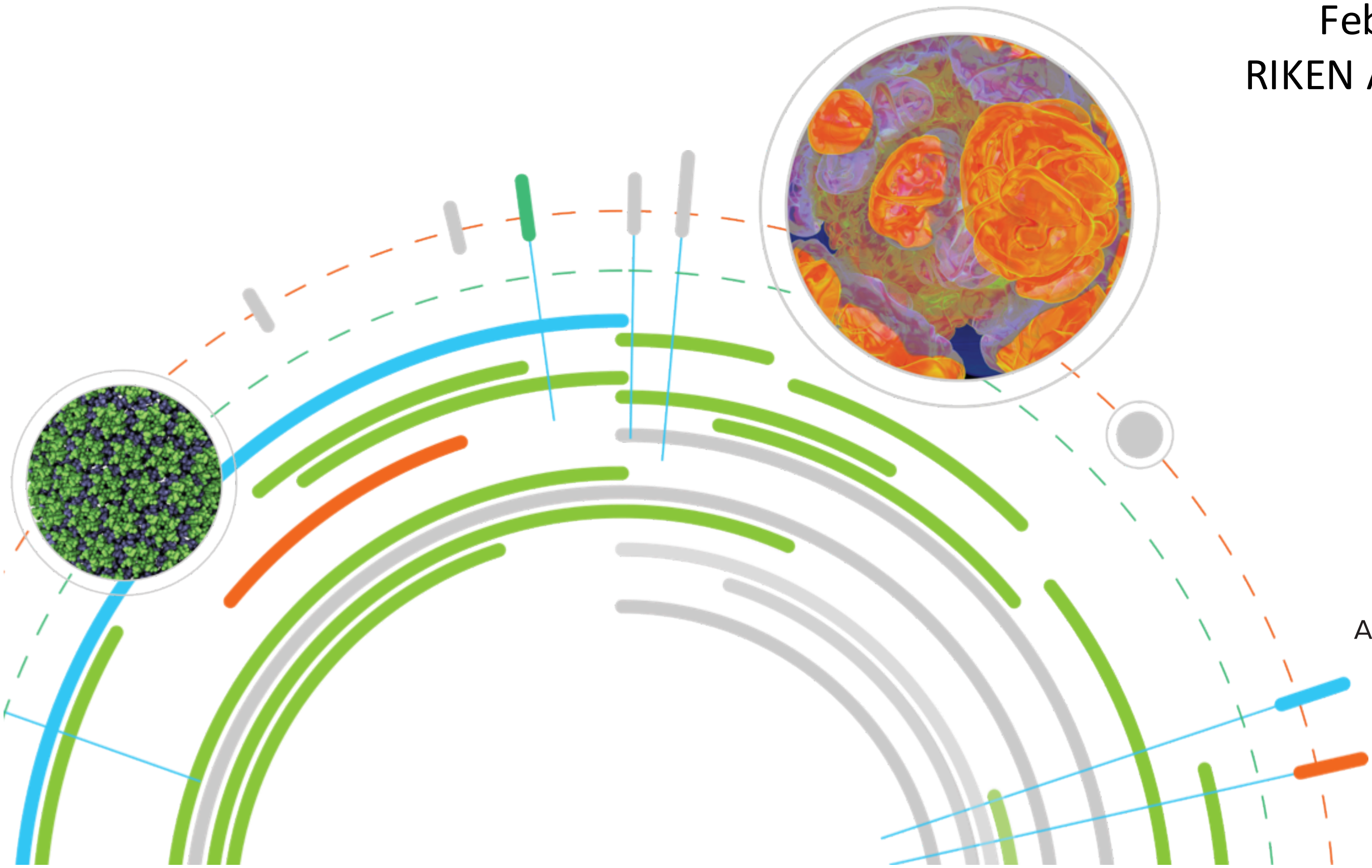
ALCF Operations Best Practices and Highlights

Mark Fahey

The 6th AICS International Symposium
Feb 22-23, 2016
RIKEN AICS, Kobe, Japan



Argonne Leadership Computing Facility

Argonne NATIONAL LABORATORY

Overview

- ALCF Organization and Structure
- IBM BG/Q Mira
- Storage/Archive
- Operational highlights
  - Scheduling
  - Monitoring
  - Operational Assessment and Job Failure Analysis
- CORAL
  - Theta and Aurora
  - What's new/what stays the same
- Collaboration opportunities

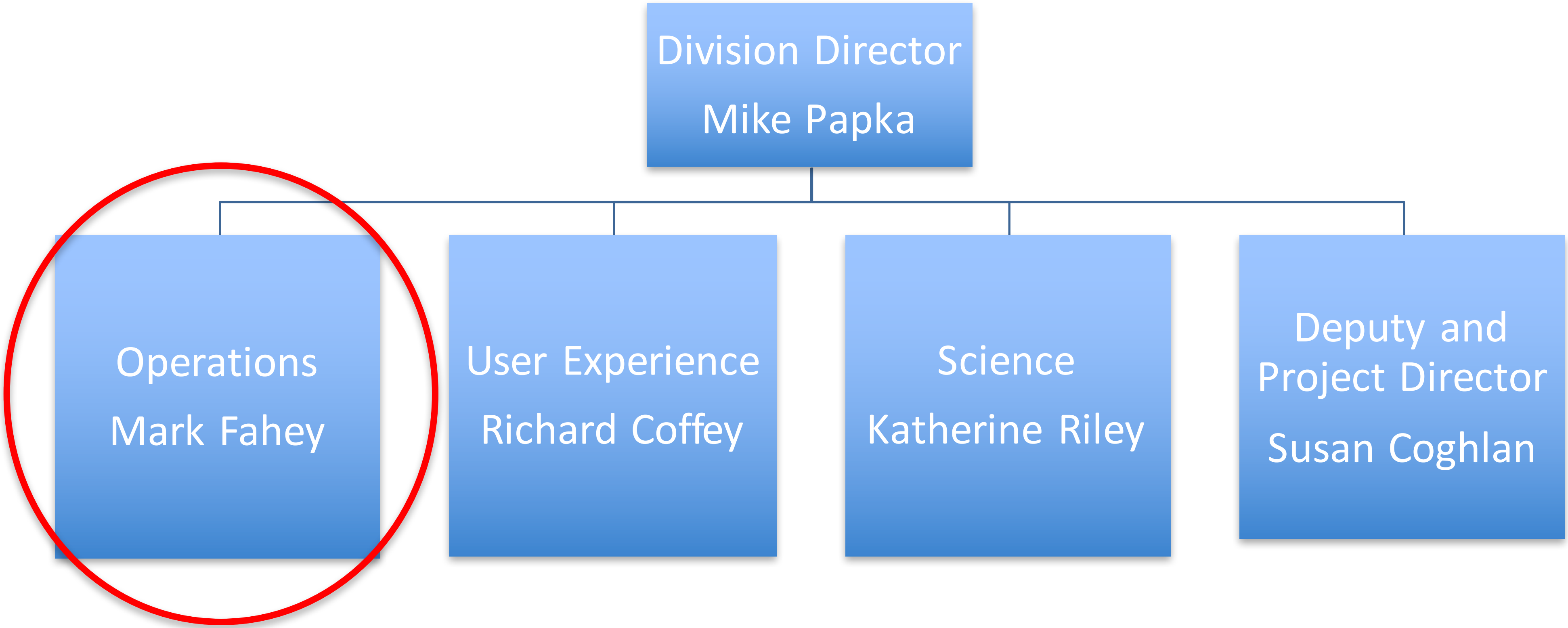
Argonne Leadership Computing Facility

- Supported by the DOE’s Advanced Scientific Computing Research program, the Argonne Leadership Computing Facility is one of two DOE Leadership Computing Facility (LCF) centers in the nation dedicated to open science.

- The LCFs deploy two diverse high-performance computer architectures that are 10 to 100 times more powerful than systems typically available for open scientific research.

- The LCF provides world-class computational capabilities to the scientific and engineering community to advance fundamental discovery and understanding in a broad range of disciplines.

ALCF Structure



Division Director

Mike Papka

Operations

Mark Fahey

User Experience

Richard Coffey

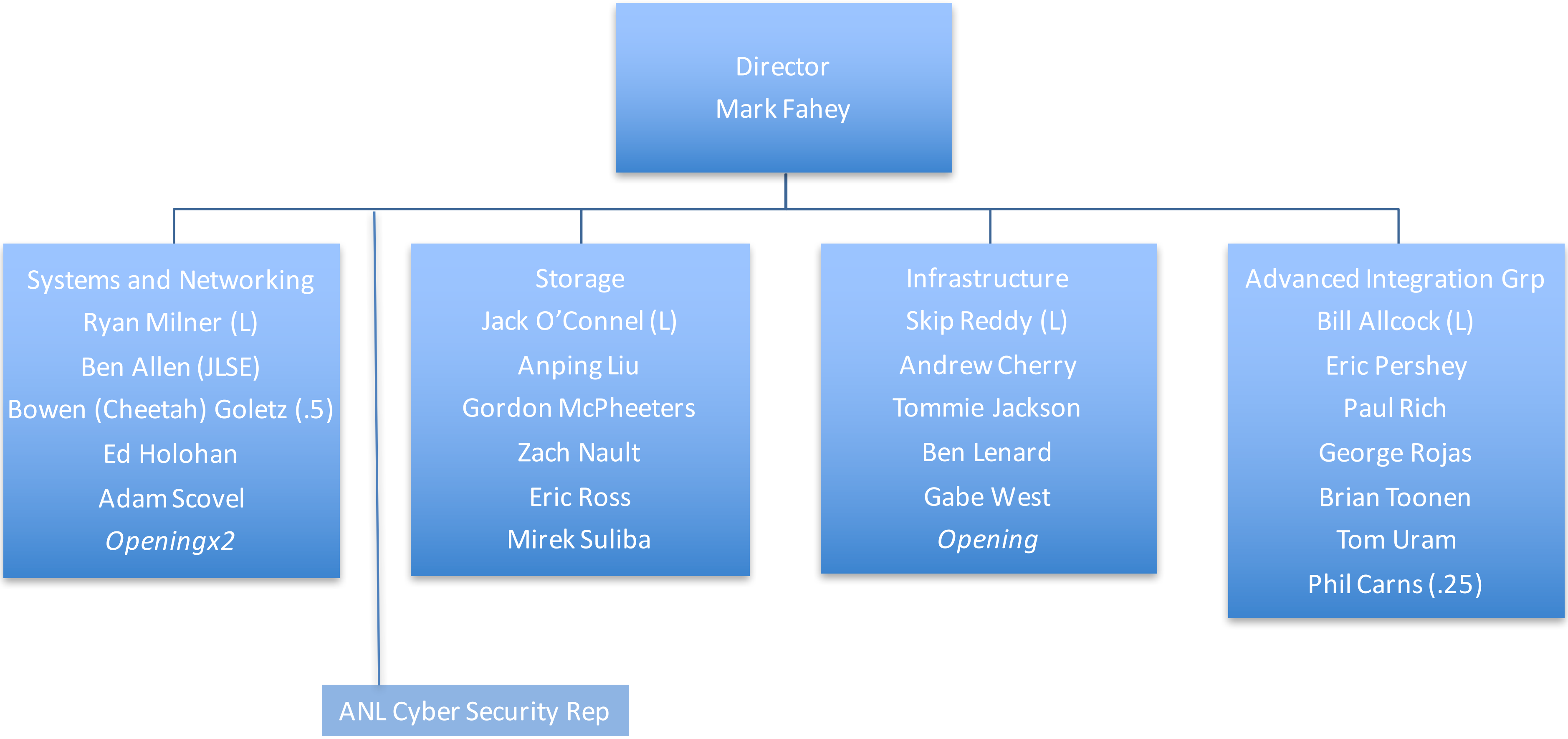
Science

Katherine Riley

Deputy and Project Director

Susan Coghlan

ALCF Operations Org Chart



Director
Mark Fahey

Systems and Networking
Ryan Milner (L)
Ben Allen (JLSE)
Bowen (Cheetah) Goletz (.5)
Ed Holohan
Adam Scovel
Openingx2

Storage
Jack O'Connel (L)
Anping Liu
Gordon McPheeters
Zach Nault
Eric Ross
Mirek Suliba

Infrastructure
Skip Reddy (L)
Andrew Cherry
Tommie Jackson
Ben Lenard
Gabe West
Opening

Advanced Integration Grp
Bill Allcock (L)
Eric Pershey
Paul Rich
George Rojas
Brian Toonen
Tom Uram
Phil Carns (.25)

ANL Cyber Security Rep

Mira – IBM Blue Gene/Q

- 49,152 nodes / 786,432 cores
- 768 TB of memory
- Peak flop rate: 10 PF
- Linpack flop rate: 8.1 PF
- 48 racks
- 16 1,600 MHz PowerPC A2 cores per node
- 5D torus interconnect
- 384 I/O nodes

Mira and her cables

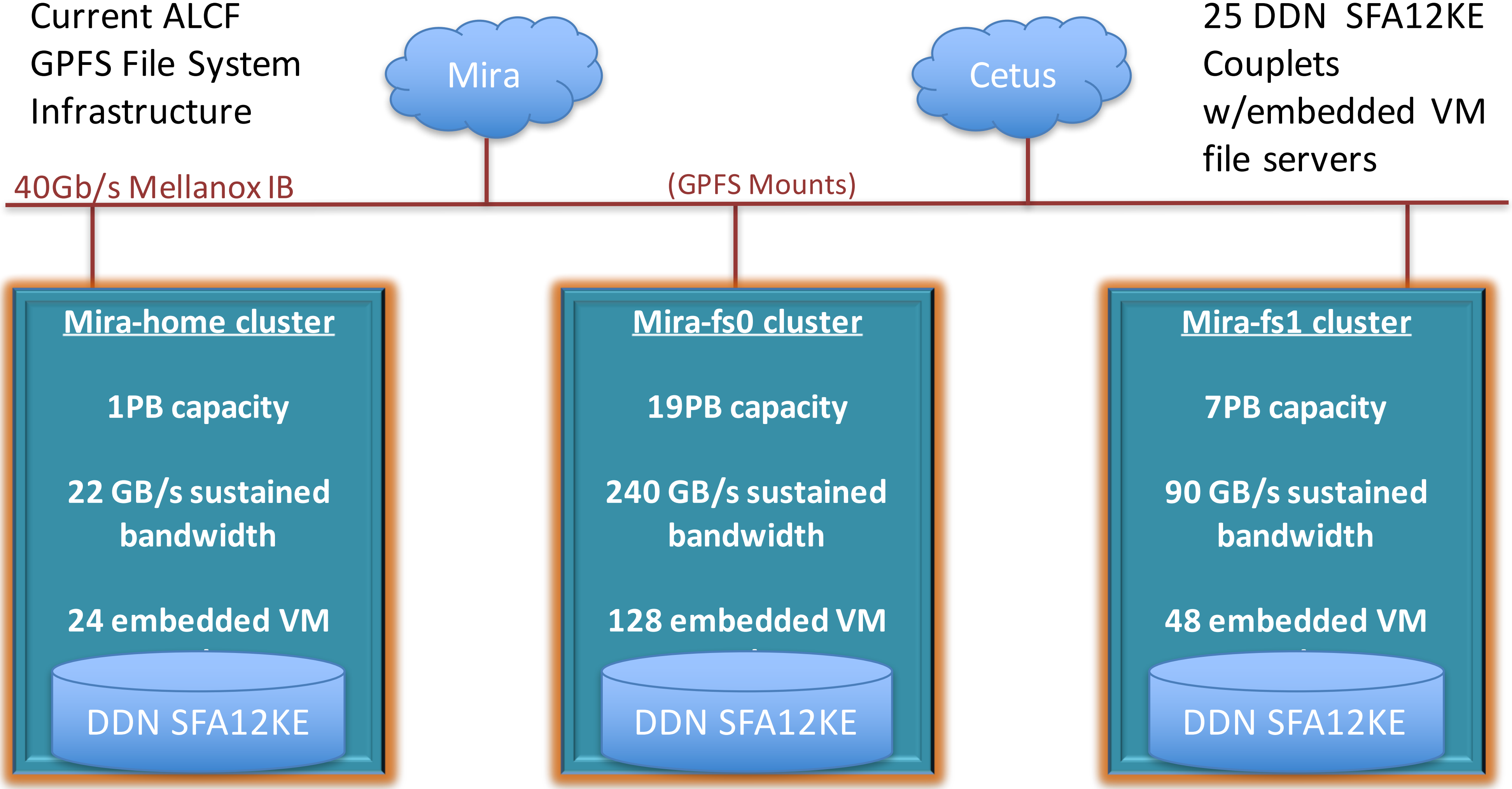
Other ALCF resources

- **Cetus** (T&D and prod.) – IBM Blue Gene/Q
  - 4,096 nodes / 65,536 cores
  - 64 TB of memory
  - 838 TF peak flop rate
- **Vesta** (T&D) – IBM Blue Gene/Q
  - 2,048 nodes / 32,768 cores
  - 32 TB of memory
  - 419 TF peak flop rate
- **Cooley** (Visualization) – Cray + NVIDIA
  - 126 nodes / 1512 x86 cores (Haswell)
  - 126 NVIDIA Tesla K80 GPUs
  - 47 TB x86 memory / 3 TB GPU memory
  - 293 TF peak flop rate

IBM Blue Gene/Q

Current ALCF
GPFS File System
Infrastructure

25 DDN SFA12KE
Couplets
w/embedded VM
file servers



40Gb/s Mellanox IB

(GPFS Mounts)

Mira-home cluster

1PB capacity

22 GB/s sustained bandwidth

24 embedded VM

Mira-fs0 cluster

19PB capacity

240 GB/s sustained bandwidth

128 embedded VM

Mira-fs1 cluster

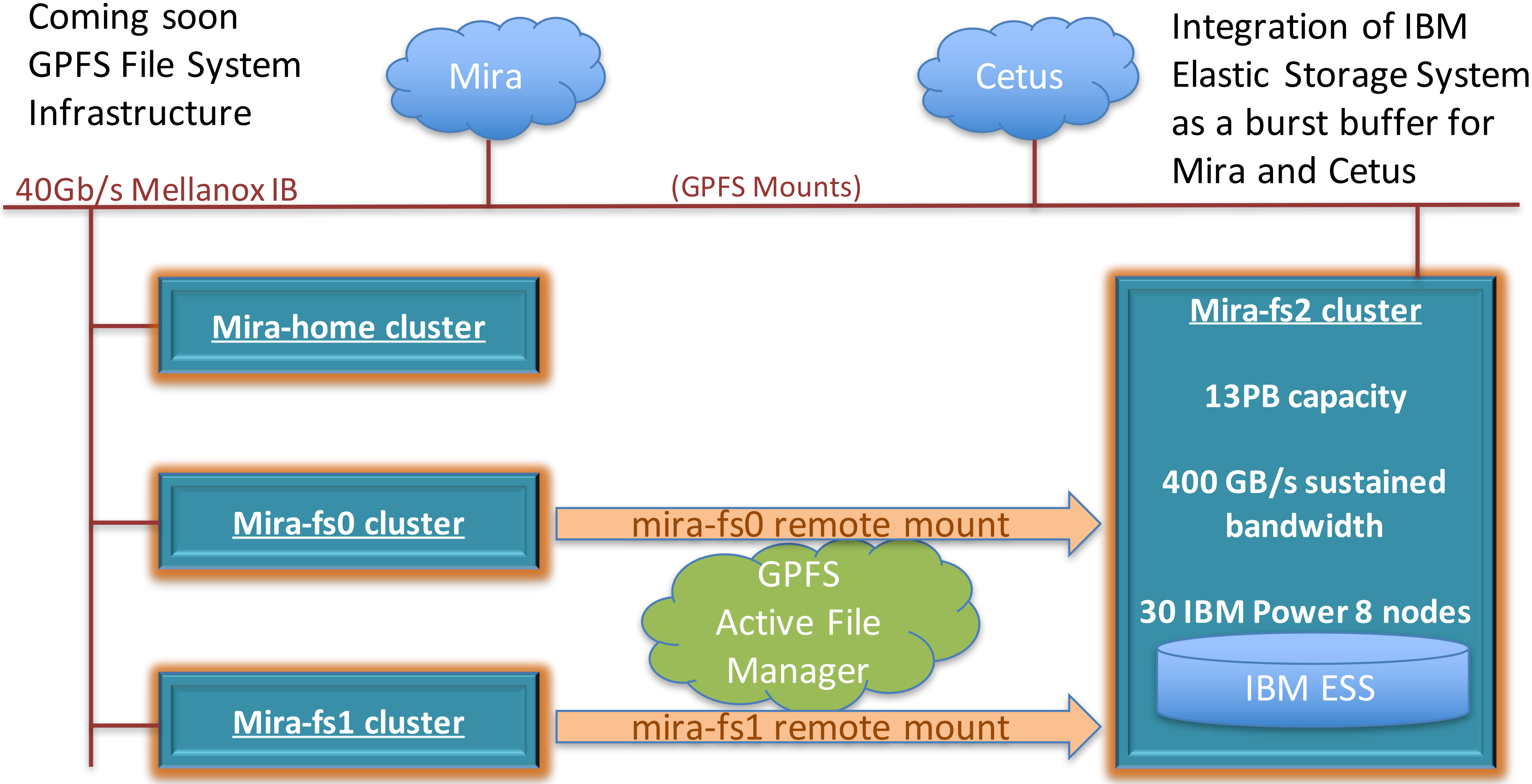
7PB capacity

90 GB/s sustained bandwidth

48 embedded VM

Coming soon
GPFS File System
Infrastructure

Integration of IBM
Elastic Storage System
as a burst buffer for
Mira and Cetus



Mira

Cetus

40Gb/s Mellanox IB

(GPFS Mounts)

Mira-home cluster

Mira-fs0 cluster

Mira-fs1 cluster

mira-fs0 remote mount

mira-fs1 remote mount

GPFS
Active File
Manager

Mira-fs2 cluster

13PB capacity

400 GB/s sustained
bandwidth

30 IBM Power 8 nodes

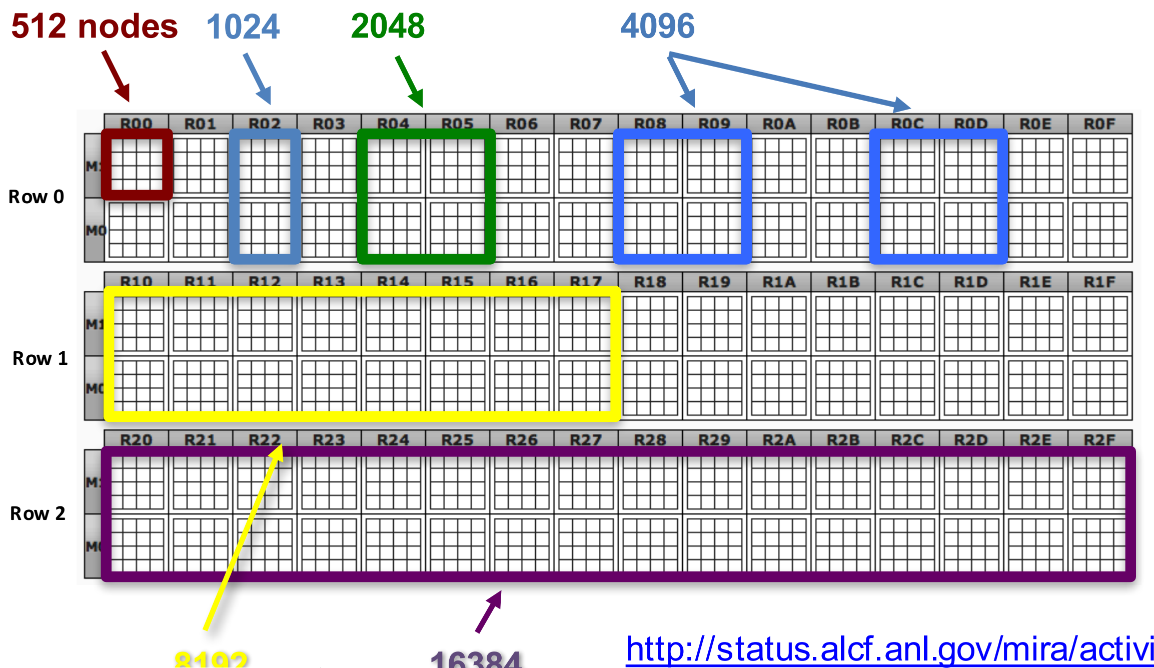
IBM ESS

In progress

Scheduling - Cobalt

- Orginally COBALT (Component-Based Lightweight Toolkit) was a set of component-based system software for system and resource management developed within Argonne's Mathematics and Computer Science Division
- Cobalt is a set of system software for high performance machines
  - The main piece is a set of resource management components for IBM BG systems and clusters.
- ALCF adopted the resource scheduling component and continued to enhance it for use within the facility
  - ALCF sees resource scheduling a major component of future facilities and its research/development efforts are focused on future needs

Mira multiple rack partitions (“blocks”)


512 nodes
1024
2048
4096
R00 R01 R02 R03 R04 R05 R06 R07 R08 R09 R0A R0B R0C R0D R0E R0F
Row 0
M1
M0
R10 R11 R12 R13 R14 R15 R16 R17 R18 R19 R1A R1B R1C R1D R1E R1F
Row 1
M1
M0
R20 R21 R22 R23 R24 R25 R26 R27 R28 R29 R2A R2B R2C R2D R2E R2F
Row 2
M1
M0
8192
16384

http://status.alcf.anl.gov/mira/activity

partlist will show you if a large free block is busy due to a wiring dependency

Nodes	A	B	C	D	E
512	4	4	4	4	2
1024	4	4	4	8	2
2048	4	4	4	16	2
4096	4/8	4	8/4	16	2
8192	4	4	16	16	2
12288	8	4	12	16	2
16384	4/8	8/4	16	16	2
24576	4	12	16	16	2
32768	8	8	16	16	2
49152	8	12	16	16	2

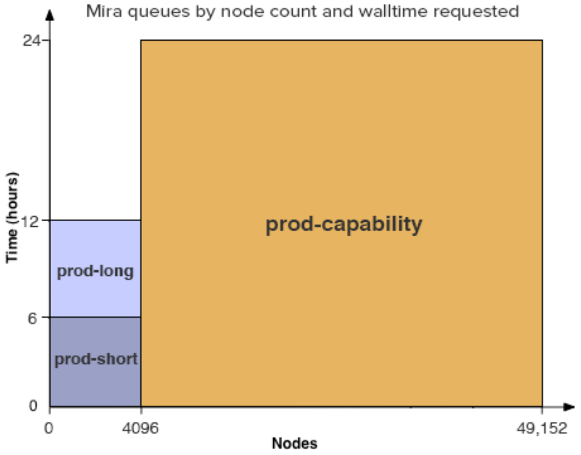
The number of large block sizes possible is:

# of nodes	# of blocks
49152	1
32768	3
24576	2
16384	9
12288	12
8192	6
4096	12
2048	24
1024	64
512	96

Mira job scheduling

- Restrictions in queues
  - **prod-long:** restricted to the row 0.
  - **prod-short**, **prod-capability:** can run in the full machine

http://www.alcf.anl.gov/user-guides/job-scheduling-policy-bgq-systems

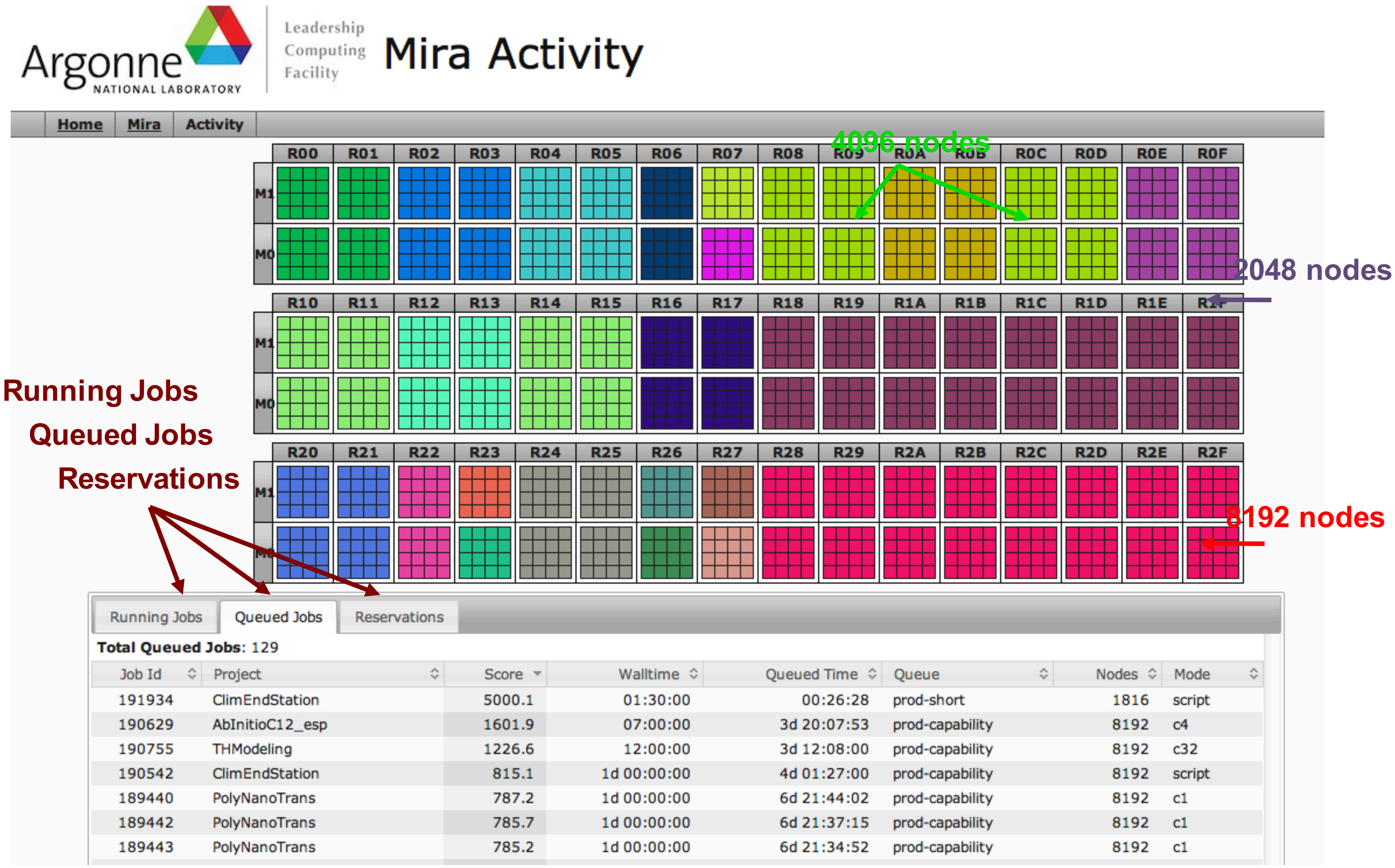


User Queued	Underlying Queue	Nodes	Wall-clock Time (hours)	Max. Running per User	Max. Queued per User
prod	prod-short	512 - 4096	0 - ≤6	5	20
	prod-long	512 - 4096	>6 - 12	5	20
	prod-capability	4097 - 49152	0 - 24	5	20
	backfill (*)	512 - 49152	0 - 6	5	20
prod-1024-torus	prod-1024-torus	1024	0 - 12	5	16
prod-32768-torus	prod-32768-torus	32768	0 - 24	1	20

(*) This queue is automatically selected based on the scheduling policy.

- **I/O to compute node ratio 1:128**

Machine status web page


Argonne NATIONAL LABORATORY
Leadership Computing Facility
Mira Activity
Home
Mira
Activity
R00 R01 R02 R03 R04 R05 R06 R07 R08 R09 R0A R0B R0C R0D R0E R0F
M1
M0
R10 R11 R12 R13 R14 R15 R16 R17 R18 R19 R1A R1B R1C R1D R1E R1F
M1
M0
R20 R21 R22 R23 R24 R25 R26 R27 R28 R29 R2A R2B R2C R2D R2E R2F
M1
M0
Running Jobs
Queued Jobs
Reservations

4096 nodes

2048 nodes

8192 nodes

Running Jobs

Queued Jobs

Reservations

Total Queued Jobs: 129

Job Id	Project	Score	Walltime	Queued Time	Queue	Nodes	Mode
191934	ClimEndStation	5000.1	01:30:00	00:26:28	prod-short	1816	script
190629	AbInitioC12_esp	1601.9	07:00:00	3d 20:07:53	prod-capability	8192	c4
190755	THModeling	1226.6	12:00:00	3d 12:08:00	prod-capability	8192	c32
190542	ClimEndStation	815.1	1d 00:00:00	4d 01:27:00	prod-capability	8192	script
189440	PolyNanoTrans	787.2	1d 00:00:00	6d 21:44:02	prod-capability	8192	c1
189442	PolyNanoTrans	785.7	1d 00:00:00	6d 21:37:15	prod-capability	8192	c1
189443	PolyNanoTrans	785.2	1d 00:00:00	6d 21:34:52	prod-capability	8192	c1

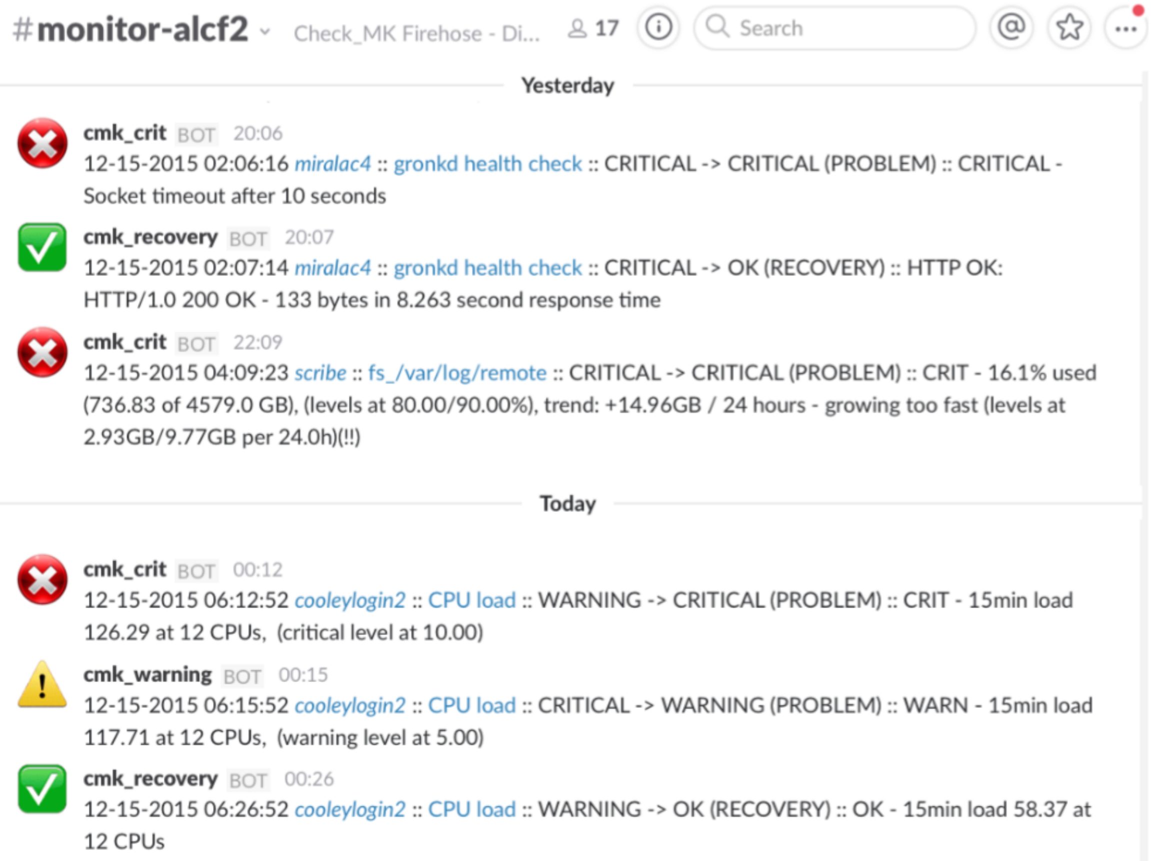
http://status.alcf.anl.gov/mira/activity

Monitoring

- Check_MK is a comprehensive Open-Source-Solution for monitoring developed around the Nagios-core
  - Allows creating rule-based configuration using Python and offloading work from the Nagios core to make it scale better, allowing more systems to be monitored from a single Nagios server
  - Checks that consist of agent-side and server-side parts
- Check_MK is monitored using monit and MRTG
- Team members are asked to subscribe to categories of alerts. Individual subscriptions are meant to ensure that notifications remain relevant for each team member.
- Stakeholders are required to tune the monitoring of hosts and services

Monitoring – Slack integration

The ALCF Check_MK instance is further customized to publish alerts to a dedicated channel using the SLACK API



Yesterday

cmk_crit BOT 20:06
12-15-2015 02:06:16 *miralac4* :: gronkd health check :: CRITICAL -> CRITICAL (PROBLEM) :: CRITICAL - Socket timeout after 10 seconds

cmk_recovery BOT 20:07
12-15-2015 02:07:14 *miralac4* :: gronkd health check :: CRITICAL -> OK (RECOVERY) :: HTTP OK: HTTP/1.0 200 OK - 133 bytes in 8.263 second response time

cmk_crit BOT 22:09
12-15-2015 04:09:23 *scribe* :: fs_/var/log/remote :: CRITICAL -> CRITICAL (PROBLEM) :: CRIT - 16.1% used (736.83 of 4579.0 GB), (levels at 80.00/90.00%), trend: +14.96GB / 24 hours - growing too fast (levels at 2.93GB/9.77GB per 24.0h)(!!)

Today

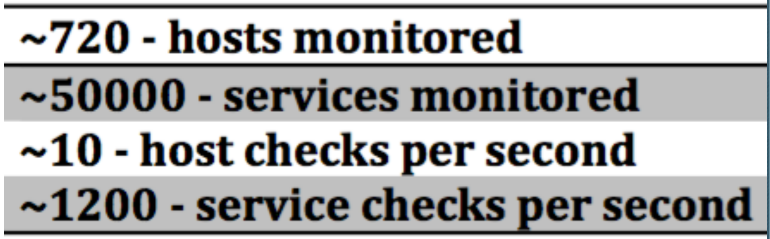
cmk_crit BOT 00:12
12-15-2015 06:12:52 *cooleylogin2* :: CPU load :: WARNING -> CRITICAL (PROBLEM) :: CRIT - 15min load 126.29 at 12 CPUs, (critical level at 10.00)

cmk_warning BOT 00:15
12-15-2015 06:15:52 *cooleylogin2* :: CPU load :: CRITICAL -> WARNING (PROBLEM) :: WARN - 15min load 117.71 at 12 CPUs, (warning level at 5.00)

cmk_recovery BOT 00:26
12-15-2015 06:26:52 *cooleylogin2* :: CPU load :: WARNING -> OK (RECOVERY) :: OK - 15min load 58.37 at 12 CPUs

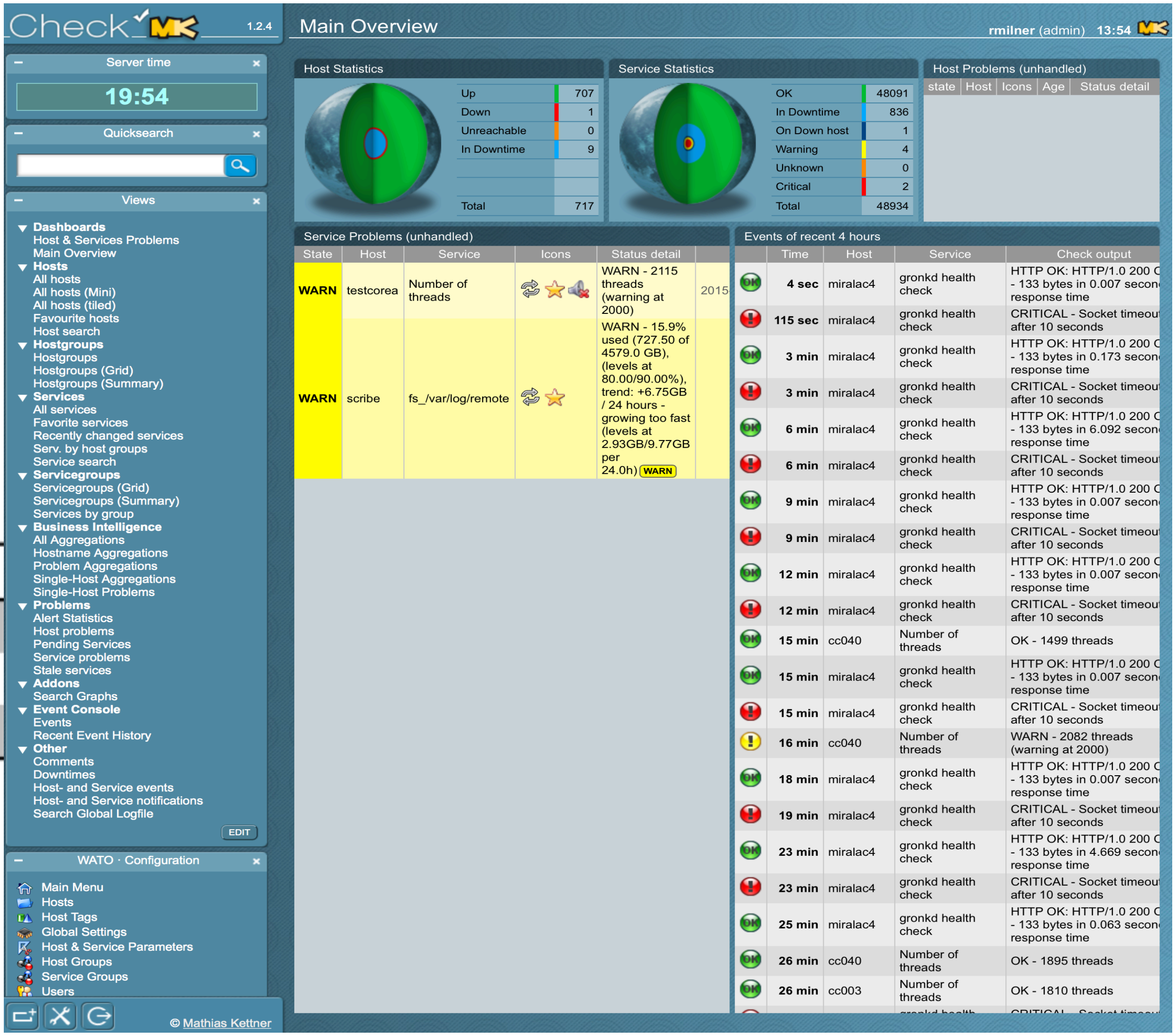
Check_MK GUI

Standard UI landing



~720	hosts monitored
~50000	services monitored
~10	host checks per second
~1200	service checks per second

ALCF monitoring statistics



Host Statistics

Up	707
Down	1
Unreachable	0
In Downtime	9
Total	717

Service Statistics

OK	48091
In Downtime	836
On Down host	1
Warning	4
Unknown	0
Critical	2
Total	48934

Host Problems (unhandled)

state	Host	Icons	Age	Status detail

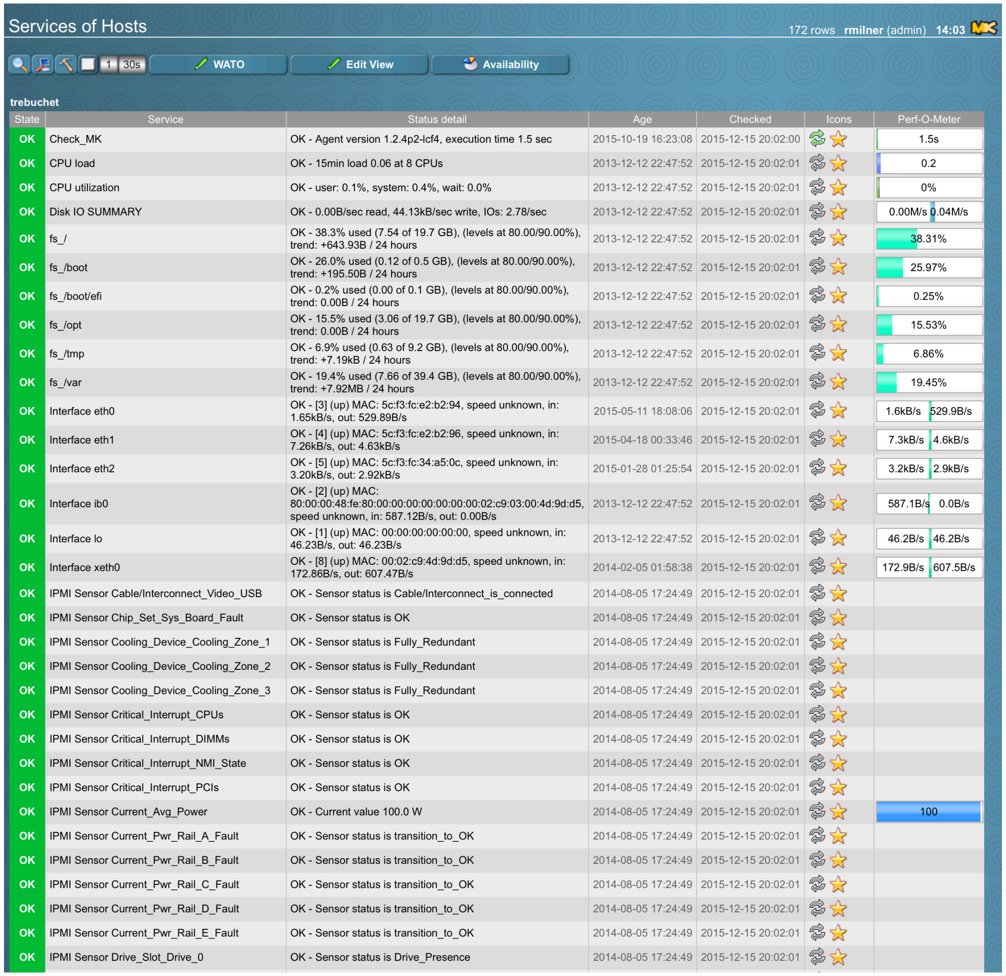
Service Problems (unhandled)

State	Host	Service	Icons	Status detail	
WARN	testcorea	Number of threads		WARN - 2115 threads (warning at 2000)	2015
WARN	scribe	fs_/var/log/remote		WARN - 15.9% used (727.50 of 4579.0 GB), (levels at 80.00/90.00%), trend: +6.75GB / 24 hours - growing too fast (levels at 2.93GB/9.77GB per 24.0h) WARN	

Events of recent 4 hours

	Time	Host	Service	Check output
OK	4 sec	miralac4	gronkd health check	HTTP OK: HTTP/1.0 200 C - 133 bytes in 0.007 secon response time
!	115 sec	miralac4	gronkd health check	CRITICAL - Socket timeout after 10 seconds
OK	3 min	miralac4	gronkd health check	HTTP OK: HTTP/1.0 200 C - 133 bytes in 0.173 secon response time
!	3 min	miralac4	gronkd health check	CRITICAL - Socket timeout after 10 seconds
OK	6 min	miralac4	gronkd health check	HTTP OK: HTTP/1.0 200 C - 133 bytes in 6.092 secon response time
!	6 min	miralac4	gronkd health check	CRITICAL - Socket timeout after 10 seconds
OK	9 min	miralac4	gronkd health check	HTTP OK: HTTP/1.0 200 C - 133 bytes in 0.007 secon response time
!	9 min	miralac4	gronkd health check	CRITICAL - Socket timeout after 10 seconds
OK	12 min	miralac4	gronkd health check	HTTP OK: HTTP/1.0 200 C - 133 bytes in 0.007 secon response time
!	12 min	miralac4	gronkd health check	CRITICAL - Socket timeout after 10 seconds
OK	15 min	cc040	Number of threads	OK - 1499 threads
OK	15 min	miralac4	gronkd health check	HTTP OK: HTTP/1.0 200 C - 133 bytes in 0.007 secon response time
!	15 min	miralac4	gronkd health check	CRITICAL - Socket timeout after 10 seconds
!	16 min	cc040	Number of threads	WARN - 2082 threads (warning at 2000)
OK	18 min	miralac4	gronkd health check	HTTP OK: HTTP/1.0 200 C - 133 bytes in 0.007 secon response time
!	19 min	miralac4	gronkd health check	CRITICAL - Socket timeout after 10 seconds
OK	23 min	miralac4	gronkd health check	HTTP OK: HTTP/1.0 200 C - 133 bytes in 4.669 secon response time
!	23 min	miralac4	gronkd health check	CRITICAL - Socket timeout after 10 seconds
OK	25 min	miralac4	gronkd health check	HTTP OK: HTTP/1.0 200 C - 133 bytes in 0.063 secon response time
OK	26 min	cc040	Number of threads	OK - 1895 threads
OK	26 min	cc003	Number of threads	OK - 1810 threads

Host and service views

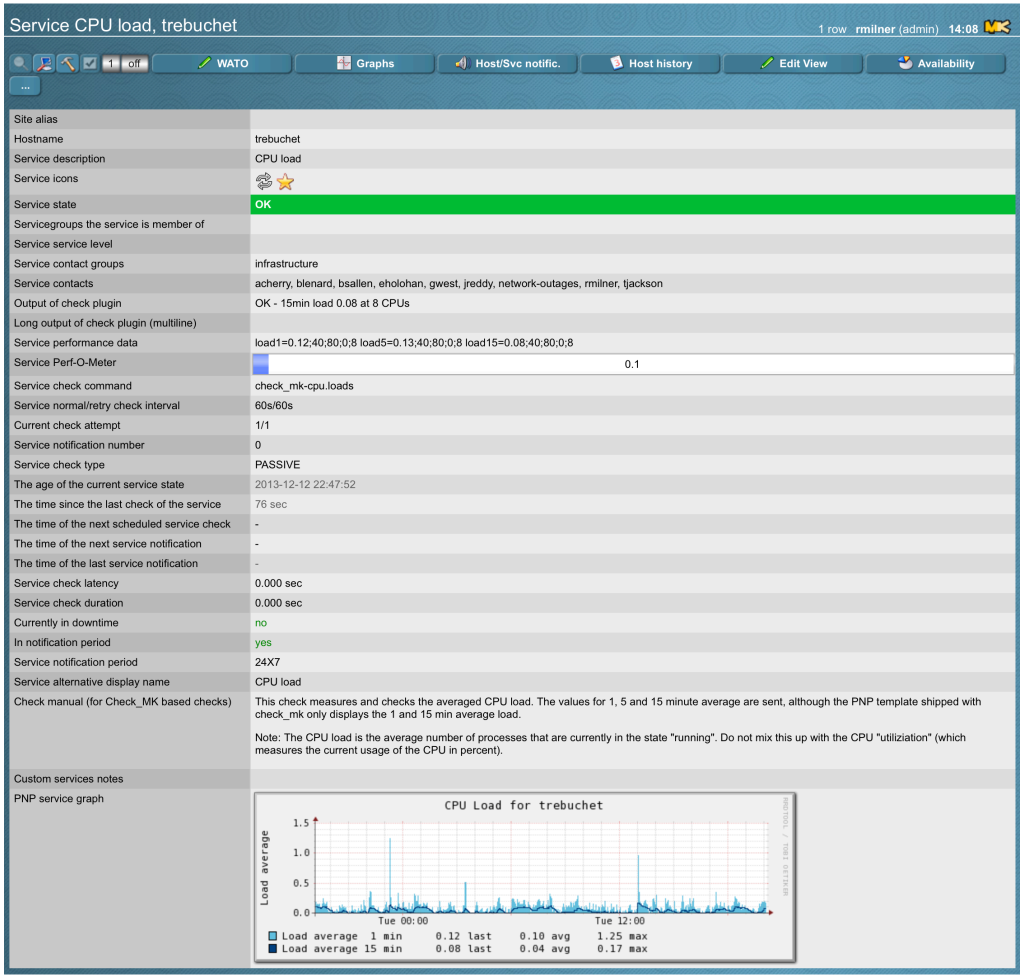


Services of Hosts

172 rows **rmilner** (admin) **14:03**

trebuchet

State	Service	Status detail	Age	Checked	Icons	Perf-O-Meter
OK	Check_MK	OK - Agent version 1.2.4p2-lcf4, execution time 1.5 sec	2015-10-19 16:23:08	2015-12-15 20:02:00	⟳ ☆	1.5s
OK	CPU load	OK - 15min load 0.06 at 8 CPUs	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	0.2
OK	CPU utilization	OK - user: 0.1%, system: 0.4%, wait: 0.0%	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	0%
OK	Disk IO SUMMARY	OK - 0.00B/sec read, 44.13kB/sec write, IOs: 2.78/sec	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	0.00M/s 0.04M/s
OK	fs_/	OK - 38.3% used (7.54 of 19.7 GB), (levels at 80.00/90.00%), trend: +643.93B / 24 hours	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	38.31%
OK	fs_/boot	OK - 26.0% used (0.12 of 0.5 GB), (levels at 80.00/90.00%), trend: +195.50B / 24 hours	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	25.97%
OK	fs_/boot/efi	OK - 0.2% used (0.00 of 0.1 GB), (levels at 80.00/90.00%), trend: 0.00B / 24 hours	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	0.25%
OK	fs_/opt	OK - 15.5% used (3.06 of 19.7 GB), (levels at 80.00/90.00%), trend: 0.00B / 24 hours	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	15.53%
OK	fs_/tmp	OK - 6.9% used (0.63 of 9.2 GB), (levels at 80.00/90.00%), trend: +7.19kB / 24 hours	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	6.86%
OK	fs_/var	OK - 19.4% used (7.66 of 39.4 GB), (levels at 80.00/90.00%), trend: +7.92MB / 24 hours	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	19.45%
OK	Interface eth0	OK - [3] (up) MAC: 5c:f3:fc:e2:b2:94, speed unknown, in: 1.65kB/s, out: 529.89B/s	2015-05-11 18:08:06	2015-12-15 20:02:01	⟳ ☆	1.6kB/s 529.9B/s
OK	Interface eth1	OK - [4] (up) MAC: 5c:f3:fc:e2:b2:96, speed unknown, in: 7.26kB/s, out: 4.63kB/s	2015-04-18 00:33:46	2015-12-15 20:02:01	⟳ ☆	7.3kB/s 4.6kB/s
OK	Interface eth2	OK - [5] (up) MAC: 5c:f3:fc:34:a5:0c, speed unknown, in: 3.20kB/s, out: 2.92kB/s	2015-01-28 01:25:54	2015-12-15 20:02:01	⟳ ☆	3.2kB/s 2.9kB/s
OK	Interface ib0	OK - [2] (up) MAC: 80:00:00:48:fe:80:00:00:00:00:00:00:00:02:c9:03:00:4d:9d:d5, speed unknown, in: 587.12B/s, out: 0.00B/s	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	587.1B/s 0.0B/s
OK	Interface lo	OK - [1] (up) MAC: 00:00:00:00:00:00, speed unknown, in: 46.23B/s, out: 46.23B/s	2013-12-12 22:47:52	2015-12-15 20:02:01	⟳ ☆	46.2B/s 46.2B/s
OK	Interface xeth0	OK - [8] (up) MAC: 00:02:c9:4d:9d:d5, speed unknown, in: 172.86B/s, out: 607.47B/s	2014-02-05 01:58:38	2015-12-15 20:02:01	⟳ ☆	172.9B/s 607.5B/s
OK	IPMI Sensor Cable/Interconnect_Video_USB	OK - Sensor status is Cable/Interconnect_is_connected	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Chip_Set_Sys_Board_Fault	OK - Sensor status is OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Cooling_Device_Cooling_Zone_1	OK - Sensor status is Fully_Redundant	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Cooling_Device_Cooling_Zone_2	OK - Sensor status is Fully_Redundant	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Cooling_Device_Cooling_Zone_3	OK - Sensor status is Fully_Redundant	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Critical_Interrupt_CPUs	OK - Sensor status is OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Critical_Interrupt_DIMMs	OK - Sensor status is OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Critical_Interrupt_NMI_State	OK - Sensor status is OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Critical_Interrupt_PCIs	OK - Sensor status is OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Current_Avg_Power	OK - Current value 100.0 W	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	100
OK	IPMI Sensor Current_Pwr_Rail_A_Fault	OK - Sensor status is transition_to_OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Current_Pwr_Rail_B_Fault	OK - Sensor status is transition_to_OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Current_Pwr_Rail_C_Fault	OK - Sensor status is transition_to_OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Current_Pwr_Rail_D_Fault	OK - Sensor status is transition_to_OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Current_Pwr_Rail_E_Fault	OK - Sensor status is transition_to_OK	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	
OK	IPMI Sensor Drive_Slot_Drive_0	OK - Sensor status is Drive_Presence	2014-08-05 17:24:49	2015-12-15 20:02:01	⟳ ☆	



Service CPU load, trebuchet

1 row **rmilner** (admin) **14:08**

Site alias	
Hostname	trebuchet
Service description	CPU load
Service icons	⟳ ☆
Service state	OK
Servicegroups the service is member of	
Service service level	
Service contact groups	infrastructure
Service contacts	acherry, blenard, bsallen, eholohan, gwest, jreddy, network-outages, rmilner, tjackson
Output of check plugin	OK - 15min load 0.08 at 8 CPUs
Long output of check plugin (multiline)	
Service performance data	load1=0.12;40;80;0;8 load5=0.13;40;80;0;8 load15=0.08;40;80;0;8
Service Perf-O-Meter	0.1
Service check command	check_mk-cpu.loads
Service normal/retry check interval	60s/60s
Current check attempt	1/1
Service notification number	0
Service check type	PASSIVE
The age of the current service state	2013-12-12 22:47:52
The time since the last check of the service	76 sec
The time of the next scheduled service check	-
The time of the next service notification	-
The time of the last service notification	-
Service check latency	0.000 sec
Service check duration	0.000 sec
Currently in downtime	no
In notification period	yes
Service notification period	24X7
Service alternative display name	CPU load
Check manual (for Check_MK based checks)	This check measures and checks the averaged CPU load. The values for 1, 5 and 15 minute average are sent, although the PNP template shipped with check_mk only displays the 1 and 15 min average load. Note: The CPU load is the average number of processes that are currently in the state "running". Do not mix this up with the CPU "utiliziation" (which measures the current usage of the CPU in percent).
Custom services notes	
PNP service graph	

CPU Load for trebuchet
Load average
1.5
1.0
0.5
0.0
Tue 00:00
Tue 12:00
Load average 1 min 0.12 last 0.10 avg 1.25 max
Load average 15 min 0.08 last 0.04 avg 0.17 max
RRDTOOL / TOBI OETIKER

Operational Assessment Process

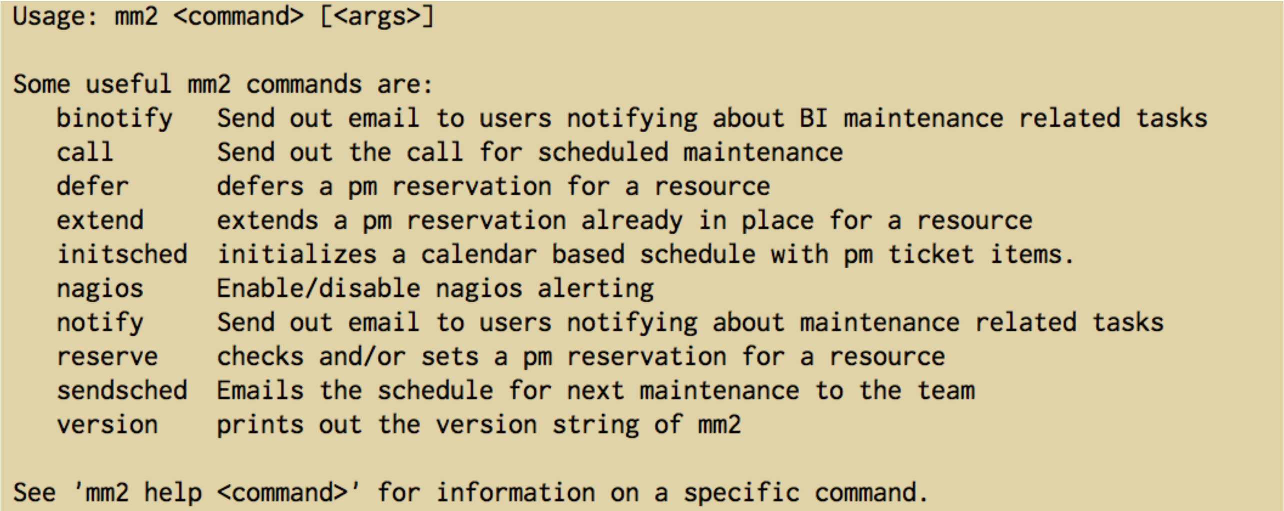
- We account for every core-second on the primary production machines
- We track the fate of every job launched on the system and classify the cause of interrupt if it does not end successfully
- Once a week, at least one member representing the major components of the machine (BG, storage, networking, infrastructure, etc.) meets to validate the previous weeks data. Most can be done automatically, but some require scrubbing logs
- All results are stored in a database and we use this information to drive where we focus our improvement efforts

Job Failure Analysis Process

- How we track availability, outages, and failure categories
- We do weekly Job Failure Analysis (JFA)
  - We do root cause analysis on every job that ran the previous week; Considered system error unless we can find explicit proof / “probable cause” it is user error.
  - On Wednesday afternoon, the Ops team gets in a room and walks through anything that wasn’t pre-classified
  - Produces Job MTTI, which internally is what we track. Also categorizes failures, which drives improvement projects.
- We account for every core-second on the machine
  - Up, scheduled down, unscheduled down; utilized or idle
  - Integers make reconciliation easy
- This software is very specific to us, but maybe someday (more on that later)
- We try to have a command for everything that also logs relevant data for later reporting
  - [begin|end]-service-action; maintman for maintenance; Scheduler reservations.
  - The Blue Gene comes with this built-in; porting to the Cray is going to be a challenge

Maintenance Manager – maintman or mm2

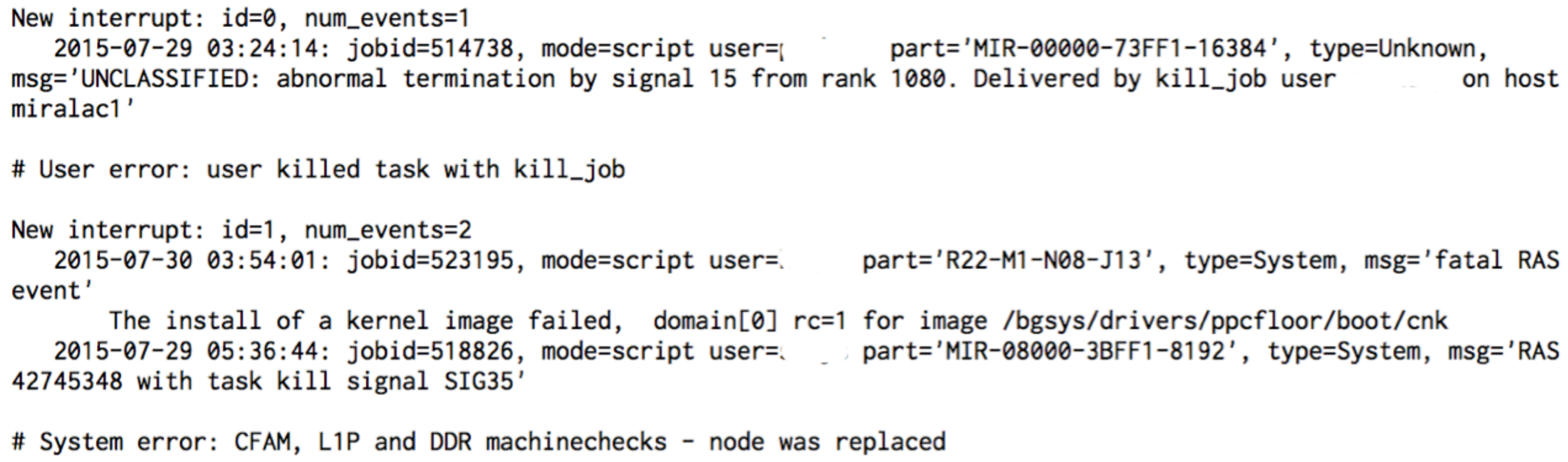
- Script that automates our maintenance processes
- In our opinion, a very nice tool
- For this discussion, what is apropos is that it writes records into our database and modifies scheduler reservations that are part of the availability calculation



```
Usage: mm2 <command> [<args>]

Some useful mm2 commands are:
   binotify   Send out email to users notifying about BI maintenance related tasks
   call       Send out the call for scheduled maintenance
   defer      defers a pm reservation for a resource
   extend     extends a pm reservation already in place for a resource
   initsched  initializes a calendar based schedule with pm ticket items.
   nagios     Enable/disable nagios alerting
   notify     Send out email to users notifying about maintenance related tasks
   reserve    checks and/or sets a pm reservation for a resource
   sendsched  Emails the schedule for next maintenance to the team
   version    prints out the version string of mm2

See 'mm2 help <command>' for information on a specific command.
```

The pre-classification script

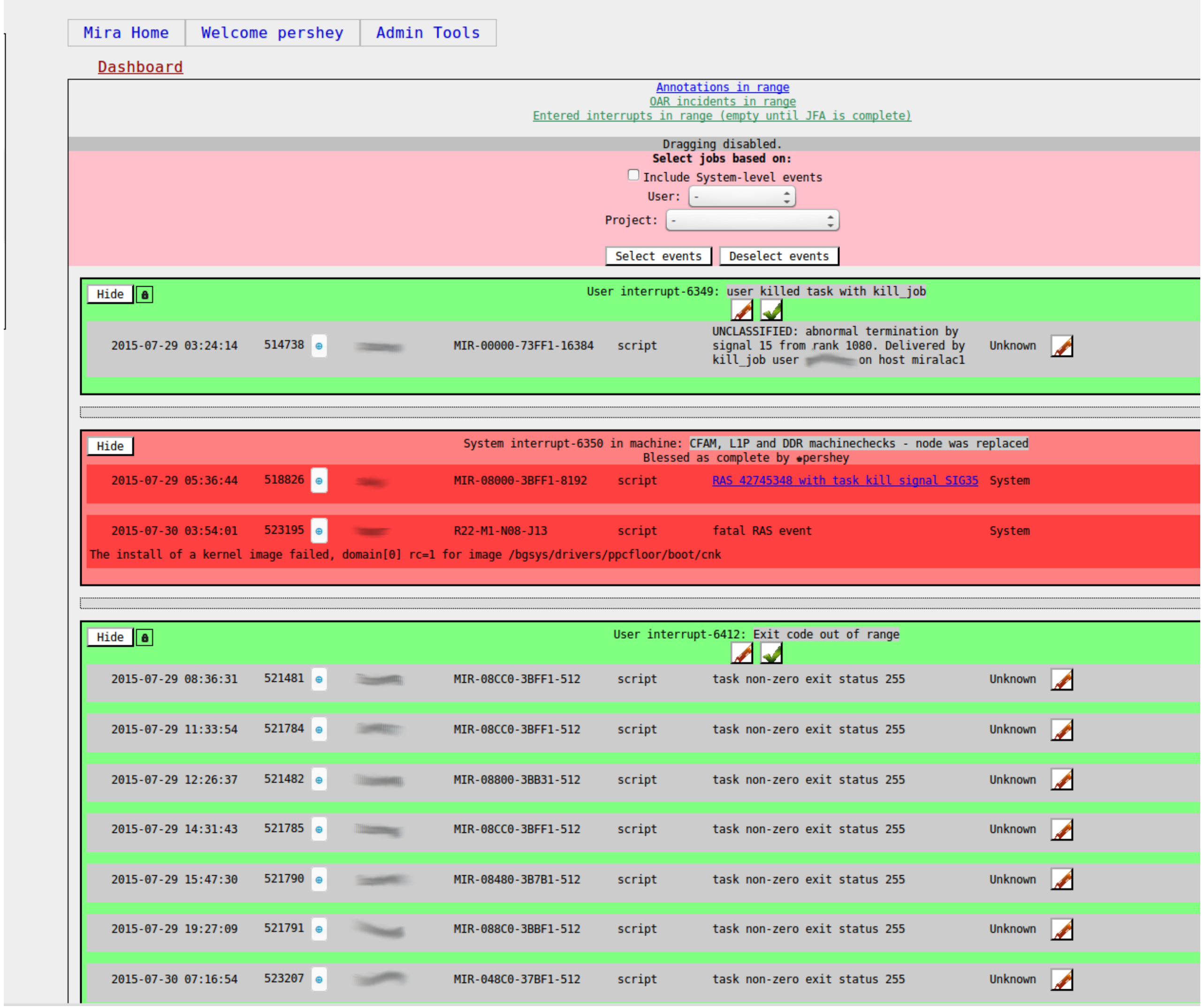
- The script is run daily, and loads interrupts to be analyzed.
- Staff can choose to do analysis / data entry ahead of the Wed meeting
- Below is an example of the output of what the script produces.
- This email is post-JFA, so it includes the resulting analysis (the comments)



```
New interrupt: id=0, num_events=1
   2015-07-29 03:24:14: jobid=514738, mode=script user=[redacted] part='MIR-00000-73FF1-16384', type=Unknown,
msg='UNCLASSIFIED: abnormal termination by signal 15 from rank 1080. Delivered by kill_job user [redacted] on host
miralac1'

# User error: user killed task with kill_job

New interrupt: id=1, num_events=2
   2015-07-30 03:54:01: jobid=523195, mode=script user=[redacted] part='R22-M1-N08-J13', type=System, msg='fatal RAS
event'
       The install of a kernel image failed,  domain[0] rc=1 for image /bgsys/drivers/ppcfloor/boot/cnk
   2015-07-29 05:36:44: jobid=518826, mode=script user=[redacted] part='MIR-08000-3BFF1-8192', type=System, msg='RAS
42745348 with task kill signal SIG35'

# System error: CFAM, L1P and DDR machinechecks - node was replaced
```

JFA Web App

- This shows all the records the pre-classification couldn't automatically identify
- Each person in the room runs this (as well as it being projected) and they can select an event that they will analyze



Dragging disabled.

Select jobs based on:

□ Include System-level events

User: -

Project: -

User interrupt-6349: user killed task with kill_job

2015-07-29 03:24:14	514738	[redacted]	MIR-00000-73FF1-16384	script	UNCLASSIFIED: abnormal termination by signal 15 from rank 1080. Delivered by kill_job user [redacted] on host miralac1	Unknown

System interrupt-6350 in machine: CFAM, L1P and DDR machinechecks - node was replaced

Blessed as complete by •pershey

2015-07-29 05:36:44	518826	[redacted]	MIR-08000-3BFF1-8192	script	RAS 42745348 with task kill signal SIG35	System
2015-07-30 03:54:01	523195	[redacted]	R22-M1-N08-J13	script	fatal RAS event	System

The install of a kernel image failed, domain[0] rc=1 for image /bgsys/drivers/ppcfloor/boot/cnk

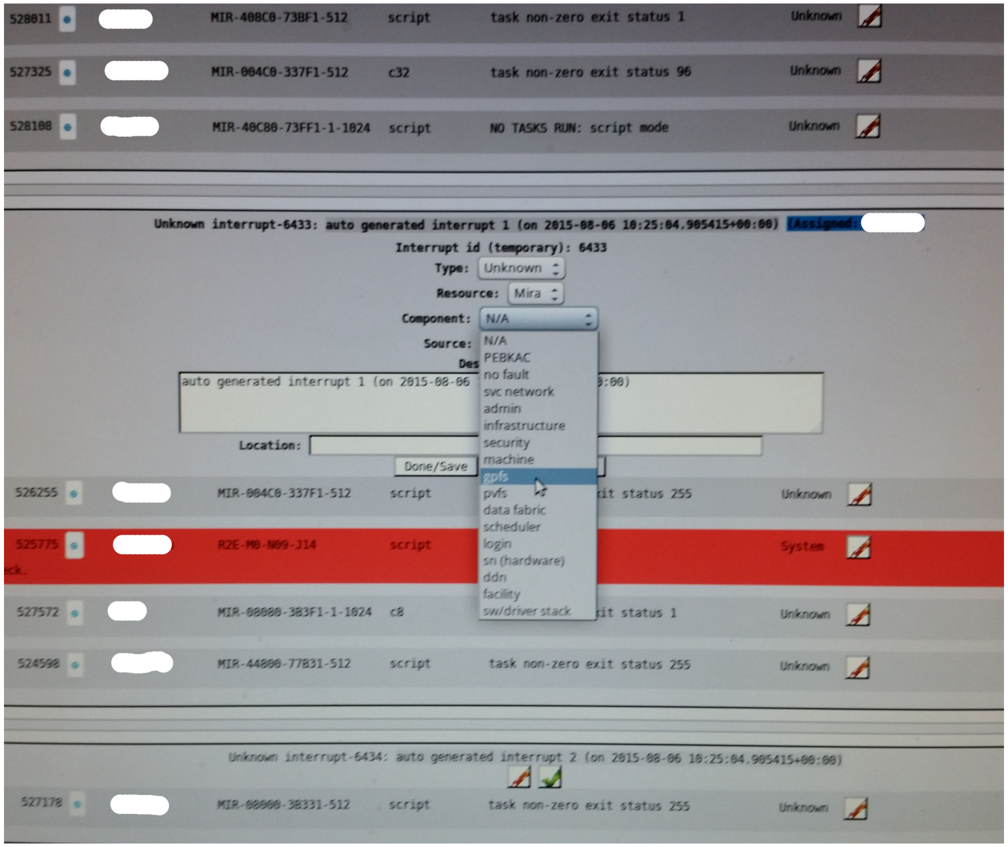
User interrupt-6412: Exit code out of range

2015-07-29 08:36:31	521481	[redacted]	MIR-08CC0-3BFF1-512	script	task non-zero exit status 255	Unknown
2015-07-29 11:33:54	521784	[redacted]	MIR-08CC0-3BFF1-512	script	task non-zero exit status 255	Unknown
2015-07-29 12:26:37	521482	[redacted]	MIR-08800-3BB31-512	script	task non-zero exit status 255	Unknown
2015-07-29 14:31:43	521785	[redacted]	MIR-08CC0-3BFF1-512	script	task non-zero exit status 255	Unknown
2015-07-29 15:47:30	521790	[redacted]	MIR-08480-3B7B1-512	script	task non-zero exit status 255	Unknown
2015-07-29 19:27:09	521791	[redacted]	MIR-088C0-3BBF1-512	script	task non-zero exit status 255	Unknown
2015-07-30 07:16:54	523207	[redacted]	MIR-048C0-37BF1-512	script	task non-zero exit status 255	Unknown

“Component” Analysis

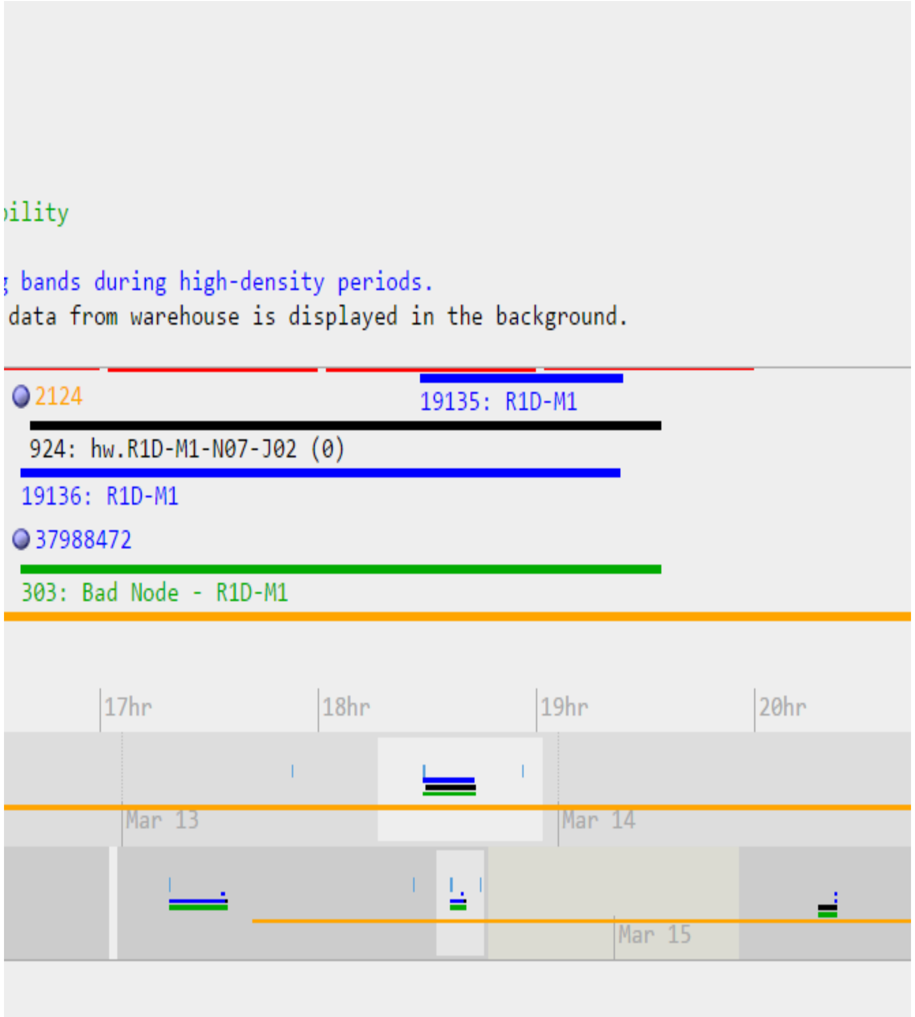
- We also classify by “component”
- This allows us to see what is giving us problems and drives improvement projects.
- One of the first real wins: We discovered that GPFS was 3x the next source of failures. We investigated and discovered we were getting timeouts and moving the management functions to dedicated nodes dropped GPFS down into the noise.

```
Component fault analysis (Recent range: prior 90 days,

Component            Count (Recent) Count (All)
-------------------- -------------- -----------
machine                          25         305
gpfs                             21          59
sw/driver stack                   3          33
ddn                               2           4
facility                          1           4
```



```
528011  [redacted]  MIR-408C0-73BF1-512      script  task non-zero exit status 1    Unknown
527325  [redacted]  MIR-004C0-337F1-512      c32     task non-zero exit status 96   Unknown
528108  [redacted]  MIR-40C80-73FF1-1-1024   script  NO TASKS RUN: script mode      Unknown

Unknown interrupt-6433: auto generated interrupt 1 (on 2015-08-06 10:25:04.905415+00:00) (Assigned: [redacted]
Interrupt id (temporary): 6433
Type: Unknown
Resource: Mira
Component: N/A
Source:
Des
auto generated interrupt 1 (on 2015-08-06                    :00)
Location:
Done/Save

526255  [redacted]  MIR-004C0-337F1-512      script  it status 255                  Unknown
525775  [redacted]  R2E-M0-N09-J14           script                                 System
ck.
527572  [redacted]  MIR-00000-3B3F1-1-1024   c8      it status 1                    Unknown
524598  [redacted]  MIR-44800-77B31-512      script  task non-zero exit status 255  Unknown

Unknown interrupt-6434: auto generated interrupt 2 (on 2015-08-06 10:25:04.905415+00:00)
527178  [redacted]  MIR-00000-3B331-512      script  task non-zero exit status 255  Unknown
```

Incident Timeline

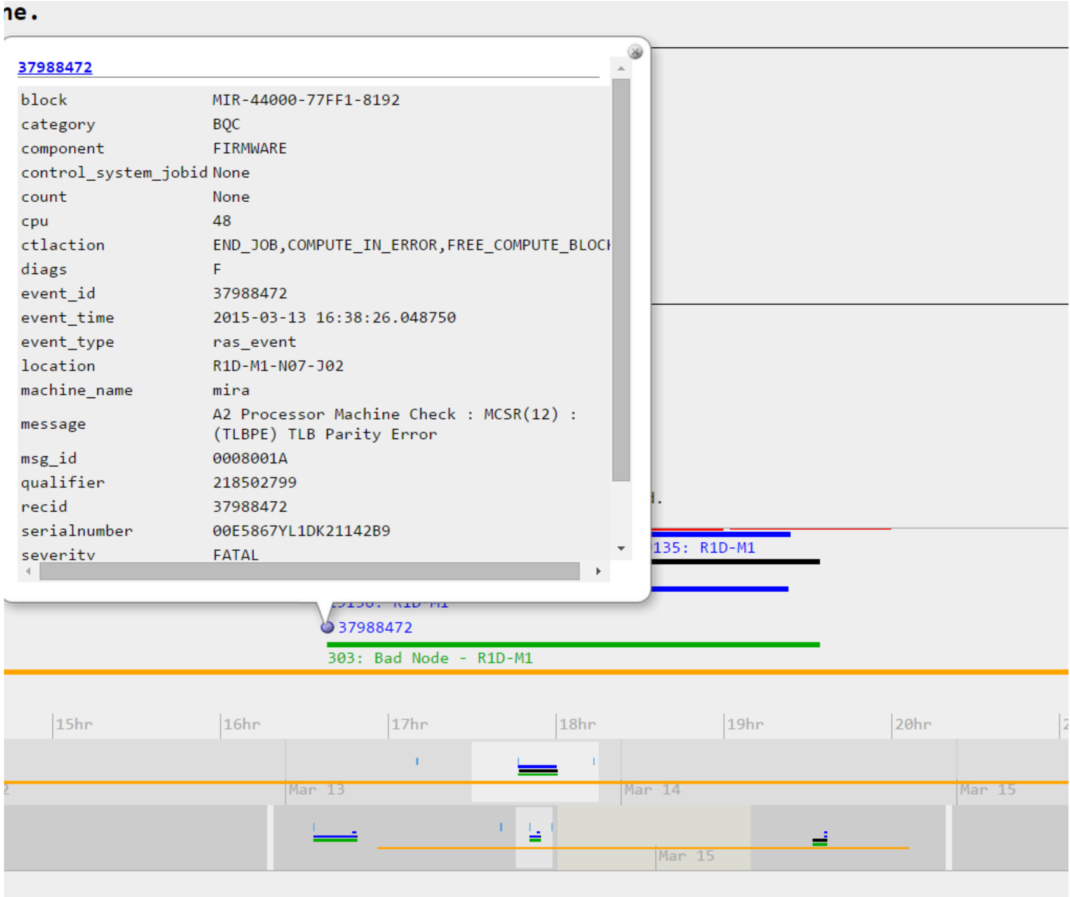


ility

bands during high-density periods.
data from warehouse is displayed in the background.

- Trying to figure out when an incident began and ended is non-trivial.
- This shows all the sources of data about a given incident.
- We take the "union" of all the events to determine the duration of the incident.

Incident Timeline – drill down



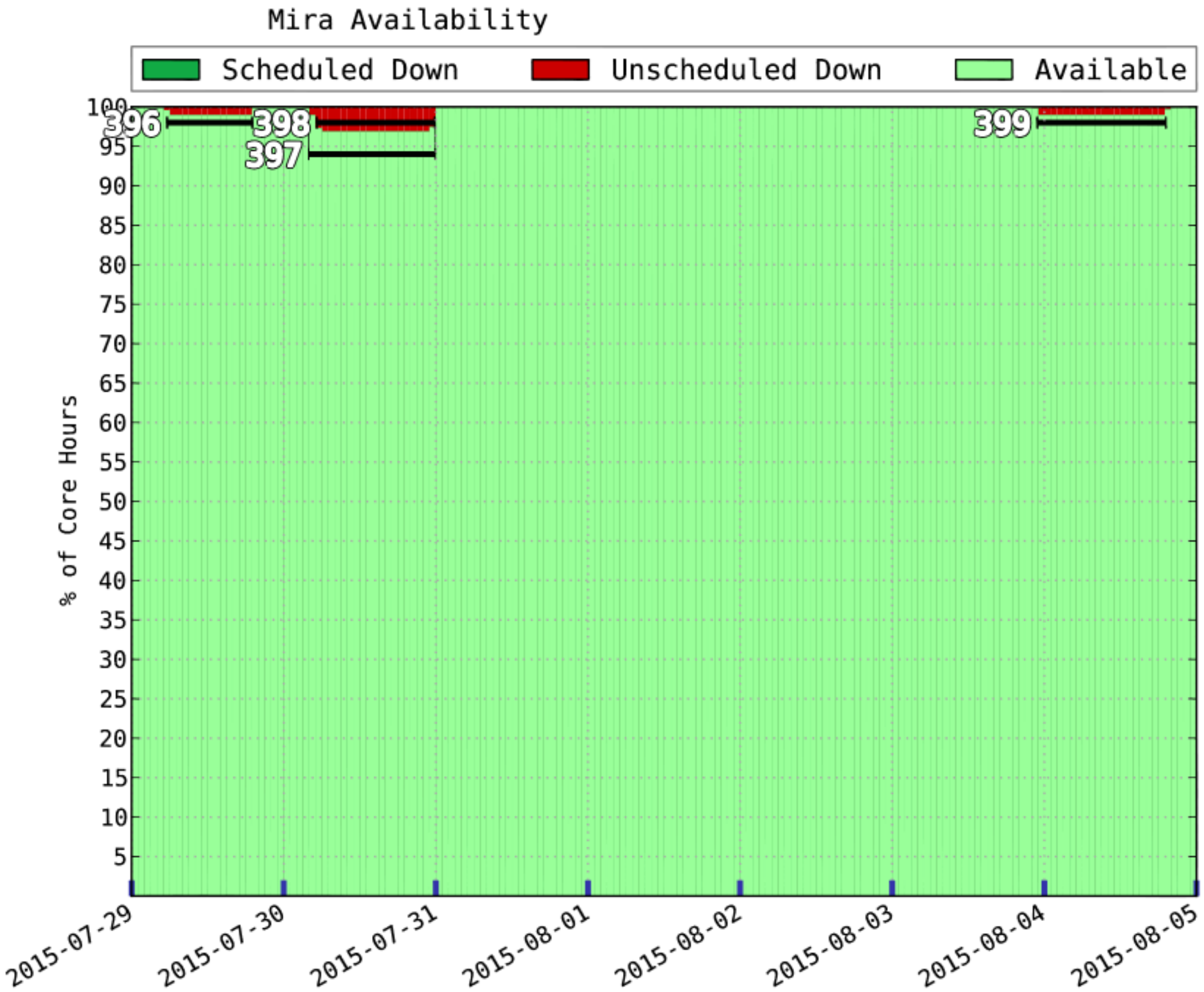
[37988472]

block	MIR-44000-77FF1-8192
category	BQC
component	FIRMWARE
control_system_jobid	None
count	None
cpu	48
ctlaction	END_JOB,COMPUTE_IN_ERROR,FREE_COMPUTE_BLOCK
diags	F
event_id	37988472
event_time	2015-03-13 16:38:26.048750
event_type	ras_event
location	R1D-M1-N07-J02
machine_name	mira
message	A2 Processor Machine Check : MCSR(12) : (TLBPE) TLB Parity Error
msg_id	0008001A
qualifier	218502799
recid	37988472
serialnumber	00E5867YL1DK21142B9
severity	FATAL

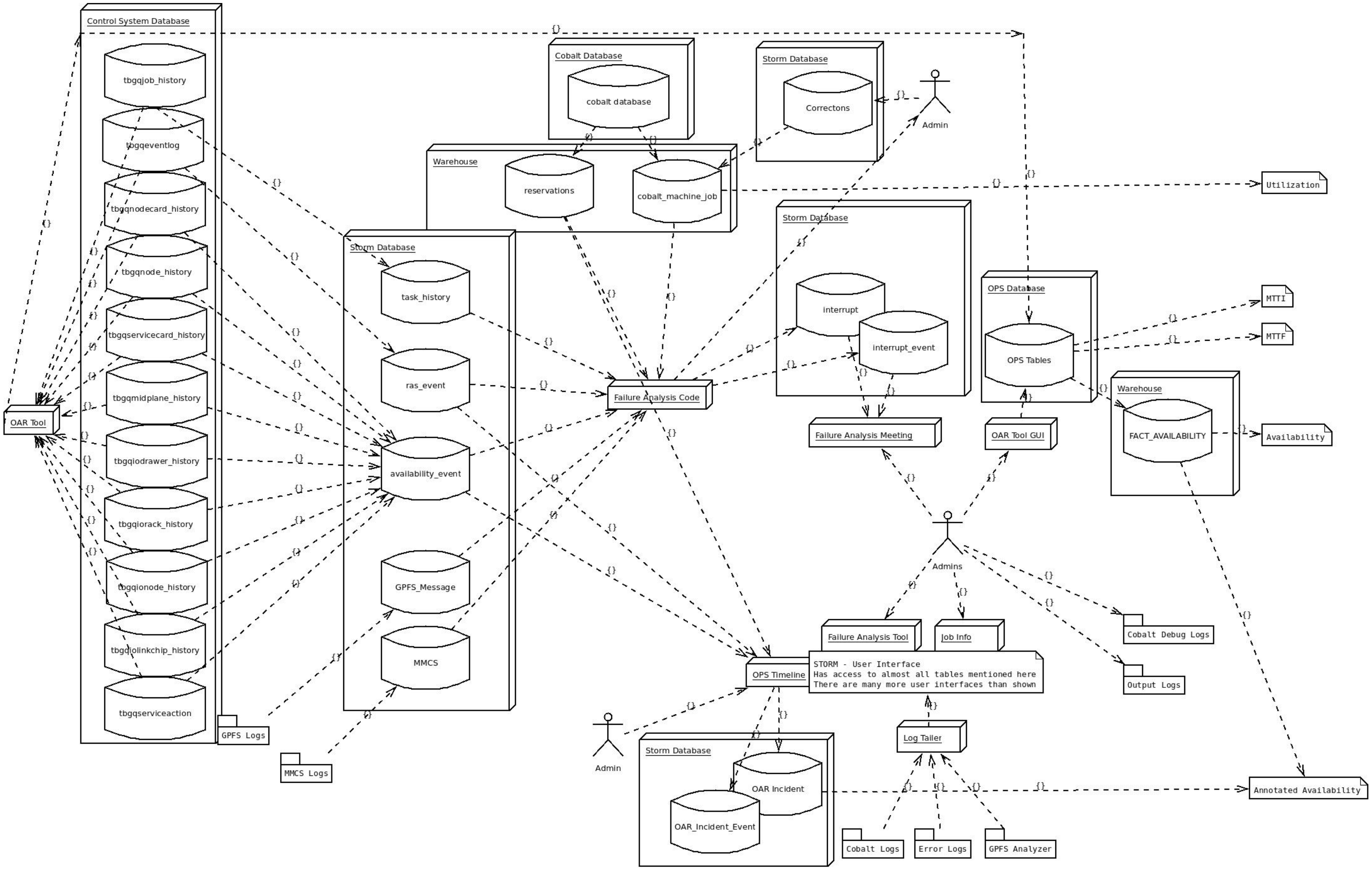
- From this screen you can drill down into the details for any entry.

Our standard availability report

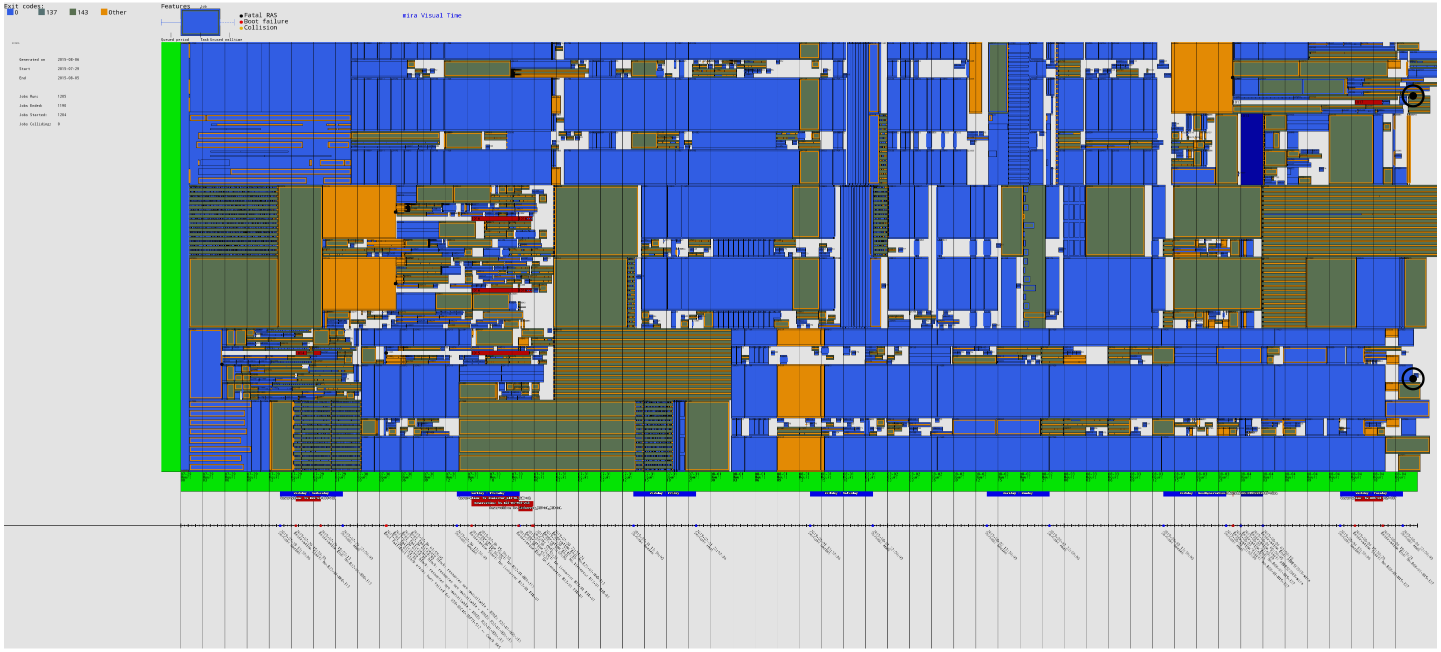
```
Generated on 2015-08-06 13:01:08

Scheduled       99.435652%
Overall         99.435652%
Start 2015-07-29
End   2015-08-04

Incidents:

  ID  Name
 396. Bad Node - R22-M1
 397. Bad Node - R22-M1
 398. Link Errors - R13-M1 to R1B-M1
 399. Bad Node - R06-M1-N15-J27
```



The (complicated) Big Picture…



Machine Time Overlay…



- Y axis are the allocable chunks of the machine (mid-planes here, nodes on the vis cluster)
- X axis is time

- Easy to see scheduling; Also helps find bugs, like two jobs running on the same resource at one time.
- There is a lot of information encoded here
- This is general; Any information you can provide that is (data, location, time) can be displayed this way; We also use this for coolant temperature, power consumption, etc..

CORAL– Collaboration of ORNL, Argonne, LLNL

- Provide the Leadership computing capabilities needed for the DOE Office of Science mission from 2018 through 2022
  - Capabilities for INCITE and ALCC science projects

- CORAL was formed by grouping the three Labs who would be acquiring Leadership computers in the same timeframe (2017-2018), benefits include:
  - Shared technical expertise
  - Decreases risks due to the broader experiences, and broader range of expertise of the collaboration
  - Lower collective cost for developing and responding to RFP

CORAL Overview

Current DOE Leadership Computers

Mira (ANL)
2012 - 2017

Sequoia (LLNL)
2012 - 2017

Titan (ORNL)
2012 - 2017

Objective - Procure 3 leadership computers to be sited at Argonne, ORNL, and LLNL in CY17-18.

Leadership Computers RFP requests >100 PF, 2 GB/core main memory, local NVRAM, and science performance 4x-8x Titan or Sequoia

Approach

Competitive process – 1 RFP (issued by LLNL) leading to 2 R&D contracts and 3 computer procurement contracts

For risk reduction and to meet a broad set of requirements, 2 architectural paths will be selected – and Argonne and ORNL must choose different architectures

Once selected, multi-year lab-awardee relationship to co-design computers

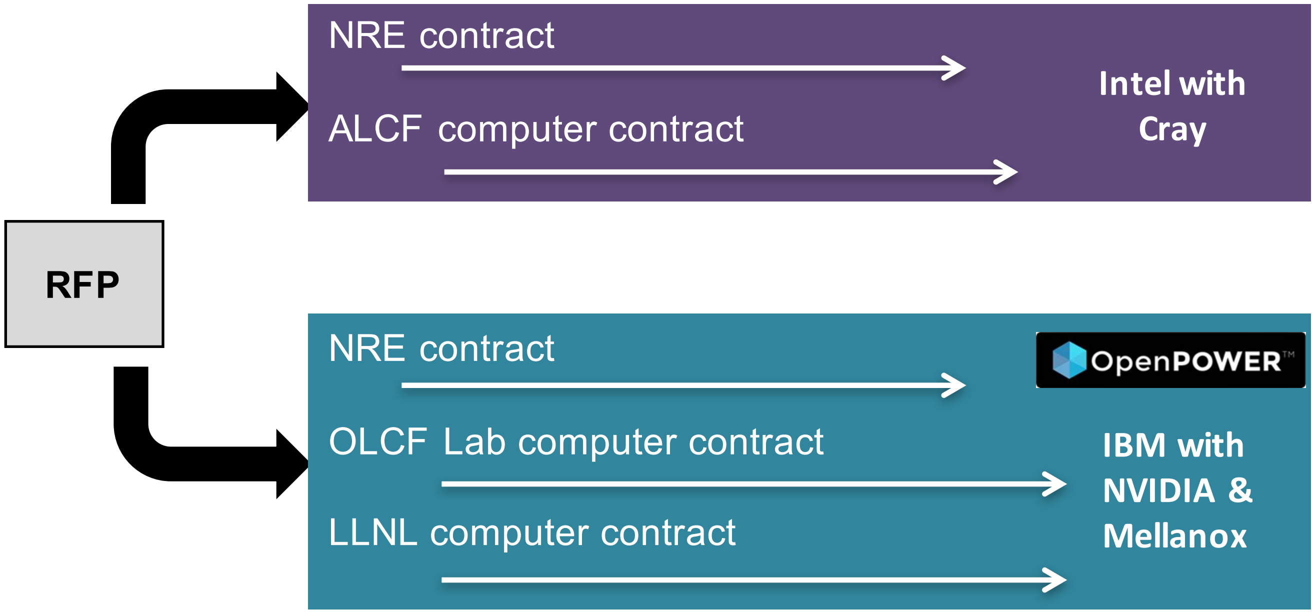
Both R&D contracts jointly managed by the 3 Labs

Each lab manages and negotiates its own computer procurement contract, and may exercise options to meet their specific needs

Understanding that long procurement lead time may impact architectural characteristics and designs of procured computers

Results of CORAL Procurement

Two Diverse Architecture Paths



RFP

NRE contract

ALCF computer contract

Intel with Cray

NRE contract

OLCF Lab computer contract

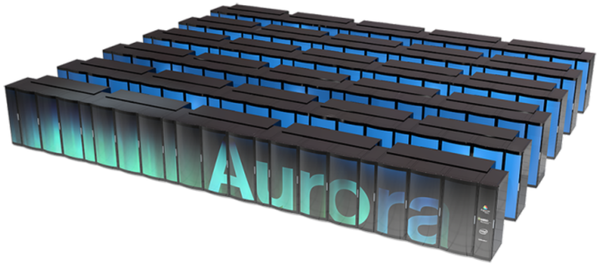
LLNL computer contract

OpenPOWER™

IBM with NVIDIA & Mellanox

2018 ALCF Leadership System

Many Core architecture



System Name: Aurora

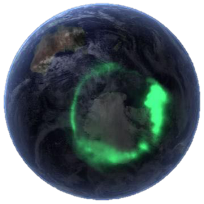
Vendor: Intel (Prime) / Cray (Integrator)

Delivery date: 2018

- Over 13X Mira's application performance
- Over 180 PF peak performance
- More than 50,000 nodes with 3rd Generation Intel® Xeon Phi™ processor
  - Code name Knights Hill, > 60 cores
- Over 7 PB total system memory
  - High Bandwidth On-Package Memory, Local Memory, and Persistent Memory
- 2nd Generation Intel® Omni-Path Architecture with silicon photonics in a dragonfly topology
- More than 150 PB Lustre file system capacity with > 1 TB/s I/O performance

2016 ALCF Theta System

Many Core architecture



Vendor: Intel (Prime) / Cray (Integrator)

- Transition and data analytics system
- Over 8.5 PF peak performance
- More than 2,500 nodes with 2nd Generation Intel® Xeon Phi™ processor
  - Code name Knights Landing, > 60 cores
- 192GB DDR4 memory and up to 16GB HBM on each node
- 128GB SSD on each node
- Cray Aries high speed interconnect in dragonfly topology
- Initial file system: 10PB Lustre file system, 200 GB/s throughput
- Cray XC system
- Cray software stack
- ~1.7 MW peak power

Systems feature summary

System Feature	Mira (2012)	Theta (2016)	Aurora (2018)
Peak Performance	10 PF	> 8.5 PF	180 PF
Number of Nodes	49,152	> 2,500	> 50,000
Aggregate HBM, local memory, and persistent mem	786 TB	> 480 TB	> 7 PB
File system capacity	26 PB	10 PB (initial)	> 150 PB
File system throughput	300 GB/s	210 GB/s (initial)	> 1 TB/s
Peak Power Consumption	4.8 MW	1.7 MW	13 MW
GFLOPS/watt	2.1	> 5	> 13
Facility Area	1,536 sq. ft.	~1,000 sq. ft.	~3,000 sq. ft.

What changes, what doesn’t

- Same
  - many core
  - GPFS
  - MPI+OpenMP
  - Cobalt scheduler

- Different
  - Network
  - RAS
  - Lustre
  - System software
  - Cray Programming Environment
  - On package, HBM memory
  - Intel x86, not powerpc
  - SSDs

Future/Opportunities

- We WILL port
  - Cobalt
  - Monitoring
  - JFA process
- Much of our reporting depends on it
- We are looking at abstractions and architectural improvements that will make this easier to use on general machines
- We have also been leaning on Cray and Intel to work with us to develop standardized interfaces and mechanisms for obtaining this kind of data
- Opportunities to share/codevelop tools

Other opportunities

- Evaluate each other's Petascale computers and software stacks
  - Scaling studies, tools, libraries, compilers
- Modeling and simulation of applications
- Community codes
- Visualization support: software, techniques
- Industry engagements

Acknowledgment

- This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Acknowledgements

- The entire ALCF team

Questions?

Mark Fahey <mfahey@anl.gov>