data(csv)	csv header	scores and thumbnails
meminfo	38	data	header	meminfo
sysclassib	38	data	header	sysclassib
vmstat	38	data	header	vmstat

metric	scores (1:bad - 10:ok)				details	plots		
	peak	usage	balance	activity		collective	per node	heat
Active	3	2	10	10	stats	sum		
ib.port_xmit_data#qib0.1	1	1	10	10	stats	rate sum		
ib.port_rcv_data#qib0.1	1	1	10	10	stats	rate sum		
ib.port_xmit_packets#qib0.1	NA	NA	10	10	stats	rate sum		
ib.port_rcv_packets#qib0.1	NA	NA	10	10	stats	rate sum		
ib.port_xmit_wait#qib0.1	NA	NA	NA	10	stats	rate sum		
ib.port_multicast_rcv_packets#qib0.1	NA	NA	10	10	stats	rate sum		
numa_hit	NA	NA	10	10	stats	rate sum		



Exploiting cheap secret sauces*

- Apache & Firefox
  - with no special plugins
- Javascript & python
  - With no proprietary libraries
- Job manager (slurm in our case)
  - Cluster instance dumps active jobid, uid in a text file on each compute node
  - Local instance controls post-processing load
- Libreoffice spreadsheet for debugging data/viz

**Replicating this is not to be made hard by fancy software dependencies!*

Sandia National Laboratories

Development status

- V3 development schedule is driven by ACES Trinity and collaboration with LANL, Cray, and OGC.
  - Copyright release in progress.
  - We anticipate initial public release this FY.
- V2 development is driven by balancing customer needs with minimizing lost work (ours and customers) when V2 is retired.
  - Support continuing until V3 is functionally equivalent and equally well-tested.

New Feature Requests (since 5/2015)

	Feature	Requestor	v2	v3
Usability	Libgenders configuration support	LLNL, SNL	✓	Q
	Allow storage options to vary across data sets	LLNL, SNL	✓	✓
	Slurm User Id association with data	SNL	✓	✓
Trinity	Cray aries network data from rhine/redwood	LANL	no	β
	Sliceable Cray Aries Viz	LANL, others	no	α
	Per-metric AMQP storage control granularity	LANL	✓	✓
Data Analysis	Per-job CSV data collection	SNL	✓	Q
	Global (collective) metrics (eg. tot. storage bw)	SNL	α	Q
	Per-job resource usage scoring	SNL	α	Q
	Splunk feed (filtering, naming, file notifices)	SNL	β	Q

✓: released to production
β: friendly user deployment
α: SNL testing

Q: on to-do queue
No: not currently planned
Dev: in development

Key Prior Requests and v2/v3 status

Feature	Requestor	v2	v3
Asynchronous Push for application data	LLNL, others	no	✓
System collection of MSR counters	NCSA, SNL	dev	dev
String data type for stack trace data	LLNL, NMSU	no	✓
GPU data	NCSA	no	dev
High-frequency or many-cores (vector support)	SNL, LANL	no	✓
Slurm Job Id association with data	SNL	✓	✓
More flexible user data association with metrics	SNL, LLNL	no	✓
Less noisy, more informative daemon logs	SNL	✓	Q

✓: released to production
β: friendly user deployment
α: SNL testing

Q: on to-do queue
No: not currently planned
Dev: in development

Job analysis development

- LDMS V2 scoring is based on a CSV file pipeline
  - In production-level testing at SNL are:
    - New job store plugin, new CSV slicing tools in C.
    - Job-end detection and analysis batch launch.
    - Prototype post-processors ported from perl/matlab/gnuplot to python (data cleanup, statistics, plotting).
  - Not yet shared with users
- V3 scoring will be based on SOS database queries
  - Most of the needed algorithmic work has been done in creating V2 pipeline and post-processors.
  - Held in the work queue.
- Database of metric invariant properties requires curation, mild per-cluster tailoring.

Summary

- Community driven development is succeeding in driving:
  - Capability
    - More metrics
    - More often
  - Usability
    - Easier configuration
    - Better robustness
- Live LDMS data in hands of admins:
  - Low-latency, comprehensive Web visualization
  - Selective Splunk feeds
  - Ease of use improvements needed
- Job-analysis pipeline built out for V2 and in test
- V3 LDMS coming soon

The End

See also:

- http://ovis.ca.sandia.gov
- 2015 proceedings LDMS presentation
- HPCMASPA proceedings from IPDPS in Chicago: https://sites.google.com/site/hpcmaspa2016/
- HPC monitoring community site:

  https://sites.google.com/site/monitoringlargescalehpcsystems/