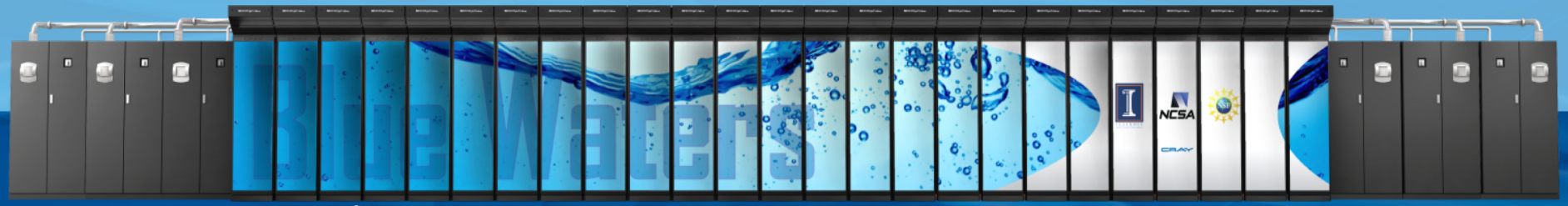
BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Holistic Measurement Driven System Assessment

Professor William Kramer, Blue Water PI
National Center for Supercomputing Applications, University of Illinois
http://bluewaters.ncsa.illinois.edu and http://portal.nersc.gov/project/m888/resilience/
wtkramer@Illinois.edu

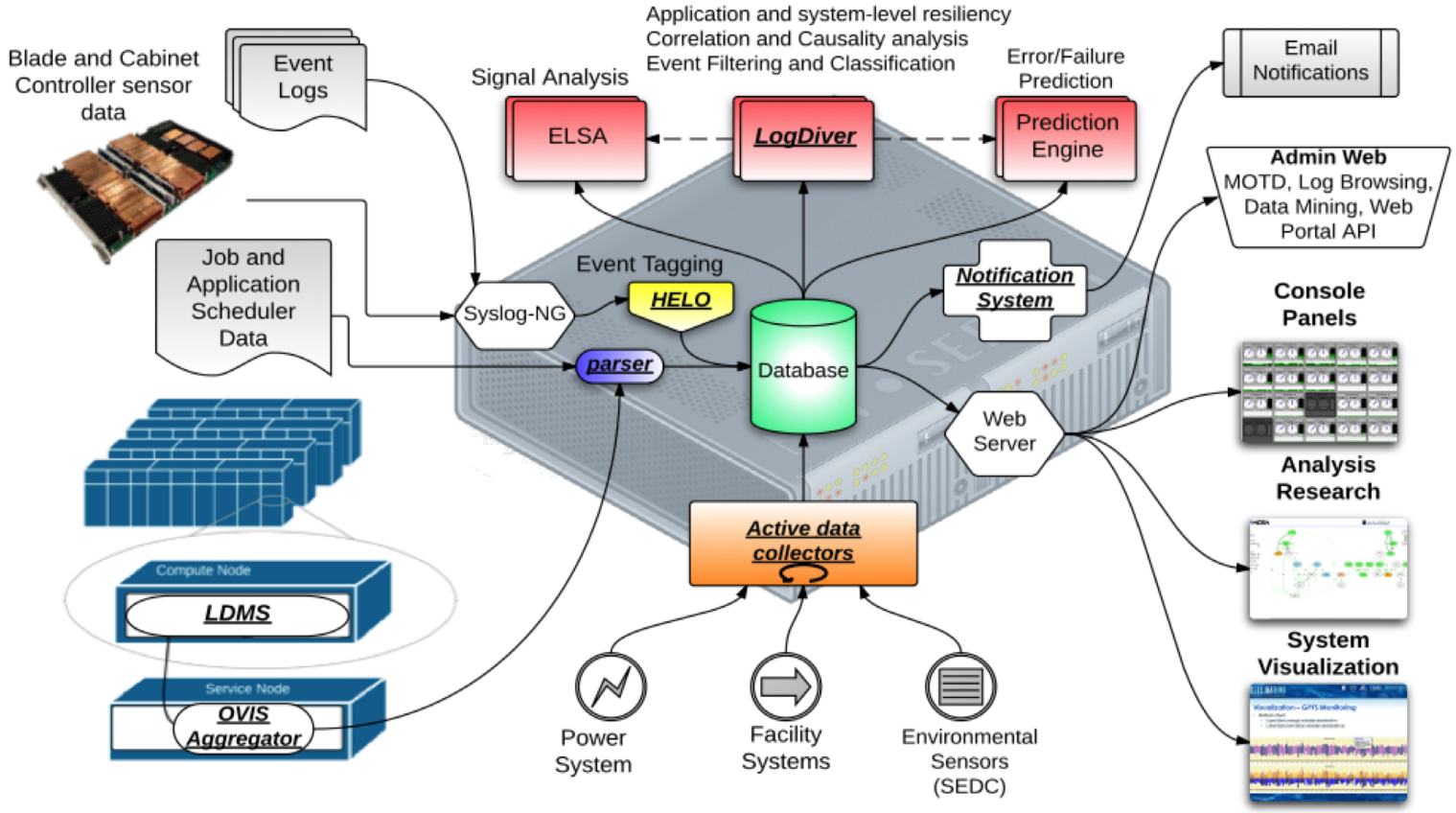


GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

CRAY

Analysis Architecture

- Integrated System Console (ISC) is a developed of the Blue Waters Deployment
- LDMS deployed at scale (> 11M data points per minute) on Petascale Systems without introducing Jitter
  - Lightweight Distributed Metric Service: A Scalable Infrastructure for Continuous Monitoring of Large Scale Computing Systems and Applications, A. Agelastos, B. Allan, J. Brandt, P. Cassella, J. Enos, J. Fullop, A. Gentile, S. Monk, N. Naksinehaboon, J. Ogden, M. Rajan, M. Showerman, J. Stevenson, N. Taerat, and T. Tucker , IEEE/ACM Int'l. Conf. for High Performance Storage, Networking, and Analysis (SC14) New Orleans, LA. Nov 2014.



On Blue Waters, has a core of about 8,100 active unique event types using HELO and ISC , with about 50,000 different event types that we have seen over the course of the project. We see around 30M log events on average per day, getting upwards of 1,470M log events on some days. This is exclusive of metric data like ovis and the plethora of other things we track. The ISC project has around 300 different tables we use to track those various other things.

- Based on NCSA's Integrated Console System (ISC), Sandia's LDMS/OVIS, Argonne's Darshan, Cray's support (e.g. performance register driver for NICs) and standard products, UTK XAltD, resoruce management logs, system event logs and other pieces
- Validated that data collection from all nodes and all NICs on 10 second resolution is without observable jitter
- Mandatory inclusion of things like Darshan for less than 1% overhead
- Able to collect and parse data in real time, including data storms (many messages during resiliency events)
- We currently hold on-line (on a modest server) about 3-4 weeks of data
- Collects about 20-25B datums per day

Data descriptions

- File stores
  - About 15T of MSR data, 35TB of data per node per day
  - Data is raw format such as counters
- Integrated System Console Database
    - ~30GB per day of node data per day, 4 days retained
    - Some data is preprocessed from counters to rates
      - Example: flop counters per core are converted to flop rate per node
      - Greatly improves query efficiency
    - All log and system data
- Access methods
  - SQL queries
  - CSV files in lustre (parallel tools to extract data)
  - Web interfaces (with image/raw data downloads)
  - Hope to begin publishing the datasets

Current Database Data

- Gemini Link Statistics
  - All 6 directions
  - Link BW, %used, average packet size, %input queue stalls, %credit stalls, …
- Gemini/NIC Statistics
  - totaloutput_optA/B, total input, FMA output, bet output
  - SMSG
    - Number tx/rx rate , Bytes tx/rx rate
  - RDMA
    - Number tx/rx rate , Bytes tx/rx rate
  - IP over Gemini
    - Transmit/Receive rate
- Application library use
- MPI I/O operations from each application (Darshan)
- Node
  - Load average
    - Latest,5min, running processes, total processes
  - Flop rate
  - Current free memory
  - GPU
    - Utilization, memory used, temperature
    - Pstate, Power Limit, Power Usage
- Filesystems
  - For each home, projects, and scratch
    - Bytes/sec Read and write
    - Rate of Opens, closes, seeks

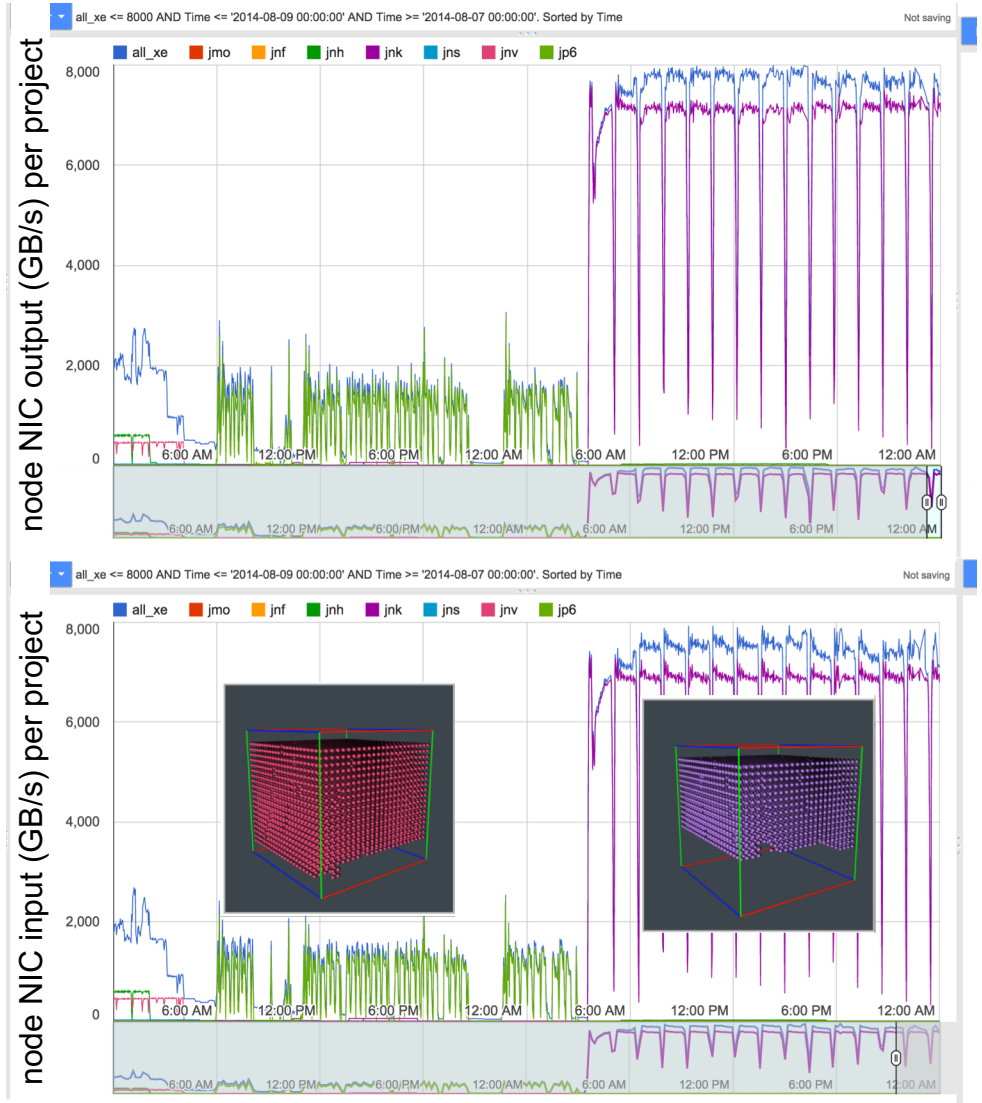
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

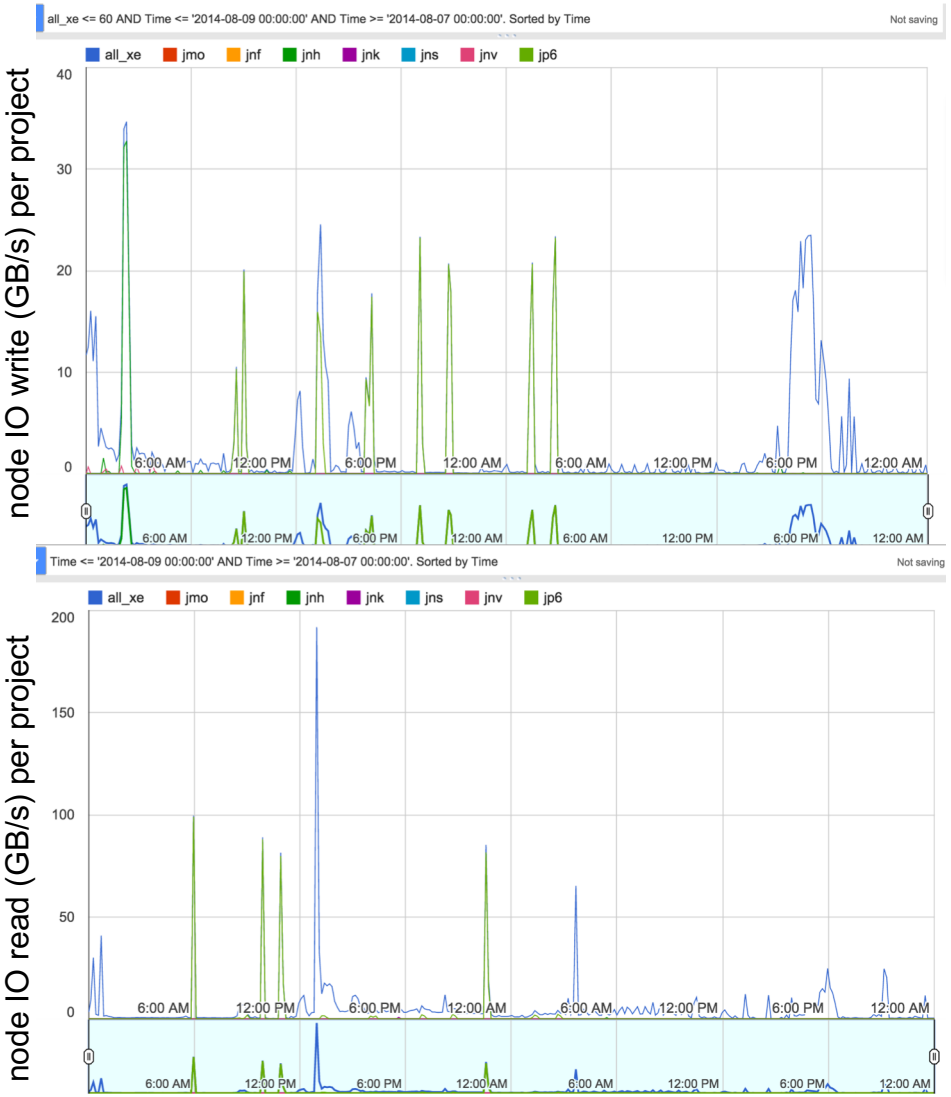
About 10.6 Trillion Datums to July 2017

Data feed	Average (Bytes/day)	Max (Bytes/day)	class
apres	30M	148M	logs
apstat	60K	62K	metrics
backup	40K	74K	metrics
ddn	43K	326K	logs
esms	1G	3G	logs
hpss	135M	3.6G	logs
hpss_core	112K	192K	metrics
ibswitch	790K	801K	logs
moab	2.5G	3G	logs
qos-ping	3.3M	3.6M	metrics
quotas-hpss	944K	950K	metrics
scheduler	76K	78K	metrics
cabinet env/pwr/temp/status	45M	45M	metrics
SEL	1K	6K	logs
sonexion	250M	3.5G	logs
sonexion perf home	4.5G	4.5G	metrics
sonexion perf projects	4.5G	4.5G	metrics
sonexion perf scratch	4.5G	4.5G	metrics
spectra	1.5K	9K	logs
Mainframe LLM	4G	120G	logs
Torque	75M	359M	logs
volkseti	19M	19M	metrics
OVIS	135G	135G	metrics

Amount of data by source as of July 2017

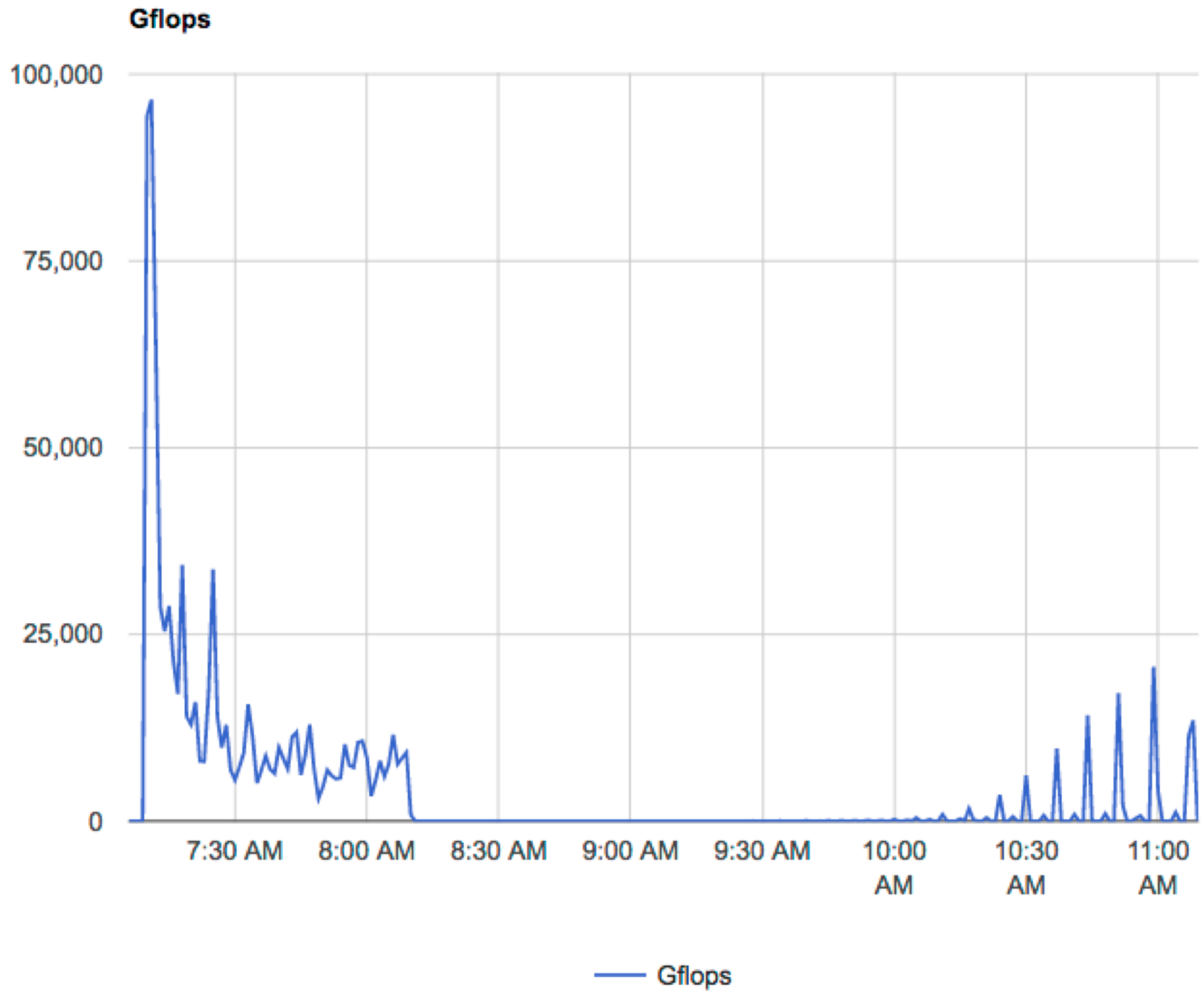
Example of External Application Monitoring @Scale





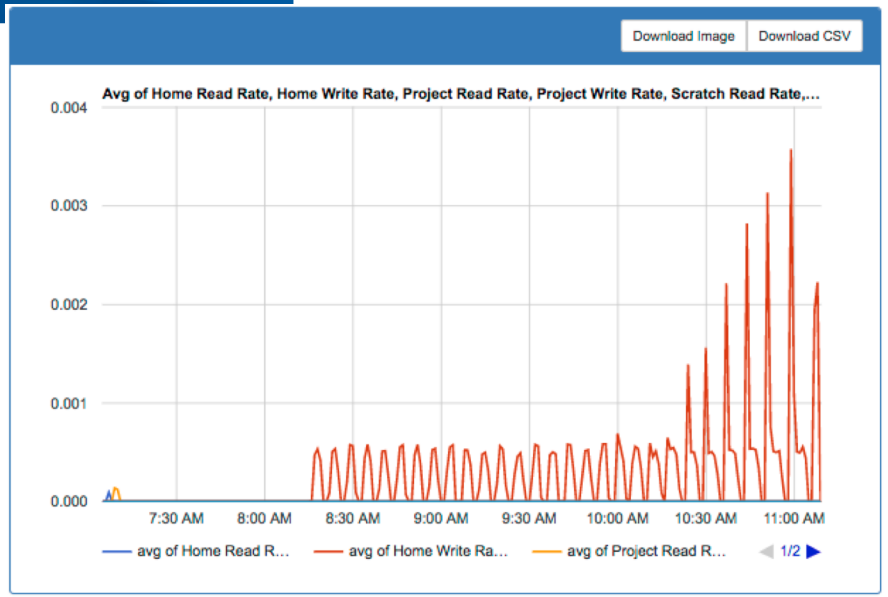
20,250 XE node P-gadget job **15,872 XE node namd job**

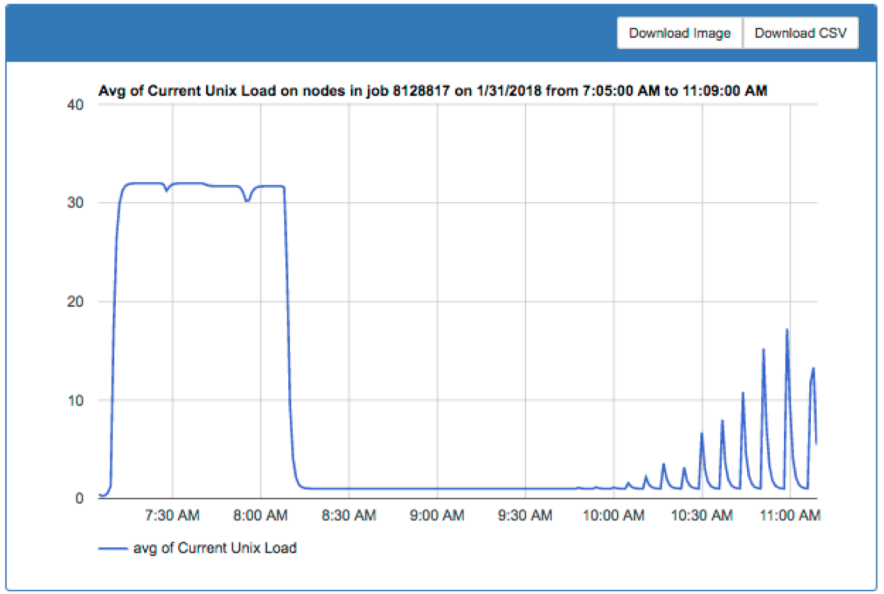
Example – Job 8128817 – Processing rate

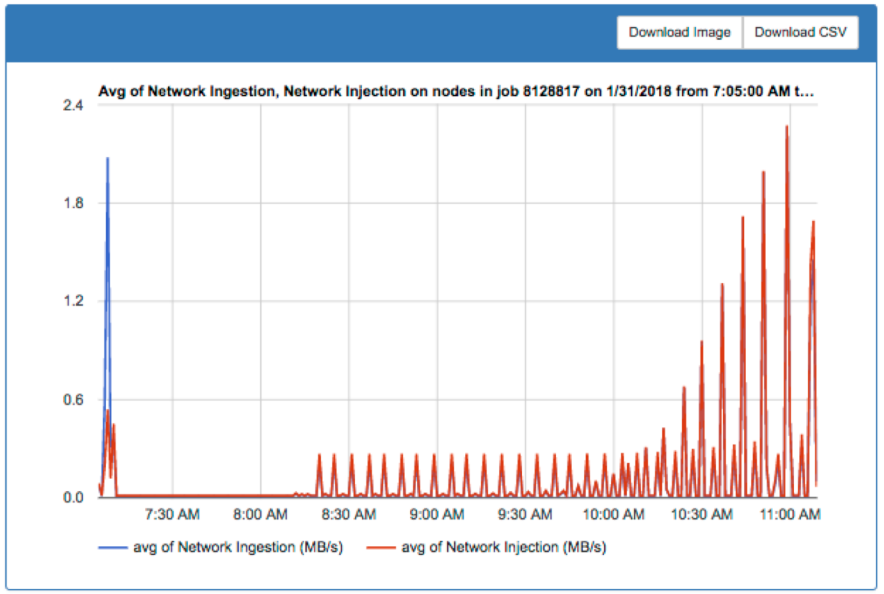


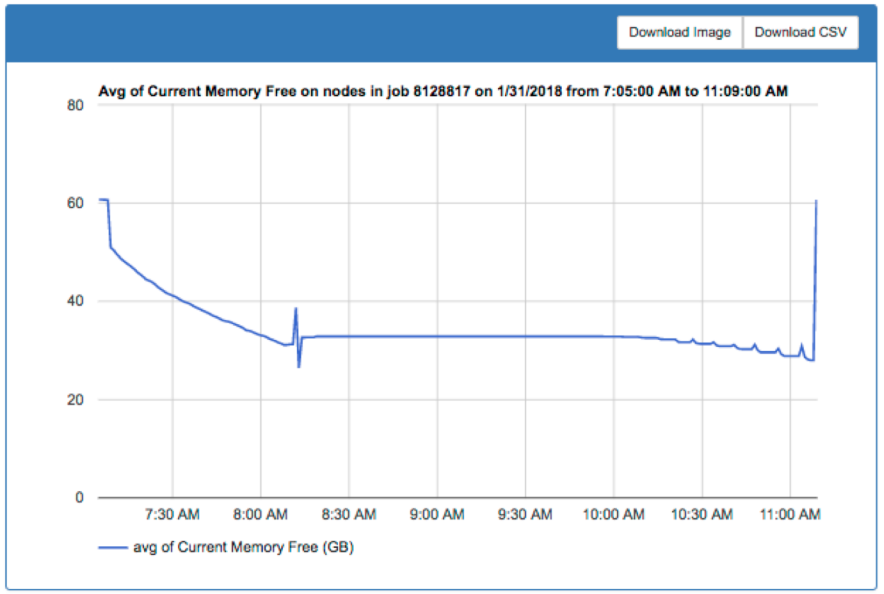
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Example – Job 8128817 – File System







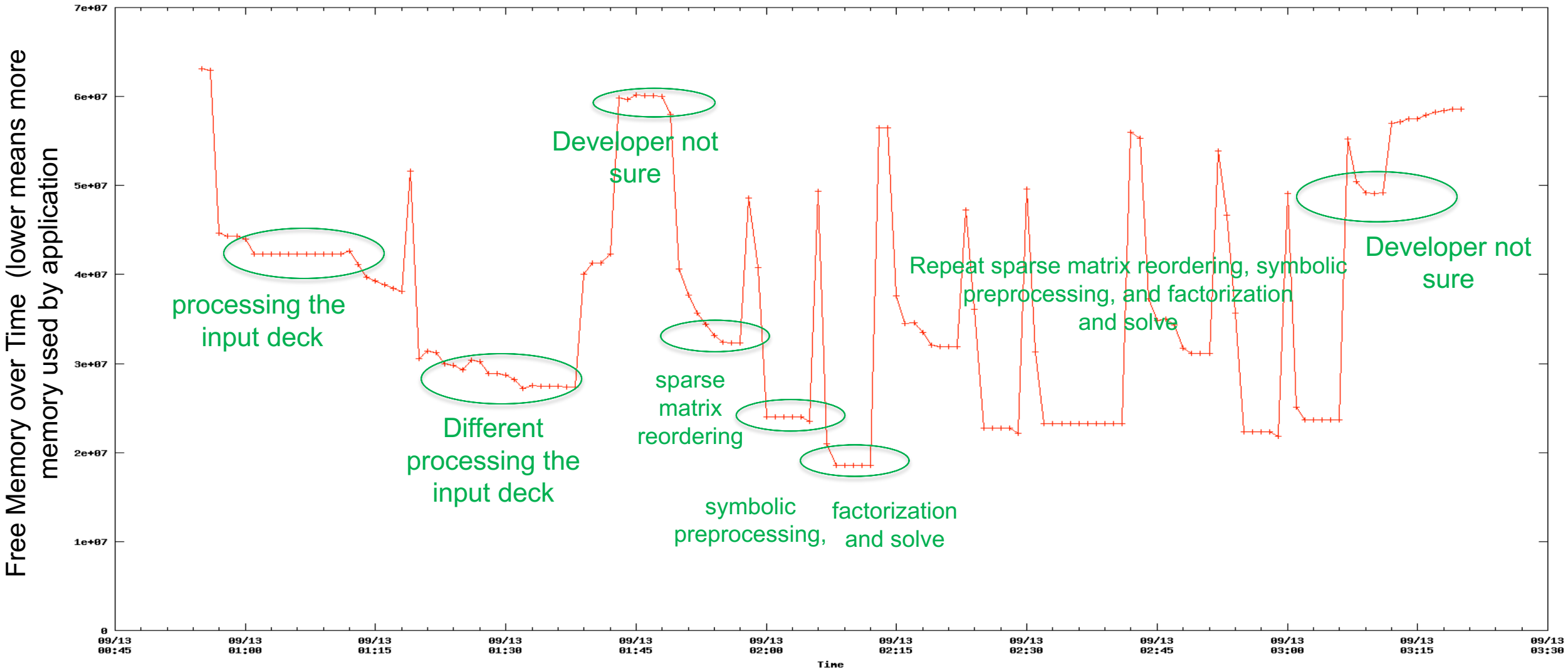


Example – Job 8128817 – L1 Cache Miss Rate

Example – Job 8128817 – Summary

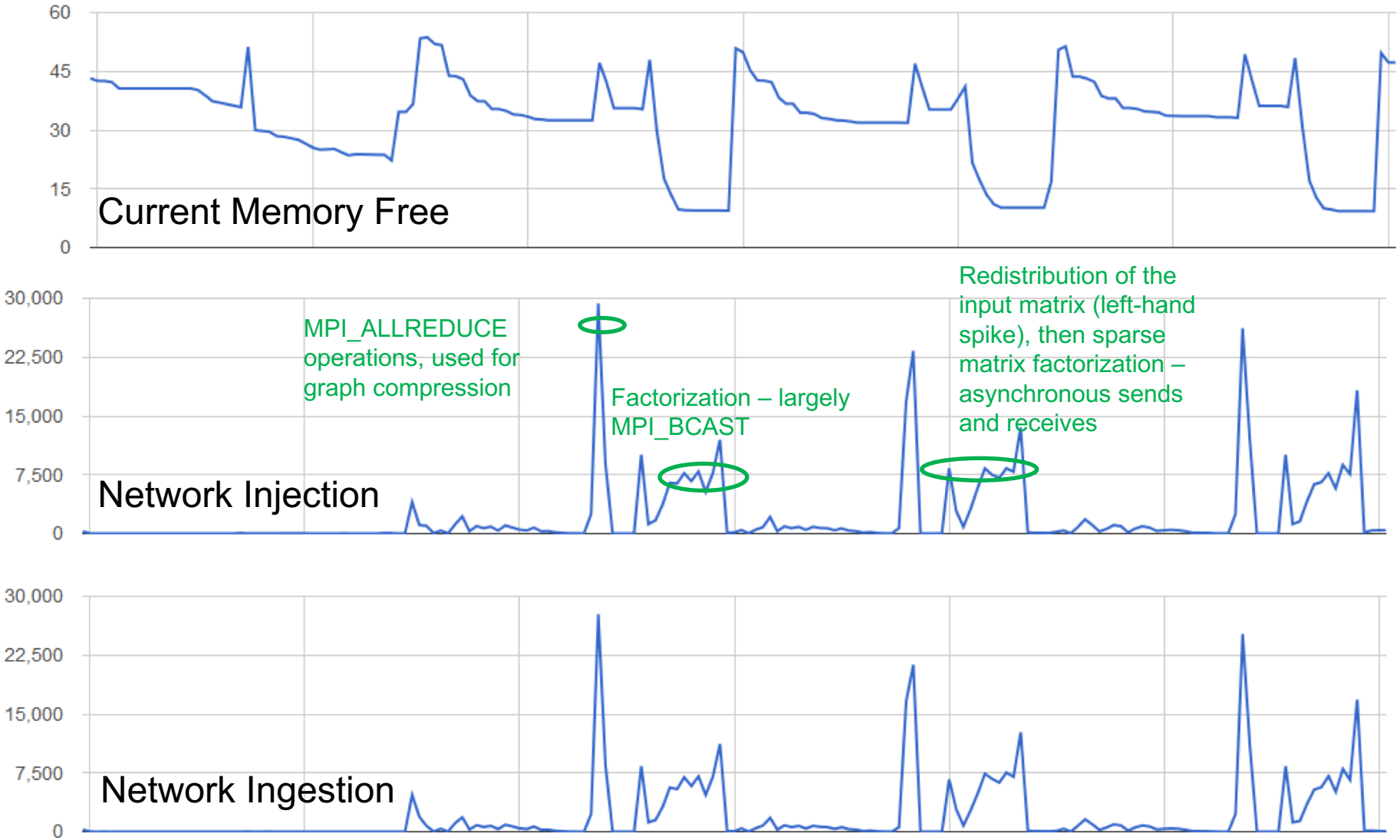
- Job 8128817 was 22,366 XE node job – the application runs regularly
- 1 MPI rank per node with 32 threads/rank
- Code uses Intel Thread Building Blocks
- Observation – Flop rate goes to nearly 0 for 1.75 hours of execution
- Looking at HMDSA node data shows
  - I/O from home (!) file system – which is only 1/10th the bandwidth of scratch
  - CPU load per node going to 1 (the single MPI rank)
  - Most network activity due to I/O, not internode communication.
  - Memory footprint constant during “idle” period.
  - L1 cache misses does not show much activity.
- Support staff contacted the User
  - “We use parallelized eigensolver for large symmetric matrices. However, it happened that matrix which had to be diagonalized was very small. Basically, massive parallelization in such case was senseless.
  - Another thing was large I/O to the home file system, as you mentioned.”
  - Team modified application to better overlap computation and I/O and use the correct storage system
- Used the wrong file system (small impact) and strong-scaled the problem out way too far without balancing other parts of the problem.

Example – Understanding LS-DYNA Performance 200M DOF and 2,048 MPI ranks



Example – Understanding LS-DYNA Performance 200M DOF and 2,048 MPI ranks

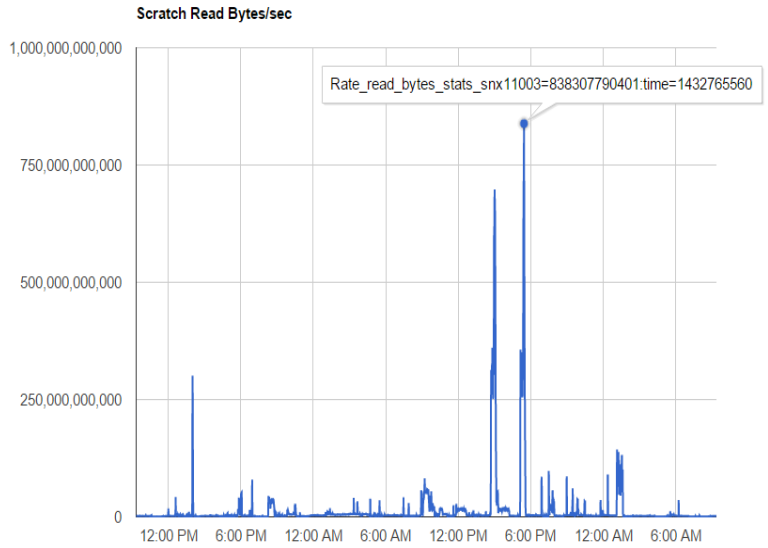
spike is MPI_ALLREDUCE. The redistribution spike is asynchronous sends and receives (I also have an MPI_ALLTOALLV variant). The factorization "ramp up", is largely MPI_BCAST.

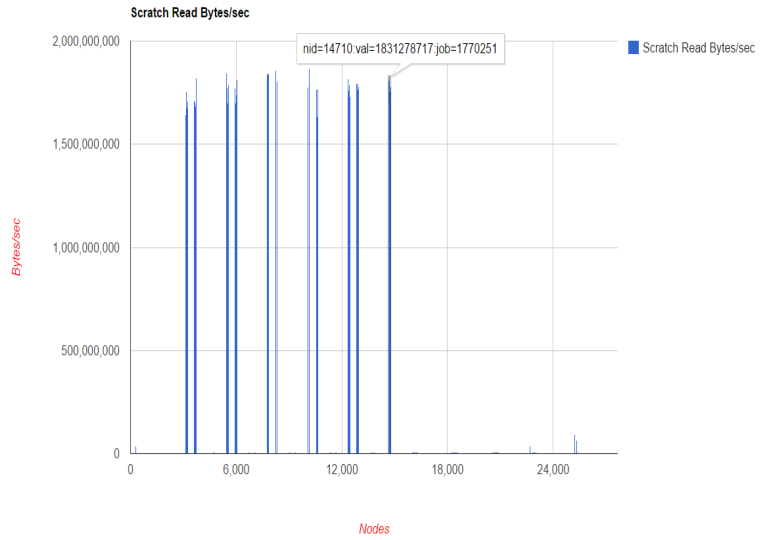


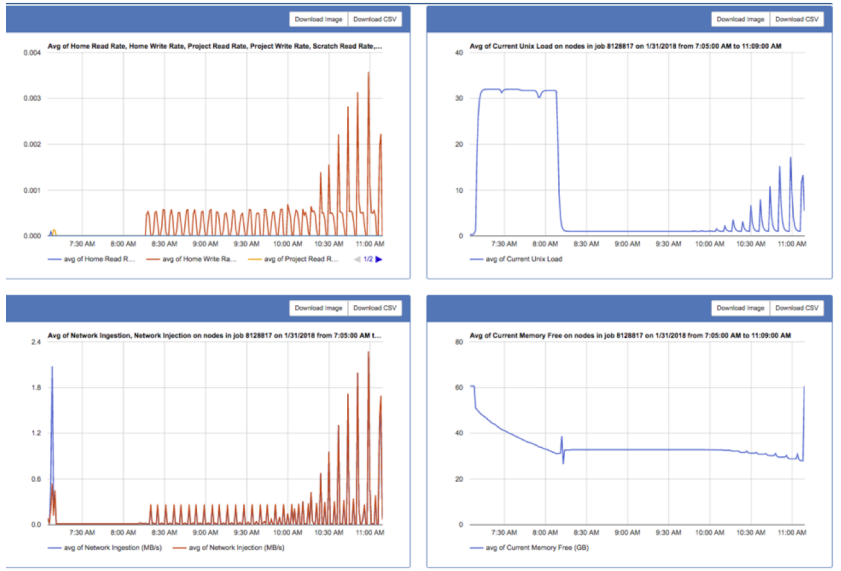
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Example - Job Analysis

- Challenge: *Diagnose system behavior anomalies caused by applications.*
- Approach: *View system aggregate metrics over time to find abnormalities. Drill down on times of interest within a metric to show contributing nodes. Overlay job/user information on data to make correlation to suspect workload. Further drill down on workload.*
- Benefit: Fast and low labor mechanism to explore metric data for diagnosing and identifying disruptive workload. Also provides mechanism to show metrics of interest for a job with a suspected problem.







The first graph looks at total system I/O and selects a point in time to reveal the nodes and job responsible for the peak in reads. From there, the suspect job is identified for deeper analysis including more metrics if necessary.

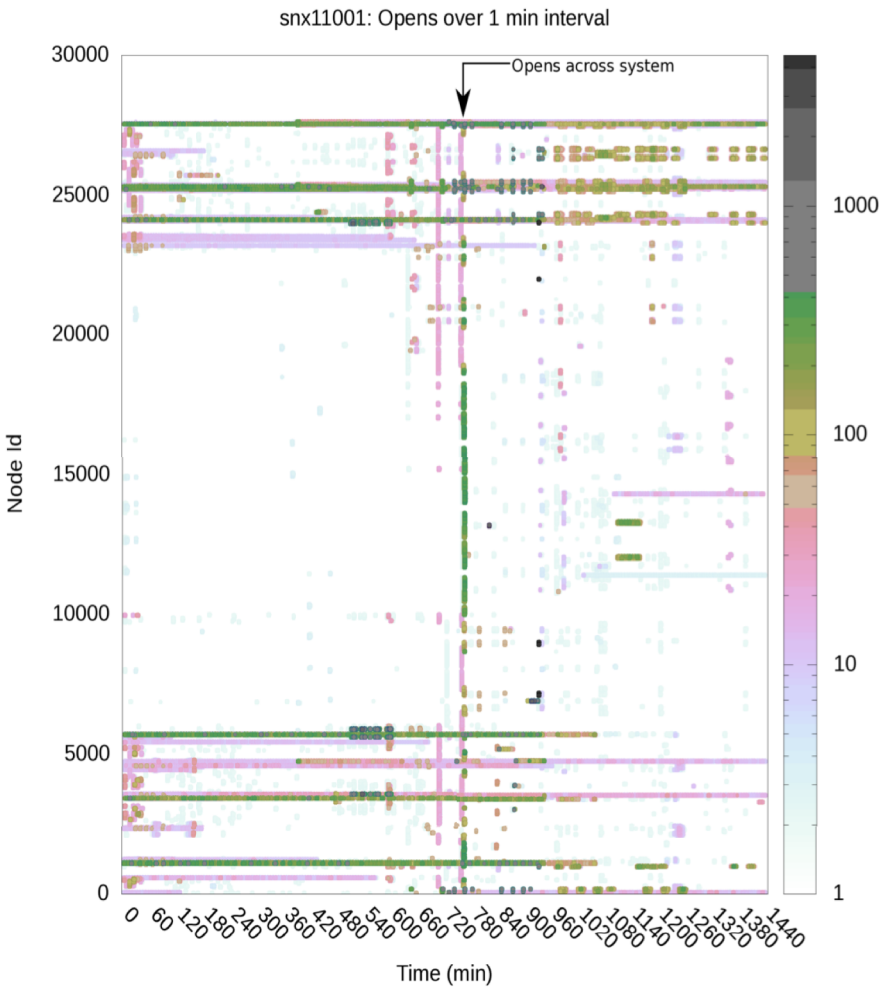
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Example - Job Analysis

- Success:
  - ***Reassemble*** *monitoring data with job and user context applied.*
  - *Quickly* ***navigate*** *multiple nodes and multiple metrics, live or historical*
  - *Quickly* ***analyze*** *system anomalies to* ***identify*** *disruptive workload.*
  - *Quickly* ***analyze*** *job anomalies to* ***identify*** *issues within a job.*
  - ***Share*** *information with user.*
- What was the most significant roadblock/gap you had/have to overcome?
  - *Large amounts of data to store and query- easy to congest backend.*
- What are your next steps?
  - *Restructure data store and data transport mechanism for capacity, resiliency, performance*
  - *Restructure front end and backend to use community supported tools if applicable*
  - *Attempt to find a seamless bridge between “hot” and “cold” spool metric data*

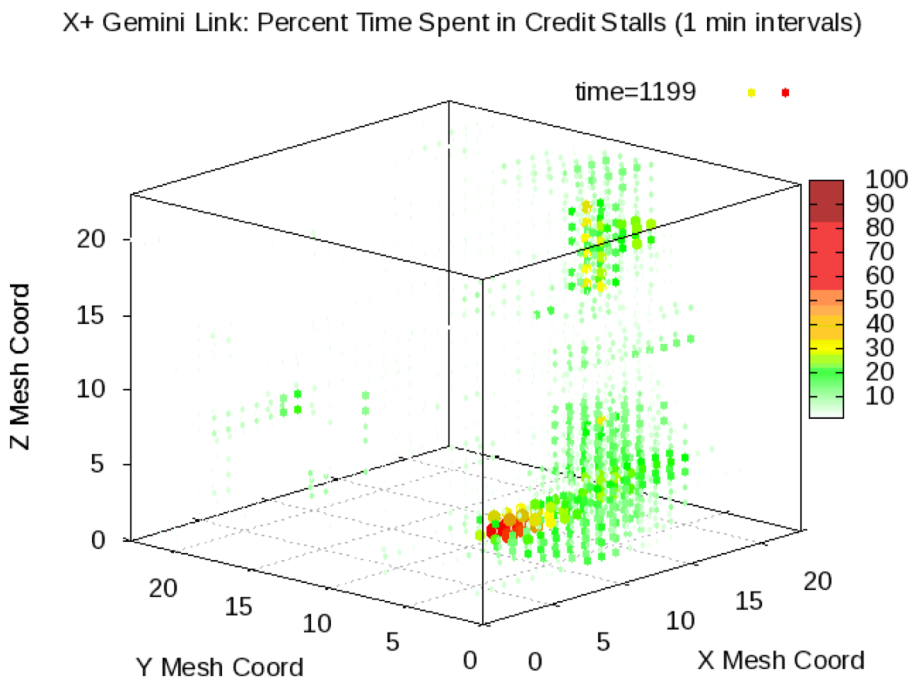
LDMS – Lightweight, High-Fidelity Data Collection, Transport, and Storage

- Provides HPC system state data unique in scope and fidelity
  - Whole system snapshots down to sub-second intervals
  - Minimal impact on platform resources
  - No measurable adverse impact on large-scale application run times
- Features:
  - Synchronous collection for coherent system snapshots
  - Minimal and efficient processing on compute resources
    - Efficient data layout and minimization of data movement
    - RDMA to pull data without involving compute resource processors
  - Aggregators on dedicated resources support high overhead tasks such as failover and in-transit analysis plugins
  - High fan in ratios (> 15000:1)

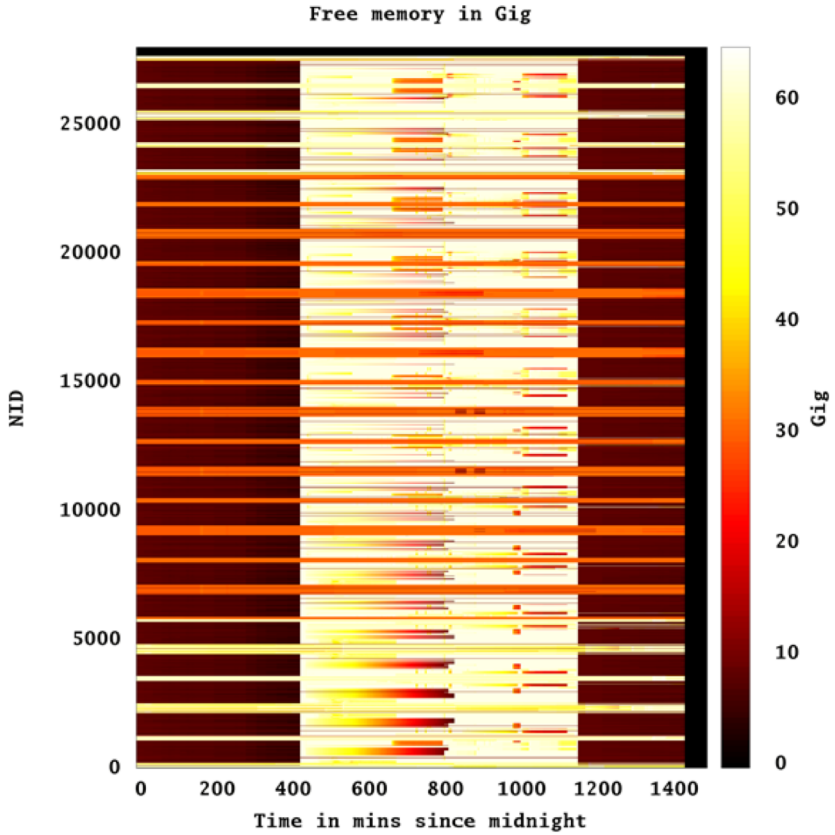


Blue Waters: One day dataset contains ~40 million data points per metric and 7.7 billion data points overall

3D Images and Graphs Add Insight



Network Contention

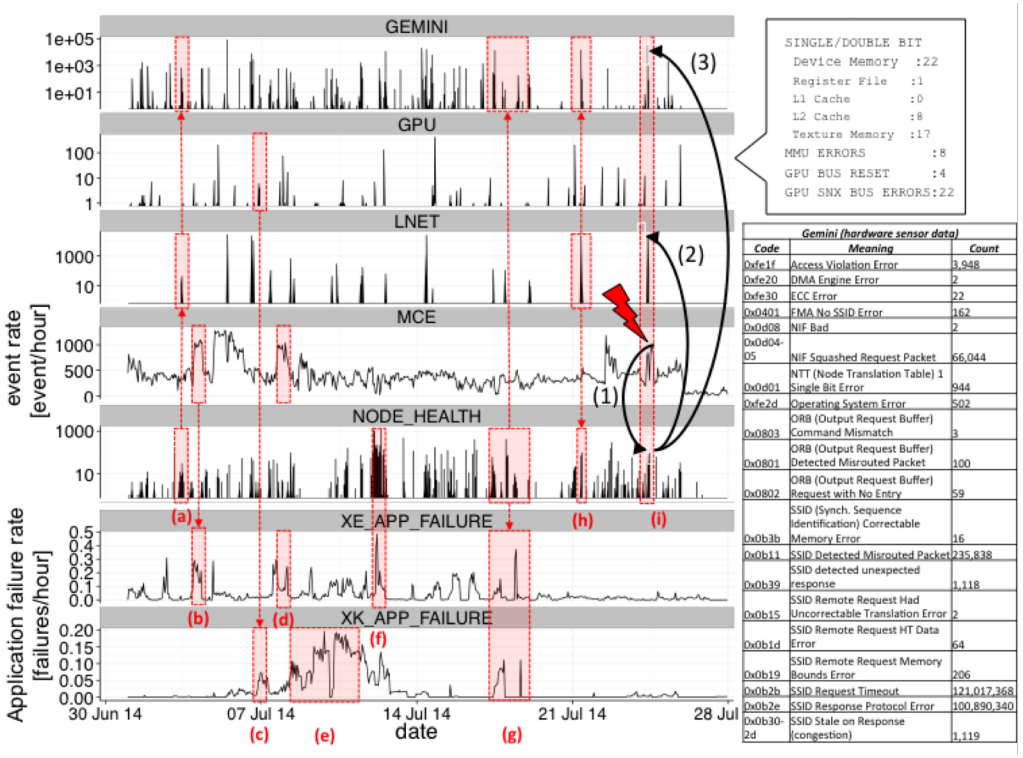


Free Memory

Example - Resiliency HDMR Insights and Results

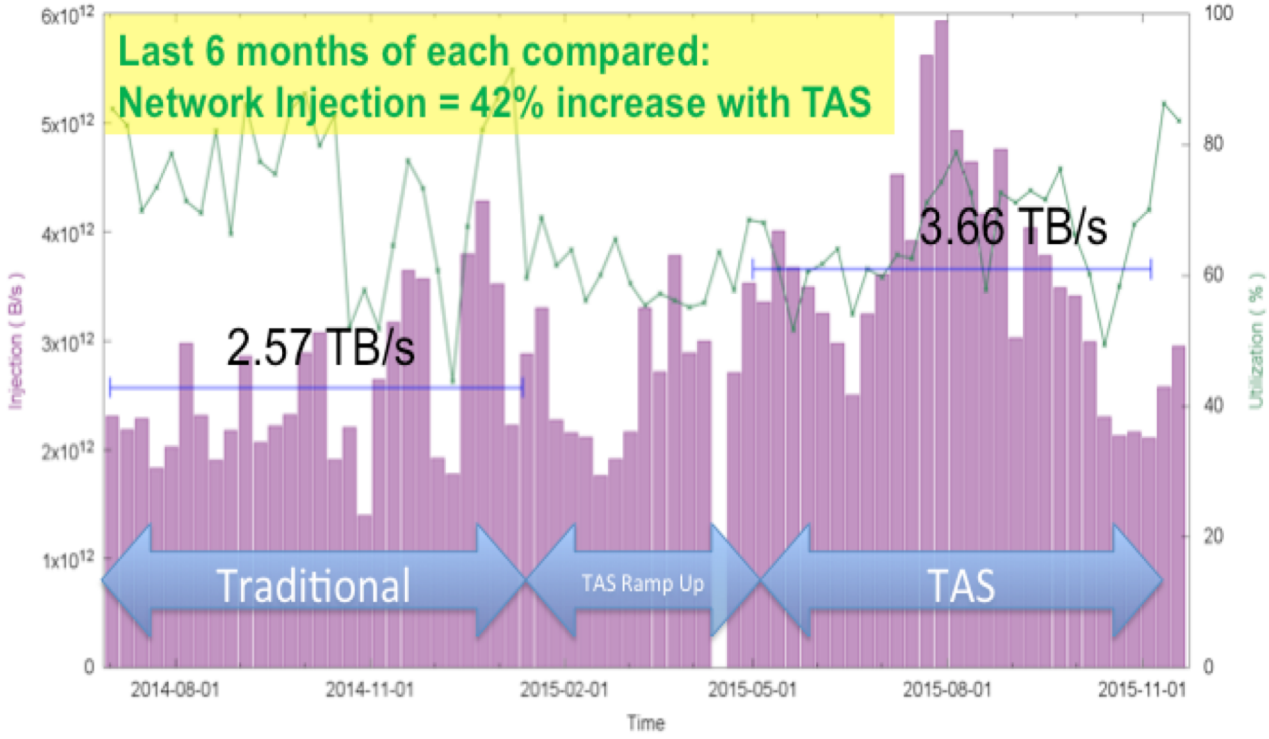
- Resiliency Insights from ISC and Logdiver
  - 99.4% of failures limited to a single blade;
  - Software errors propagate 20 times more often than hardware failures;
  - DDR5 ECC is 100x more prone to uncorrected errors then DDR3 with x8 Chipkill;
  - Software accounts for 53% of repair hours;
  - Hardware failure rates decline over time but software does not;
    - Node failure rates ½ of the first year (<2 per day in last quarter)
    - Last quarter non CPUs or GPUs failed
    - System wide MTBF now 90-180 days
  - 74% of system wide outages are due to software;
  - 50% of these are during failover;
  - Filesystem and interconnect are prime contributors;
  - Failure of failover causes a significant number of system wide outages;
  - Application failure increases with increasing duration of failover time, mostly for those applications that do not use fault tolerant communication frameworks (.e.g. not Cray MPI)

Examples of Results: Improving systems though understanding of root causes of faults, failure propagation, and performance degradation



Gemini (hardware sensor data)

Code	Meaning	Count
0xfe1f	Access Violation Error	3,948
0xfe20	DMA Engine Error	2
0xfe30	ECC Error	22
0x0401	FMA No SSID Error	162
0x0d08	NIF Bad	2
0x0d04-05	NIF Squashed Request Packet	66,044
0x0d01	NTT (Node Translation Table) 1 Single Bit Error	944
0xfe2d	Operating System Error	502
0x0803	ORB (Output Request Buffer) Command Mismatch	3
0x0801	ORB (Output Request Buffer) Detected Misrouted Packet	100
0x0802	ORB (Output Request Buffer) Request with No Entry	59
0x0b3b	SSID (Synch. Sequence Identification) Correctable Memory Error	16
0x0b11	SSID Detected Misrouted Packet	235,838
0x0b39	SSID detected unexpected response	1,118
0x0b15	SSID Remote Request Had Uncorrectable Translation Error	2
0x0b1d	SSID Remote Request HT Data Error	64
0x0b19	SSID Remote Request Memory Bounds Error	206
0x0b2b	SSID Request Timeout	121,017,368
0x0b2e	SSID Response Protocol Error	100,890,340
0x0b30-2d	SSID Stale on Response (congestion)	1,119



Analysis of data for root cause fault analysis

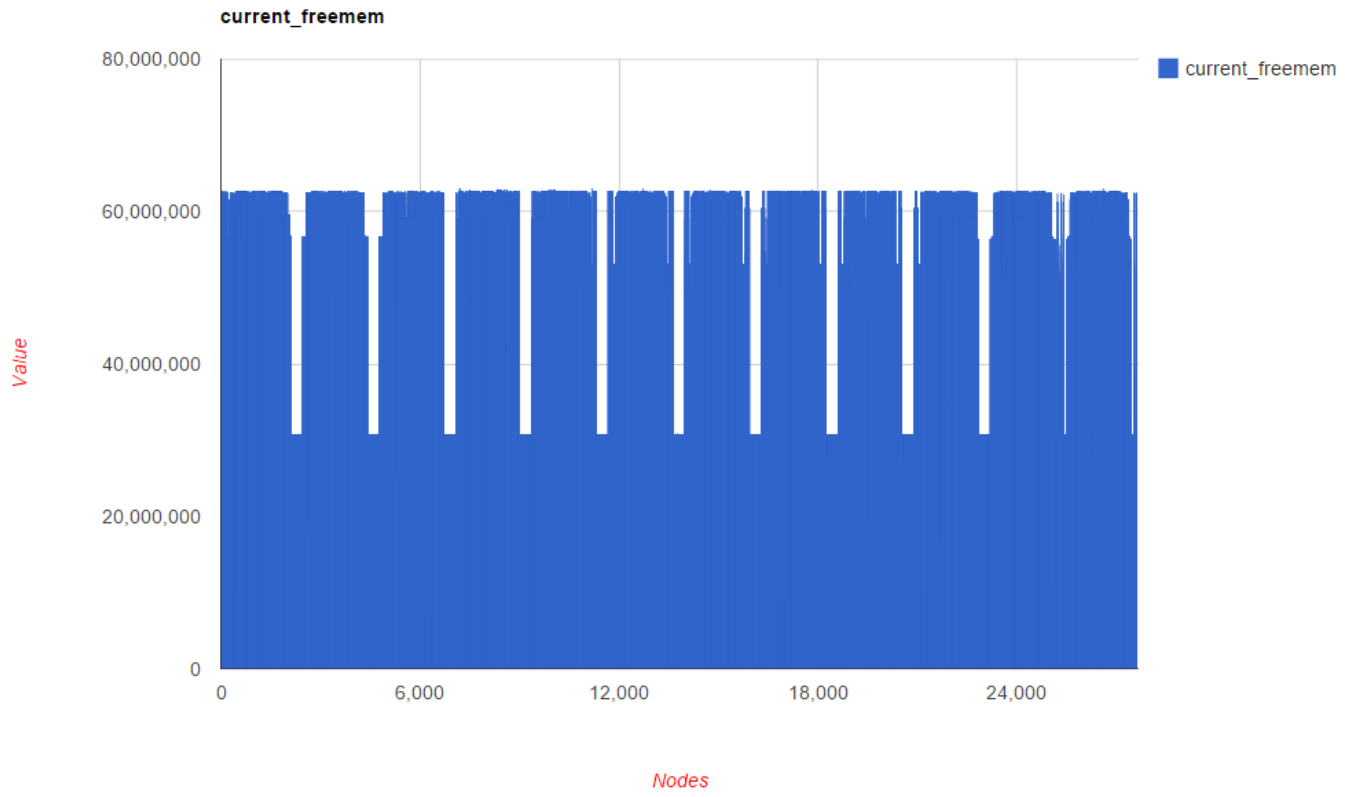
One root cause of significant performance degradation addressed by “topologically aware scheduling”

BW Example - Similar Application Set Comparisons over a 16 month period

Dates Compared	From: Traditional ~6 months (July 1 2014 - Jan 13 2015) To: TAS ~10 months (Jan 15– Nov 5 2015)
Representation	92 comparable job sets 16 projects 29 distinct partners 228 MNH of allocation
Application Runtime	TAS improved by 16%
Application Runtime Consistency (CV)	TAS improved CV by 63%
Network Injection by app	TAS improved by 19% weighted average by node*hrs run

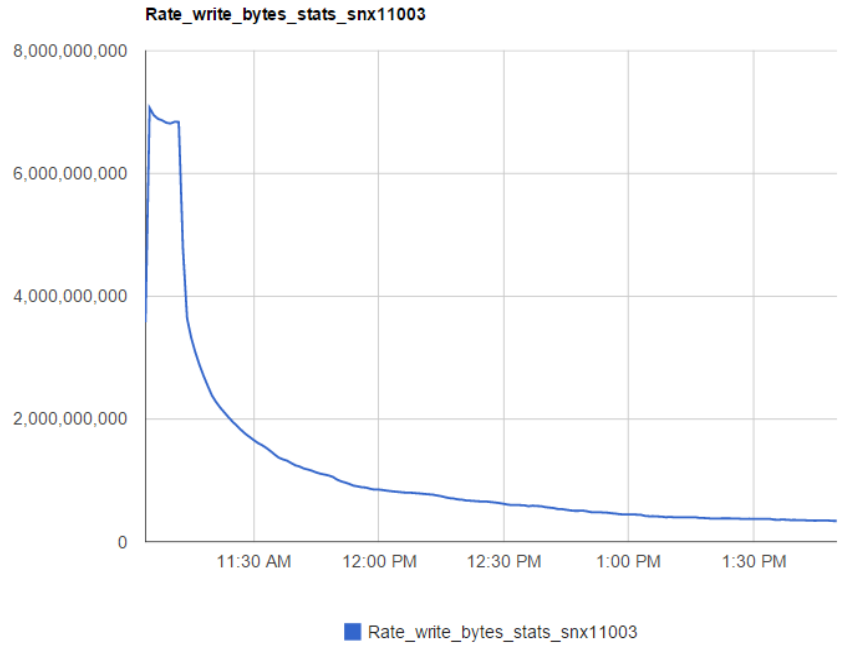
NSF Blue Waters Review Panel Dec, 2016 –
"This analysis is unique in that it is done based on data from a real full-size top-of-its-class system running real workload in comparison to similar work which almost always relies on simulation data."

Single Metric in Time (Free Memory)

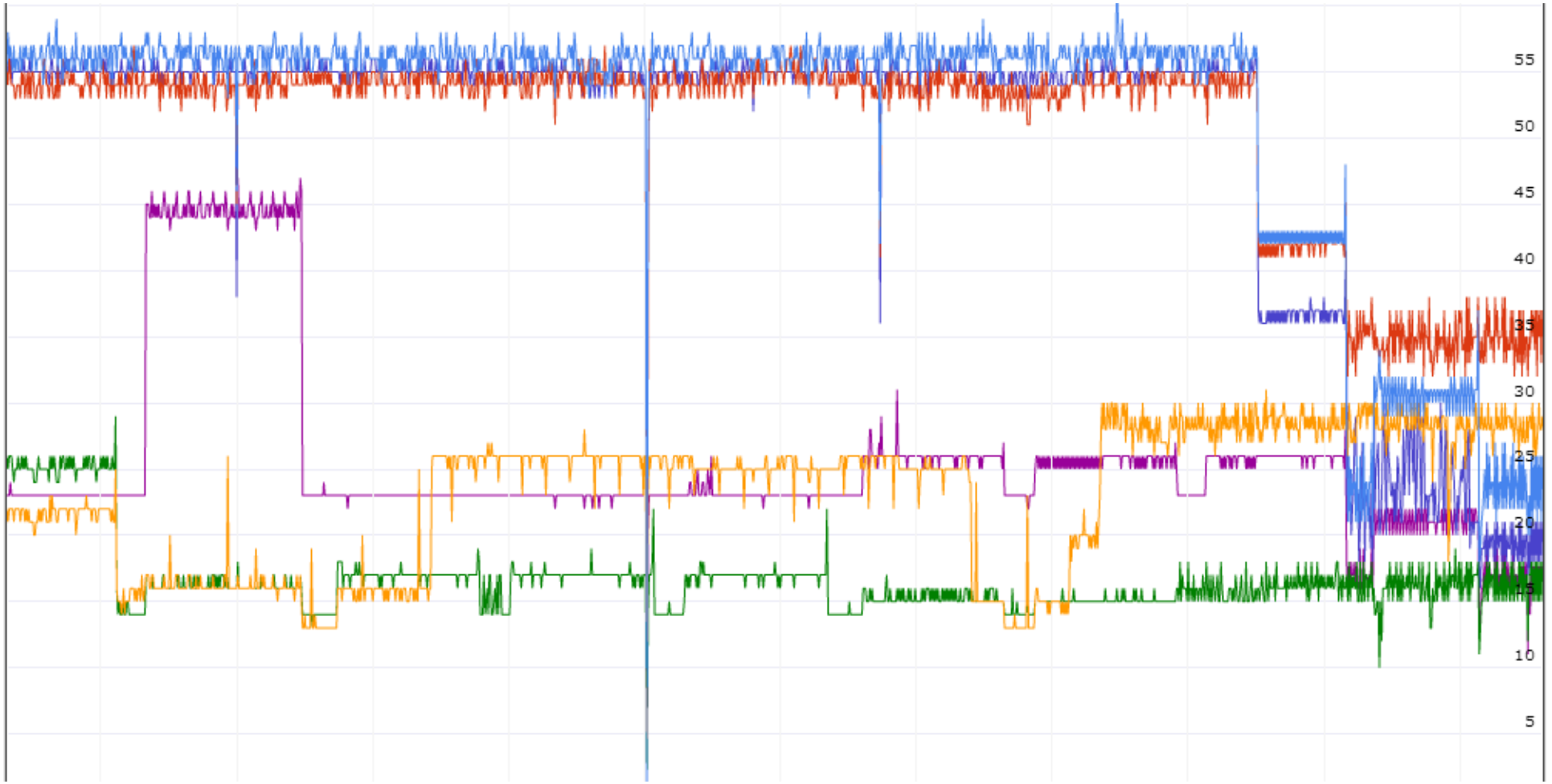


Job Write Performance

No data reduction selected via calc= , using SUM
Job 1622393 is still running
Start=Tue, 28 Apr 2015 11:03:15 -0500
End =Tue, 28 Apr 2015 13:51:32 -0500
Data Query took 10 seconds



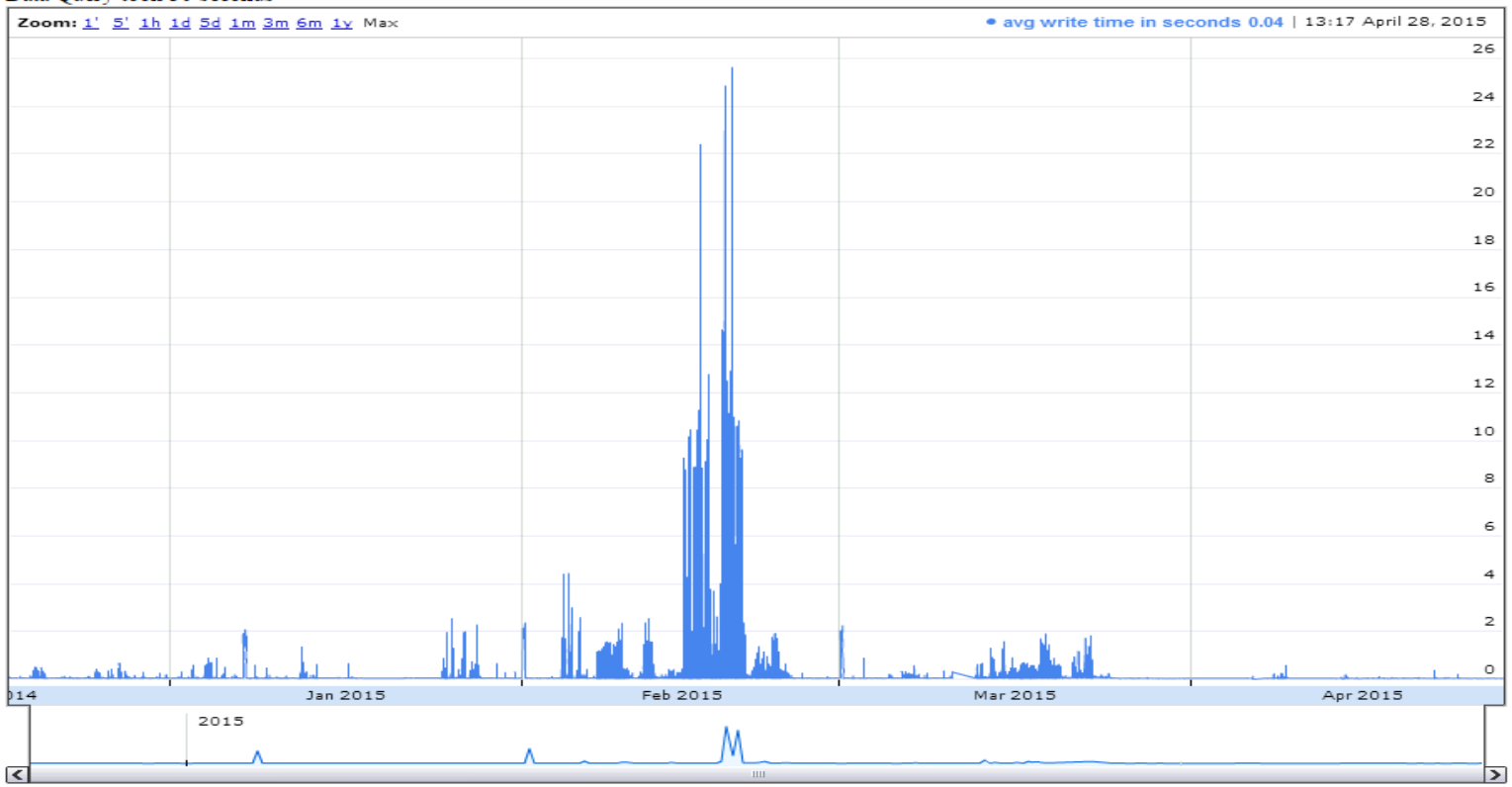
Maximal value across entire system of percent time spent in input queue stall



Network quiesce event just before 10am – all counter to 0

Real World Problem Example

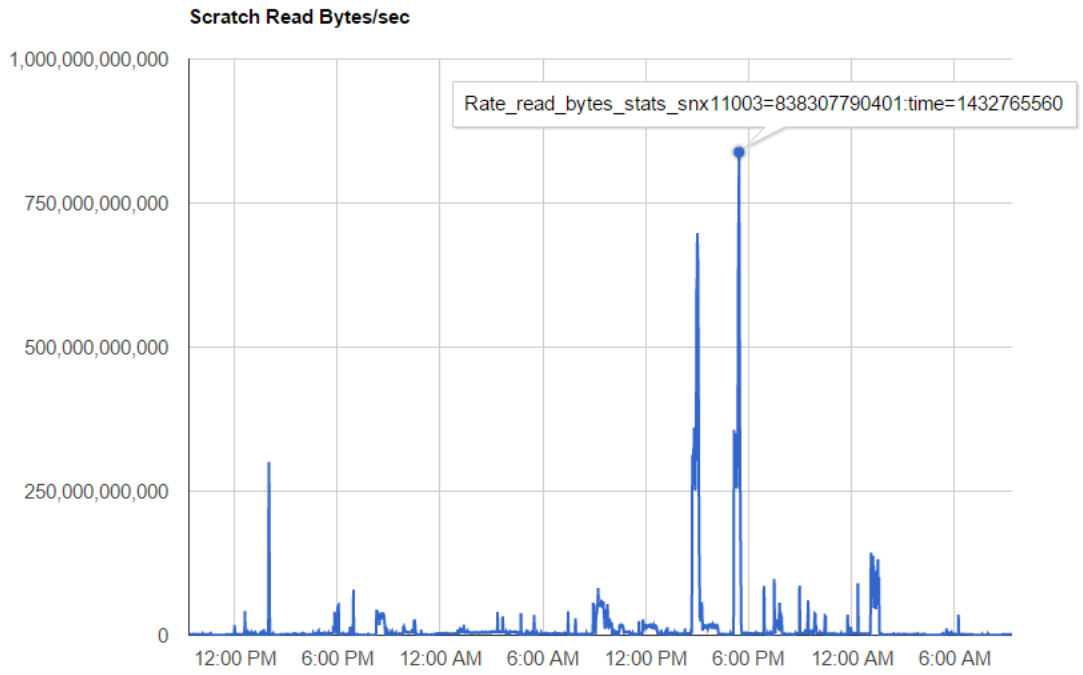
Displaying data for h2ologin
Test Number not set via testnum=, using 2 (write)
Displaying data for home filesystem
No duration Time set via length= , showing data through current time
Start=Mon, 01 Dec 2014 12:14:15 -0600
End =Tue, 28 Apr 2015 13:17:11 -0500
Data Query took 30 seconds



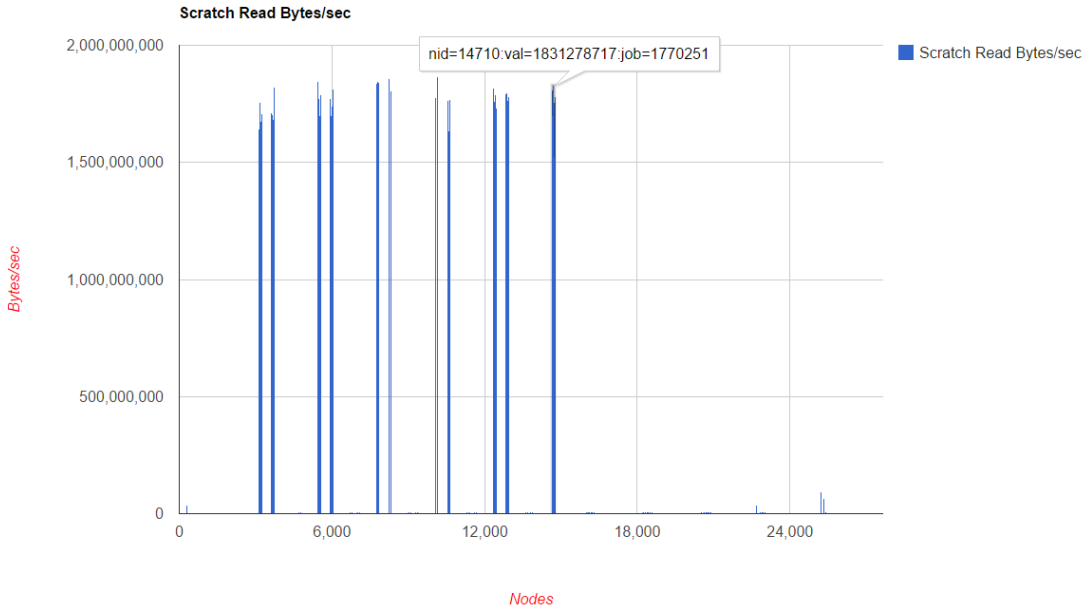
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Real World Problem Example

Sum of scratch reads across entire system



Individual node behavior at the selected time



Next Steps for HMDSA

- The LDMS/OVIS infrastructure has been tested and operates on multiple generations of Cray Systems (XE/XK, XC40) as well as generic Linux clusters at the DOE labs
  - Needs to be optimized and validated on other architectures ECP is interested in – at increasing scale
  - May need to work with other vendors for low level data access to data
- Need to implement ISC components for other architectures, resource managers and log formats
  - Improve and simplify data access for both administrators and users
  - Need to improve data handling for longer term at hand data
- Need to create Intelligent agents to speed up identification of performance impacting events
  - Need training sets, AI implementations, validation, ….

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Monitoring, Performance and Resiliency Related Papers

- S. Leak, A. Greiner, A. Gentile, and J. Brandt, "Supporting Failure Analysis with Discoverable, Annotated Log Datasets," accepted to Cray Users Group (CUG), Stockholm, Sweden. May 2018.
- "Network Congestion in Supercomputers," in submission (double-blind).
- "Data-driven Application-oriented Reliability Model of a High-Performance Computing System", in submission (double blind).
- S. Jha et al., "Holistic Measurement-Driven System Assessment," 2017 IEEE International Conference on Cluster Computing (CLUSTER), Honolulu, HI, 2017, pp. 797-800. doi: 10.1109/CLUSTER.2017.124
- V. Formicola, S. Jha, F. Deng, D. Chen, A. Bonnie, M. Mason, A. Greiner, A. Gentile, J, Brandt, L. Kaplan, J. Repik, J. Enos, M. Showerman, Z. Kalbarczyk, W. Kramer, and R. Iyer. "Data-Driven Understanding of Fault Scenarios and Impacts Through Fault Injection: Experimental Campaign in Cielo." Cray Users Group (CUG), May 2017. Highlight slide
- J. Brandt, E. Froese, A. Gentile, L. Kaplan, B. Allan, and E. Walsh, "Network Performance Counter Monitoring and Analysis on the Cray XC Platform," In Proc. Cray User's Group (CUG), London, England, April 2016.
- A. DeConinck, A. Bonnie, K. Kelly, S. Sanchez, C. Martin, M. Mason, J. Brandt, A. Gentile, and B. Allan, "Design and Implementation of a Scalable Monitoring System for Trinity," In Proc. Cray User's Group (CUG), London, England, April 2016.
- S. Sanchez, A. Bonnie, G. Van Heule, C. Robinson, A. DeConinck, K. Kelly, Q. Snead, and J. Brandt, "Design and Implementation of a Scalable HPC Monitoring System," In Workshop on Monitoring and Analysis for High Performance Computing Systems Plus Applications (HPCMASPA) in conjunction with IEEE Int'l. Parallel and Distributed Processing Symposium (IPDPS) Chicago, IL, USA, May 2016.
- S. Jha, V. Formicola, C. Di Martino, Z. Kalbarczyk, W. Kramer, and R. Iyer, "Analysis of Gemini Interconnect Recovery Mechanisms: Methods and Observations," In Proc. Cray User's Group (CUG), London, England. April 2016.
- C. Keywhan, V. Formicola, Z. Kalbarczyk, R. Iyer, A. Withers, and Adam J. Slagell. "Attacking supercomputers through targeted alteration of environmental control: A data driven case study." In Communications and Network Security (CNS), 2016 IEEE Conference on, pp. 406-410. IEEE, 2016.
- S. Jha, V. Formicola, C. Di Martino, M. Dalton, W. Kramer, Z. Kalbarczyk, and R. Iyer. "Resiliency of HPC Interconnects: A case study of interconnect failures and recovery in Blue Waters." in *IEEE Transactions on Dependable and Secure Computing*, doi: 10.1109/TDSC.2017.2737537.
- Martino, Catello Di, Saurabh Jha, William Kramer, Zbigniew Kalbarczyk, and Ravishankar K. Iyer. "Logdiver: A tool for measuring resilience of extreme-scale systems and applications." In Proceedings of the 5th Workshop on Fault Tolerance for HPC at eXtreme Scale, pp. 11-18. ACM, 2015.
- Eric Heien, Derrick Kondo, Ana Gainaru, Dan LaPine, Bill Kramer, and Franck Cappello: "Modeling and Tolerating Heterogeneous Failures in Large Parallel Systems", ACM Press, Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC '11), pp 45:1-45:11, Seattle, Washington, U.S.A., 2011, doi:10.1145/2063384.2063444
- Ana Gainaru, Franck Cappello, Joshi Fullop, Stefan Trausan-Matu, and William Kramer: "Adaptive Event Prediction Strategy with Dynamic Time Window for Large-Scale HPC Systems", ACM Press, Managing Large-scale Systems via the Analysis of System Logs and the Application of Machine Learning Techniques (SLAML '11), pp 4:1-4:8, Cascais, Portugal, 2011, doi:10.1145/2038633.2038637
- Gainaru, Ana and Cappello, Franck and Snir, Marc and Kramer, William: "Fault Prediction Under the Microscope: A Closer Look into HPC Systems", IEEE Computer Society Press, Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '12), pp 77:1-77:11, Salt Lake City, Utah, U.S.A., 2012
- Ana Gainaru, Franck Cappello, and William Kramer: "Taming of the Shrew: Modeling the Normal and Faulty Behaviour of Large-Scale HPC Systems", IEEE, 2012 IEEE 26th International Parallel and Distributed Processing Symposium, pp 1168-1179, Shanghai, China, 2012, doi:10.1109/ipdps.2012.107
- Saurabh Jha, Jim Brandt, Ann Gentile, Zbigniew Kalbarczyk, Greg Bauer, Jeremy Enos, Michael Showerman, Larry Kaplan, Brett Bode, Annette Greiner, Amanda Bonnie, Mike Mason, William Kramer and Ravishankar Iyer: "Holistic Measurement Driven System Assessment", IEEE, 2017 IEEE International Conference on Cluster Computing (CLUSTER), pp 797-800, Honolulu, Hawai'i, U.S.A., 2017, doi:10.1109/CLUSTER.2017.124
- Di Martino, Catello, William Kramer, Zbigniew Kalbarczyk, and Ravishankar Iyer. "Measuring and understanding extreme-scale application resilience: A field study of 5,000,000 hpc application runs." In Dependable Systems and Networks (DSN), 2015 45th Annual IEEE/IFIP International Conference on, pp. 25-36. IEEE, 2015.
- C. Di Martino, Z. Kalbarczyk, R. Iyer, "Measuring the Resiliency of Extreme-Scale Computing Environments," in *Principles of Performance and Reliability Modeling and Evaluation: Essays in Honor of Kishor Trivedi on His 70th Birthday*, L. Fiondella, A. Puliafito, Eds., Springer International Publishing AG Switzerland, pp. 609–655, 2016.
- Baler: Deterministic, lossless log message clustering tool. N. Taerat, J. Brandt, A. Gentile, M. Wong, and C. Leangsuksun. In: Computer Science - Research and Development. Volume 26, Numbers 3-4, 285-295, DOI: 10.1007/s00450-011-0155-3. Int'l. Supercomputing Conference (ISC). Hamburg, Germany. June 2011.
- New Systems, New Behaviors, New Patterns: Monitoring Insights from System Standup. J. Brandt, A. Gentile, C. Martin, J. Repik, and N. Taerat Workshop on Monitoring and Analysis for High Performance Computing Systems Plus Applications (HPCMASPA) at IEEE Int'l. Conf. on Cluster Computing (CLUSTER) Chicago, IL. Sept 2015.
- A. Agelastos, B. Allan, J. Brandt, P. Cassella, J. Enos, J. Fullop, A. Gentile, S. Monk, N. Naksinehaboon, J. Ogden, M. Rajan, M. Showerman, J. Stevenson, N. Taerat, and T. Tucker. "Lightweight Distributed Metric Service: A Scalable Infrastructure for Continuous Monitoring of Large Scale Computing Systems and Applications." IEEE/ACM Int'l. Conf. for High Performance Storage, Networking, and Analysis (SC14)New Orleans, LA. Nov 2014.
- Michael Showerman, Jeremy Enos, Joseph Fullop, Paul Cassella, Nichamon Naksinehaboon, Narate Taerat, Thomas Tucker, James Brandt, Ann Gentile, and Benjamin Allan. "Large Scale System Monitoring and Analysis on Blue Waters using OVIS." Proc. Cray Users Group, 2014.
- Di Martino, Catello, F. Baccanico, W. Kramer, J. Fullop, J, Z Kalbarczyk, and R Iyer, Lessons Learned From the Analysis of System Failures at Petascale: The Case of Blue Waters, The 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2014)}, June 23-26 2014
- Gainaru Anna, Franck Cappello, Bill Kramer, Event log mining tool for large scale HPC systems, Proceedings of Europar 2011, August 29-September 2, 2011, Bordeaux France.
- B. D. Semeraro, Robert Sisneros, Joshi Fullop, and Gregory H. Bauer: "It Takes a Village: Monitoring the Blue Waters Supercomputer", IEEE, 2014 IEEE International Conference on Cluster Computing (CLUSTER), pp 392-399, 2014, doi:10.1109/cluster.2014.6968671
- R. Sisneros, K. Chadalavada: "Toward Understanding Congestion Protection Events on Blue Waters Via Visual Analytics", presented at CUG 2014, Lugano, Switzerland, 2014
- J. Fullop, R. Sisneros: "A Diagnostic Utility For Analyzing Periods Of Degraded Job Performance", presented at CUG 2014, Lugano, Switzerland, 2014
- https://www.osti.gov/servlets/purl/1368706

Community Interaction

- HMDR participates in the Organizing and Program Committees of the Workshop on Monitoring and Analysis for HPC Systems Plus Applications Series at IEEE Cluster (2014-Current).
- HMDR hosts a Community Vendor-Neutral Website: Monitoring Large-Scale HPC Systems: https://sites.google.com/site/monitoringlargescalehpcsystems/.
- HMDR hosts a regular BoF Series at Supercomputing and CUG. SC14-17 and CUG 16-18.
- HMDR provided leadership and participation in the Cray System Monitoring Working Group (SMWG) -- an international group of Cray sites seeking to improve monitoring and analysis on large-scale platforms.
- Birds-of-a-Feather Sessions, FRESCO: An Open Failure Data Repository for Dependability Research and Practice, SC 2015.

Related Web Sites

- Bluewaters.ncsa.Illinois.edu
- https://github.com/ovis-hpc/ovis/wiki
- http://portal.nersc.gov/project/m888/resilience/~

The HMDSA Team (so far)

University of Illinois

- Greg Bauer – NCSA/BW
- Brett Bode – NCSA/BW
- Jeremy Enos – NCSA/BW
- Ravi Iyer - ECE/CSL
- Zbigniew Kalbarczyk - ECE/CSL
- Bill Kramer – NCSA/BW/CS
- Aaron Saxton – NCSA/BW
- Mike Showerman – NCSA/BW

Sandia National Laboratory

- James Brandt
- Ann Gentile

CS – Computer Science Department
CSL - Coordinated Systems Laboratory
ECE - Electrical and Computer Engineering Department
NCSA - National Center for Scientific Applications
BW – Blue Waters Project

This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.