BLUE WATERS

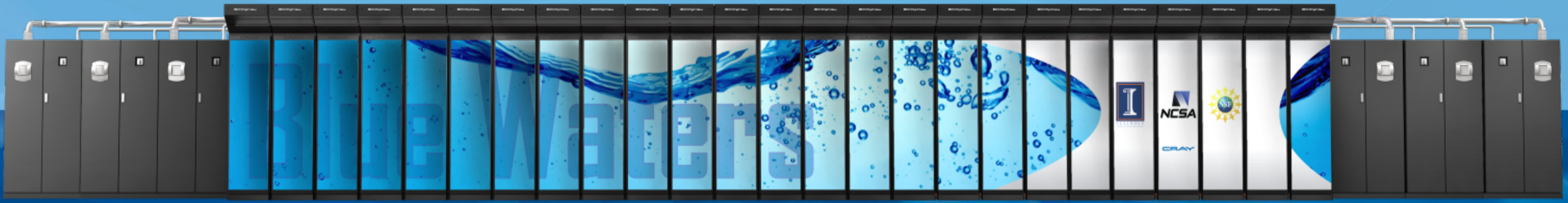
SUSTAINED PETASCALE COMPUTING

Failure and Resiliency in the Shadow of Extreme Scale – Will our Current Assumptions Take Us in the Right Direction?

Professor William Kramer

National Center for Supercomputing Applications, University of Illinois

http://bluewaters.ncsa.illinois.edu



NCSA NSF GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

Agenda

- Overarching Comments
- HMDR Project
- Compare how we have progress
  - Current - Blue Waters Insights and Examples
  - Past – NERSC Insights and Examples
- Blue Waters Monitoring Infrastructure
- Lessons, Needs and Implications for Extreme Scale

WE CANNOT MANAGE WHAT WE CANNOT MEASURE

This is true in general, but especially true in very large, vey complex systems.

THE BEHAVIOR WE MEASURE IS THE BEHAVIOR WE GET

True with people and systems.

With people, measuring is the primary motivator and rewards are enhancers

With systems, we have be the ones that interpret and take action based on the measurements.

WE HAVE TO TAKE HOLISTIC APPROACHES TO @SCALE SYSTEM AND SERVICES MEASUREMENT AND IMPROVEMENT

Holistic – “characterized by comprehension of the parts of something as intimately interconnected and explicable only by reference to the whole” – Google

Holistic – “relating to or concerned with complete systems rather than with individual parts” – Merriiam-Webster Dictonary

What People Want from an HPC System

- Performance – How fast will a system process their work if everything is perfect
- Effectiveness – What is the likelihood they can get the system to do their work
- Reliability – The system is available to do work and operates correctly all the time
- Consistency/(un)Variability – How often will the system process their work as fast as it can
- Usability – How easy is it for them to get the system to go as fast as possible

PERCU

CURRENT AND PAST INSIGHTS

Two Time Periods to Compare

- 2000 – 2008 - NERSC systems
  - 2002 - IBM SP-3 – Seaborg
    - POWER 3+ 375 MHz
    - IBM "Colony" interconnect
    - 416 nodes with 16 cores per node – 6,656 cores
    - 6.7 TBs of Memory (1 GB/core)
  - 2007 - Cray XT4 - Franklin
    - AMD Opteron 2.6 GHz - Dual-Core
    - SeaStar Torus Interconnect
    - 9,660 nodes2 cores pre node - 19,320 cores
    - 38.6 TBs of Memory (2 GB/core)
    - 356 TB of disk
- 2013-2016 - NCSA Blue Waters
  - The largest Cray systems every built
  - 84% XE6 blades, 16% XK7 blades

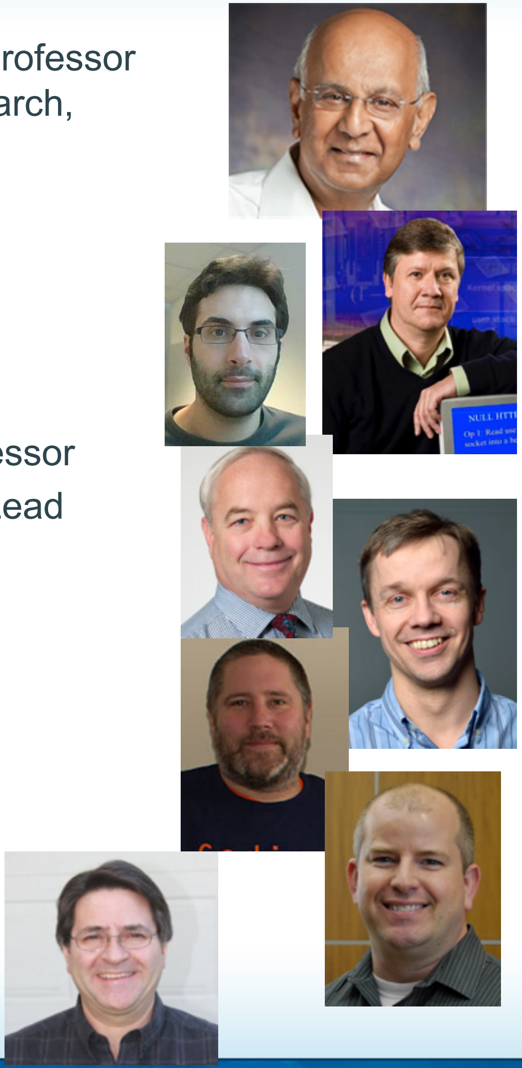
Holistic Measurement Driven Resiliency (HMDR) Project Goals

Overall Goal - We will determine fault → error → failure paths in extreme scale systems of today to help improve those and future systems.

- HMDR is collecting and analyzing a rich set of log and metric data from three generations of large scale HPC platforms
  - Blue Waters at NCSA, Hopper (XE6) and Edison (XC30) at NERSC , Cielo (XE6) at LANL and their successors, Trinity (XC40) and Cori
- Provide a rich understanding of failure modes, root causes, detection/mitigation mechanisms, costs, and impact on applications.
- Provide a deeper understanding of fundamental fault mechanisms and fault/error propagation in current and future systems
- Facilitate other resiliency research by providing annotated datasets from modern extreme-scale systems
- Address a number of multi-generational extreme-scale architectures, including next-generation advanced technologies
- Use fault injection experiments in combination with field data analytics to identify fault/error propagation and improve diagnosability and detectability of faults/errors
- FOR ECP: Enhance our current software tools to support scalable automated: measurement collection, transport, analysis, and fault categorization for enabling application resilience to fault and contention based degradation and failure at Extreme Scale and beyond

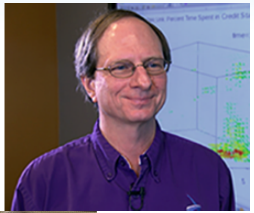
Project Participants

- UIUC:
  - Prof. Ravishankar Iyer - the George and Ann Fisher Distinguished Professor of Engineering AND Leads the DEPEND group focuses on the research, design, and validation of highly available, reliable, and trustworthy computing systems and networks.
  - Dr. Zbigniew Kalbarczyk – Principle Research Scientist in the UI Coordinated Systems Laboratory
  - Dr. Valerio Formicola – visiting scholar and post-doc
- NCSA
  - Prof. William Kramer - Director Blue Waters and CS Research Professor
  - Jeremy Enos – Blue Waters System Management & Development Lead
  - Joseph Fullop – Integrated System Console Lead
  - Mike Showerman – Blue Waters System Resource Manager
  - Graduate students
- Cray:
  - Larry Kaplan – Chief Software Architect – Cray Inc.



Project Participants

- SNL:
  - James Brandt - Distinguished Member of the Computer Science R&D Staff – HPC Monitoring Lead, leads software development effort OVIS/LDMS
  - Dr. Ann Gentile - Principal Member of the Computer Science R&D Staff, SNL Trinity Operations Lead
- NERSC:
  - Dr. Nicholas J. Wright - Advanced Technologies Group Leader and and works in performance modeling and characterization
  - Dr. James Botts – Computational Systems Group
  - Tina Butler. – Computational Systems Group – lead for NERSC 4 and NERSC 6
- LANL:
  - Jim Lujan – Project director of Trinity and Crossroads
  - Cindy Martin – HPC Operations Group Leader

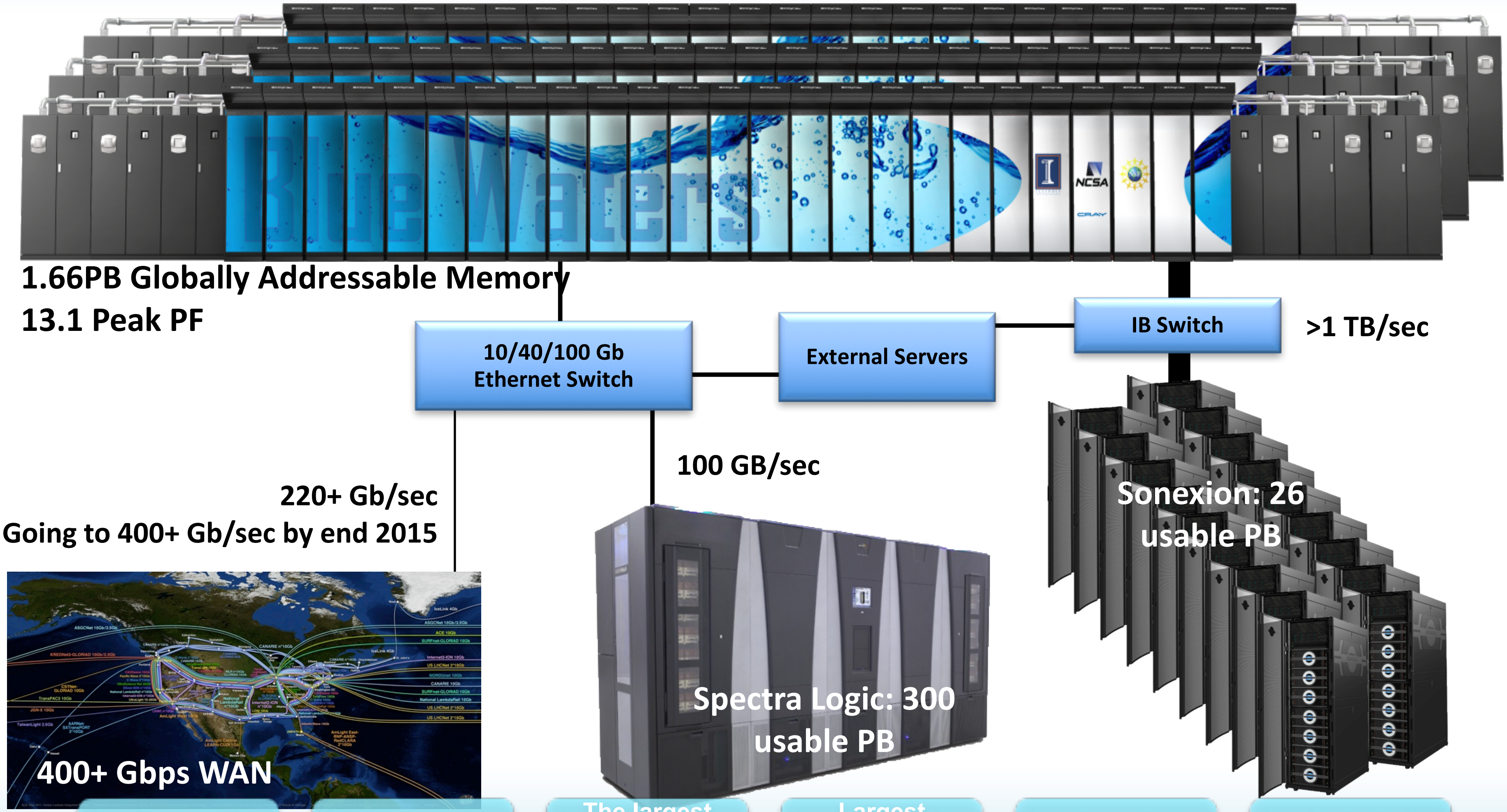


The Blue Waters System

- Comprehensive development, deployment and service phases with co-design and other aspects
- The Blue Waters system is a top ranked system in all aspects of its capabilities.
- Diverse Science teams are able to make excellent use of those capabilities due to the system's flexibility and emphasis on sustained performance.
    - 45% larger than any system Cray has ever built
    - Peak Performance and delivered cycles are approximately the same as the aggregate of all the NSF XSEDE resources.
    - Ranks in the top 5 systems in the world in peak performance – despite being over two years old
    - Largest memory capacity (1.66 PetaBytes) of any HPC system in the world! One of the fastest file systems (>1 TB/s) in the world!
    - Largest nearline tape system (>250 PB) in the world
    - Fastest external network capability (370-470 Gb/s) of any open science site.

Blue Waters Computing System

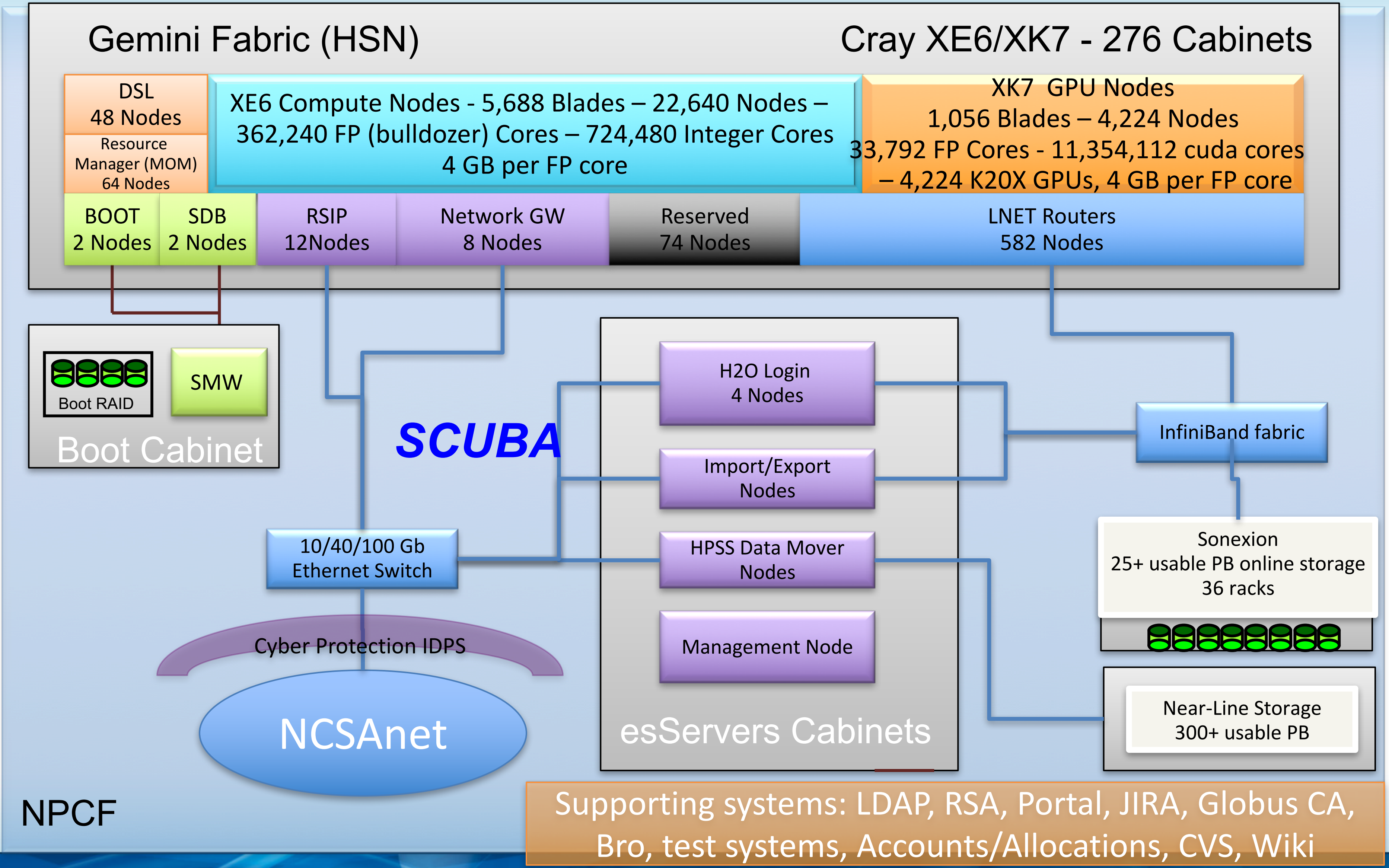
bluewaters.ncsa.illinois.edu



1.66PB Globally Addressable Memory

13.1 Peak PF

- Full Scale use across all ranges of research
- Measured, Sustained 1.3 PF/s over 14 benchmarks
- The largest System Cray has ever built – 45% larger than Titan
- Largest Memory of any system in open science – 1.66 PB
- Most networked facility in open science
- Not listed on the Top500 on purpose



Gemini Fabric (HSN)

Cray XE6/XK7 - 276 Cabinets

DSL
48 Nodes

Resource Manager (MOM)
64 Nodes

XE6 Compute Nodes - 5,688 Blades – 22,640 Nodes –
362,240 FP (bulldozer) Cores – 724,480 Integer Cores
4 GB per FP core

XK7 GPU Nodes
1,056 Blades – 4,224 Nodes
33,792 FP Cores - 11,354,112 cuda cores
– 4,224 K20X GPUs, 4 GB per FP core

BOOT
2 Nodes

SDB
2 Nodes

RSIP
12Nodes

Network GW
8 Nodes

Reserved
74 Nodes

LNET Routers
582 Nodes

Boot RAID

SMW

Boot Cabinet

SCUBA

H2O Login
4 Nodes

Import/Export
Nodes

HPSS Data Mover
Nodes

Management Node

esServers Cabinets

InfiniBand fabric

Sonexion
25+ usable PB online storage
36 racks

Near-Line Storage
300+ usable PB

10/40/100 Gb
Ethernet Switch

Cyber Protection IDPS

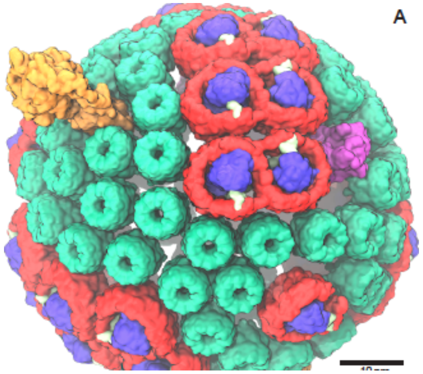
NCSAnet

NPCF

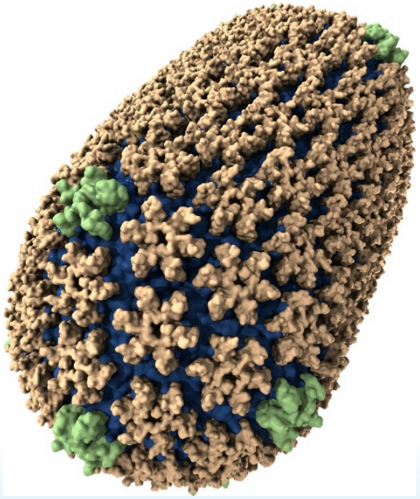
Supporting systems: LDAP, RSA, Portal, JIRA, Globus CA, Bro, test systems, Accounts/Allocations, CVS, Wiki

Schulten - The Computational Microscope

"We were challenged with describing an extremely large structure. … at the very moment when Blue Waters was available. Five years ago, this breakthrough simulation of the HIV virus wouldn't have happened."

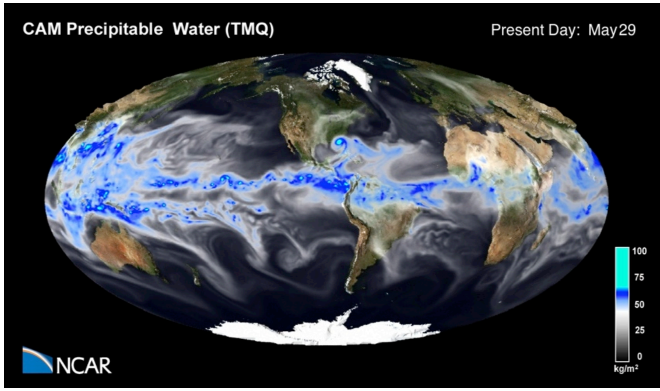


- Challenge Goals
  - First ever atomic-level structure of a native, mature HIV capsid to help scientists understand better how the HIV capsid infects the host cells and could lead to new HIV therapies.
  - The first all-atom model of a cellular organelle. – the Chromatophore which allow the bacteria to absorb sunlight and turn it into chemical fuel that drives many processes in the cell. The chromatophore is composed of about 200 proteins and carries out about 20 processes.
- Usage/Accomplishments
  - Explored the interactions of the full HIV capsid with small molecules for potential drug therapy, Together with experimental collaborators, were able to describe the action of cyclophilin A on the capsid to help scientists understand better how the HIV capsid infects the host cells
  - Complete Chromatophore simulation at the full organelle structure
- Blue Waters Help
  - Enabled graphics driver support on XK nodes that supported work that lead to SC'14 Visualization and Data Analytics Showcase award winner.
  - Performance, Compiler and runtime tuning
  - Topology study using shapes to identify ideal node allocation.



Wuebbles, Washington, et. al. – Climate Change Uncertainties

Blue Waters allows multiple, high resolution runs, 150+ years past and 100 years future, to characterize uncertainty.



- Challenge Goal
  - Validated the effects of very high resolution (10-30 km horizontal resolution) in coupled climate models.
- Usage/Accomplishments
  - 3 present day AMIP (1979-2010) experiments were conducted using CAM5 at 0.25º resolution with different atmosphere/ocean coupling. “After examining the simulations in detail we believe the modified coupling approach (flux calculations on the higher-resolution atmosphere grid) is correct, while the current default coupling is demonstrably unphysical in situations with strong wind curvature.
  - WRF with a resolution of ~1 degree, and dynamically downscale the data using weather research forecasting model (WRF) so we can view predicted atmospheric variables at 12 km resolution
  - Climate-Weather Research Forecasting model (CWRF, Liang et al. 2012) to examine uncertainties in the treatment of cloud, aerosol and radiative transfer processes
- PRAC 338 Million core hours

CURRENT

Example HDMR Insights and Results

- 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. TuckerIEEE/ACM Int'l. Conf. for High Performance Storage, Networking, and Analysis (SC14) New Orleans, LA. Nov 2014.
- Software installed and in use on all current systems within the HDMR collaboration
- Initial log data templates defined and being replicated
- 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; …

Insights credited to the HMDR Project Team

Example HDMR Insights and Results (cont)

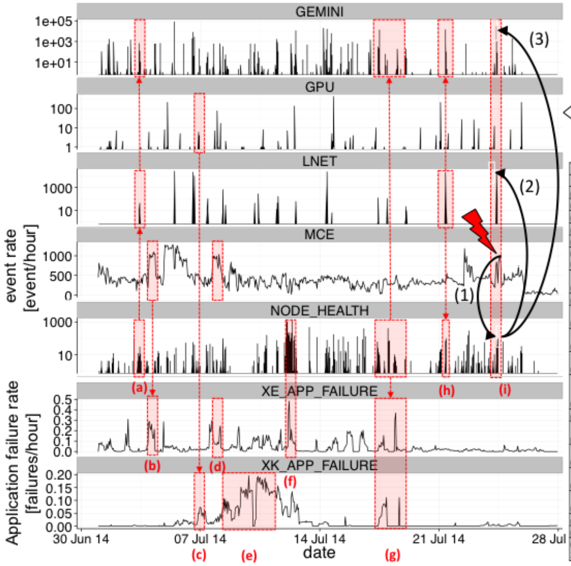
- Insights (cont)
  - 74% of system wide outages are due to software;
  - 50% of these are during failover;
  - filesystem and interconnect are prime contributors;
  - failure of failover causes a significant number of system wide outages;
  - application failure increases with increasing duration of failover time.
    - Catello Di Martino, Zbigniew Kalbarczyk, William Kramer, Ravishankar Iyer, “Measuring and understanding extreme-scale resilience: A field study of 5,000,000 HPC application runs,” 45th IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2015, pp. 25–36, 22–25 June 2015. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7266835
    - 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

Insights credited to the HMDR Project Team

Example HDMR Insights and Results (cont)

- HMDR Web portal created and published
  - http://portal.nersc.gov/project/m888/resilience/
- Software released
  - Blue Waters ISC posted on github - https://github.org/ncsa/isc
  - LDMS and OVIS available – https://gethub.org/ovis-hpc

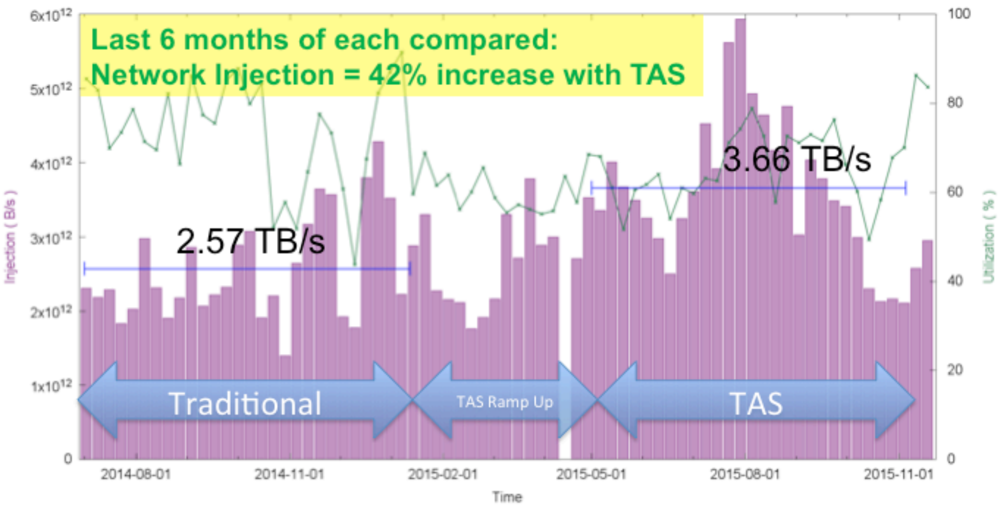
Examples of Results: Improving systems through understanding of root causes of faults, failure propagation, and performance changes



```
SINGLE/DOUBLE BIT
 Device Memory  :22
 Register File  :1
 L1 Cache       :0
 L2 Cache       :8
 Texture Memory :17
MMU ERRORS         :8
GPU BUS RESET      :4
GPU SNX BUS ERRORS:22
```

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 - Published

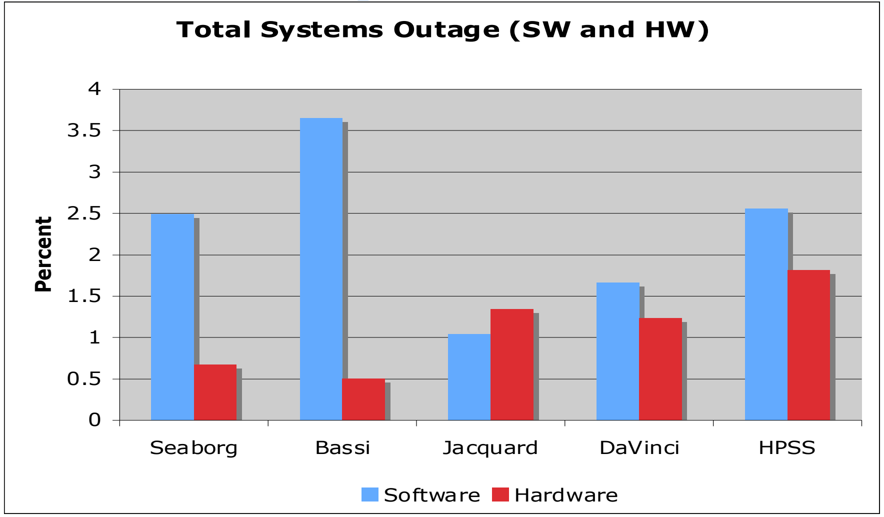


One root cause of significant performance degradation addressed by “topologically aware scheduling” – Publication in preparation

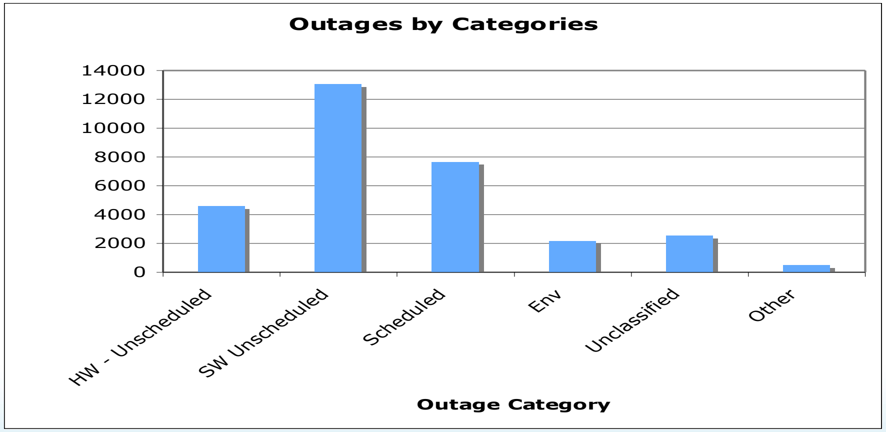
BACK TO THE PAST

Reliability is the Key Issue

- 6 Years of NERSC System

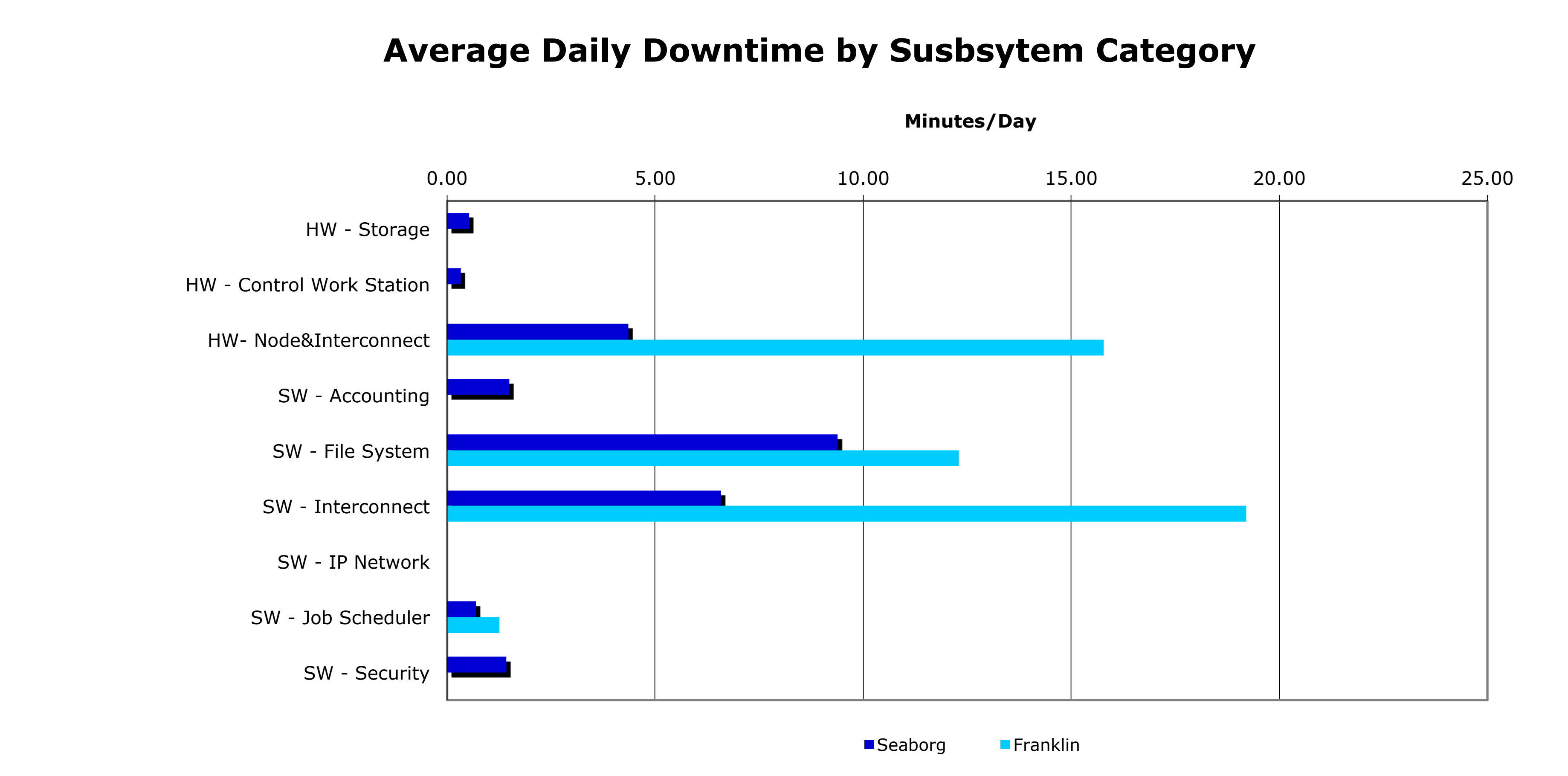


- 4 Months of XT-4 Franklin

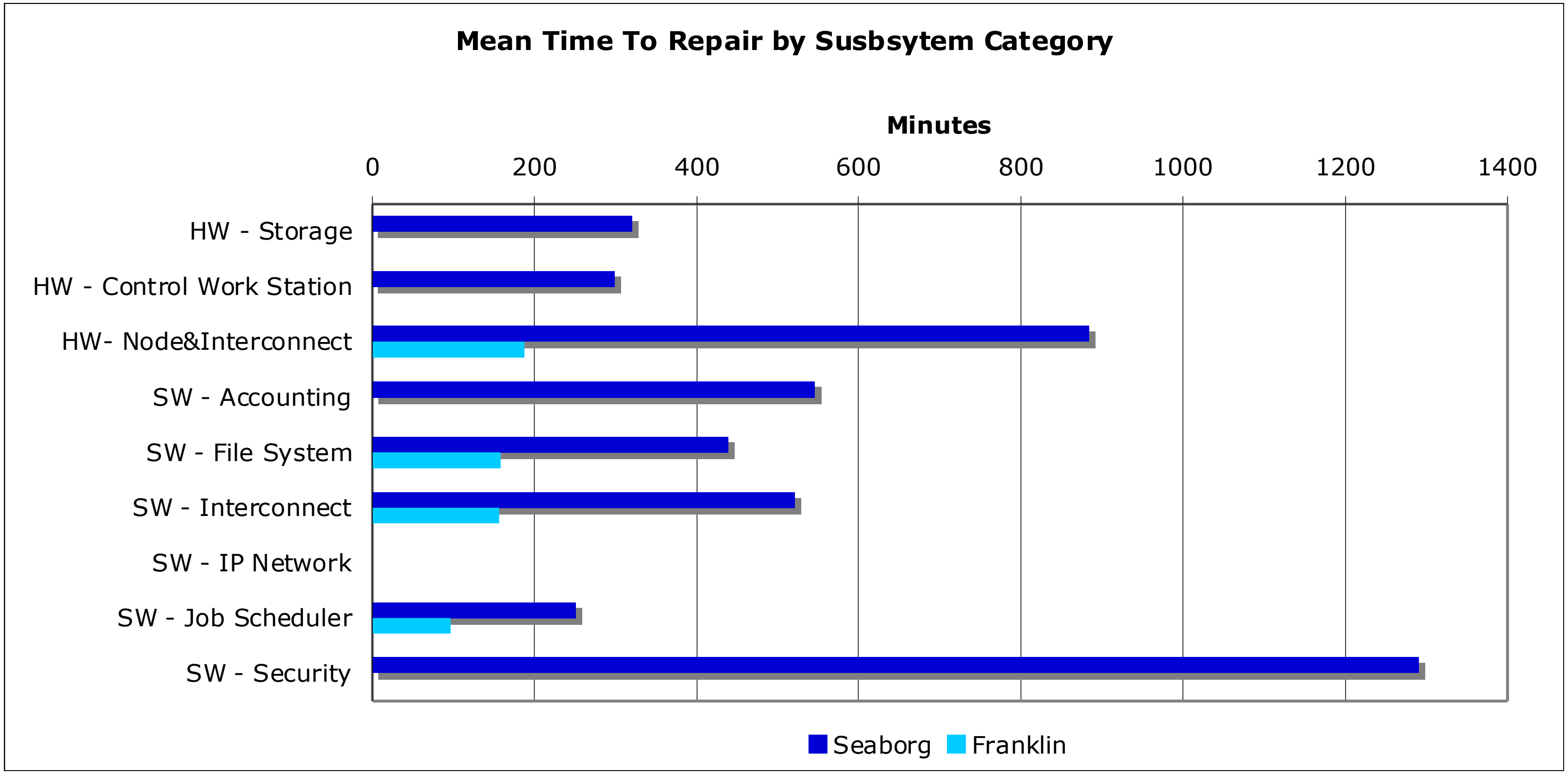


6 Year Charts Courtesy of Akbar Mokhtarani

Software Reliability



Software Reliability



Reliability - Even If a System Stays Up- Will Jobs Succeed?

Job Failure Error Categories for the NERSC Cray XT-4	Number of Jobs	Percent of Jobs
SUCCESS - Job clearly succeeds	117,884	66.2%
WALLTIME - Job ran to the wall clock time limit - some user causes - some system causes	12,614	7.1%
WIDTH - A mismatch between job request and aprun command request -– normally a user error	0	0.0%
NODEFAIL – A node assigned to the job failed or crashed – possibly hardware	192	0.1%
UNEX - This error indicates MPI buffers need to be increased	75	>0.05%
ENOENT – A requested executable file does not exist	1,148	0.6%
LIBSMA - An error within the SHMEM communication library	70	>0.05%
SIGTERM - Job received a Terminate Signal that could have been from the user or the system	58	>0.05%
NOAPRUN - The batch job did not appear to execute an aprun, usually due to a scripting error	6,516	3.7%
NOTRACE –Process accounting data could not be traced to identify the aprun associated with this job. The job did execute an aprun but the parent process id was 1 so it could not be properly matched. Usually a job was killed or a system wide failure	11,389	6.4%
QUOTA - Job exceeded a File System quota	2,865	1.6%
ATOMIC – The job failed due to a software problem when using parts of the SHMEM library (fixed)	4	>0.05%
UNKNOWN - The status of the job completion was non-determinate. aprun command had a non-zero exit code may be due to a system problem or due to some user action that prevents recording the exit status	25,318	14.2%
Total	178,133	

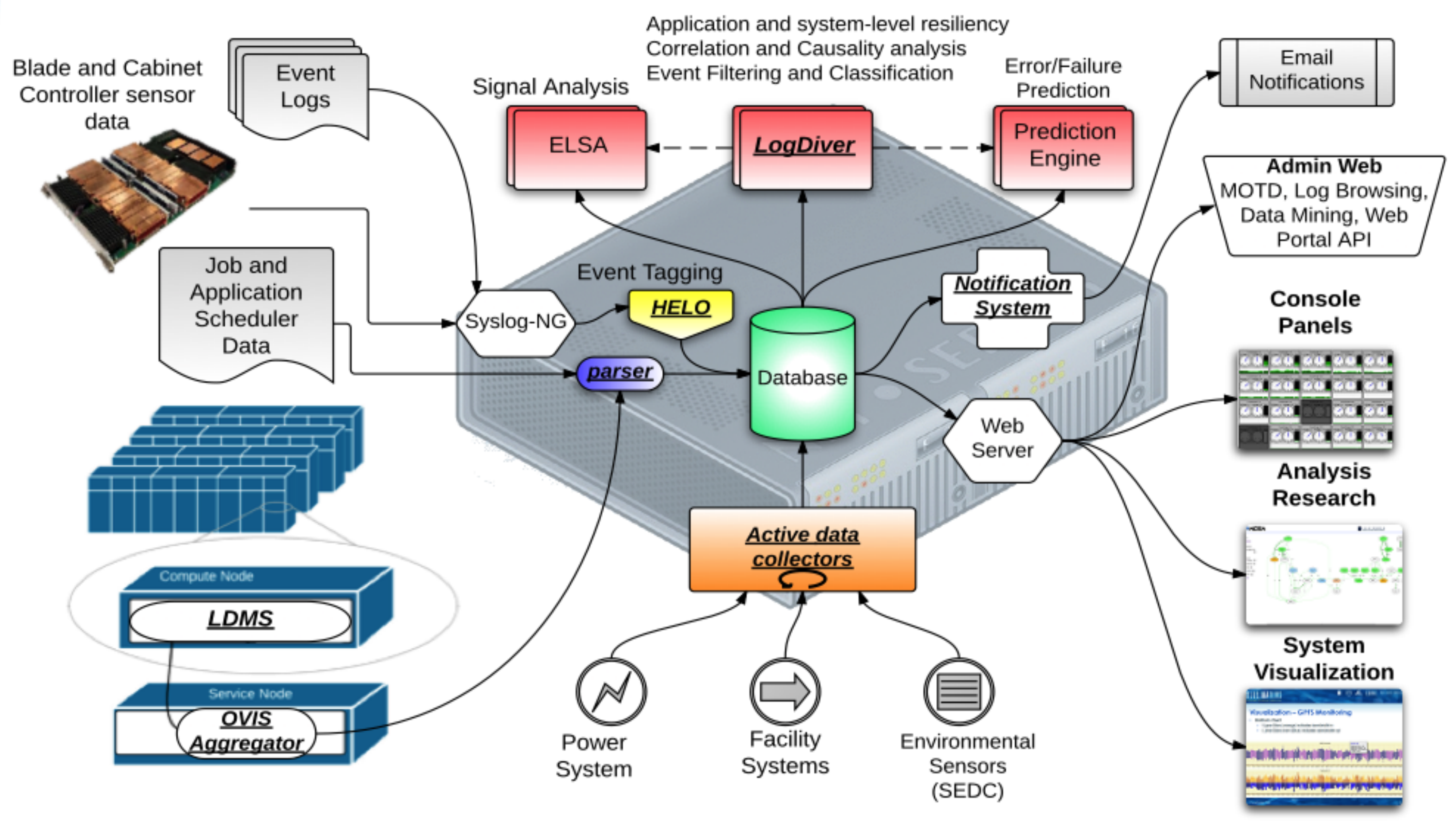
COLLECTING AND ANALYZING SYSTEM MONITORING DATA

Current Database Data

- Node
  - Load average
    - Latest,5min, running processes, total processes
  - 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

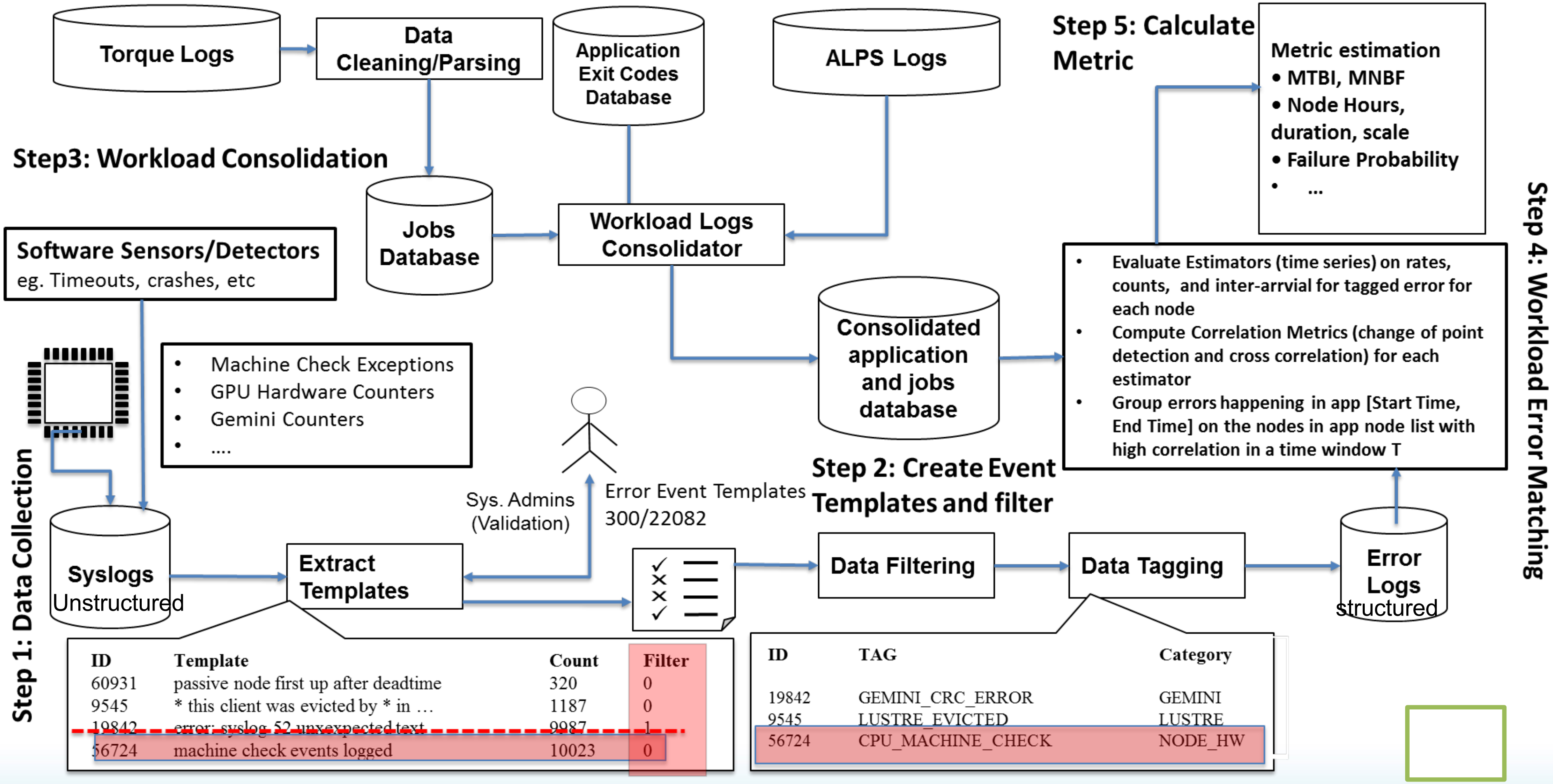
- Gemini Link Statistics
  - All 6 directions
  - Link BW, %used, avg packet size, %input queue stalls, %credit stalls
- Gemini/NIC Statistics
  - totaloutput_optA/B, total input, fma output, bet output
  - SMSG
    - Number tx/rx rate , Bytes tx/rx rate
  - RDMA
    - Number tx/rx rate , Bytes tx/rx rate
  - IP over Gemini
    - Tx/rx rate

Analysis Architecture



On Blue Waters, today 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.

LogDiver workflow



Torque Logs
Data Cleaning/Parsing
Application Exit Codes Database
ALPS Logs
Step 5: Calculate Metric
Step3: Workload Consolidation
Jobs Database
Workload Logs Consolidator
Consolidated application and jobs database
Software Sensors/Detectors
eg. Timeouts, crashes, etc
Syslogs
Unstructured
Extract Templates
Sys. Admins (Validation)
Error Event Templates 300/22082
Step 2: Create Event Templates and filter
Data Filtering
Data Tagging
Error Logs
structured
Step 1: Data Collection
Step 4: Workload Error Matching

Metric estimation
- MTBI, MNBF
- Node Hours, duration, scale
- Failure Probability
- …

- Machine Check Exceptions
- GPU Hardware Counters
- Gemini Counters
- ….

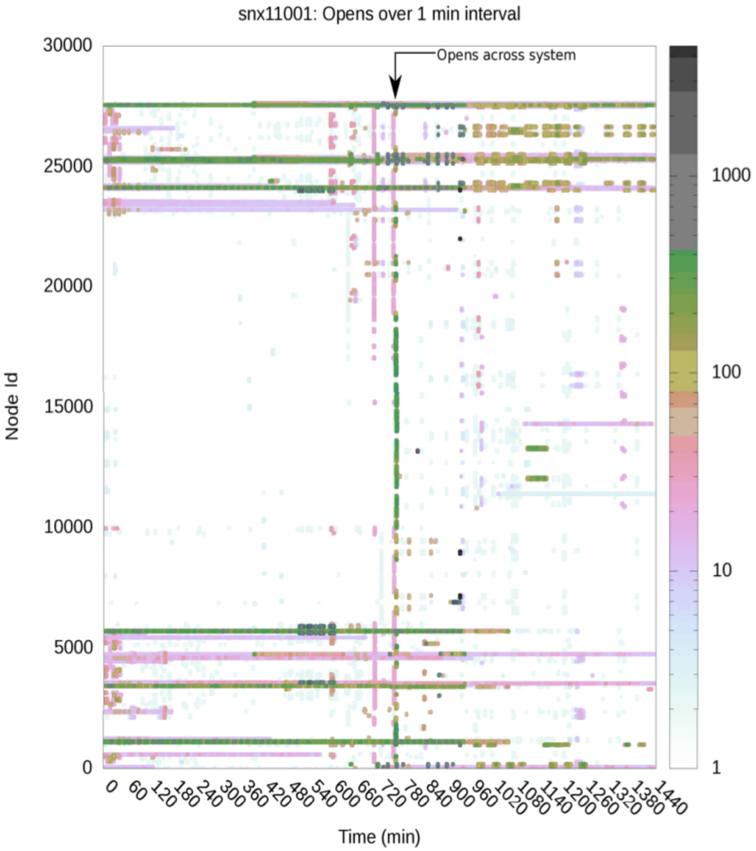
- Evaluate Estimators (time series) on rates, counts, and inter-arrvial for tagged error for each node
- Compute Correlation Metrics (change of point detection and cross correlation) for each estimator
- Group errors happening in app [Start Time, End Time] on the nodes in app node list with high correlation in a time window T

ID	Template	Count	Filter
60931	passive node first up after deadtime	320	0
9545	* this client was evicted by * in …	1187	0
19842	error: syslog 52 unexpected text	9987	1
56724	machine check events logged	10023	0

ID	TAG	Category
19842	GEMINI_CRC_ERROR	GEMINI
9545	LUSTRE_EVICTED	LUSTRE
56724	CPU_MACHINE_CHECK	NODE_HW

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

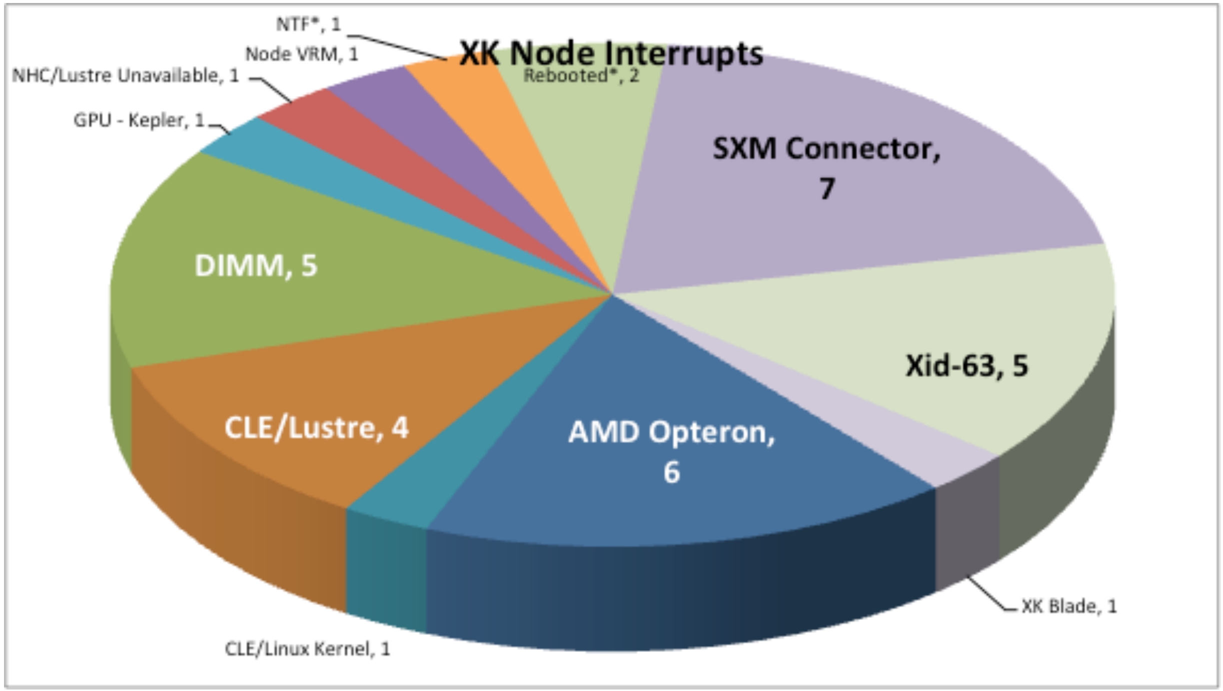
- Provides HPC system state data unique in scope and fidelity
  - Whole system snapshots down to sub-second intervals
  - Minimal impact on platform resources
  - No measurable adverse impact on large-scale application run times
- Features:
  - Synchronous collection for coherent system snapshots
  - 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)



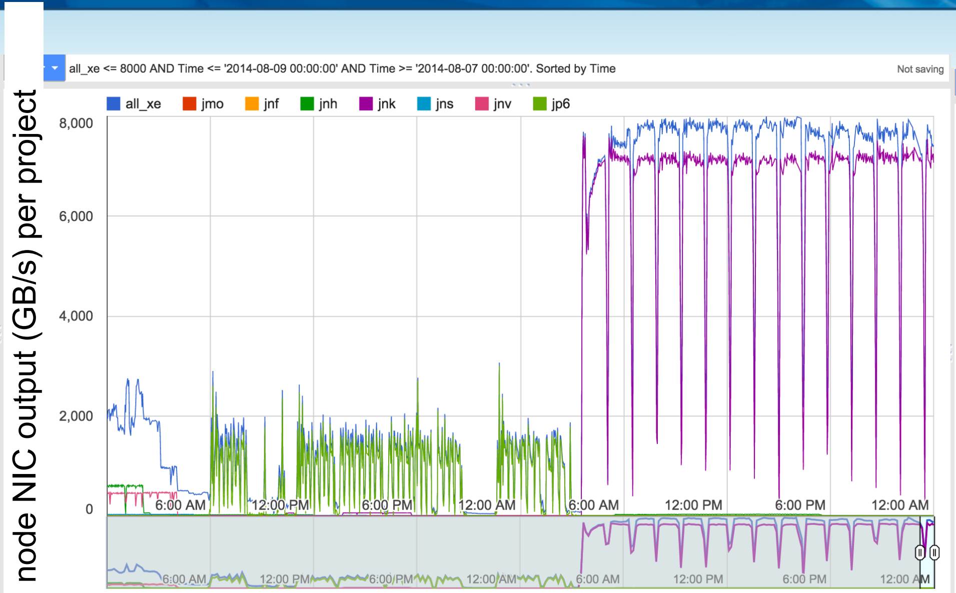
Blue Waters: One day dataset contains ~40 million data points per metric and 7.7 billion data points overall

Slide credited the HMDR Project Team and LDMS/Ovis team

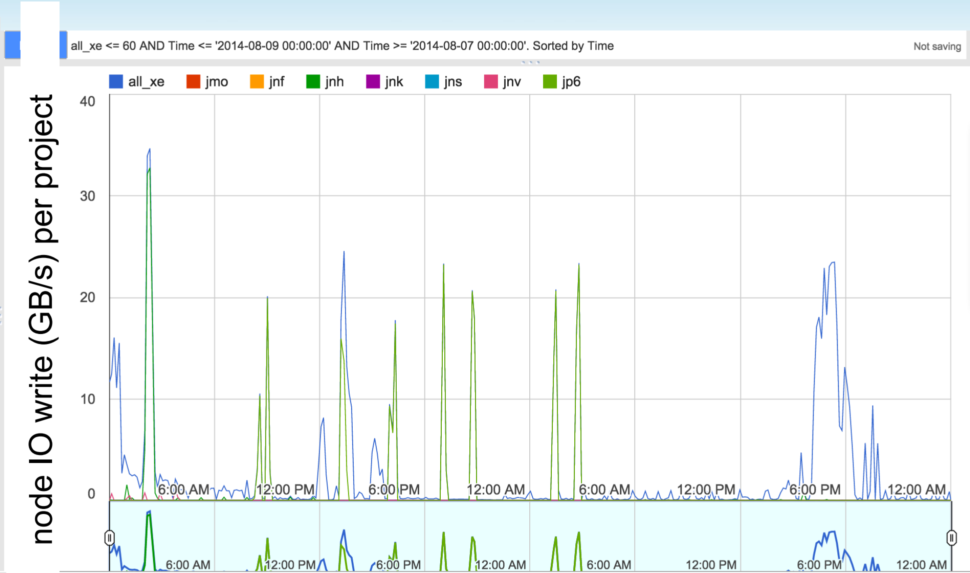
Examples of Data: Node Failure Causes on Blue Waters



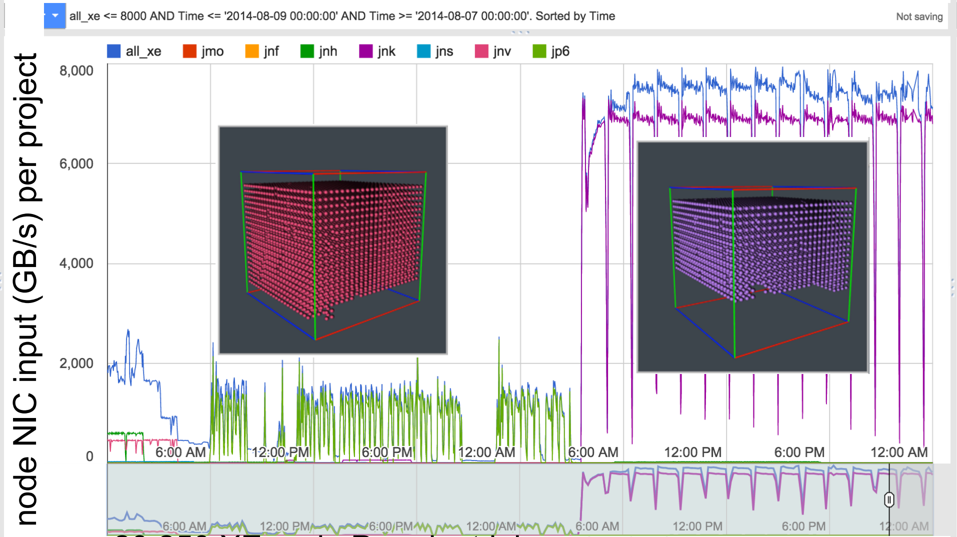
Example Views of Job Behavior



all_xe <= 8000 AND Time <= '2014-08-09 00:00:00' AND Time >= '2014-08-07 00:00:00'. Sorted by Time

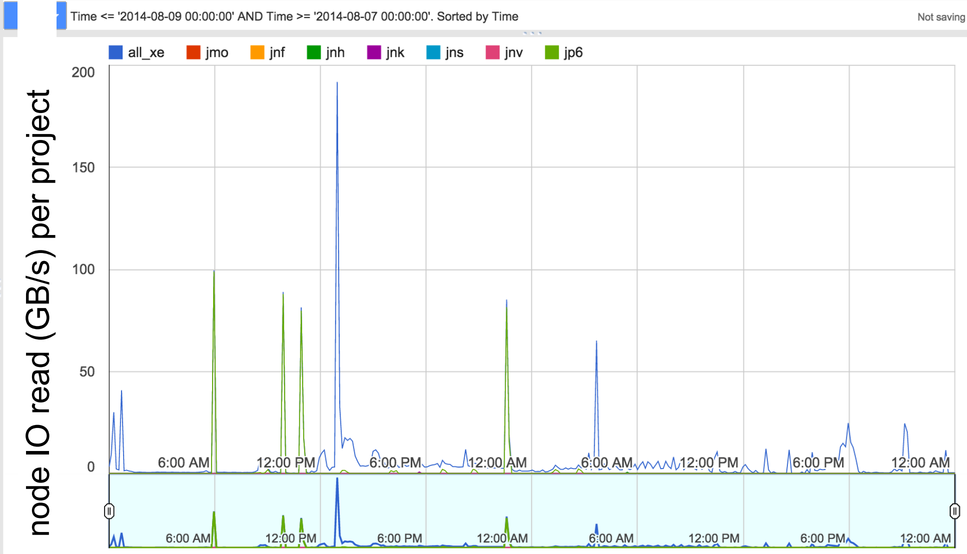


all_xe <= 60 AND Time <= '2014-08-09 00:00:00' AND Time >= '2014-08-07 00:00:00'. Sorted by Time



all_xe <= 8000 AND Time <= '2014-08-09 00:00:00' AND Time >= '2014-08-07 00:00:00'. Sorted by Time

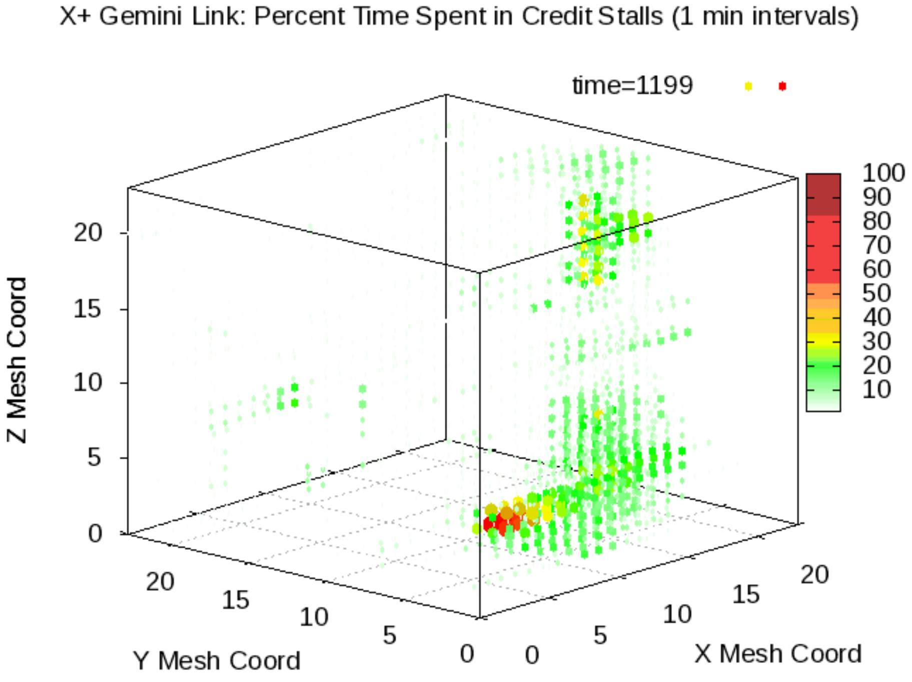
20,250 XE node P-gadget job 15,872 XE node namd job

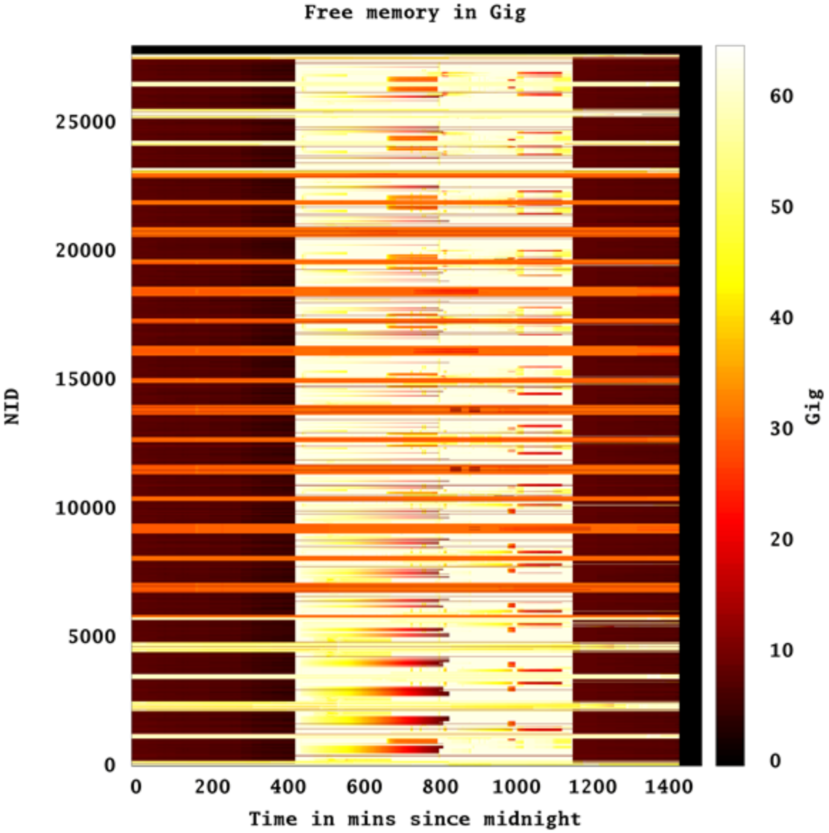


Time <= '2014-08-09 00:00:00' AND Time >= '2014-08-07 00:00:00'. Sorted by Time

Data credited the HMDR Project Team, the LDMS/Ovis Team and NCSA Blue Waters Team

3D Animations and Graphs



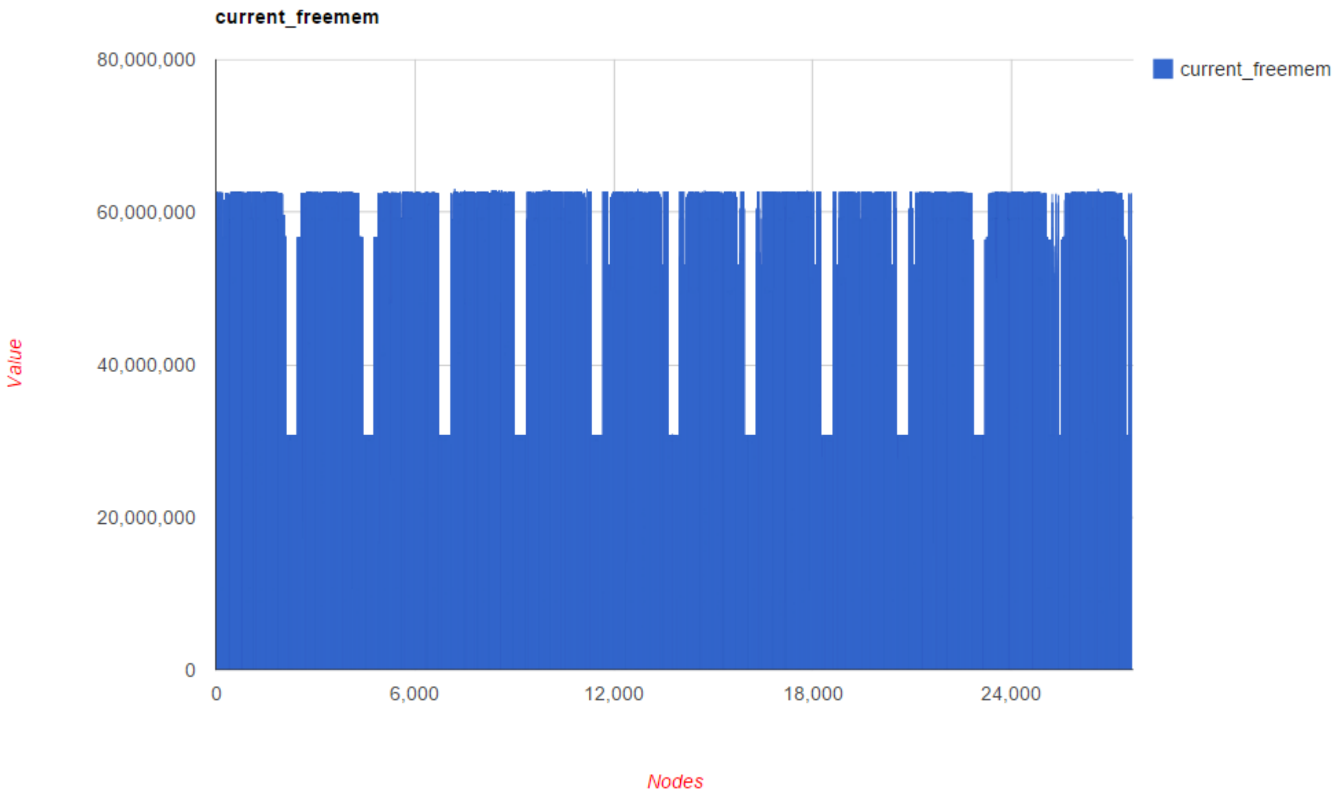


Network Contention

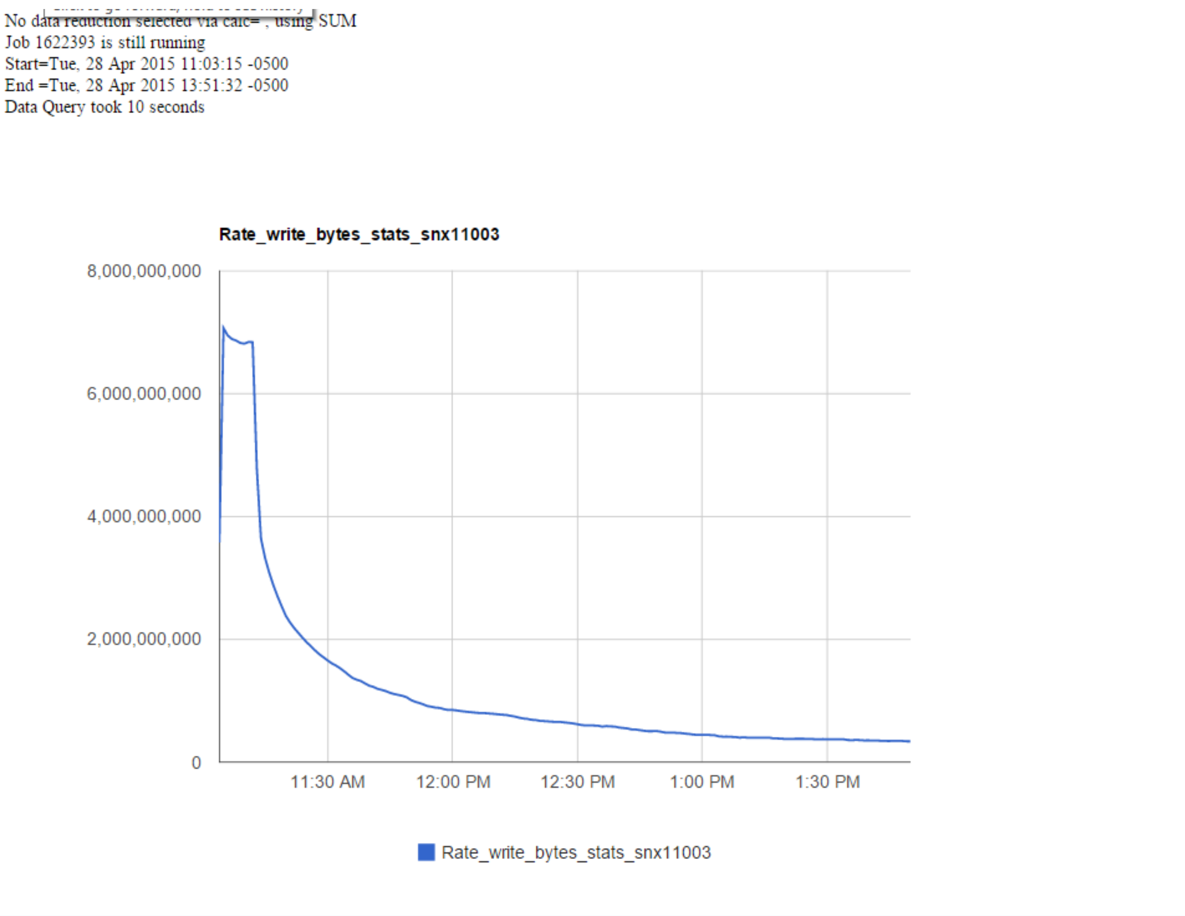
Free Memory

Data credited the HMDR Project Team, the LDMS/Ovis Team and NCSA Blue Waters Team

Single Metric in Time (Free Memory)



Job Write Performance

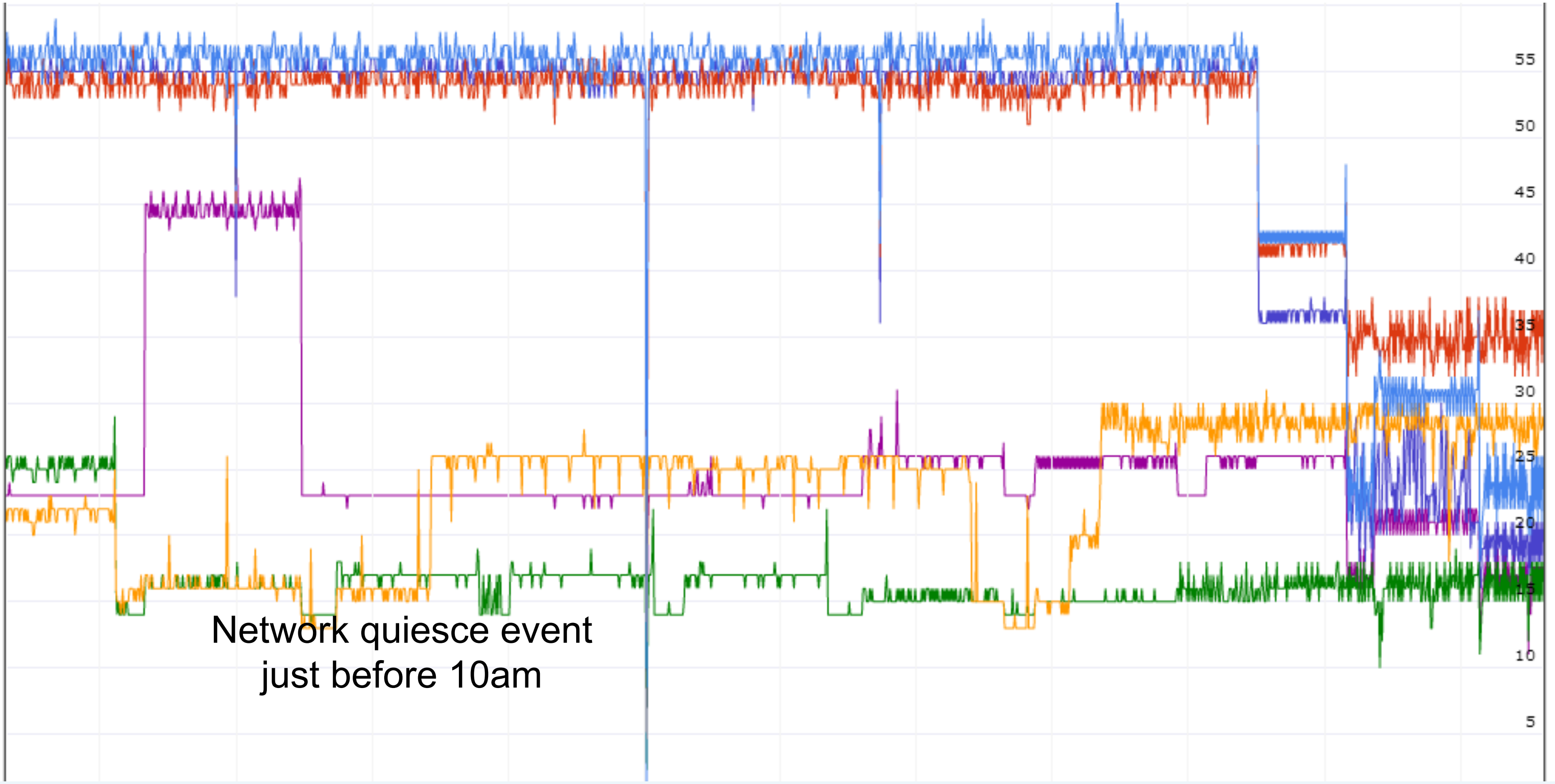


```
No data reduction selected via calc= , using SUM
Job 1622393 is still running
Start=Tue, 28 Apr 2015 11:03:15 -0500
End =Tue, 28 Apr 2015 13:51:32 -0500
Data Query took 10 seconds
```

Rate_write_bytes_stats_snx11003

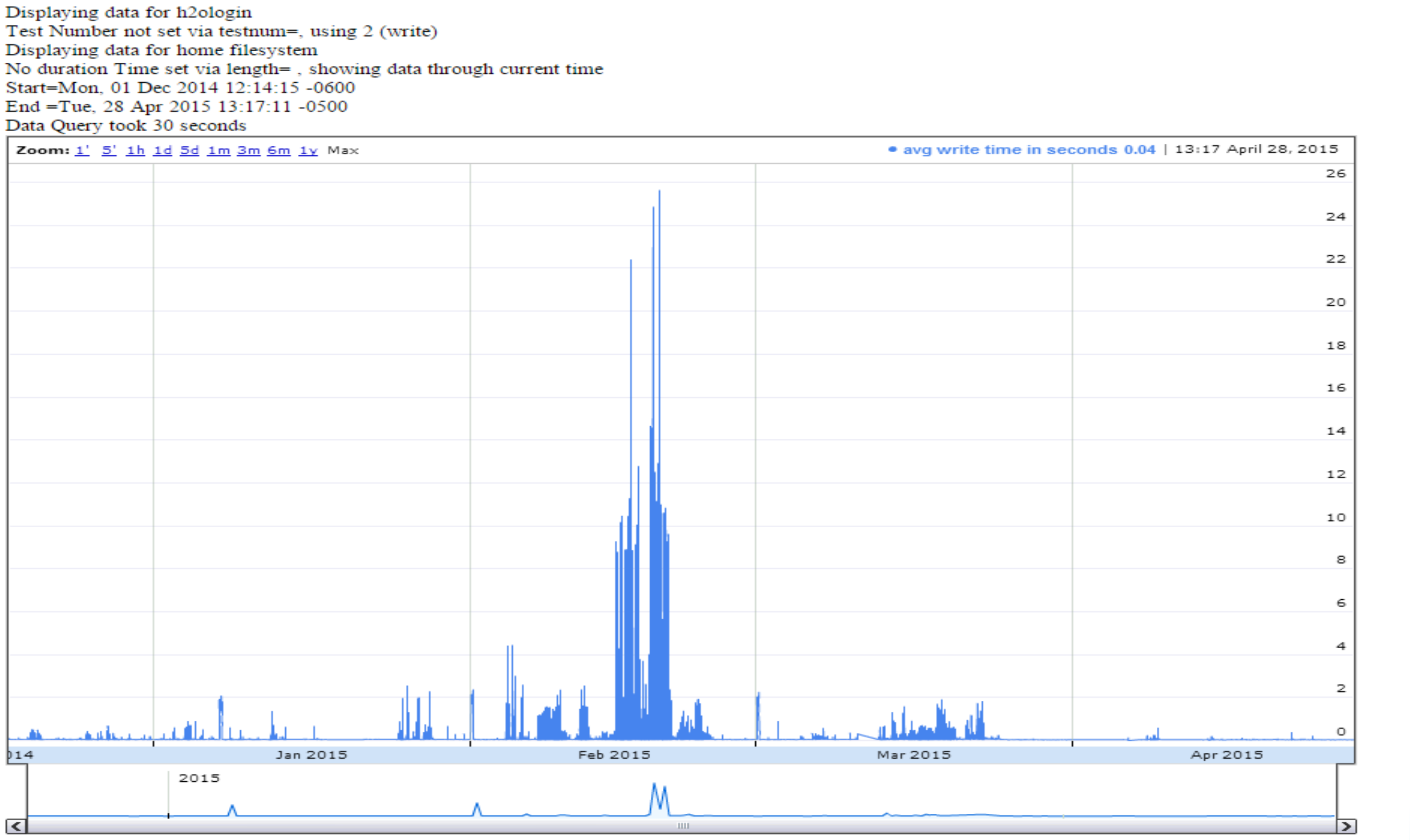
Data credited the HMDR Project Team, the LDMS/Ovis Team and NCSA Blue Waters Team

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



Network quiesce event just before 10am

Real World Problem Example



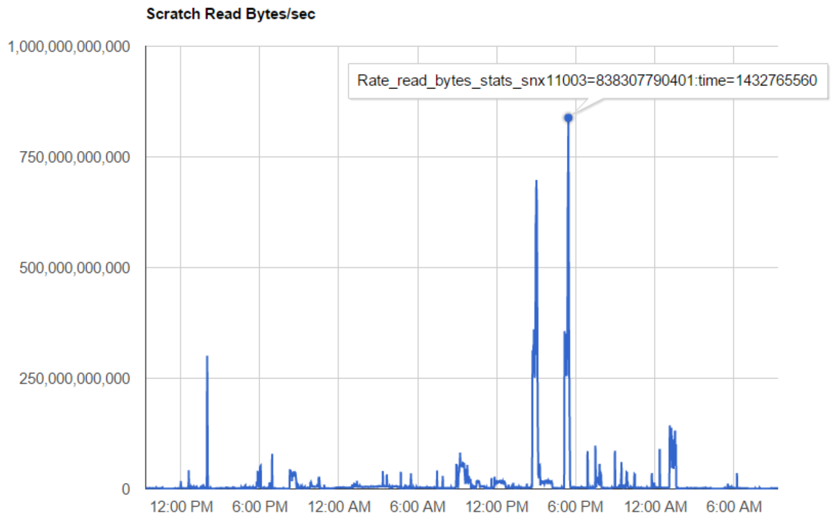
Displaying data for h2ologin
Test Number not set via testnum=, using 2 (write)
Displaying data for home filesystem
No duration Time set via length= , showing data through current time
Start=Mon, 01 Dec 2014 12:14:15 -0600
End =Tue, 28 Apr 2015 13:17:11 -0500
Data Query took 30 seconds

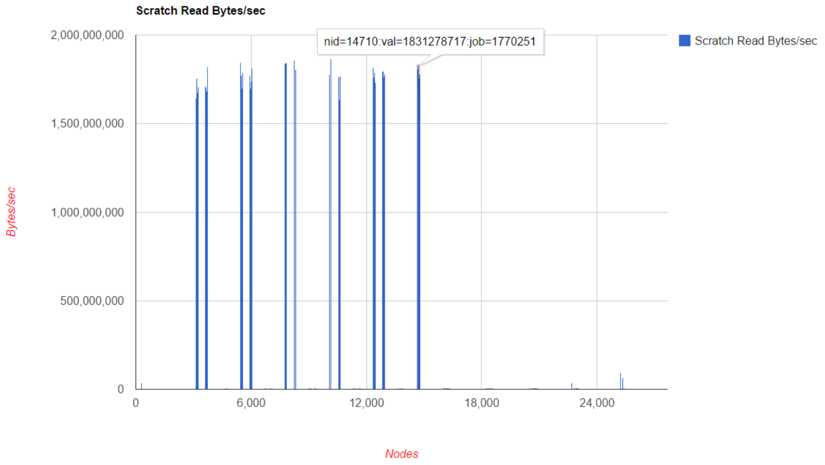
Zoom: 1' 5' 1h 1d 5d 1m 3m 6m 1y Max
● avg write time in seconds 0.04 | 13:17 April 28, 2015
26
24
22
20
18
16
14
12
10
8
6
4
2
0
014
Jan 2015
Feb 2015
Mar 2015
Apr 2015
2015

Data credited the HMDR Project Team, the LDMS/Ovis Team and NCSA Blue Waters Team

Sum of scratch reads across entire system

Individual node behavior at the selected time





EXTREME SCALE VISION – HOLISTIC MEASUREMENT DRIVEN SYSTEM ASSESSMENT (HMDSA)

Extreme Scale Vision – Holistic Measurement Driven System Assessment (HMDSA)

- A complete, holistic performance and resiliency monitoring, run-time analysis, and system software/application interactive infrastructure for Extreme Scale platforms that is
  - Platform independent with architecture specific probes
  - Provides near realtime situational data for applications and system services
  - Provides extensive post-processing of monitoring data
  - Automatically catalogues and processes notifications and data
- The data and analysis needed for HMDR is expandable to other areas for understanding complex (aka Extreme Scale) systems.
  - Reliability and Resiliency
  - Fault root causes
  - Failover
  - Holistic system performance and productivity
    - “Black Box” Analysis
  - Sub-system specific improvements
  - Application based assessments with large numbers of application runs and data points
- From the “user” perspective, performance inconsistency and resiliency and failovers are all on a continuum of making systems more productive for their work

Extreme Scale HMDSA Tools Challenges

- Distributed parallel system level monitoring framework
  - New technology data source samplers
  - New network technology transport plugins
- Parallel Post-processing log and numeric data Analysis
  - New technology related characterization and anomaly detection
  - New technology related root cause and propagation analysis
    - New fault to failure propagation paths
    - New performance degradation mechanisms
- Streaming analysis, filter, and feedback plugins to monitoring framework
  - Develop hooks and mechanisms to enable low latency stack level and application appropriate feedback, of fault/failure/performance degradation situations
- Already addressing scalability with parallelism incorporated into tool design

Vision for HMDSA Activities for Extreme Scale

A complete, holistic performance and resiliency monitoring, run-time analysis, and system software/application interactive infrastructure for Extreme Scale platforms that is

- Assesses all PERCU areas
- Platform independent with architecture specific probes
- Provides near realtime situational data for applications and system services
- Provides extensive post-analysis of monitoring data
- Automatically catalogues and processes notifications and data
- Creates a suite of analysis tools for understanding the huge data volumes monitoring and managing Extreme Scale systems will create

What Should We Monitor and Act On for Petascale and Extreme Scale

- We have to collect, assess and analzye the measures that will tell us where a system is Performant, Effective, Reliable/resilient, Consistent and Usable.
  - Control Metrics – tens
  - Operational Metrics – millions to billions
- We have to be able to answer the relationship questions about these areas
  - When does a slowdown in performance become a failure
  - How to tell a system failure from an application load
  - …

HMDSA Today and for Extreme Scale

- Tools Today
  - Right fidelity system-level monitoring – LDMS/OVIS, ISC, HELO
  - Run-time and post-processing analysis – HELO, Logdiver, LDMS plugins
    - Root cause analysis of failures including performance degradation
  - Run-time feedback and system reconfiguration (e.g., topologically aware scheduling) - ISC
- Tools for Extreme Scale System
  - Insights with Petascale and Trans-petascale systems can inform the design and development and operation of Extreme Scale systems.
  - Streaming distributed parallel system-level diagnosis and system/application feedback to enable real time and non-real time reconfiguration for malleable applications and resource management
  - Provide longer term guidance for improvements in resiliency and failover and also for reducing system bottlenecks that limit application performance and time to solution
  - HDMR can evolve to HDMSA

What are our Practical Issues

- System logging inconsistency
  - System and subsystem logs change – we spend lots of time just collecting and parsing logs
  - We need more common and self documenting formats
- We still use email messages for most information exchange
- Lack of common definitions causes confusion
  - We need a well understood, flexible, common taxonomy for discussion and understanding
- We may not be attacking the most important issues
- …..

Summary

- As a community, we have made tremendous progress in being able to collect “Petascale” metrics
  - Orders of Magnitude More
  - Little or now interference with application performance
- We struggle to understand what we collect
- We need a much larger focus and better understanding of software failures since they have the largest impacts
- We need to understand the relationships and how to handle multiple, simultaneous failures and faults
- We need to move from forensic analysis to situational awareness for the Extreme Scale

Acknowledgements

- 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.
- This research is part of the *Holistic Measurement Driven Resilience: Combining Operational Fault and Failure Measurements and Fault Injection for Quantifying Fault Detection and Impact* Project which is supported by the Department of Energy Office of Science ASCR awards (OCI-0725070 and ACI-1238993) and the state of Illinois.

Partial List of Other references

- Brett Bode, Michelle Butler, Thom Dunning, William Gropp, Torsten Hoe- fler, Wen-mei Hwu, and William Kramer (alphabetical). The Blue Waters Super-System for Super-Science. Contemporary HPC Architectures, Jeffery Vetter editor. Sitka Publications, November 2012.Edited by Jeffrey S . Vetter, Chapman and Hall/CRC 2013, Print ISBN: 978-1-4665-6834-1, eBook ISBN: 978-1-4665-6835-8
- Kramer, William, Michelle Butler, Gregory Bauer, Kalyana Chadalavada, Celso Mendes, Blue Waters Parallel I/O Storage Sub-system, High Performance Parallel I/O, Prabhat and Quincey Koziol editors, CRC Publications, Taylor and Francis Group, Boca Raton FL, 2015, Hardback Print ISBN 13:978-1-4665-8234-7.
- *Understanding the System Wide Impacts of Topology Aware Scheduling and Extreme Scale Systems* – in preparation
- *Resiliency Challenges in HPC interconnects: Understanding Failures of Failovers in Gemini Networks* – in preparation
- *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.Cappello, Franck, Al Geist, William Gropp, Sanjay Kale, Bill Kramer, Marc Snir, *Toward Exascale Resilience: 2014 update*, Supercomputing Frontiers and Innovations, Jack Dongarra and Vladimir Voevodin editors, June 2014.
- *Large-Scale Persistent Numerical Data Source Monitoring System Experiences*,, J. Brandt, A. Gentile, M. Showerman, J. Enos, J. Fullop, and G. BaueWorkshop on Monitoring and Analysis for High Performance Computing Systems Plus Applications (HPCMASPA) at IEEE Int'l. Parallel and Distributed Processing Symposium (IPDPS) Chicago, IL. May 2016.
- Large Scale System Monitoring and Analysis on Blue Waters Using OVIS -- Best Paper Finalist
- M. Showerman, J. Enos, J. Fullop (NCSA), P. Cassella (Cray), N. Naksinehaboon, N. Taerat, T. Tucker (OGC), J. Brandt, A. Gentile, and B. Allan (SNL)
- Cray User's Group (CUG), Lugano, Switzerland. May 2014. Gainaru, Ana, Franck Cappello, Marc Snir, William Kramer - *Failure prediction for HPC systems and applications: current situation and open issues*, International Journal of High Performance Computing, 2013.
- Dongarra, Jack et al., *The International Exascale Software Project Roadmap*. Int. Journal of High Performance Computing Applications 25(1): 3-60 (2011)
- Cappello, Franck, Al Geist, William Gropp, L. Kale, Bill Kramer, and Marc Snir. *Toward Exascale Resilience*. Int. Journal of High Performance Computing Applications, 23(4):374{388, 2009, http://hpc.sagepub.com/content/23/4.toc

Partial List of Other references

- Cappello, Franck, Al Geist, William Gropp, L. Kale, Bill Kramer, and Marc Snir. *Toward Exascale Resilience*. Int. Journal of High Performance Computing Applications, 23(4):374{388, 2009, http://hpc.sagepub.com/content/23/4.toc
- Kramer William, and David Skinner, *An Exascale Approach to Software and Hardware Design,* International Journal of High Performance Computing Applications November 2009 23: 389-391, doi:10.1177/1094342009347768, http://hpc.sagepub.com/content/23/4.toc
- William Kramer and David Skinner, *Consistent Application Performance at the Exascale*, International Journal of High Performance Computing Applications November 2009 23: 392-394, doi:10.1177/1094342009347700, http://hpc.sagepub.com/content/23/4.toc
- 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
- Mendes, Celso, Greg Bauer, William Kramer, Robert Feidler, *Expanding Blue Waters with Improved Acceleration Capability*, 2014 Cray User Group Proceedings, Lugano, Switzerland, May 5-8, 2014
- Mendes, Celso L., Brett Bode, Gregory H. Bauer, Joseph R. Muggli, Cristina Beldica and William T. Kramer, *Blue Waters Acceptance: Challenges and Accomplishments*, Cray User Group 2013, May 10, 2013, Napa California.
- Gainaru Ana, F. Cappello, M. Snir, B. Kramer, *Failure Prediction for HPC systems and applications: current situation and open issues*, International Journal of High Performance Computing Applications, SAGE, 2013
- Gainaru, Ana, Franck Cappello, Marc Snir, William Kramer*, Fault prediction under the microscope: A closer look into HPC systems.* ACM/IEEE SC12, November 12-15, 2012, Salt Lake City, UT