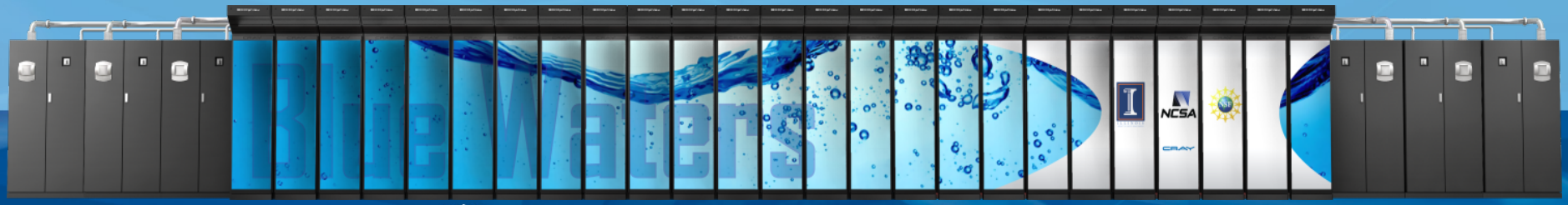
BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Expected and Unexpected Challenges to Extreme Scale Reliability

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

Abstract

Extreme scale systems of today and tomorrow have on the order of a million to ten million processing elements, tens of millions of memory components and kilometers of interconnection cables. But these extreme scale systems, such as Blue Waters, are executing potentially billions of lines of software at any given instant in time. Studies of reliability traditionally focus on the many hardware components, their failure rates and the steps an application might take to mitigate such failures. While hardware failures are important to address, it is increasingly obvious many, some argue most, of system failures are software based. Equally concerning is that recovery time (MTTR) from software errors takes longer and means the probability of double and triple faults having to be addressed simultaneously makes recovery and resiliency much more challenging. This talk will exam recent trends in reliability and performance analysis using most the data collected over more than two years of Blue Waters service. It will draw insights as to the failure causes and possible solutions to make systems and applications more resilient. It will also offer comments on how to use today's insights into designing and implementing better systems and applications in the future.

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

When does performance/consistency issues become a reliability issue?

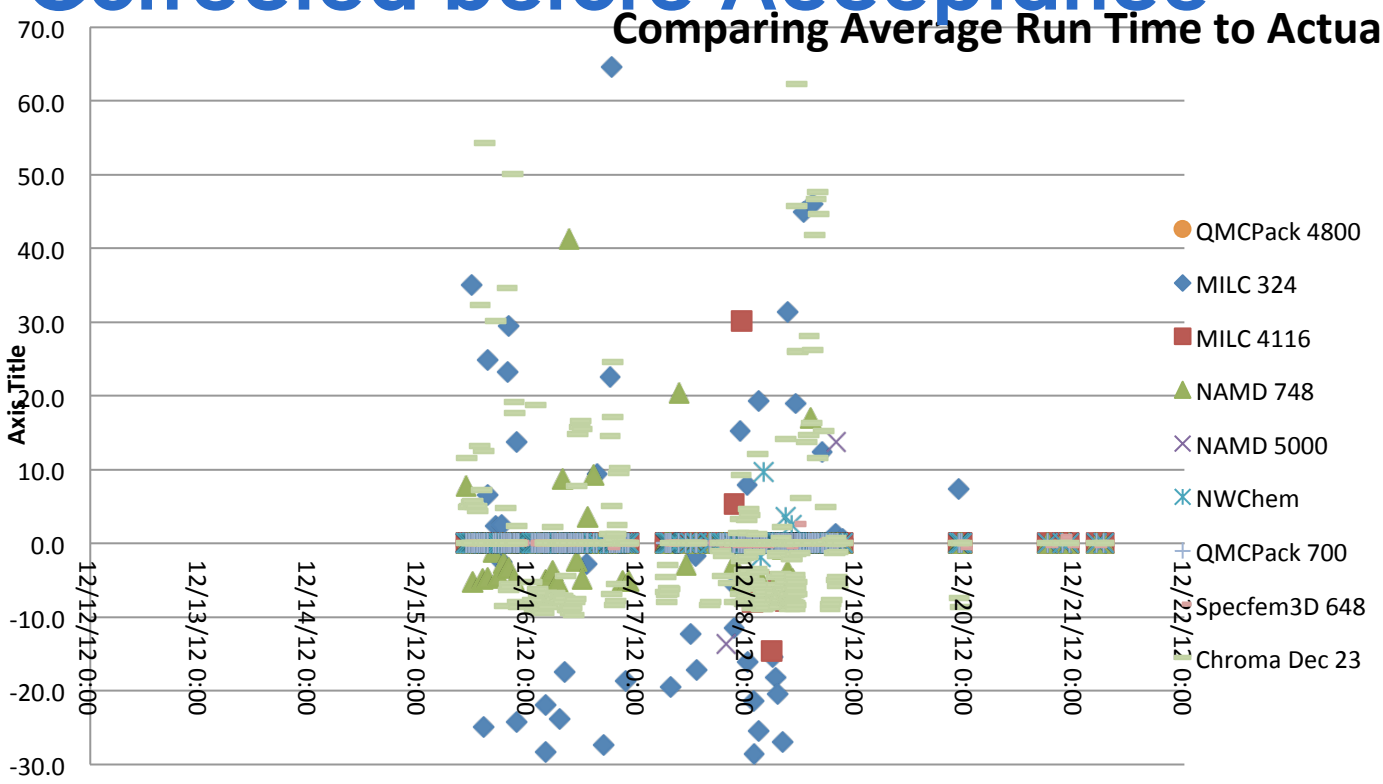
Have to be clear on Definitions

- Fault, Error, Failure
- Root Cause vs Probable Cause
  - Most logs/documents list probable cause
    - Some SW guess
    - Some Human Expert
  - Root Cause is
    - Very difficult to determine
    - Very expensive to determine
    - Takes Long time to determine
- Examples
  - Consistency Team
  - Lustre Design and Metadata Impacts
  - Via failures in processors
  - Lustre Metadata Performance

Interconnect Congestion Fault Found and Corrected before Acceptance

Application	Partition	Time Period Start	Time Period End	Size (Nodes)	Runs	Job Time COV	Aprun Time COV	Comment
QMPACK	XE	11/15 13:48	11/18 14:28	4800	34	0.32%	0.34%	
QMPACK	XE	11/15 13:47	11/18 14:33	700	34	0.11%	2.85%	aprun COV 0.34% if one outlier removed
Chroma	XK	11/13 04:16	11/30 19:22	768	588	17.9%	23.4%	November Runs
QMCPACK	XK	11/15 08:55	11/18 09:36?		46	16.5%	16.5%	All data, bimodal operation
QMCPACK	XK	11/15 08:55	11/18 09:36?		35	0.24%	0.27%	Low series data
QMCPACK	XK	11/15 08:55	11/18 09:36?		11	0.19%	0.38%	High series data
MILC	XE	12/17 21:15	12/18 11:23	4116	6	16.2%	18.4%	
MILC	XE	12/15 11:36	12/19 23:49	324	53	26.0%	38.0%	
NAMD	XE	12/15 10:54	12/18 21:36	5000	4	13.0%	12.8%	
NAMD	XE	12/15 10:28	12/18 15:25	768	33	9.90%	9.57%	
NWCHEM	XE	12/18 02:58	12/18 10:22	5000	6	6.50%	6.70%	
Chroma	XE	12/15 10:40	12/19 22:58	768	231	13.0%	12.1%	Job Time COV is 31.4% if assumed crashed run included
SPECFEM3D	XE	12/15 12:59	12/16 00:24	5419	7	1.78%	1.83%	3600 Steps
SPECFEM3D	XE	12/17 21:15	12/18 21:22	5419	7	0.77%	0.78%	7200 Steps
SPECFEM3D	XE	12/16 17:47	12/21 11:06	675	22	0.66%	0.67%	
HPL	XK	12/15 13:01	12/20 12:08		1051	17.9%	18.6%	Odd Toggling behavior
Chroma	XK	12/15 10:41	12/18 11:10	768	63	12.5%	9.6%	December Runs
NAMD	XK	12/15 08:55	12/15 10:39	768	3	9.1%	9.1%	Very small sample size

Interconnect Congestion Fault Found and Corrected before Acceptance



Took a team of >20 approximately 6 months to identify and create solutions

System is now highly consistent

Other Examples

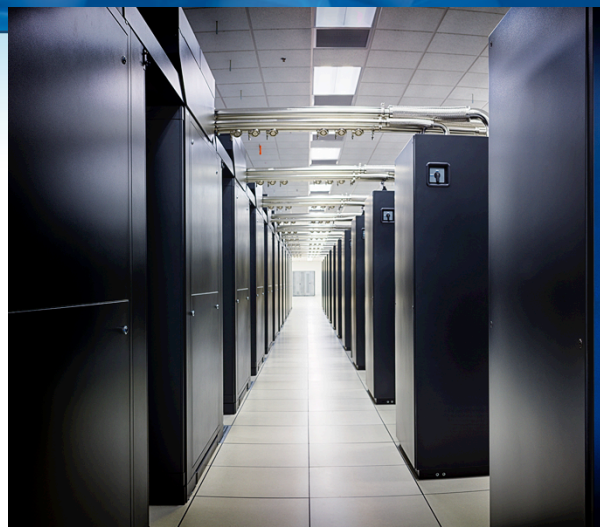
- Node failures caused by Chip Via Degradation
  - In Silicon chips, a via is a metal pathway between one layer of the chip to the another
  - Took team of 100's about 3 months to identify and solve
  - Determined via candidates
  - Sliced chips, electron micrgraphs of good and bad
  - Determined unexpected degradation of material
  - Team of 100's took about 4 months
  - Correction was to slightly redesign the fabrication process
- Connector
  - Solder failures for the PCIe connector for GPU to CPU connection was causing node failures
  - Team of Material Scientists, HW designers, …
  - Took months to diagnose
  - Solder components made it brittle with heat. Connector expanded and contracted with heat. Since connecter was very tightly screwed down tightly so the only way to relieve the stress was to cause solder cracks.
  - Solution is to replace the connector and use screws with ones that allow a little expansion
- Lustre
  - Metadata aging gradually erodes metadata performance
  - Functions of file system size and activity
  - Mitigations – but root cause (design error) took > 6 months, multiple long term experiments, teams of experts

One Set of Definition

- Failure (No Interrupt): failures that are naturally tolerated by the architecture, i.e., do not cause node/system downtime or trigger failover actions (e.g., cooling hardware or performance problems).
- Interrupt (Failover): a critical problem that is successfully handled by the automatic failover mechanisms.
- Link and Node Failure (Job Failed): the failure of one or more nodes and one or more links that cause the failure of one or more user jobs. *AKA Node Failure*
- Link Failure (No Job Failed): Failures of a single or multiple node(s) that cause a link failure that is successfully handled by the automatic network failover mechanisms;
- Link Failure (Job Failed): Link failures that cause job loss.
- Single/Multiple Node Failure: Failures of single or multiple node(s) that do require repair, but do not impact core system functionalities. – *AKA Node Repair Action*
- Interruption (System-Wide Outage): The whole systemis unavailable, e.g., because of a system-wide repair or restart. A system-wide outage occurs if specific requirements cannot be met, such as: i) the ability to access all data blocks of all files in the file system; ii) user ability to log in; iii) full interconnection between nodes; iv) access to external storage server (esDM, or external data mover); v) ability to support user applications submission, scheduling, launch, and/or completion; vi) ability to recover from file system or network failures through automated failover operations; and vii) performance (e.g., network bandwidth or file system throughput) above acceptable levels.
- Complexity due to “contractual” definitions and definitions in practice

BLUE WATERS

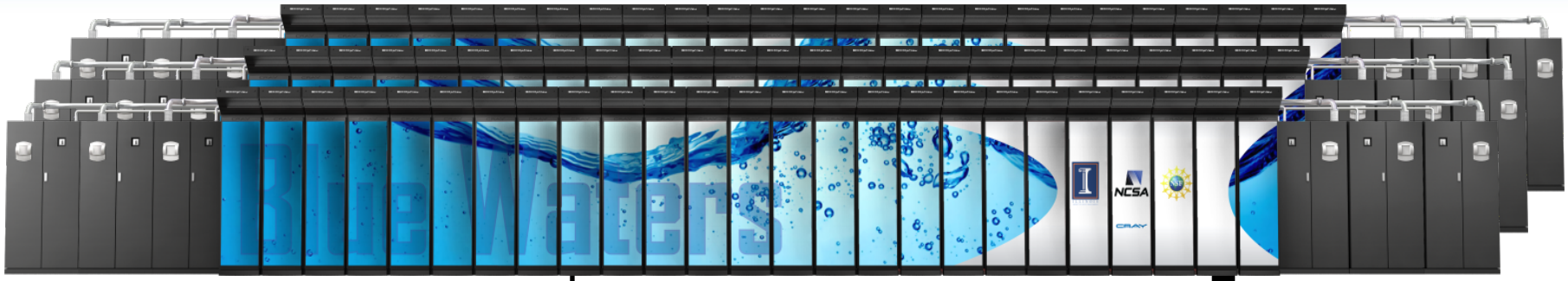
The Blue Waters System



- 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

10/40/100 Gb Ethernet Switch

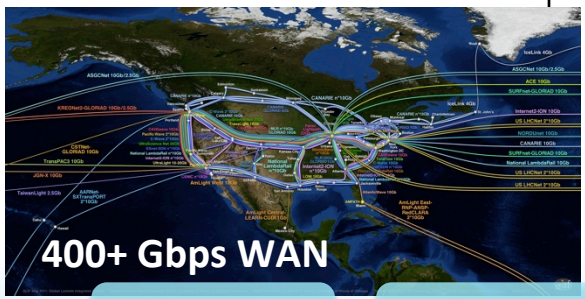
External Servers

IB Switch

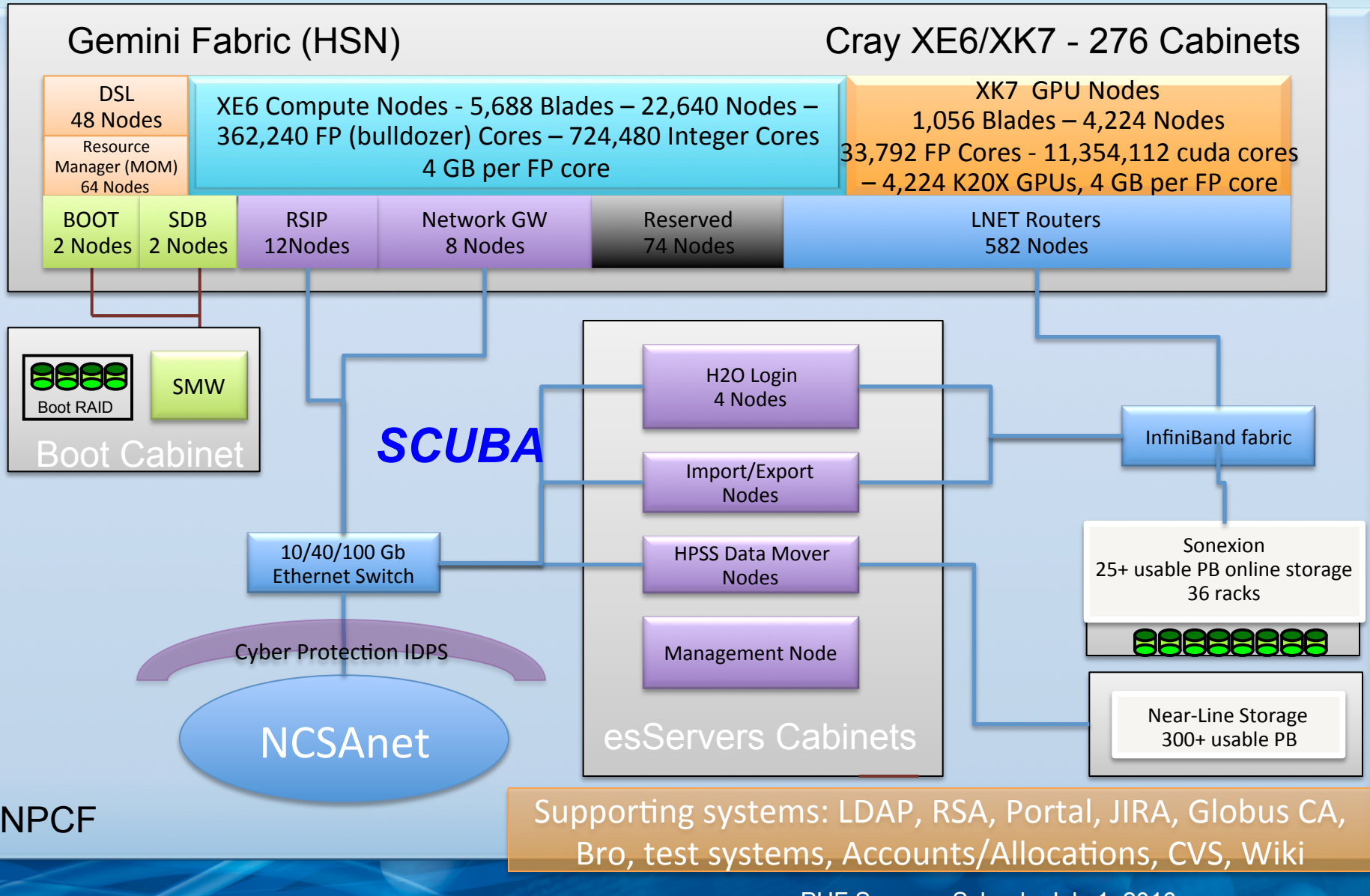
>1 TB/sec

220+ Gb/sec
Going to 400+ Gb/sec by end 2015

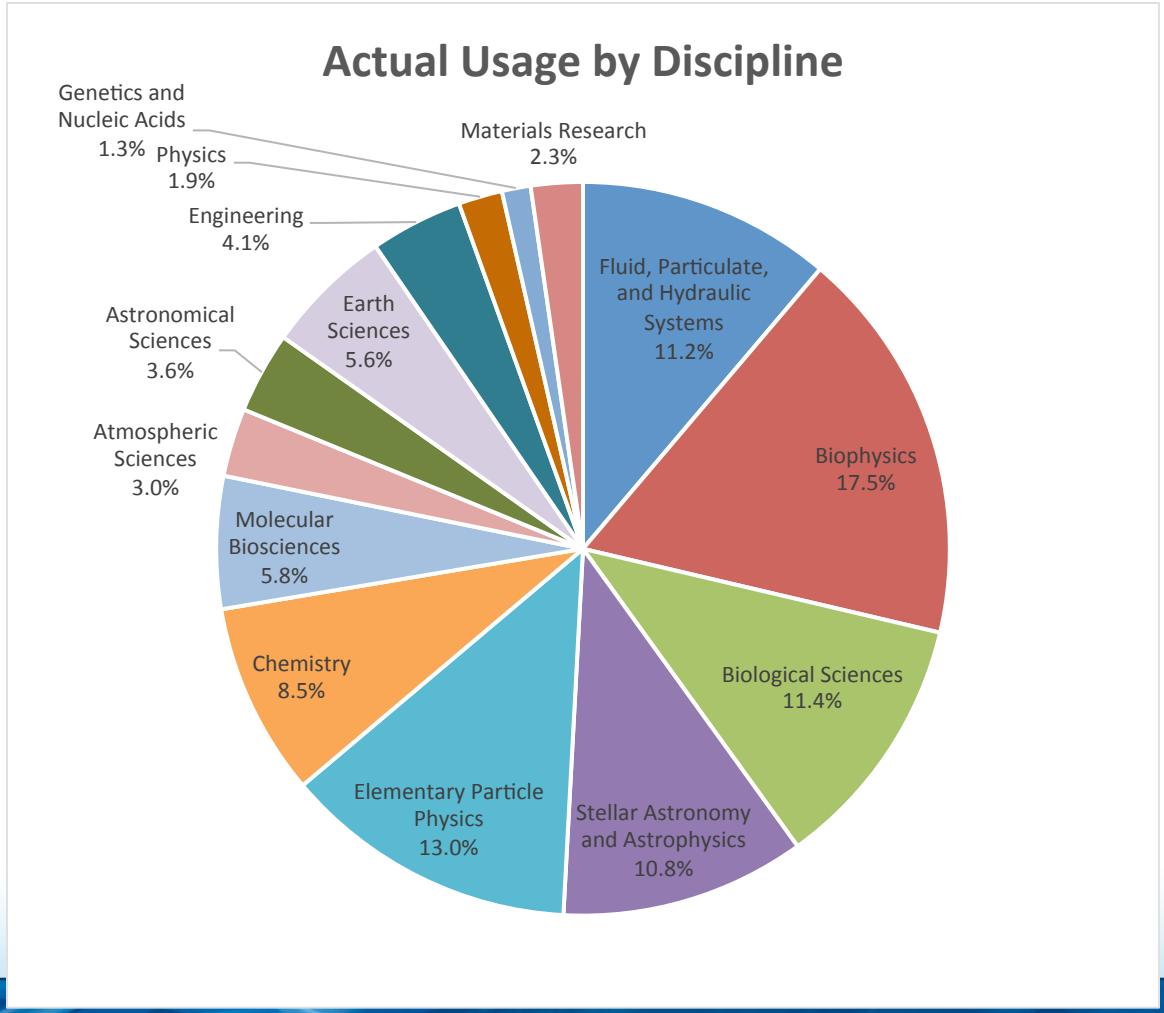
100 GB/sec



- 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

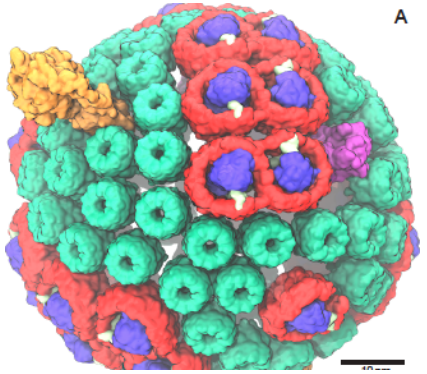

Gemini Fabric (HSN)
Cray XE6/XK7 - 276 Cabinets
DSL
48 Nodes
Resource Manager (MOM) 64 Nodes
XE6 Compute Nodes - 5,688 Blades – 22,640 Nodes – 362,240 FP (bulldozer) Cores – 724,480 Integer Cores
4 GB per FP core
XK7 GPU Nodes
1,056 Blades – 4,224 Nodes
33,792 FP Cores - 11,354,112 cuda cores – 4,224 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
10/40/100 Gb Ethernet Switch
Sonexion
25+ usable PB online storage
36 racks
Cyber Protection IDPS
NCSAnet
Near-Line Storage
300+ usable PB
NPCF
Supporting systems: LDAP, RSA, Portal, JIRA, Globus CA, Bro, test systems, Accounts/Allocations, CVS, Wiki

Q1 2016 Actual Usage by Discipline

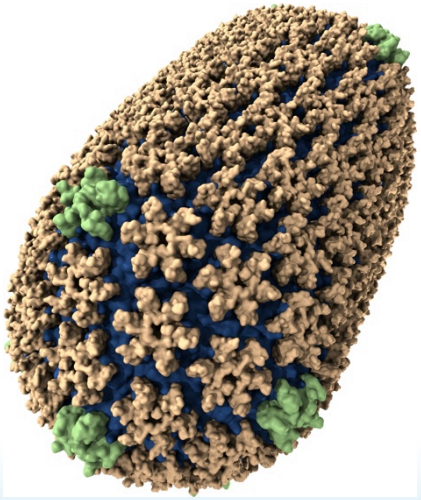


Schulten - The Computational Microscope

"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."

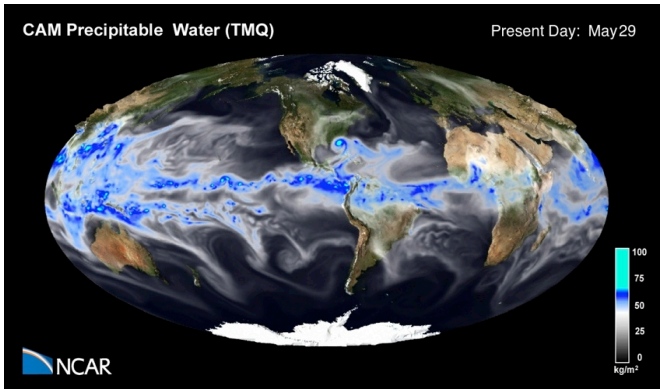


- Challenge Goals
  - First ever atomic-level structure of a native, mature HIV capsid to help scientists understand better how the HIV capsid infects the host cells and could lead to new HIV therapies.
  - The first all-atom model of a cellular organelle. – the Chromatophore which allow the bacteria 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 0.25° 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 INSIGHTS

Example Insights and Results

- Software is the cause of 74.4% of system-wide outages (SWOs).
- Close to two-thirds (62%) of software caused SWOs resulted from failure/ inadequacies of the failover procedures, such as those invoked to handle Lustre failures.
- Hardware is highly resilient to errors. Out of 1,544,398 analyzed machine check exceptions, only 28 (0.003%) resulted in uncorrectable errors, showing the value of the adopted protection mechanisms (Chipkill and ECC).
- The GPU accelerator DDR5 memory, protected only with ECC, is 100 times more sensitive to uncorrectable errors than DDR3 node RAM is.
- 99.4% of failures limited to a single blade;
- Software errors propagate 20 times more often than hardware failures;
- Software Errors accounts for 53% of repair hours;
- Hardware failure rates decline over time but software does not
- failure of failover causes a significant number of system wide outages;
- application failure increases with increasing duration of failover time.

Insights credited to the HMDR Project Team

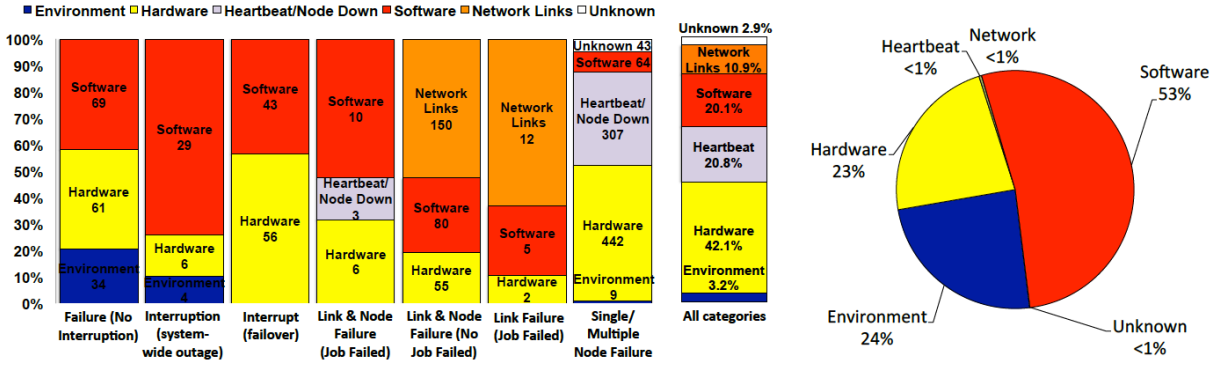
Failures During First 261 days

- While hardware is the primary contributor to the total number of failures (42%), failures due to hardware causes are responsible for only 23% of the total repair time.
- Software is the major contributor to the total repair time (53%), despite being the cause of only 20% of the total number of failures.
- The error protection mechanisms do a very good job in containing failures within boundaries of a single node.
  - Nearly 97.6% of failures due to hardware root causes do not propagate outside the boundary of a single node.
- Lustre distributed file system dominates (46%) the software root causes of failures.
  - Long recovery time (15-30 minutes) and the chance of a second error during this period make the Lustre highly vulnerable. As a result, the
- Lustre operates at its limits and has a hard time to keep up with handling the data volumes processed by BW.

Failures During First 261 days

- System-wide Outages (SWO) characterization
  - During the measurement period (i.e., 261 days) the system suffered 39 SWO (i.e., each 6 days and 15h), which gives availability of 97.76%.
- Software is a major cause (74.4%) of SWOs and contributes to 68% of the total system downtime. Close to two-third (62%) of software-caused SWO are due the failure of the failover procedure invoked to handle Lustre problems.
  - Failure of failure had maximum time period before being declared an outage
- A single node failure causing a SWO is a rare event and occurs in 0.7% of all reported single node failures
- Error correcting codes (SEC-DED and Chipkill) techniques correct 99.997% of the memory and processor errors. Only 28 out of 1,544,398 single and multiple bit(s) errors are uncorrectable. While this is somewhat

Failures During First 261 days



(a) (b)

Fig. 2: Breakdown of the failure categories (a) and distribution of the cumulative node repair hours (b) across hardware, software, network, unknown, heartbeat, and environment root causes types.

TABLE III: Breakdown of the top-3 hardware and software failures root-causes

	Failure (no interrupt)		Interrupt		Interrupt (Failover)		Link Failure (User Job Failed)		Link & Node Failure (User Job Failed)		Single/Multiple Node Failure	
HW	PSU	20	EPO	1	Disks	45	Optic	12	GPU	2	Processor	160
	IPMI	15	Compute Blade	2	IPMI	5	RAM	9	Gemini ASIC	1	RAM	158
	Fan Tray Assy	14	Storage Module	2	Storage Module	2	Gemini Voltage Regulator	8	Compute Blade	2	GPU	38
SW	Moab/TORQUE	33	Lustre	18	Lustre	29	Lustre	2	Lustre	8	Lustre	30
	CLE/Kernel	17	Moab/TORQUE	6	Sonexion/Storage	8			CLE/Kernel	1	CLE/Kernel	16
	warm swap	5	Gemini	3	CLE/	4					Sonexion/Storage	5

Failures During First 261 days

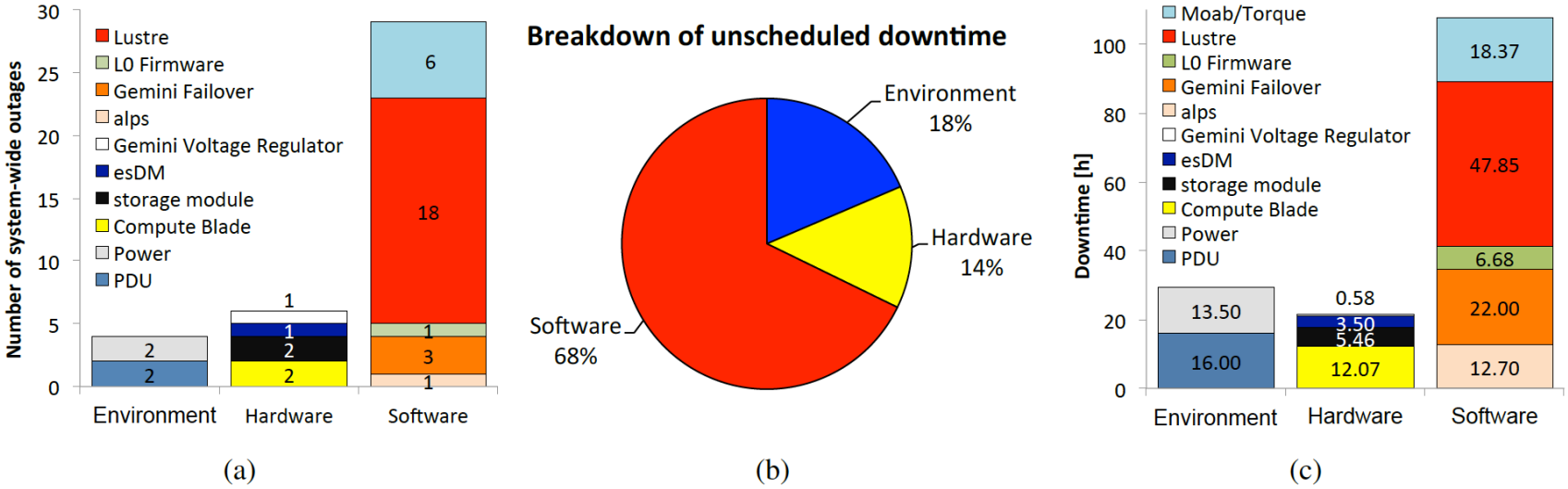


Fig. 4: Breakdown of the SWO root-causes (a) and repair times (b)

Lustre Failures in first 261 days

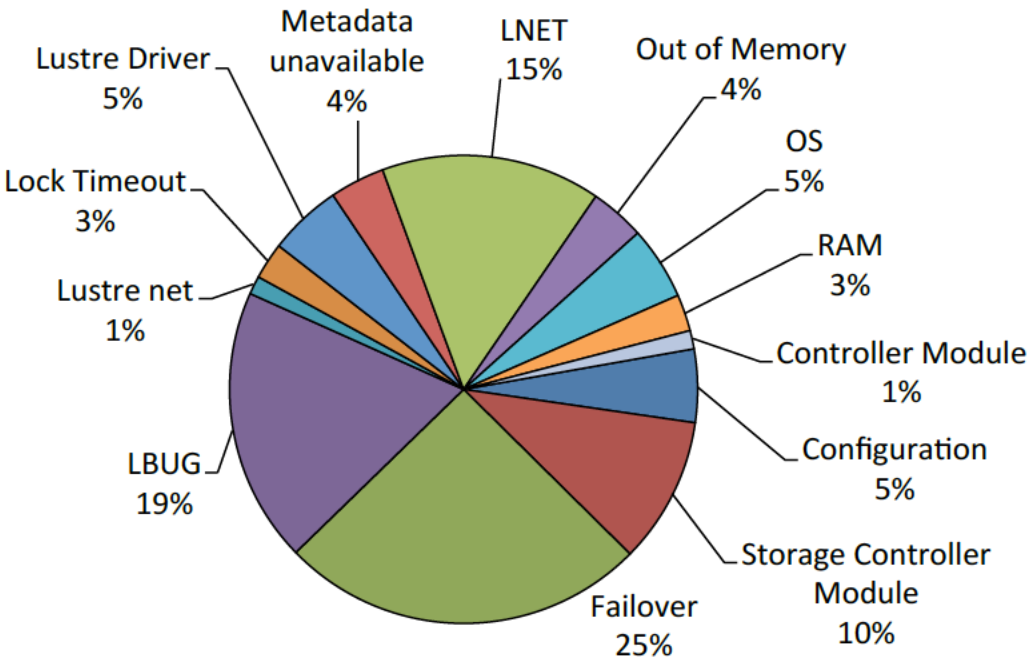
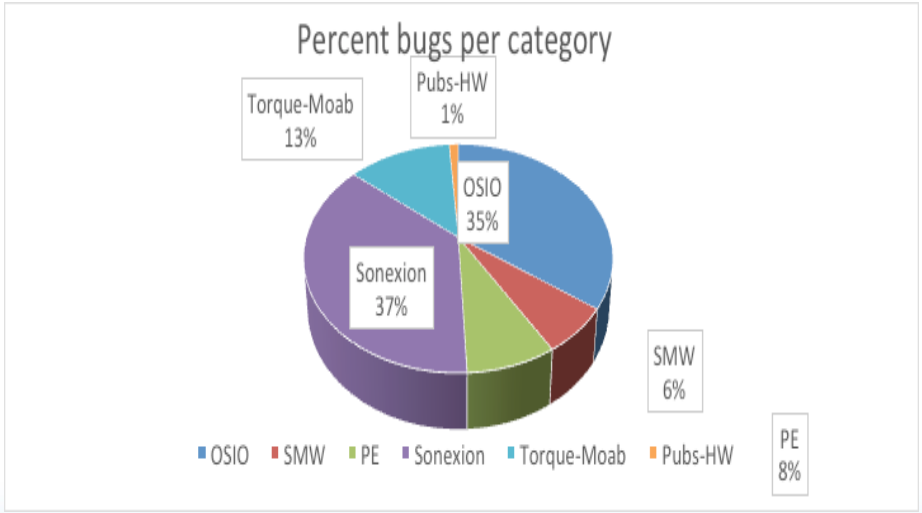


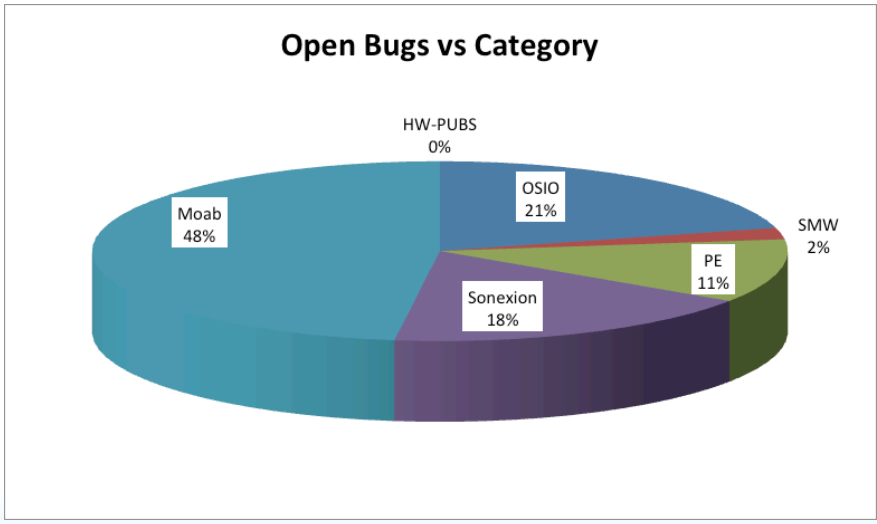
Fig. 3: Breakdown of the Lustre Failures

Number of Reported/Open Bugs in Two Periods indicate improved reliability

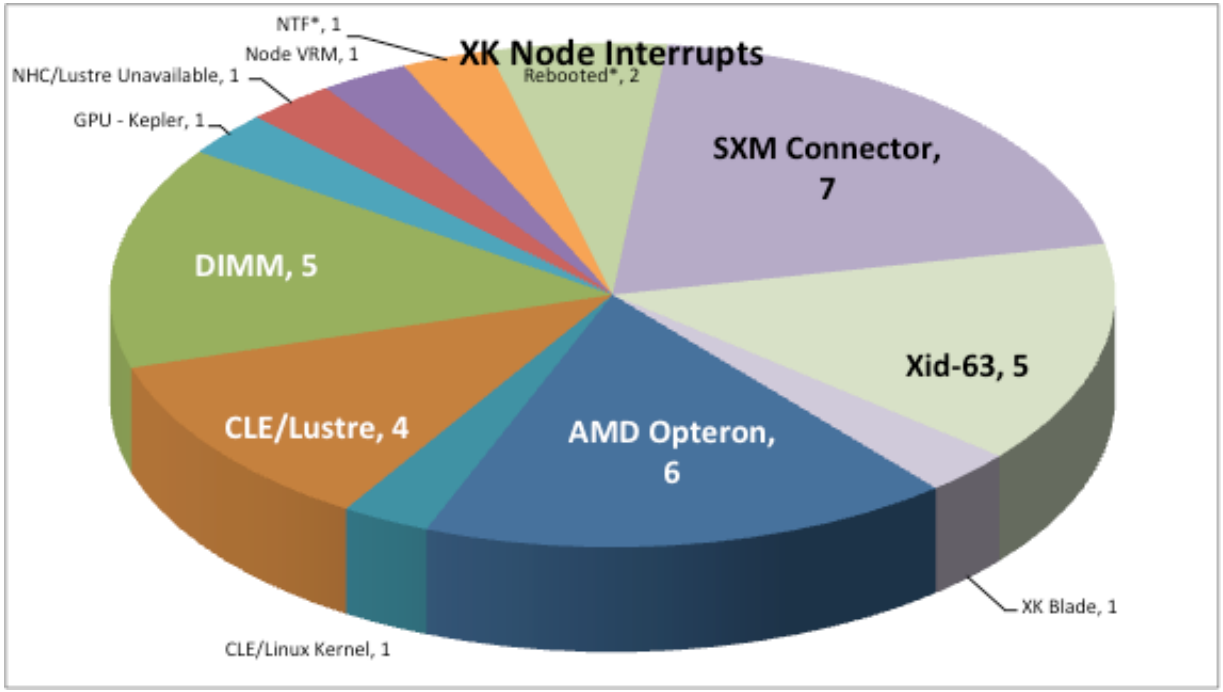
Q2 2013 Total - 268
Critical - 0
Urgent - 28
Major - 212
Minor - 28

Q1 2016 Total - 94
Critical - 0
Urgent - 17
Major - 62
Minor - 15

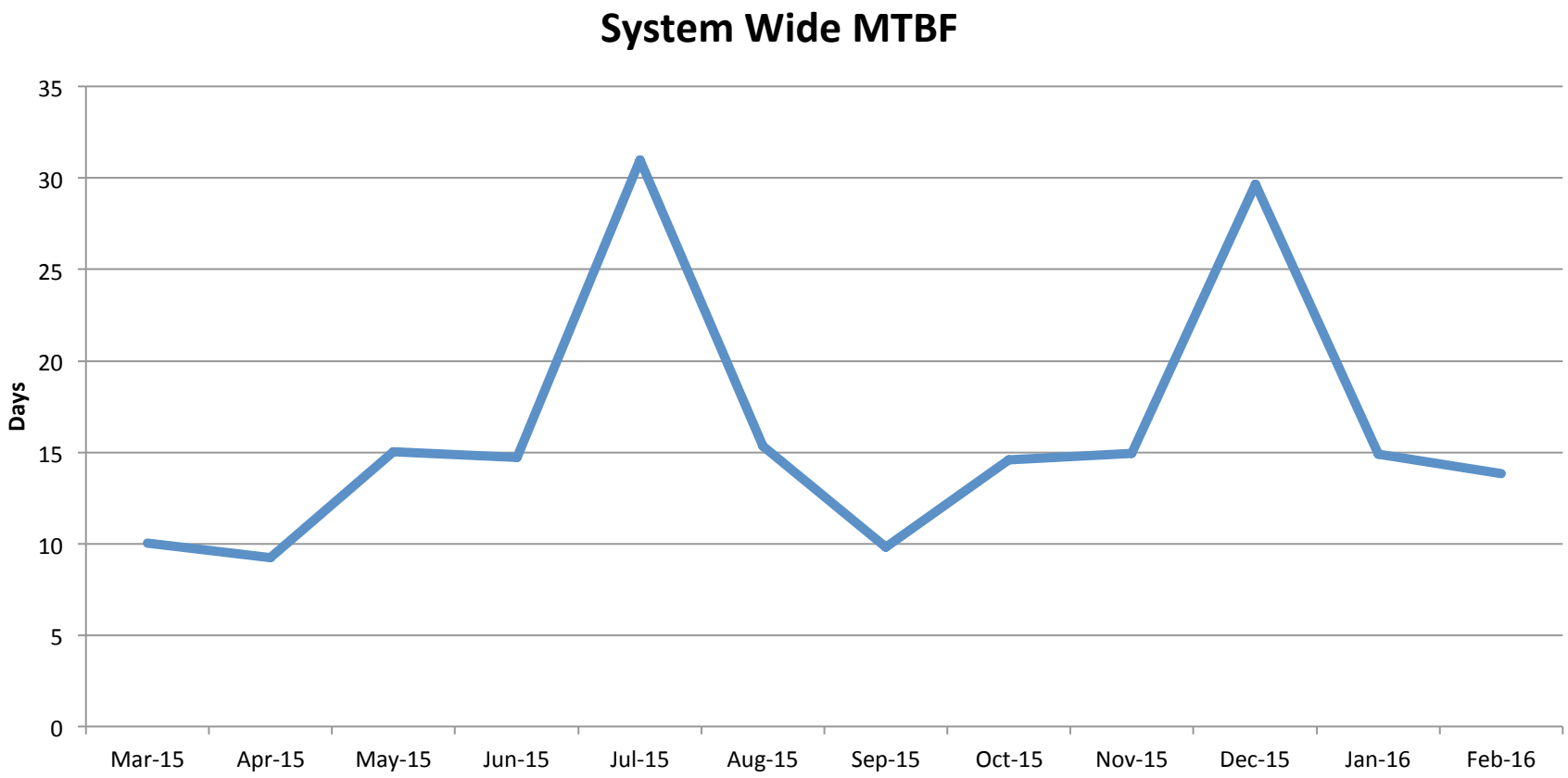




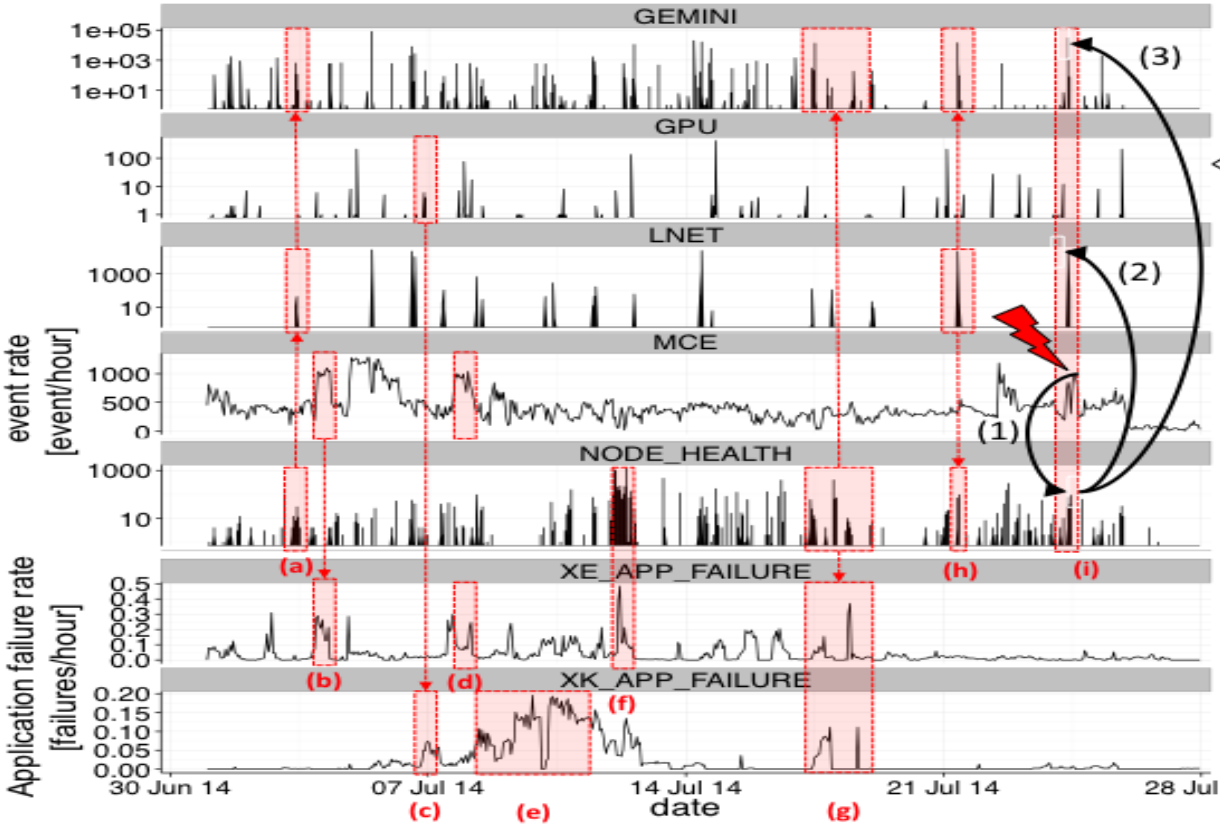
Examples of Data: Node Failure Causes on Blue Waters



Blue Waters System Wide MTBF for any cause



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

Failover-induced application failure

- 392 failover events caused by system-level failures (e.g., file system errors)
  - 5% of the node hours the system executes failover procedures
  - 94% of system-level failovers are successful
- 37% applications failed during failovers
  - 6x increase in application failure rate during failovers as compared with normal execution

date	apid	type node	jobID	application	Cores / Node	Nodes
8/12/2013 18:17	2062620	xk	425045	namd2	16	16

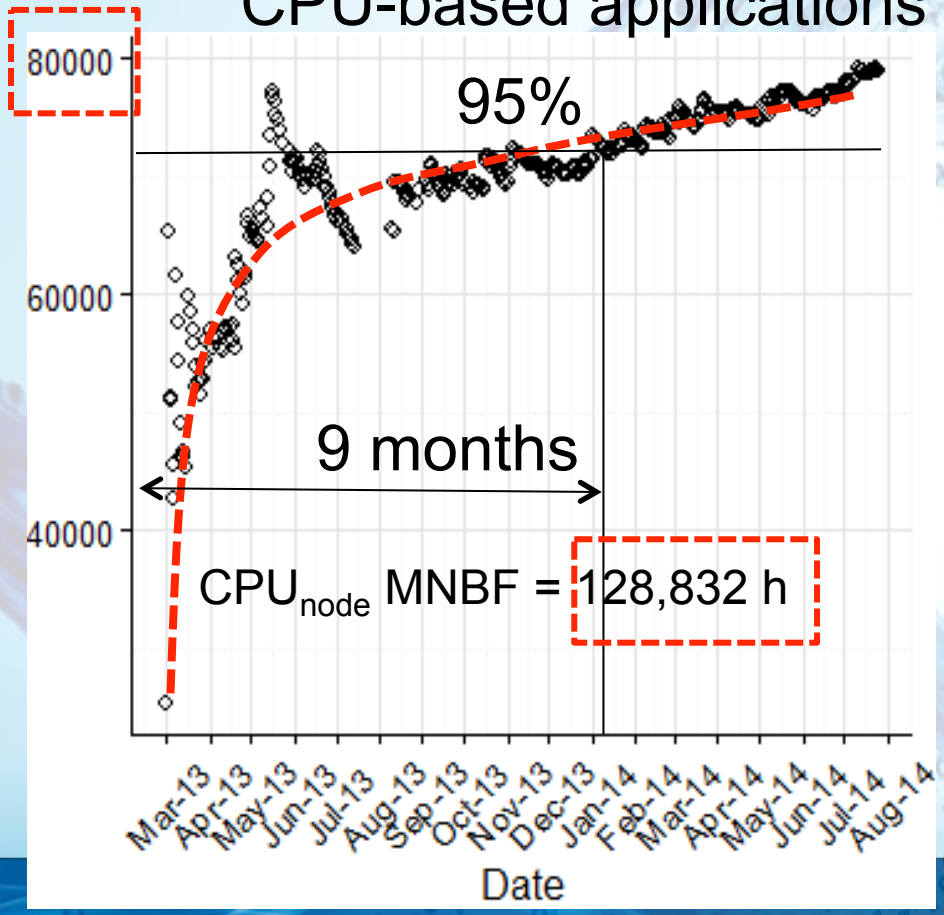
application duration (h - node hours)	job duration (h - node hours)	Recovery	# Error events	Exit Reason	Exit Code (meaning)	job Exit Status (code)
5.4 – 86.4	12.0- 191.4	YES	847,713	walltime	143 (SIGTERM)	walltime (-10)

Error Trace
LUSTRE_INTERRUPTED_SYSCALL-> **LUSTRE_LBUG**-> LNET_PROTOCOL_ERROR-> **LUSTRE_OST_DEVICE_BUSY**-> LUSTRE_TRANSPORT_END_SHUTDOWN-> **LUSTRE_EVICT-> LUSTRE_WAITING_FAILOVER_END-**> LNET_TIMEOUT-> LNET_NO_ROUTERS-> LNET_STALE-> LNET_NETWORK_ERROR-> LNET_CONNECTION_FAILED-> **LUSTRE_WAITING_FAILOVER_END**

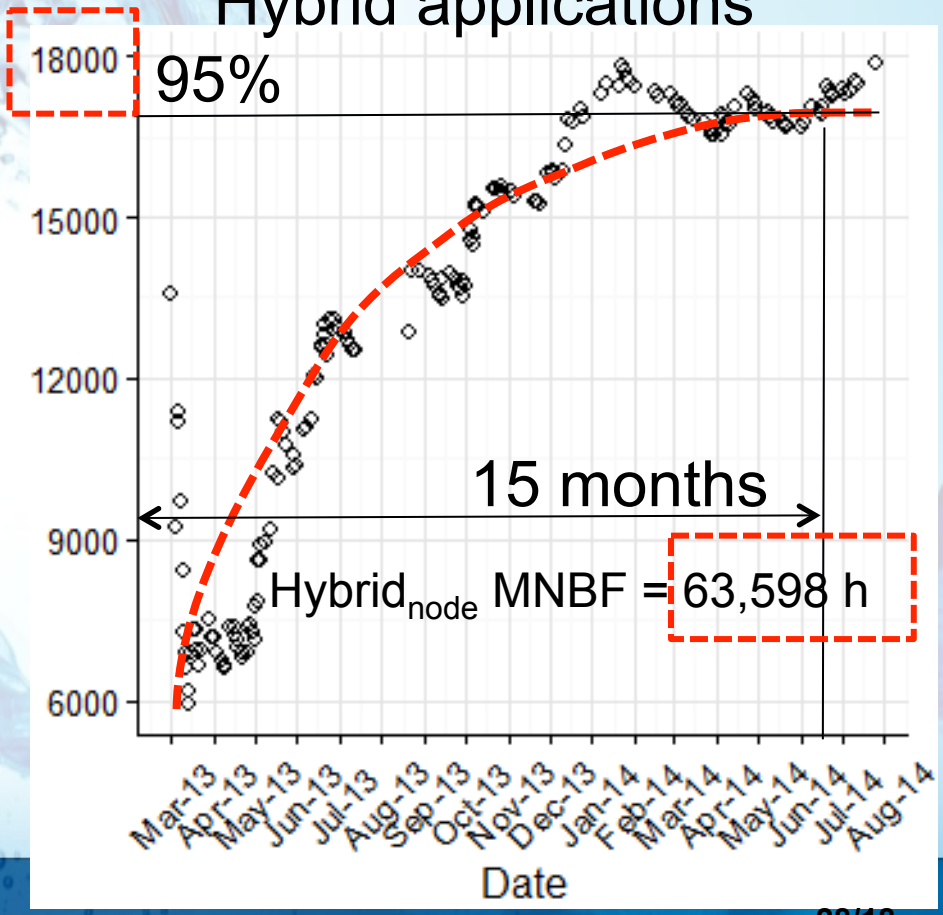
Platform and application resiliency Year 1

- MNBF [node hours] = #used Node Hours/#app. Failures
  - Resiliency of the platform taking into account computed node hours

CPU-based applications

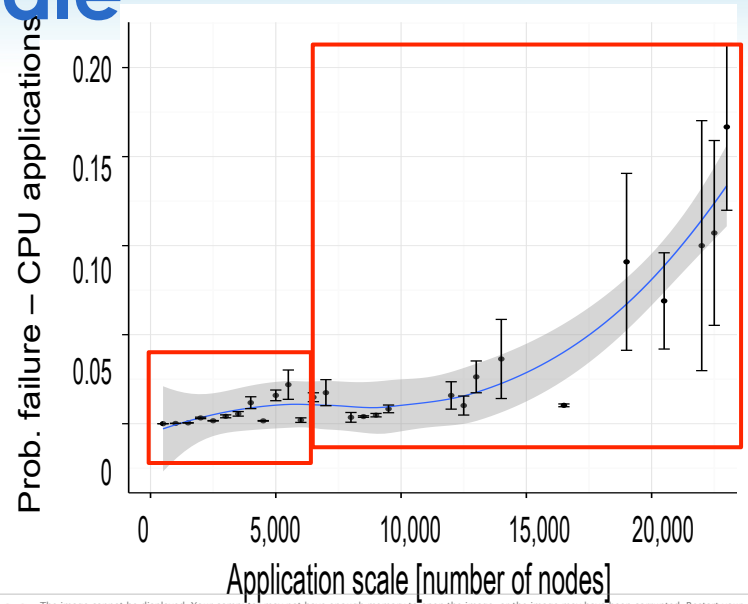


Hybrid applications



Resiliency Vs. Scale

- Probably of a app. failure from sys. errors incereases with scale of total node hours running
  - With and without checkpoint/restart
- Similar behavior of CPU and GPU app.:
  1. P(app. failure) constant (~10^{-3}) from small scale (1 node) to 25% of the system
  2. At full scale, P(app. failure) increases sharply to 0.18 for XE based and 0.1 for XK apps
     1. Remember – full scale for XK nodes is only 20% of full scale for XE nodes



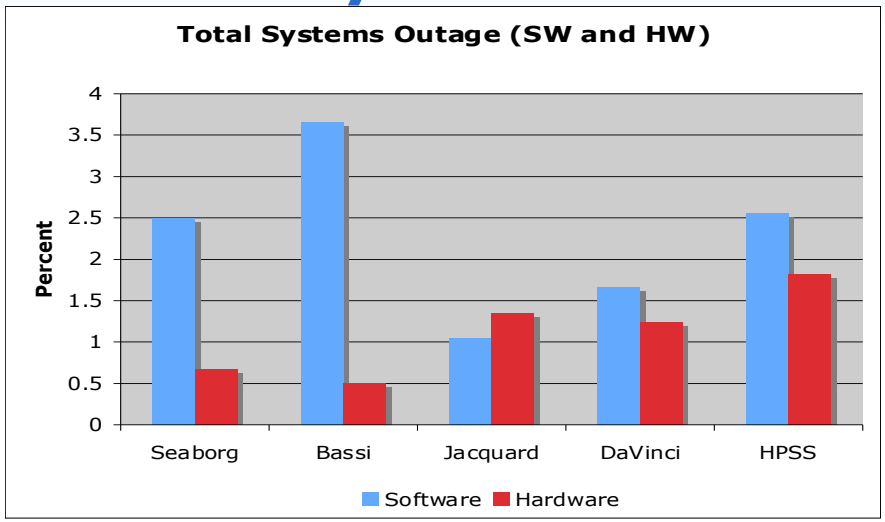
BACK TO THE PAST

Two Time Periods to Compare

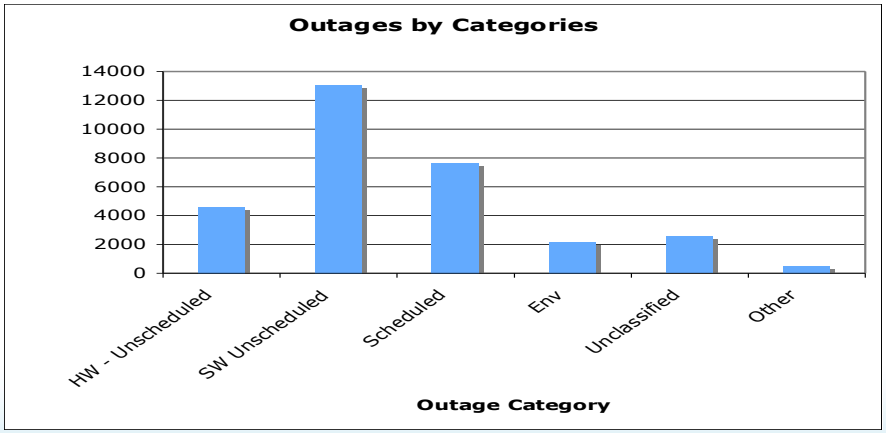
- 2000 – 2008 - NERSC systems
  - 2002 - IBM SP-3 – Seaborg
    - POWER 3+ 375 MHz
    - IBM “Colony” interconnect
    - 416 nodes with 16 cores per node – 6,656 cores
    - 6.7 TBs of Memory (1 GB/core)
  - 2007 - Cray XT4 - Franklin
    - AMD Opteron 2.6 GHz - Dual-Core
    - SeaStar Torus Interconnect
    - 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

Reliability is the Key Issue

- 6 Years of NERSC System

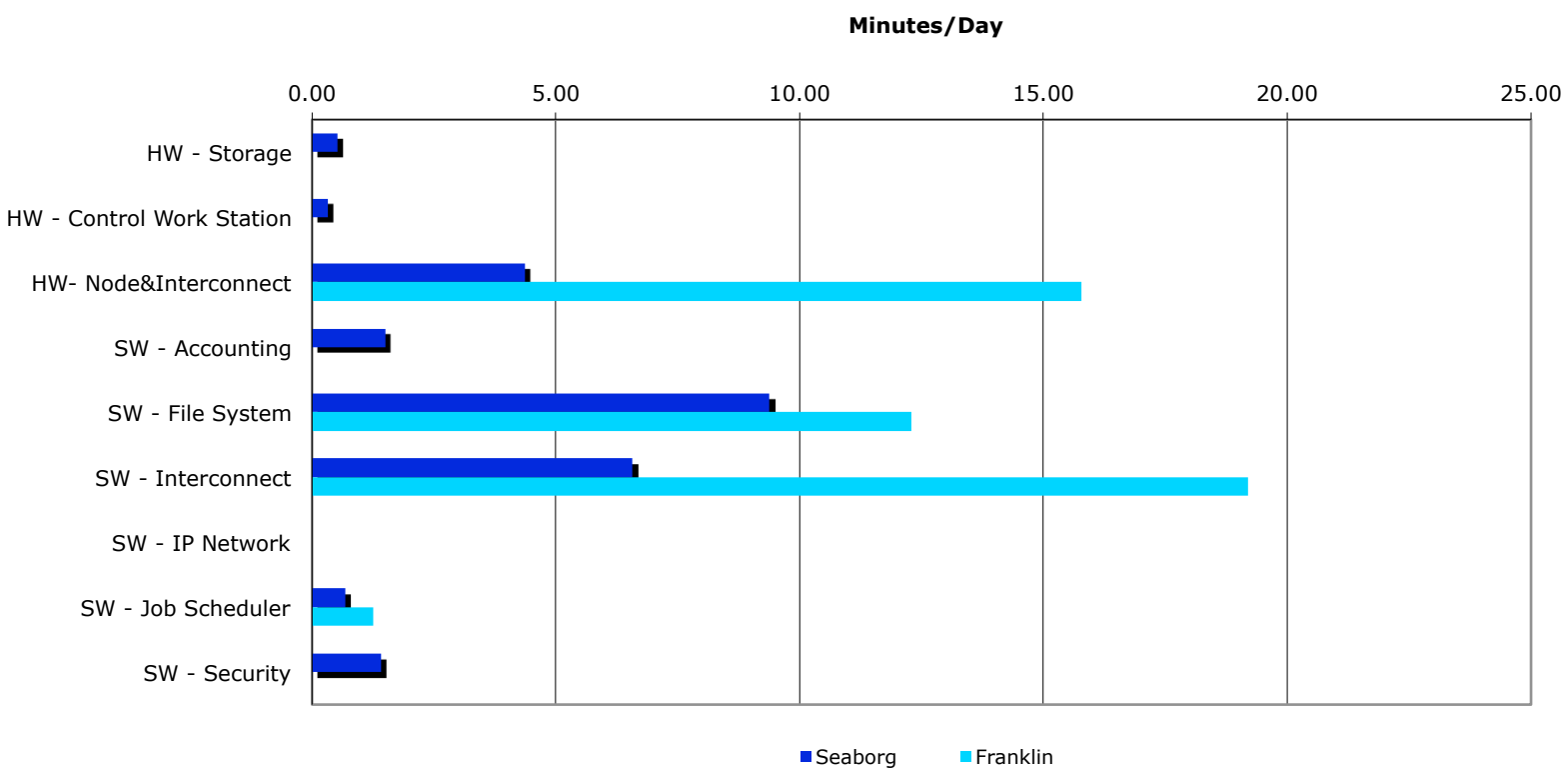


- 4 Months of XT-4 Franklin

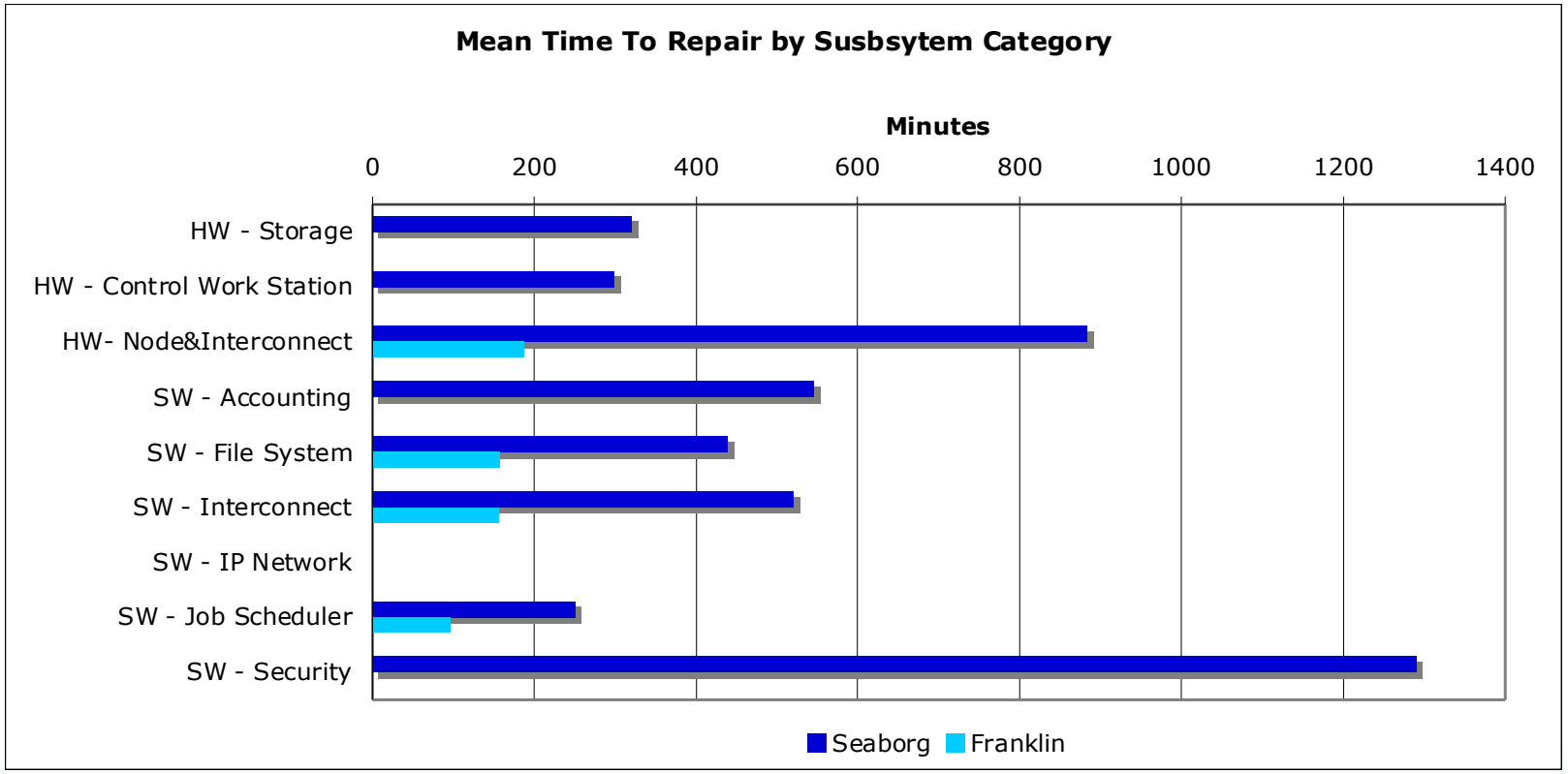


6 Year Charts Courtesy of Akbar Mokhtarani

Software Reliability



Software Reliability



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

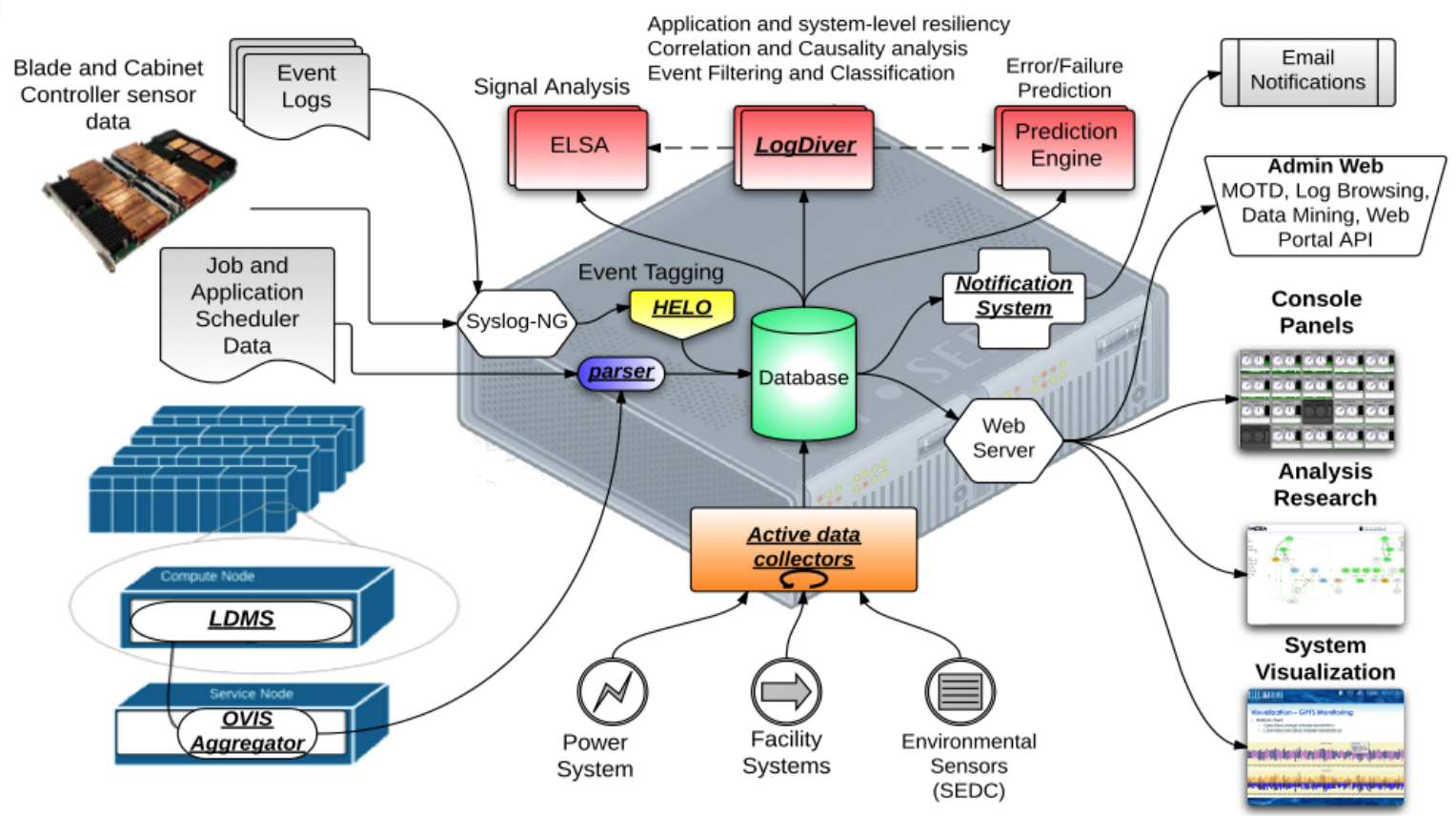
Job Failure Error Categories for the NERSC Cray XT-4	Number of Jobs	Percent of Jobs
SUCCESS - Job clearly succeeds	117,884	66.2%
WALLTIME - Job ran to the wall clock time limit - some user causes - some system causes	12,614	7.1%
WIDTH - A mismatch between job request and aprun command request -– normally 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

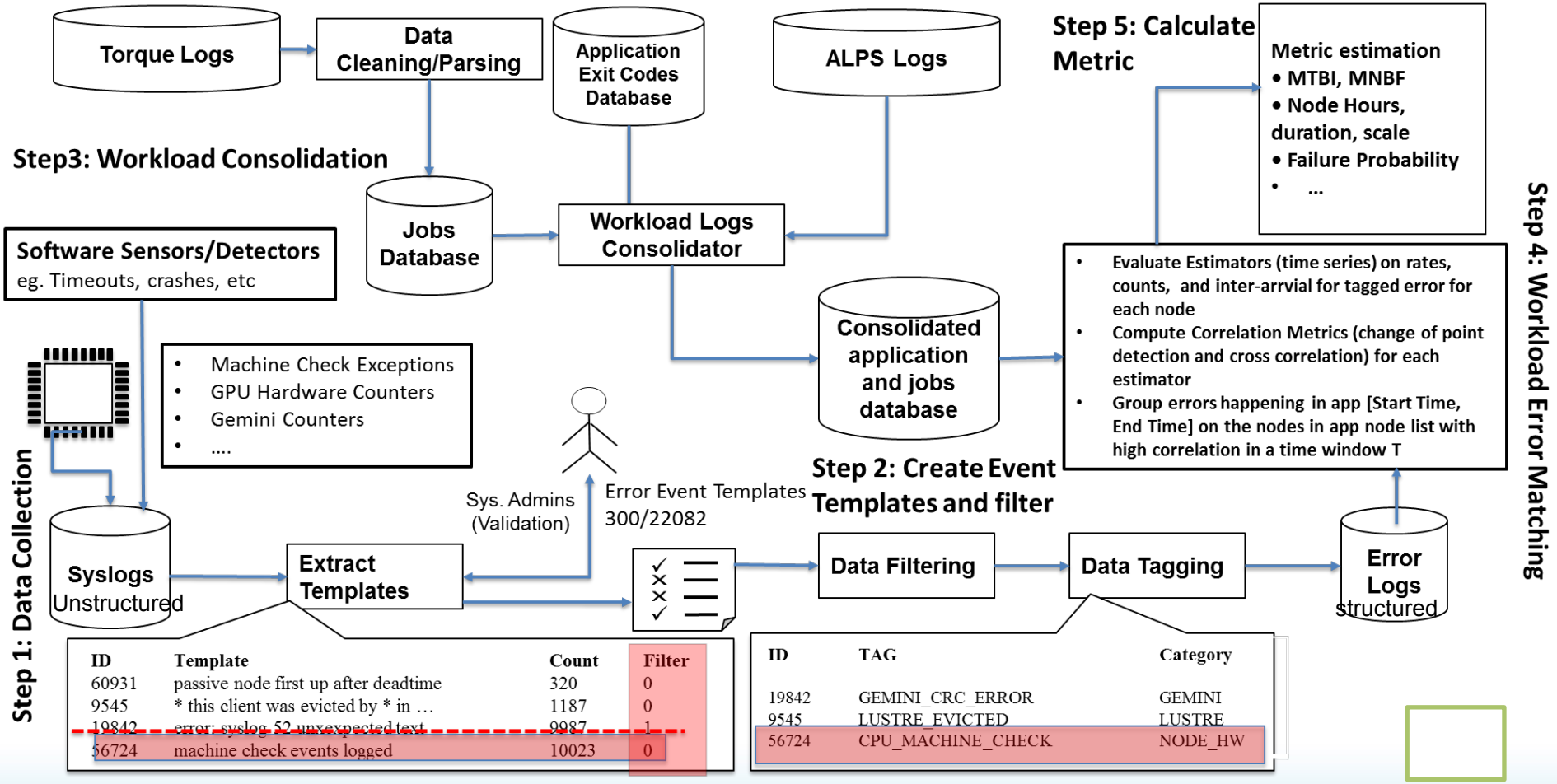
- Node
  - Load average
    - Latest,5min, running processes, total processes
  - Current free memory
  - GPU
    - Utilization, memory used, temperature
    - Pstate, Power Limit, Power Usage
- Filesystems
  - For each home, projects, and scratch
    - Bytes/sec Read and write
    - Rate of Opens, closes, 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

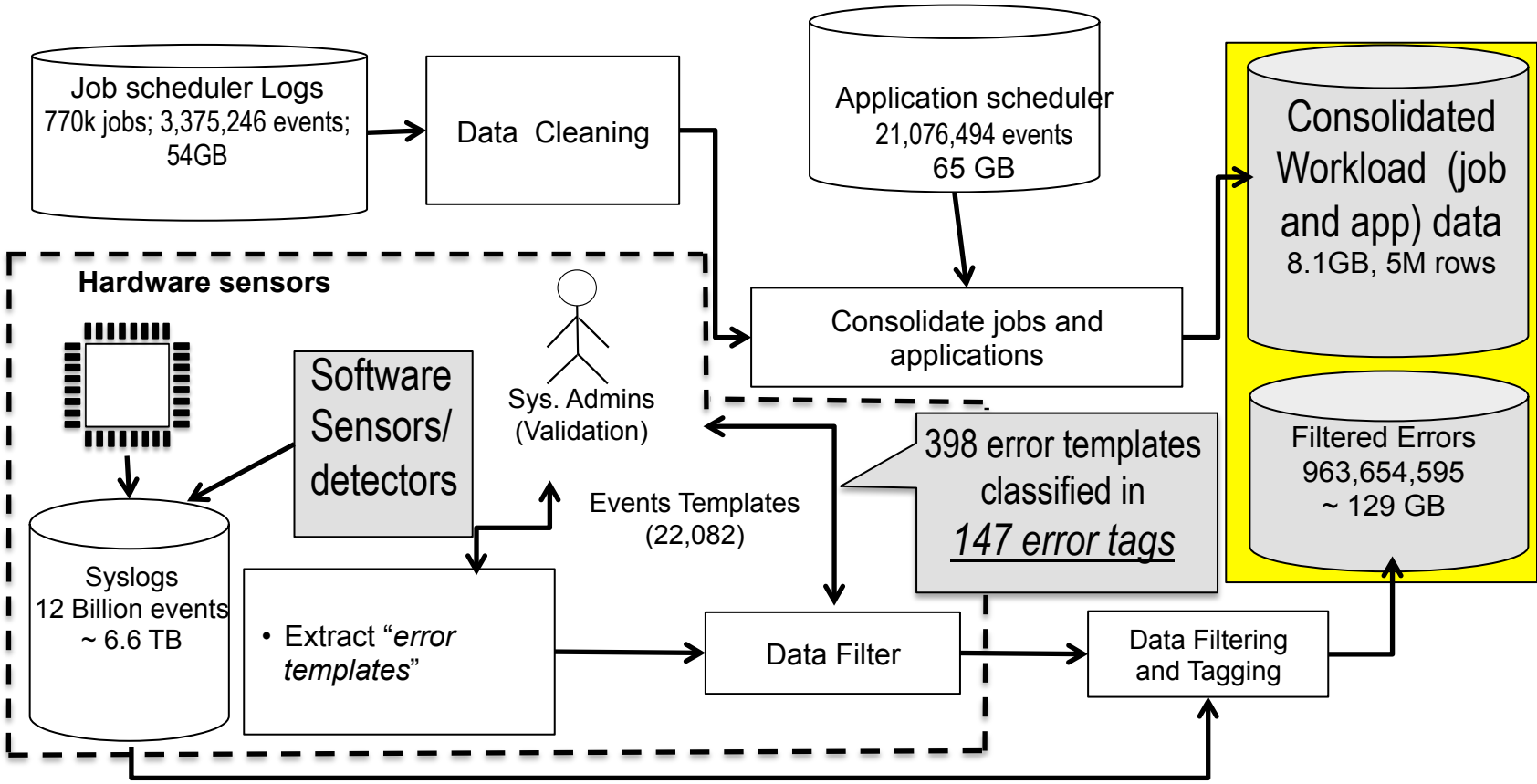


On Blue Waters, today has a core of about 8,100 active unique event types using HELO and ISC , with about 50,000 different event types that we have seen over the course of the project. We see around 30M log events on average per day, getting upwards of 1,470M log events on some days. This is exclusive 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



The LogDiver workflow (data consolidation) for Applications Study



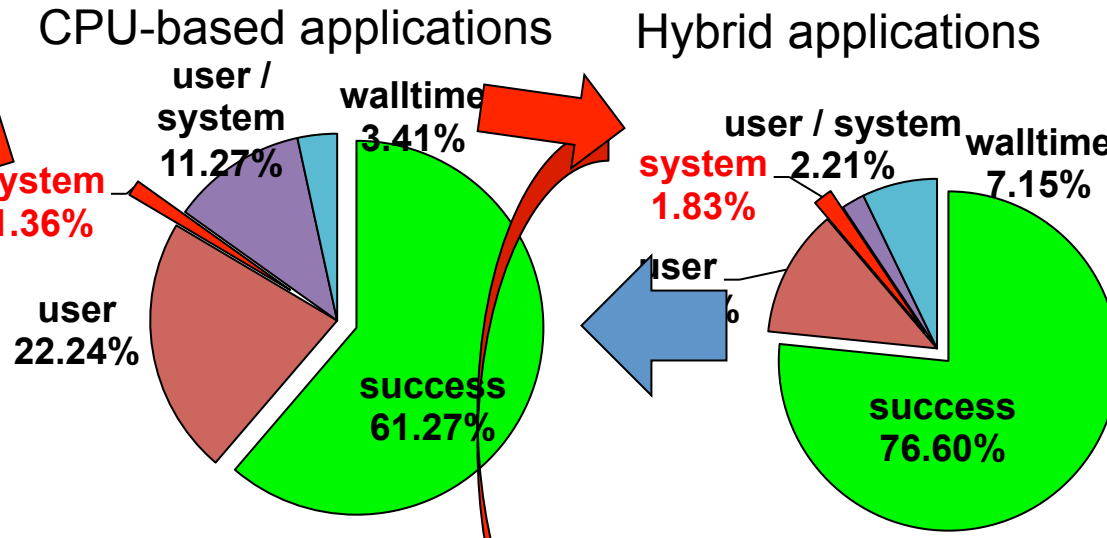
node 21877: unaligned access to address 0x27692A2F
node 11923: unaligned access to address 0xF22331CA

template → node *******: unaligned access to address **********

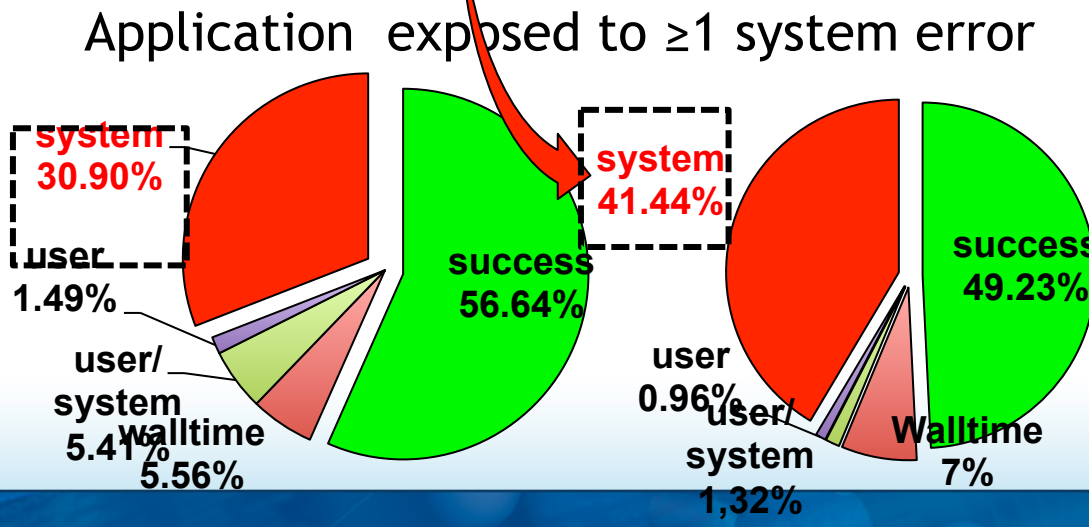
Size: 60GB (100x reduction)

System Errors Propagating to Applications

- 256 exit codes disambiguated and decoded by LogDiver
  - E.g., code 143 (term)

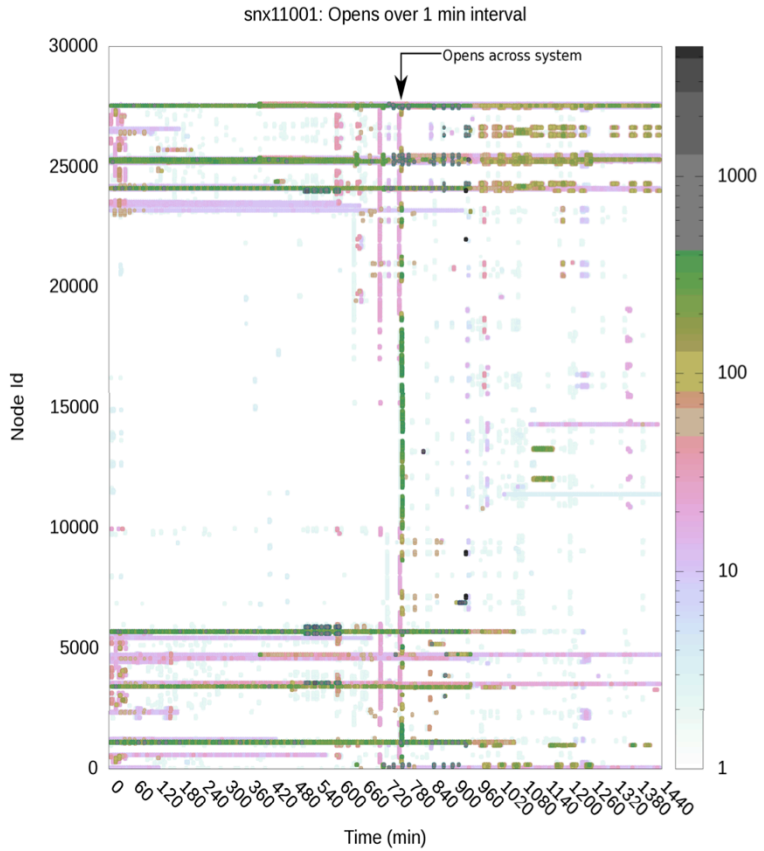


- Application-level resiliency features only provide limited protection from system errors



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)

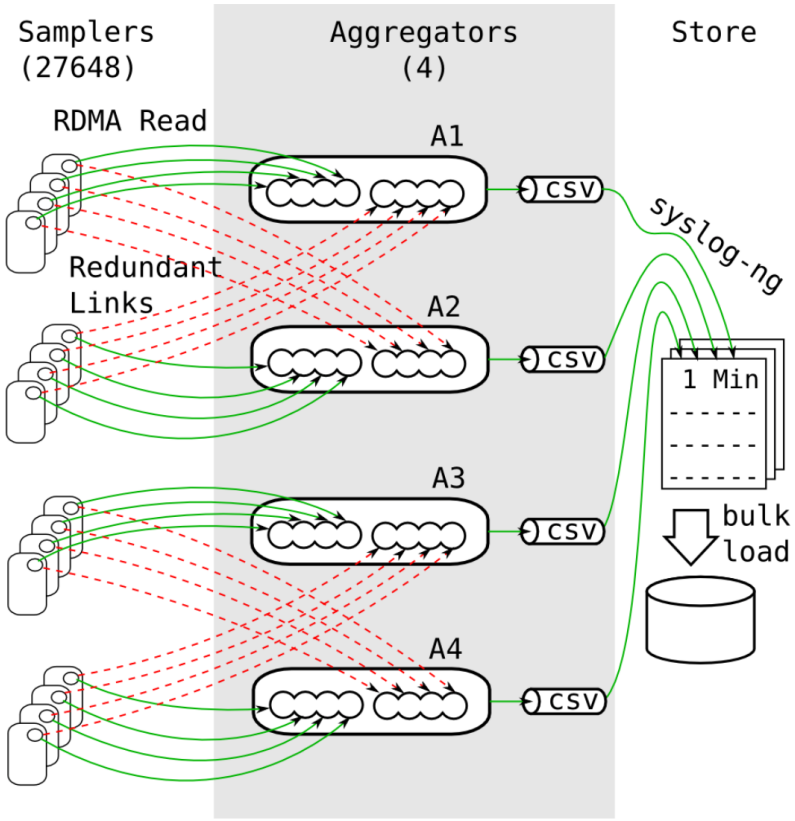


Blue Waters: One day dataset contains ~40 million data points per metric and 7.7 billion data points overall

Slide credited the HMDR Project Team and LDMS/Ovis team

LDMS Configuration on Blue Waters

Configuration has multipel connections (dashed arrows) to each sampler *ldmsd* for fast failover capability. Rather than writing directly to local stable storage, the aggregators each write a CSV file to a local named pipe. Data from this pipe is forwarded by syslog-ng to the ISC where it is bulk loaded into a database.

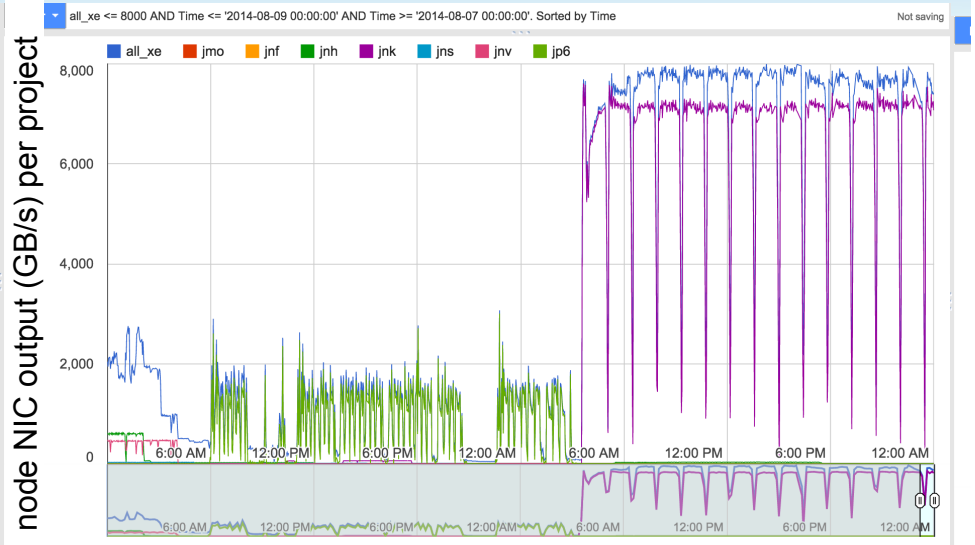


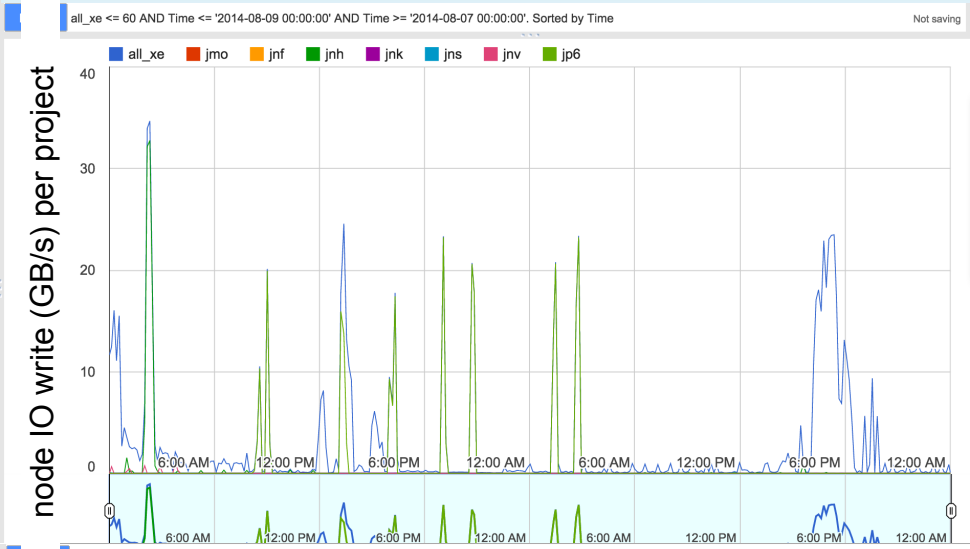
Source and Amount of Data Collected – Q1 2016

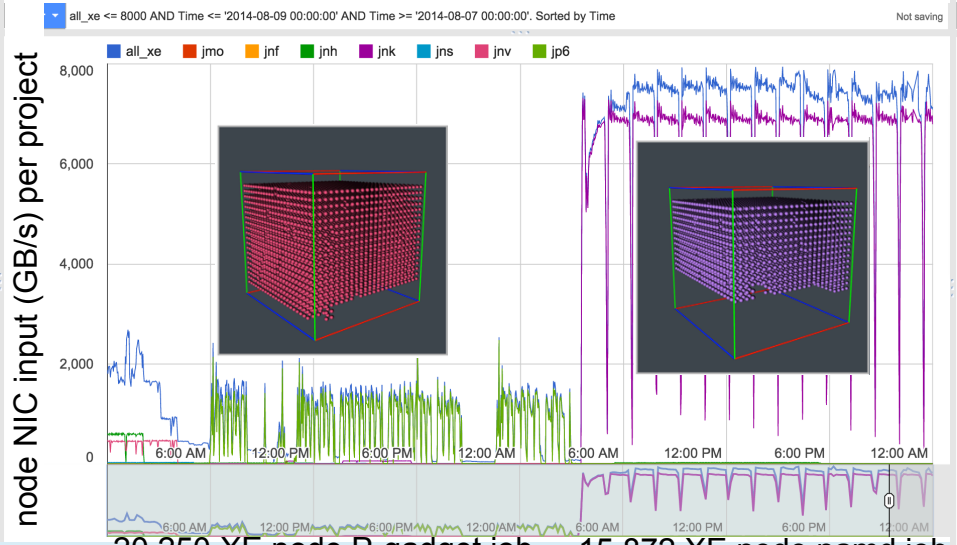
Data feed	Average (Bytes/day)	Max (Bytes/day)	class
apres	30M	148M	logs
apstat	60K	62K	metrics
backup	40K	74K	metrics
ddn	43K	326K	logs
esms	1G	3G	logs
hpss	135M	3.6G	logs
hpss_core	112K	192K	metrics
ibswitch	790K	801K	logs
moab	2.5G	3G	logs
qos-ping	3.3M	3.6M	metrics
quotas-hpss	944K	950K	metrics
scheduler	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	42G	49G	metrics

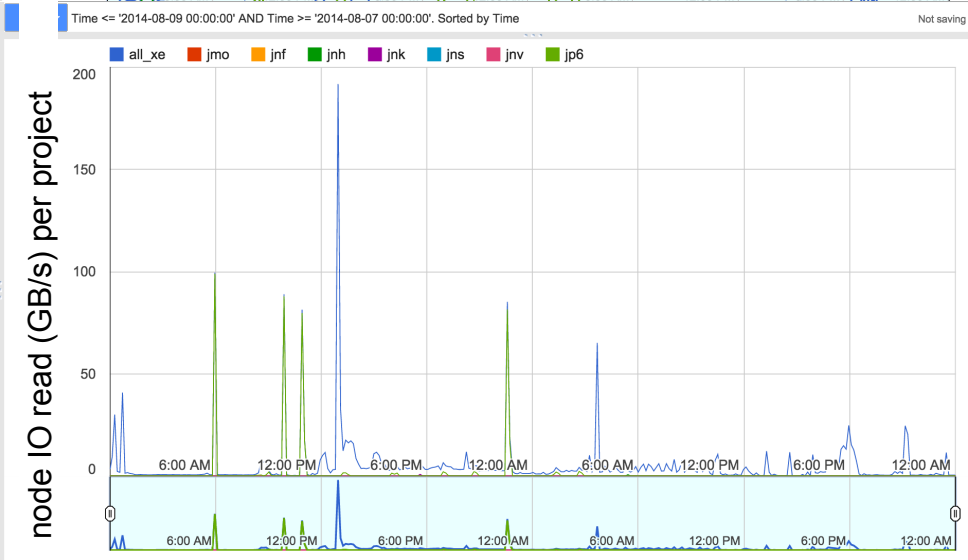
EXAMPLES OF ADDITIONAL THINGS YOU CAN DO WITH THE DATA

Example Views of Job Behavior





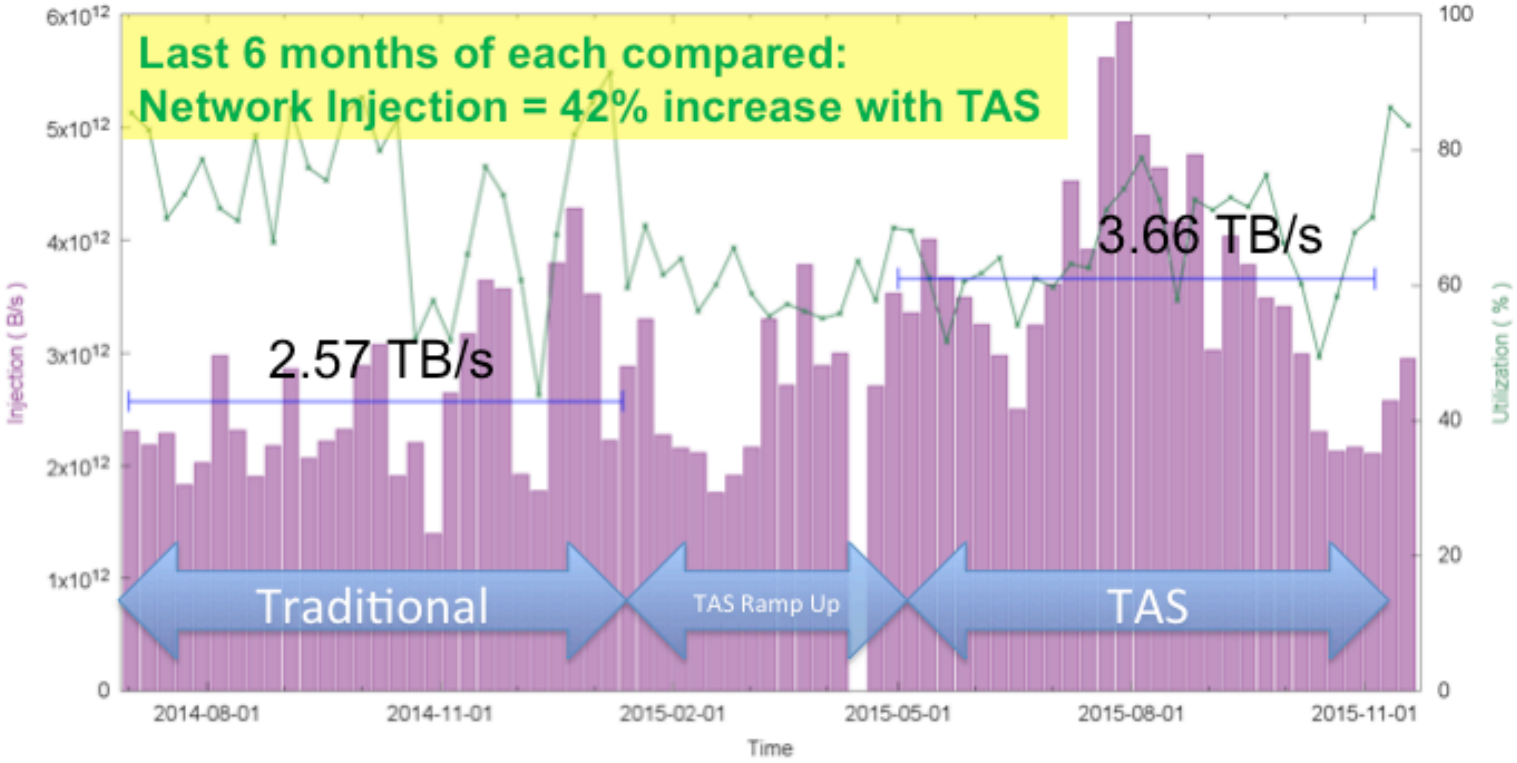




20,250 XE node P-gadget job 15,872 XE node namd job

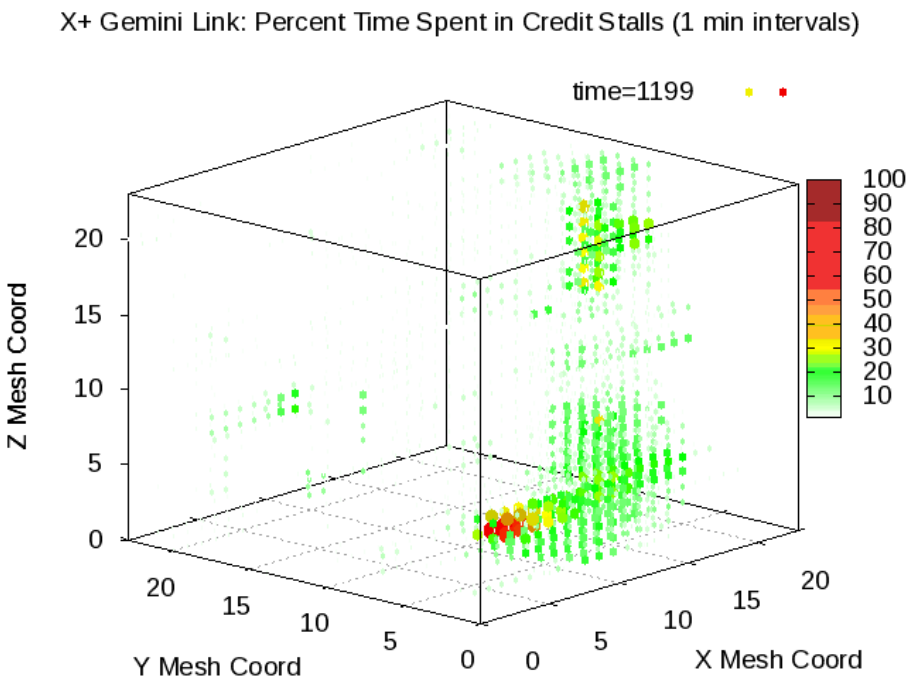
Data credited the HMDR Project Team, the LDMS/Ovis Team and NCSA Blue Waters Team

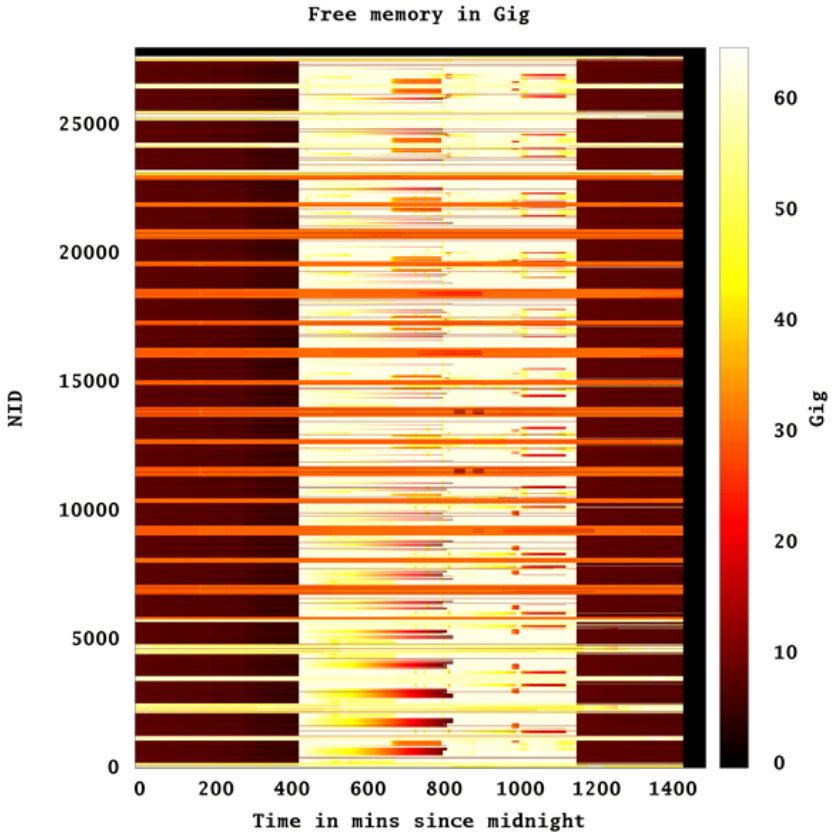
Examples of Results: Improving systems through understanding of root causes of faults, failure propagation, and performance changes



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

3D Animations and Graphs





Network Contention

Free Memory

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

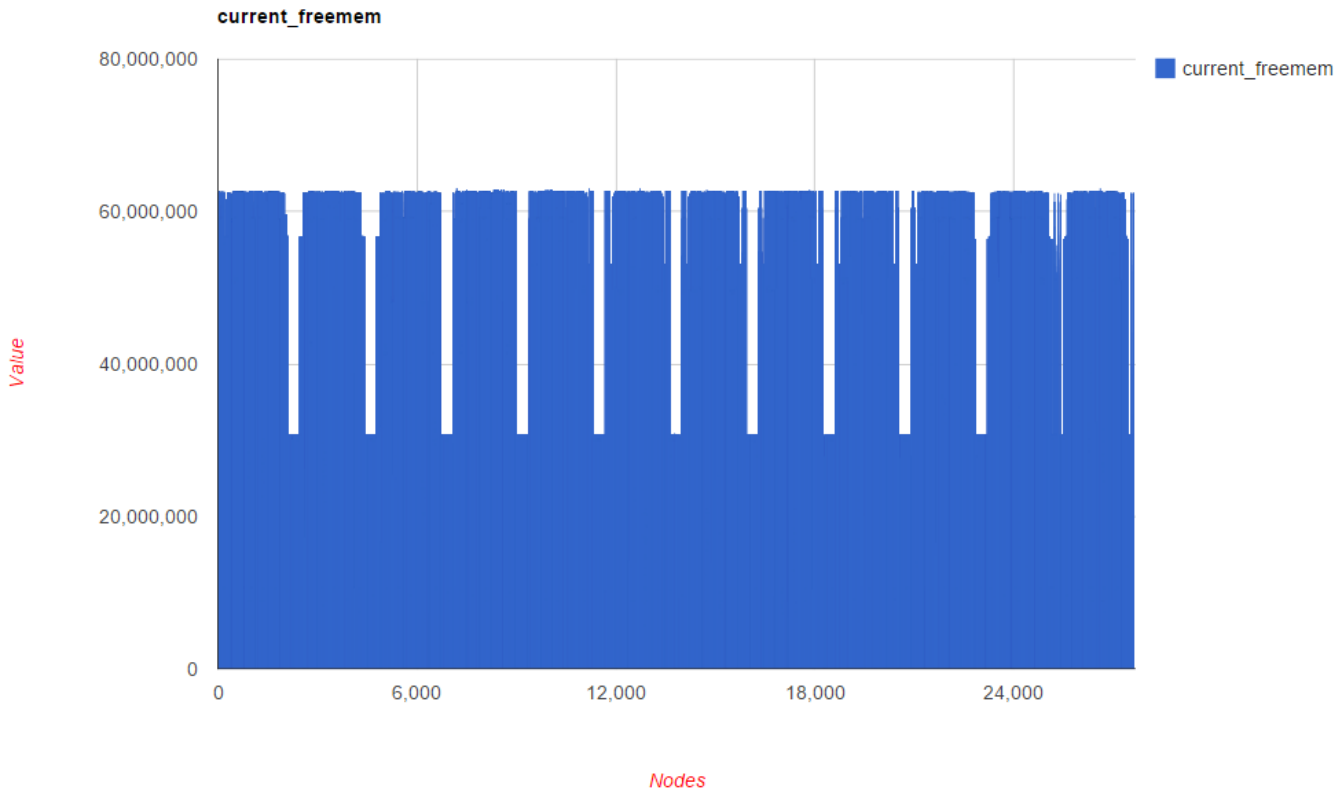
Dates Compared	From: Traditional ~6 months (July 1 2014 - Jan 13 2015) To: TAS ~10 months (Jan 15– Nov 5 2015)
Representation	92 comparable job sets 16 projects 29 distinct partners 228 MNH of allocation
Application Runtime	TAS improved by 16%
Application Runtime Consistency (CV)	TAS improved CV by 63%
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."

Publication of this and similar results under way – slide courtesy of J. Enos – not for distribution or publication

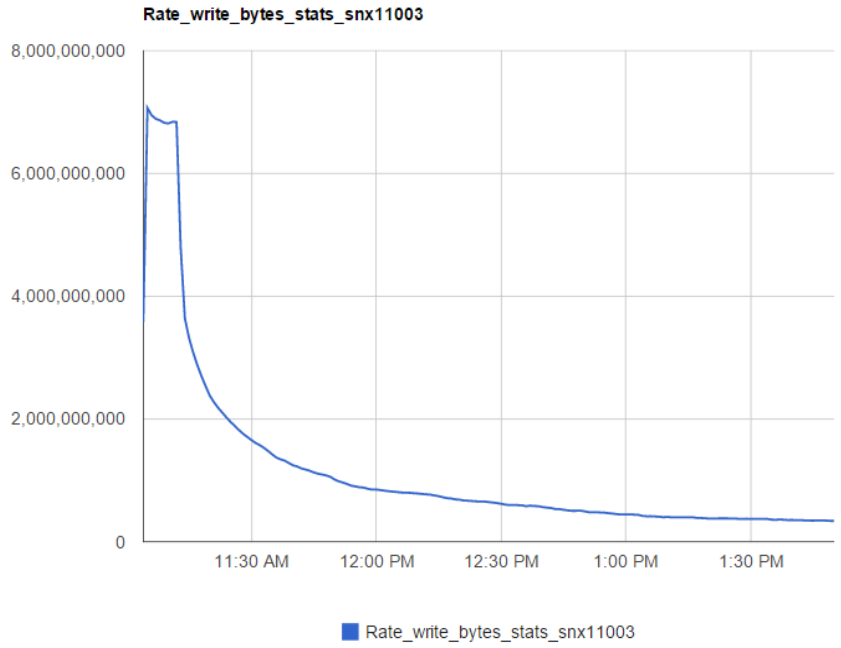
Data credited the NCSA Blue Waters Team

Single Metric in Time (Free Memory)

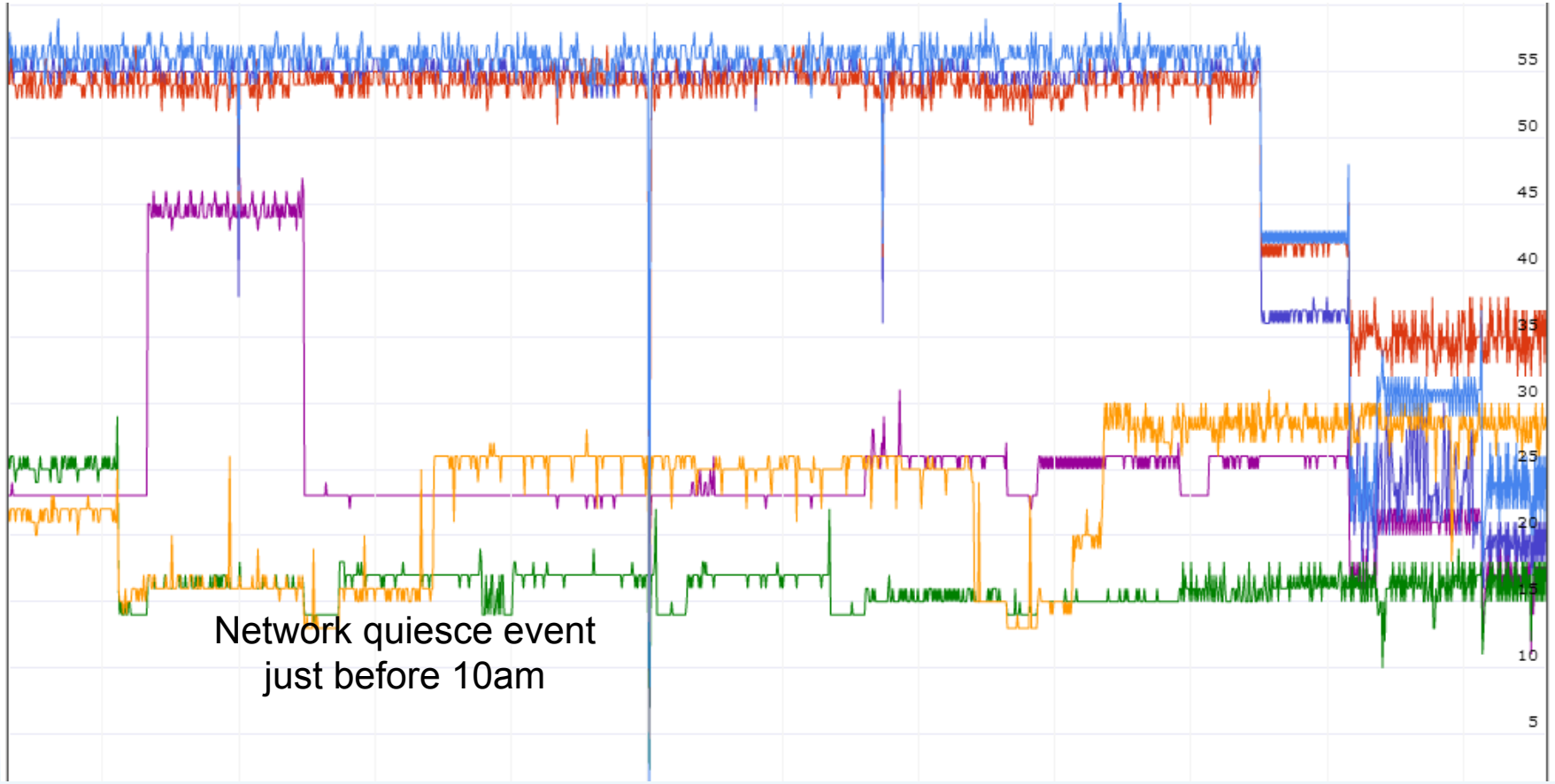


Job Write Performance

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

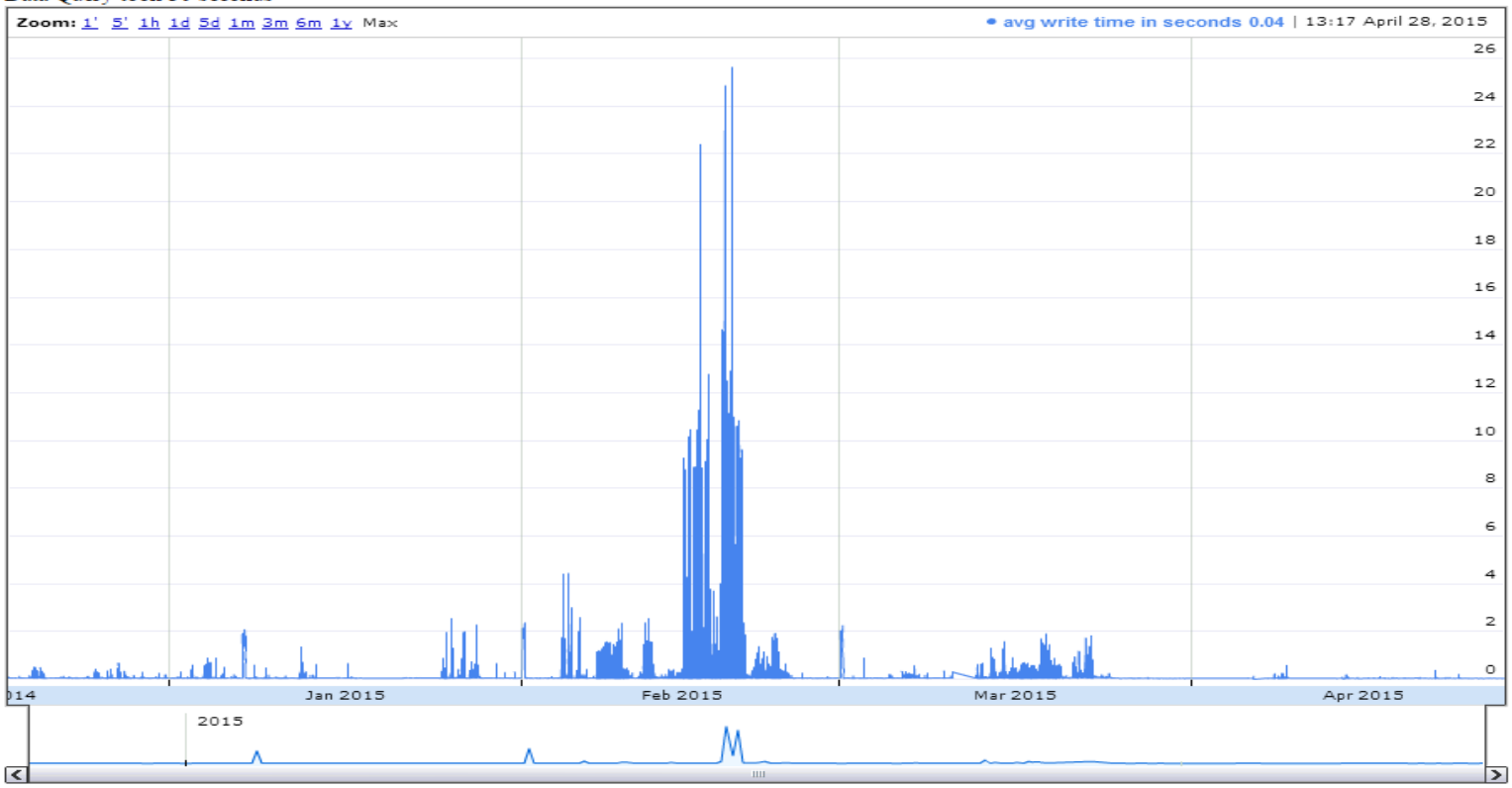


Maximal value across entire system of percent time spent in input queue stall

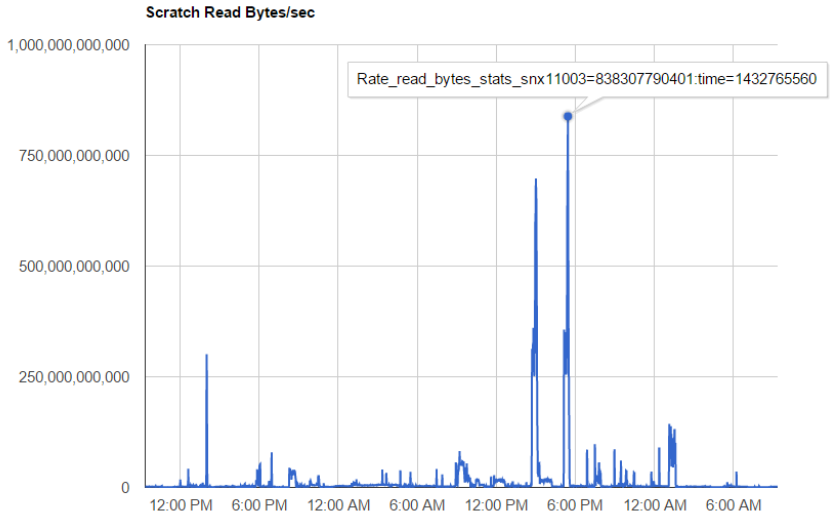


Real World Problem Example

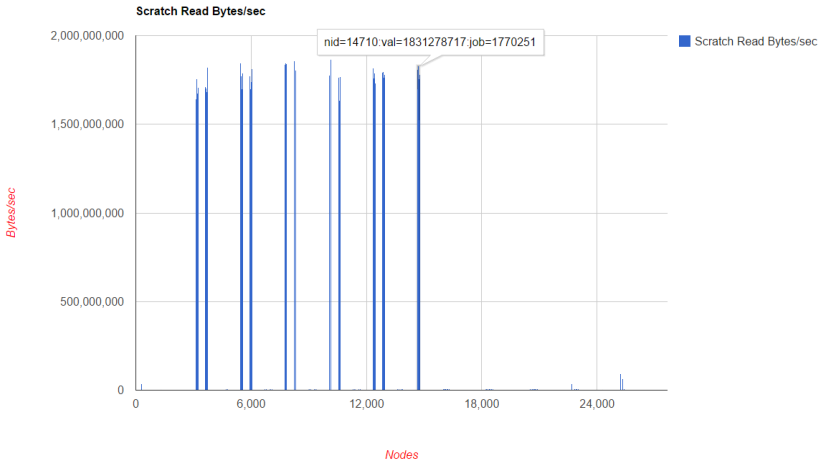
Displaying data for h2ologin
Test Number not set via testnum=, using 2 (write)
Displaying data for home filesystem
No duration Time set via length= , showing data through current time
Start=Mon, 01 Dec 2014 12:14:15 -0600
End =Tue, 28 Apr 2015 13:17:11 -0500
Data Query took 30 seconds



Sum of scratch reads across entire system



Individual node behavior at the selected time



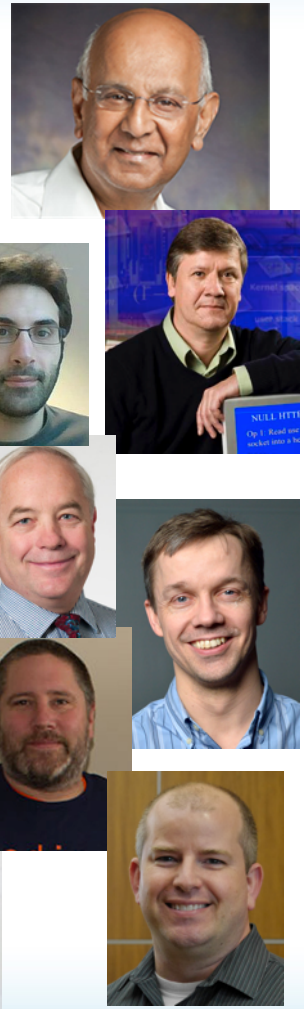
Holistic Measurement Driven Resiliency (HMDR) Project Goals

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

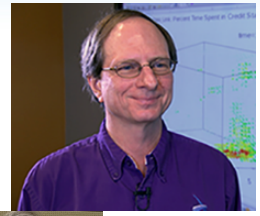
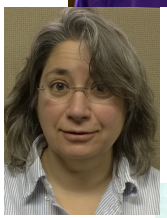
Project Participants

- 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.

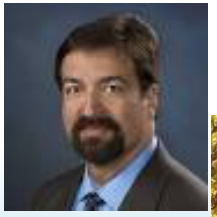


Project Participants

- 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






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. Tucker
    IEEE/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
- 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

Insights credited to the HMDR Project Team

Conclusions

- Reliability changes over time – almost always in the positive direction
  - Need to take great care in understanding when and how long studies are done
  - Ability to repair hardware while in full service has increased greatly – making HW errors less impactful
- Software errors are more impactful than hardware
  - There is much less study of SW errors in systems
  - Much less ability to do SW repairs while in full service
- Failover is often second in design and seldom designed for performance

Summary

- As a community, we have made tremendous progress in being able to collect “Petascale” and use metrics
  - Orders of Magnitude More data
  - 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 Hoefler, 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
- S. Jha, V. Formicola, Z. Kalbarczyk, C. Di Martino, W. Kramer, R. Iyer, "*Understanding Gemini Interconnect Failovers on Blue Waters: Methodology and Tools*," Cray User Group, 2016
- C. Di Martino, S. Jha, W. Kramer, Z. Kalbarczyk, R. K. Iyer, "*LogDiver: A Tool for Measuring Resilience of Extreme-Scale Systems and Applications,*" Proceedings of the 5th Workshop on Fault Tolerance for HPC at eXtreme Scale, FTXS '15, Portland, OR, USA, pp.11–18, June 15–19, 2015
- C. Di Martino, Z. Kalbarczyk, W. Kramer, R. Iyer, "*Measuring and Understanding Extreme-Scale Application Resilience: A Field Study of 5,000,000 HPC Application Runs*," 2015 45th IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2015, Rio de Janeiro, Brazil, pp. 25–36, June 22–25, 2015