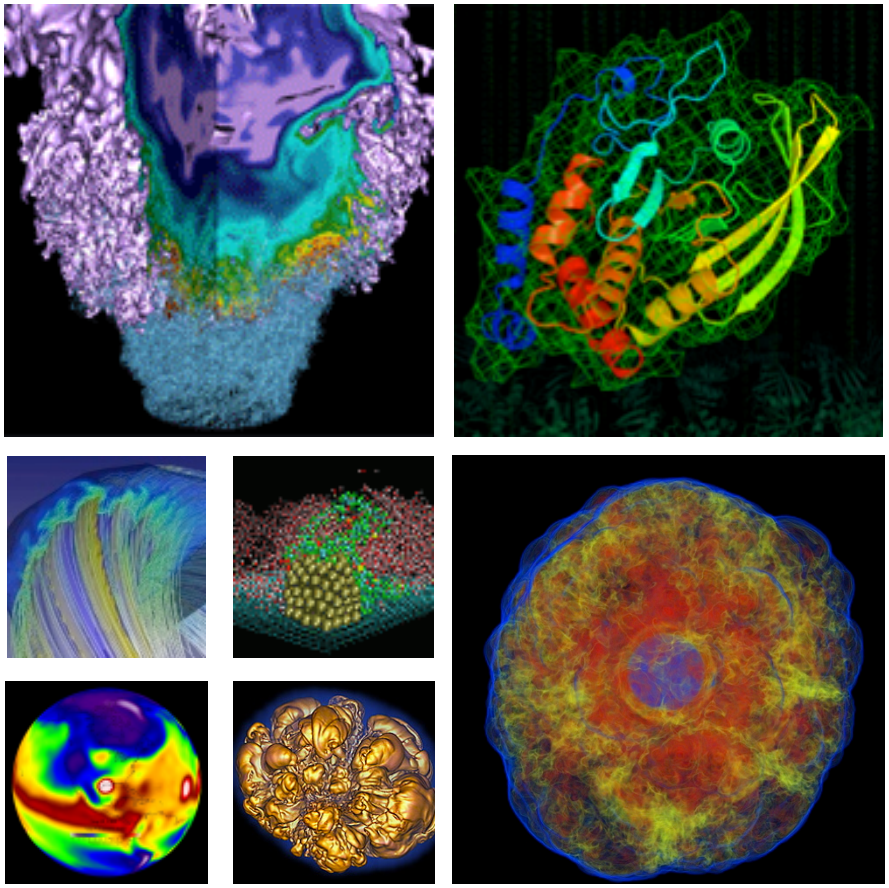
I/O Performance on Cray XC30



Zhengji Zhao[1], Doug Petesch[2], David Knaak[2], and Tina Declerck[1]

[1] NERSC
[2] Cray, Inc

Cray User Group Meting
May 7, 2014

NERSC 40 YEARS at the FOREFRONT

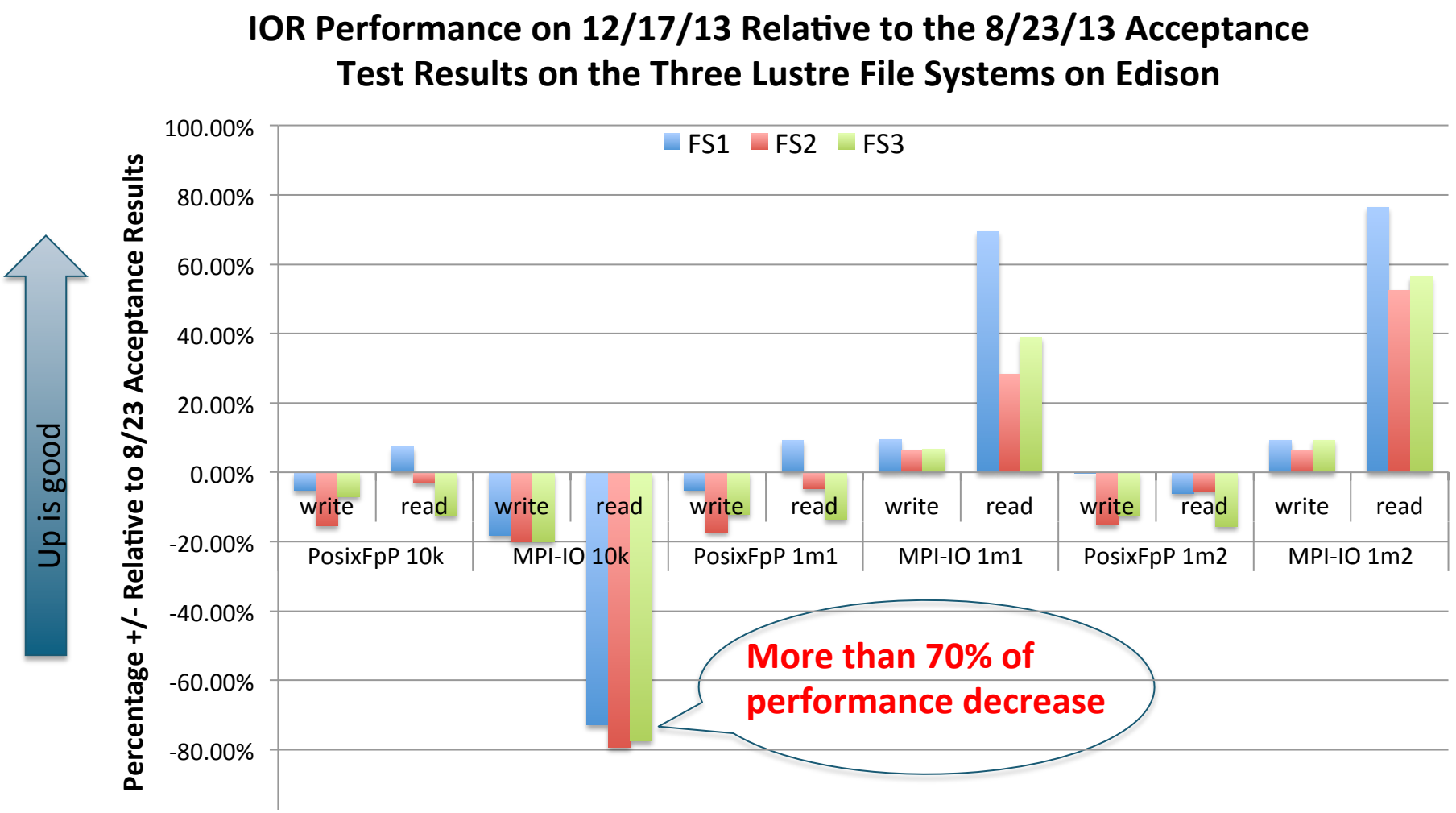
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Acknowledgement

- Mark Swan at Cray for the LMT data extractions
- Steve Luzmoor, Patrick Farrell at Cray who helped resolving the bug 809189.
- Marcus Petschlies, a NERSC user, for providing IOBUF test data with a QLUA code.
- Harvey Wasserman at NERSC for valuable discussion and help
- Shane Canon at NERSC, for providing Edison file system usage figures.
- Nathan Wichmann at Cray for doing the Edison acceptance tests.
- Jeff Broughton, NERSC-7 project manager, for his support including granting the dedicated system time for this investigation.
- Cray onsite and NERSC system staff for their support to use the system in dedicated mode

Motivation

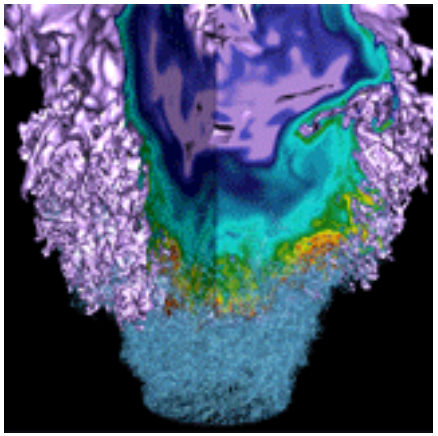


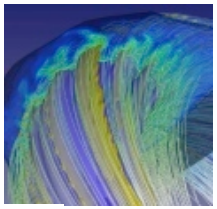
About 50% of all I/Os on Hopper, NERSC's large Cray XE system, were unaligned, and/or small I/Os with transfer sizes that are much smaller than the Lustre block size.

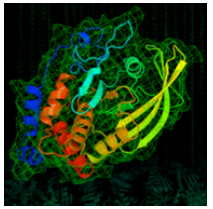
Agenda

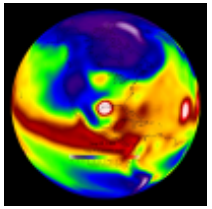
- Edison and Lustre file system overview
- Benchmark codes and tests
- I/O performance at acceptance tests
- I/O performance change over time
- I/O performance monitoring in production environment
- Summary

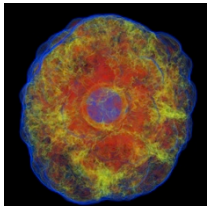
Edison and Lustre File System Overview

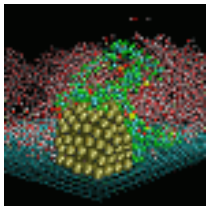












NERSC 40 YEARS at the FOREFRONT

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Edison, a Cray XC30, is the Newest Supercomputer at NERSC

- First Cray XC30
- Peak Flops (PF) 2.57
- Compute Nodes 5,576
- CPU Cores (*Total / Per-node*) 133,824/ 24
- Intel Ivy Bridge 12-core, 2.4GHz processors
- Memory (TB) (*Total / Per-node*) 357 / 64
- Memory (Stream) BW (TB/s) 498.4
- Memory BW/node* (GB/s) 89
- Aries interconnect with Dragonfly topology for great scalability
- Peak Bisection BW (TB/s) 23.7 TB/s
- File system(s) 7.56 PB @ 168 GB/s
- 3 Lustre file systems with Sonexion storage system, configured as 2:2:3 for capacity and bandwidth
- Access to NERSC's GPFS global file system via DVS
- 12 x 512GB login nodes to support visualization and analytics
- Ambient cooled for extreme energy efficiency
- Power (MW Linpack) 1.9

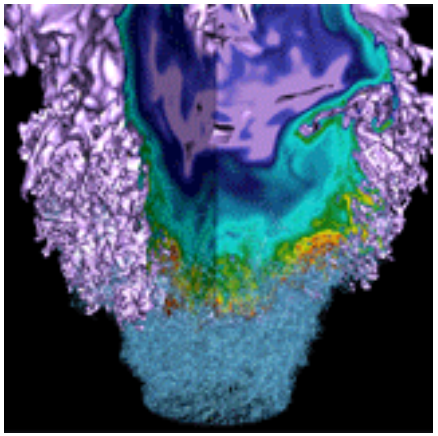
Lustre File Systems (Sonexion 1600)

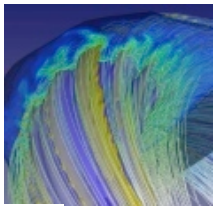
	Size (PB)	Agg. Peak I/O Bandwidth (GB/s)	No. 0f SSUs	No. of OSSs	No. of OSTs
FS1	2.1	48	12	24	96
FS2	2.1	48	12	24	96
FS3	3.2	72	18	36	144

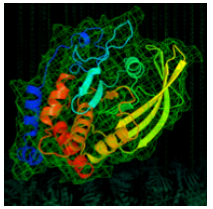
SSU Configuration:

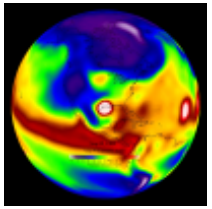
- Each SSU has 8 Lustre OSTs, 2 OSSs. Each OSS serves 4 OSTs.
- Each OST contains 8 data disks and 2 parity disks (dual-ported 3.5 inch 3TB NL-SAS 7,200 RPM disk drives) configured as a RAID 6 array
- Two dual-ported 3.5 inch 100GB SSDs drives, are configured as a shared RAID 1 array, partitioned and used for the MDRAID and the file system journals.
- Two spare 3TB NL-SAS disk drives

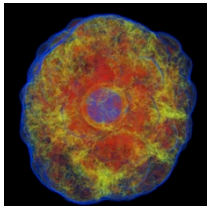
Benchmark Codes and Tests

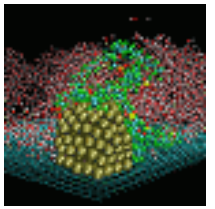












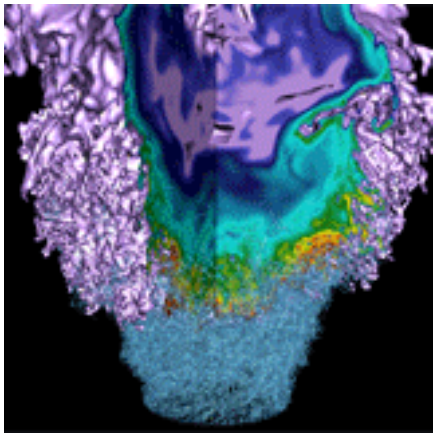
IOR

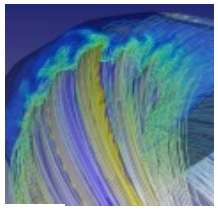
- **IOR**
  - http://www.nersc.gov/systems/nersc-8-procurement/trinity-nersc-8-rfp/nersc-8-trinity-benchmarks/ior/
  - Measures file system I/O performance at both Posix and MPI-IO levels
- **Instrumented IOR provided by Doug Petesch**
  - Reports bandwidth over time during a run
- **IOBUF library**
  - Cray provided I/O buffering library that can intercepts I/O system calls such as read and open and adds a layer of buffering, thus improving program performance by enabling asynchronous prefetching and caching of file data.
  - Used in the multiple IOR tests, especially in the MPI-IO 10k and Posix1m2 tests

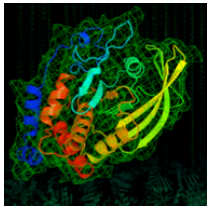
IOR Benchmark tests

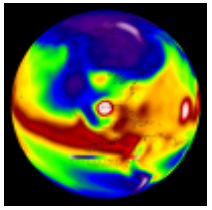
	Posix FpP 10k,1m1,1m2			MPI-IO			MPI-IO 1m1, 1m2		
	FS1	FS2	FS3	FS1	FS2	FS3	FS1	FS2	FS3
Cores used	768	768	1152	2304	2304	4608	2304	2304	4608
Nodes used	32	32	48	96	96	144	96	96	144
Aggr. File Size (TB)	3.1	3.1	4.6	9.2	9.2	13.8	9.2	9.2	13.8
No. of Files	768	768	1152	1			1		
IOBUF_PARAMS	count=2:size=32m:direct			count=1:size=1000000: prefetch=0			IOBUF was not used		
MPIIO Hints				cb_romio_read=disable cb_romio_write=disable			cb_romio_read=enable cb_romio_write=enable		
Lustre Striping	lfs setstripe -s 1m -c 1			lfs setstripe -s 1m -c -1			lfs setstripe -s 4m -c -1		

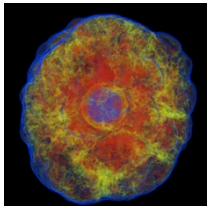
I/O Performance at Acceptance Tests (8/23/2013)

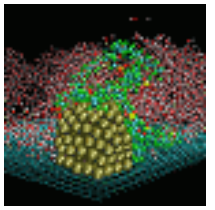






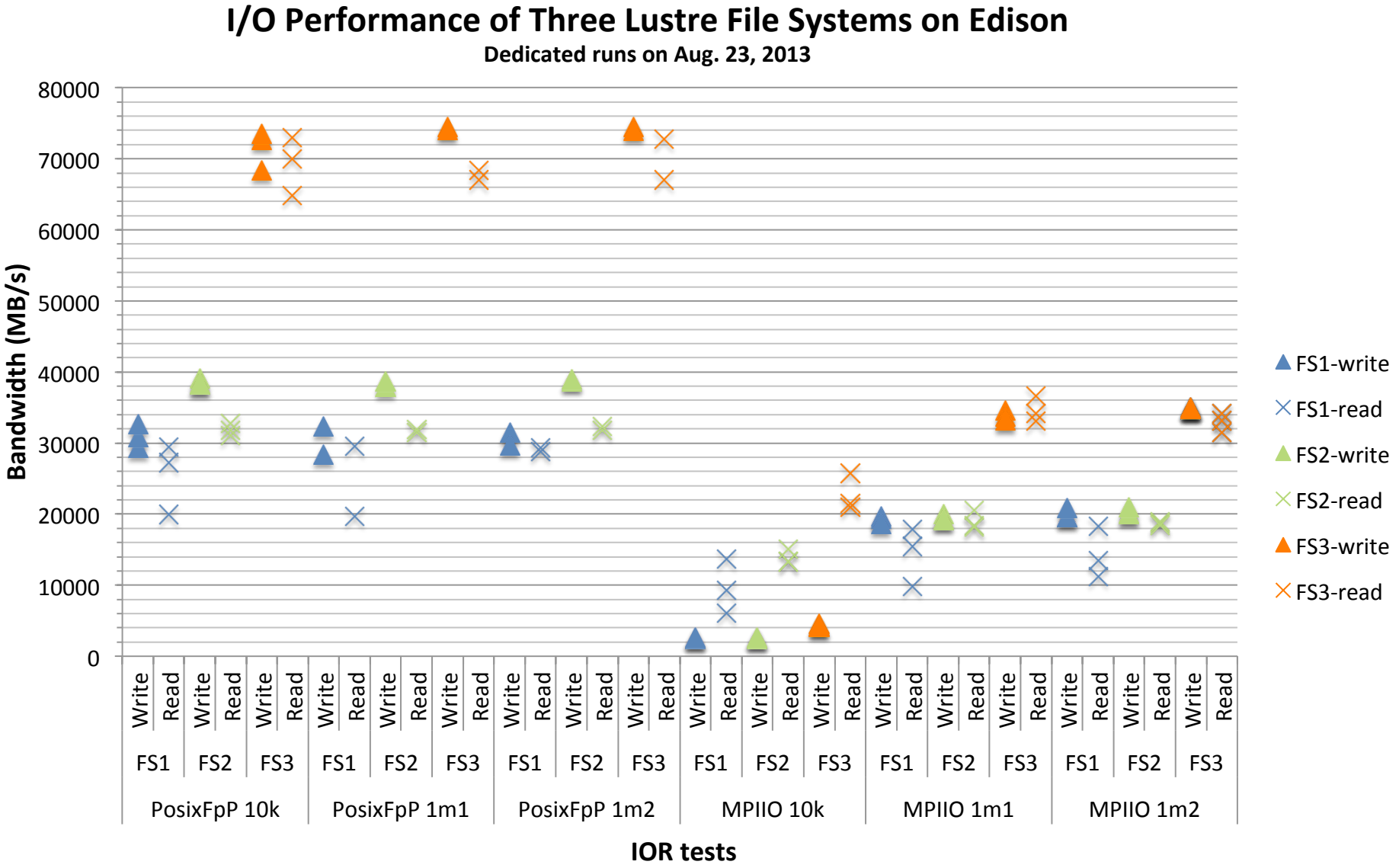




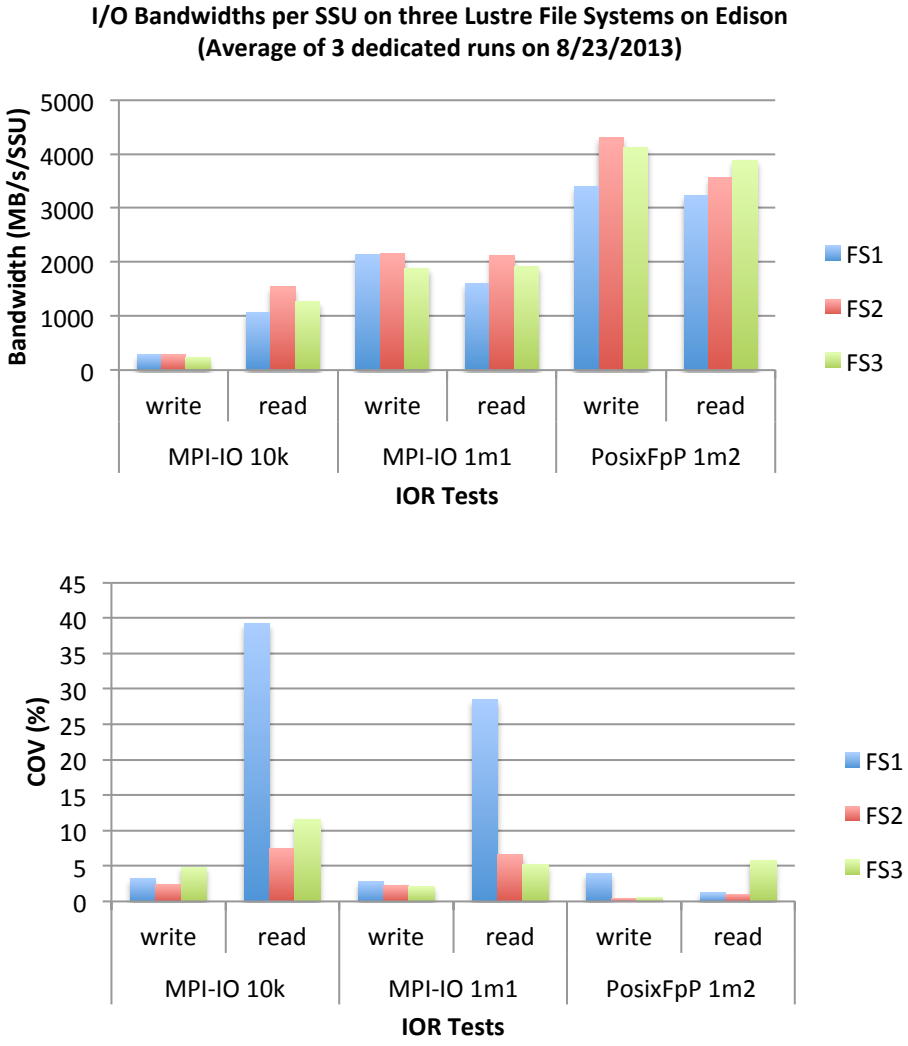


NERSC 40 YEARS at the FOREFRONT

I/O Acceptance Tests on Aug, 2013

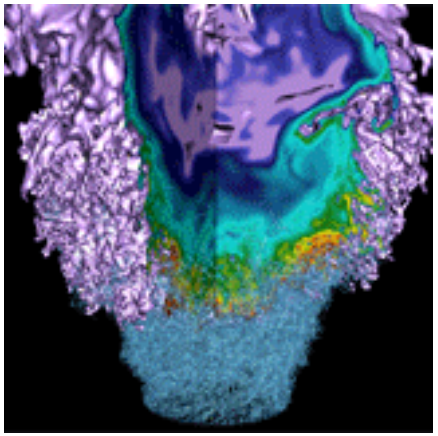


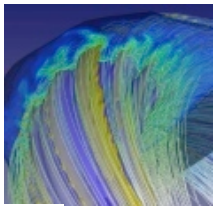
I/O Acceptance tests --continued

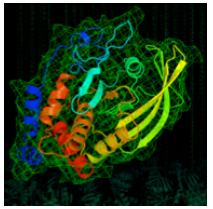


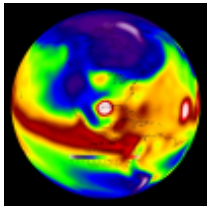
- Among the three file systems, FS2 and FS3 were almost clean (1% full); FS1 was 30% full.
- Max write/read rate per SSU is about 4GB/s. The performance scales almost linearly to 144 OSTs on the clean file systems.
- There was up to 40% performance variation on FS1 even with the dedicated runs; while on the other two clean file systems the variation was about 0-12%.
- The fragmentation and the physical position of files relative to the slower or faster end of the disk drive may contribute to the dedicated I/O performance variation.

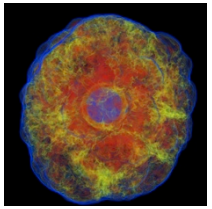
I/O Performance Change Over Time

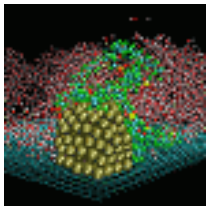












NERSC 40 YEARS at the FOREFRONT

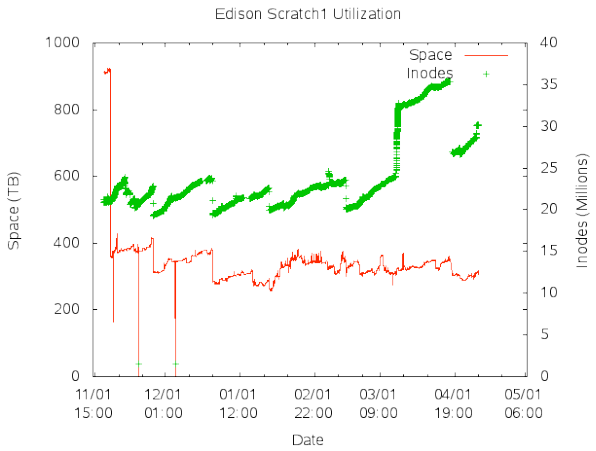
U.S. DEPARTMENT OF ENERGY | Office of Science

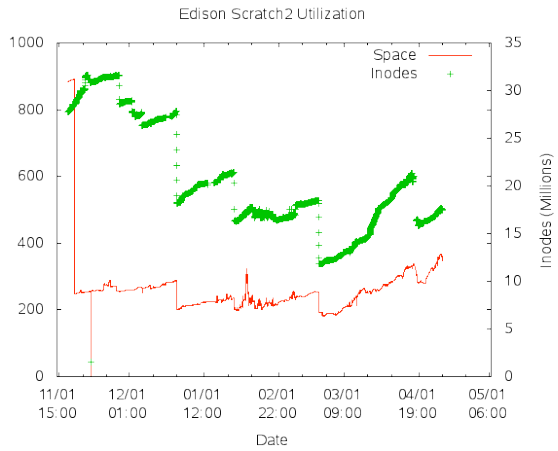
BERKELEY LAB
Lawrence Berkeley National Laboratory

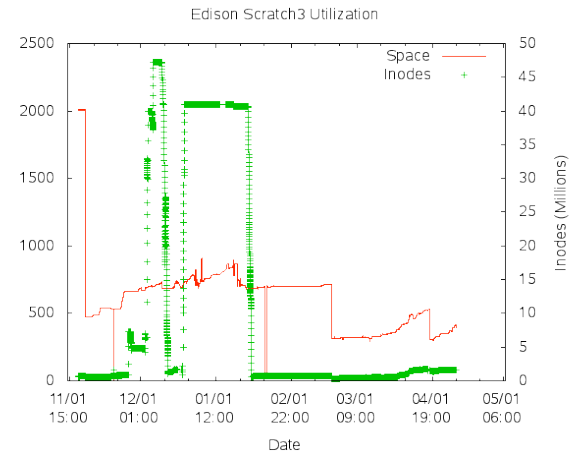
File System Hardware and Software Upgrades and File System Usage

	FS1	FS2	FS3	CLE/Lustre upgrades
Aug 1, 2013	72 OSTs	72 OSTs	144 OSTs	5.0.UP03/2.3.0
Dec 6, 2013				5.1.UP00/2.4.0
Dec 16,2013		96 OSTs		
Jan 17, 2014	96 OSTs			
Mar 11, 2014				5.1.UP01/2.4.1

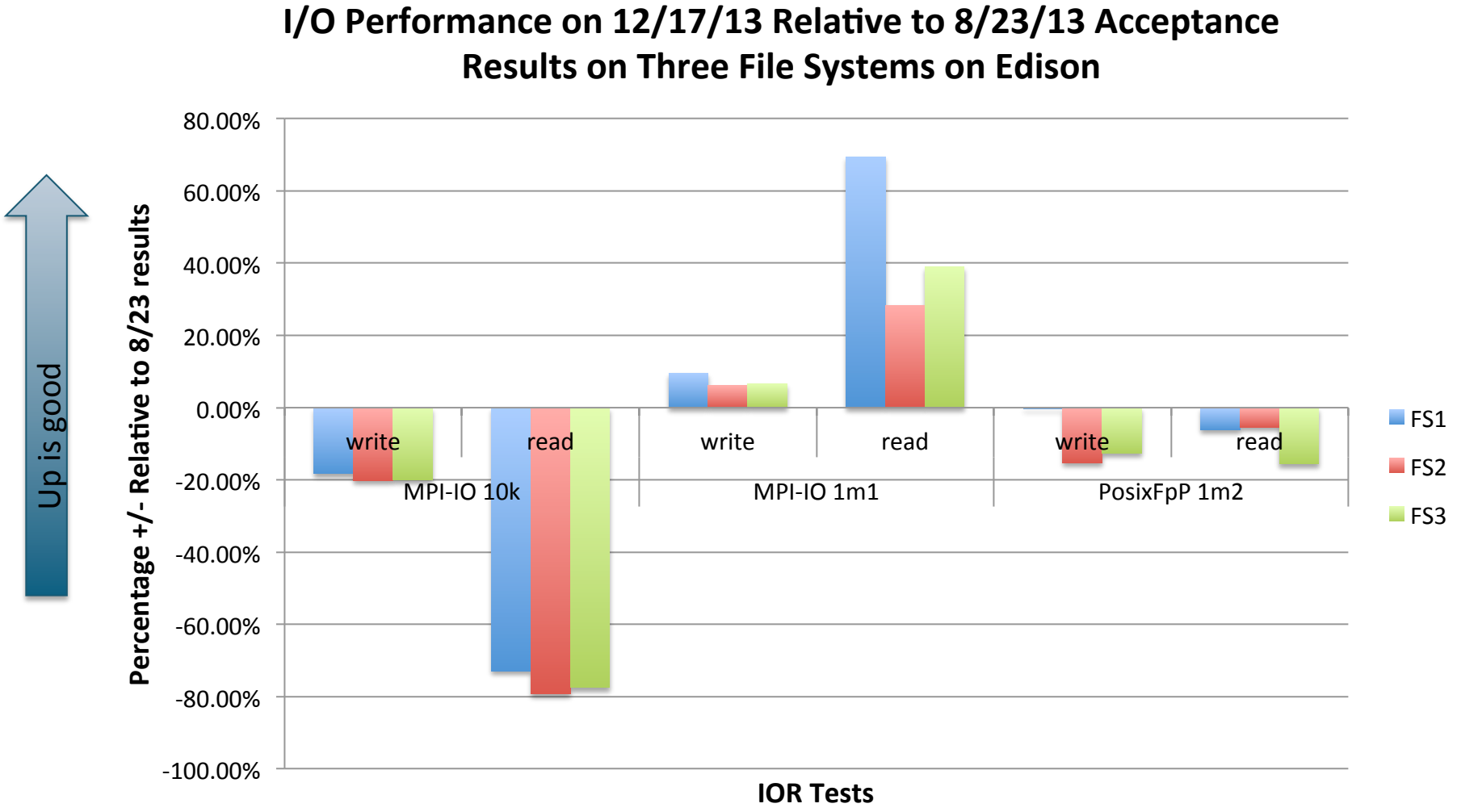
	July 10, 2013	Nov 27, 2013	Dec 16, 2013	Apr 24, 2014
CDT	1.06	1.10	1.11	1.15







MPI-IO 10k read rates decreased by more than 70% in Dec, 2013

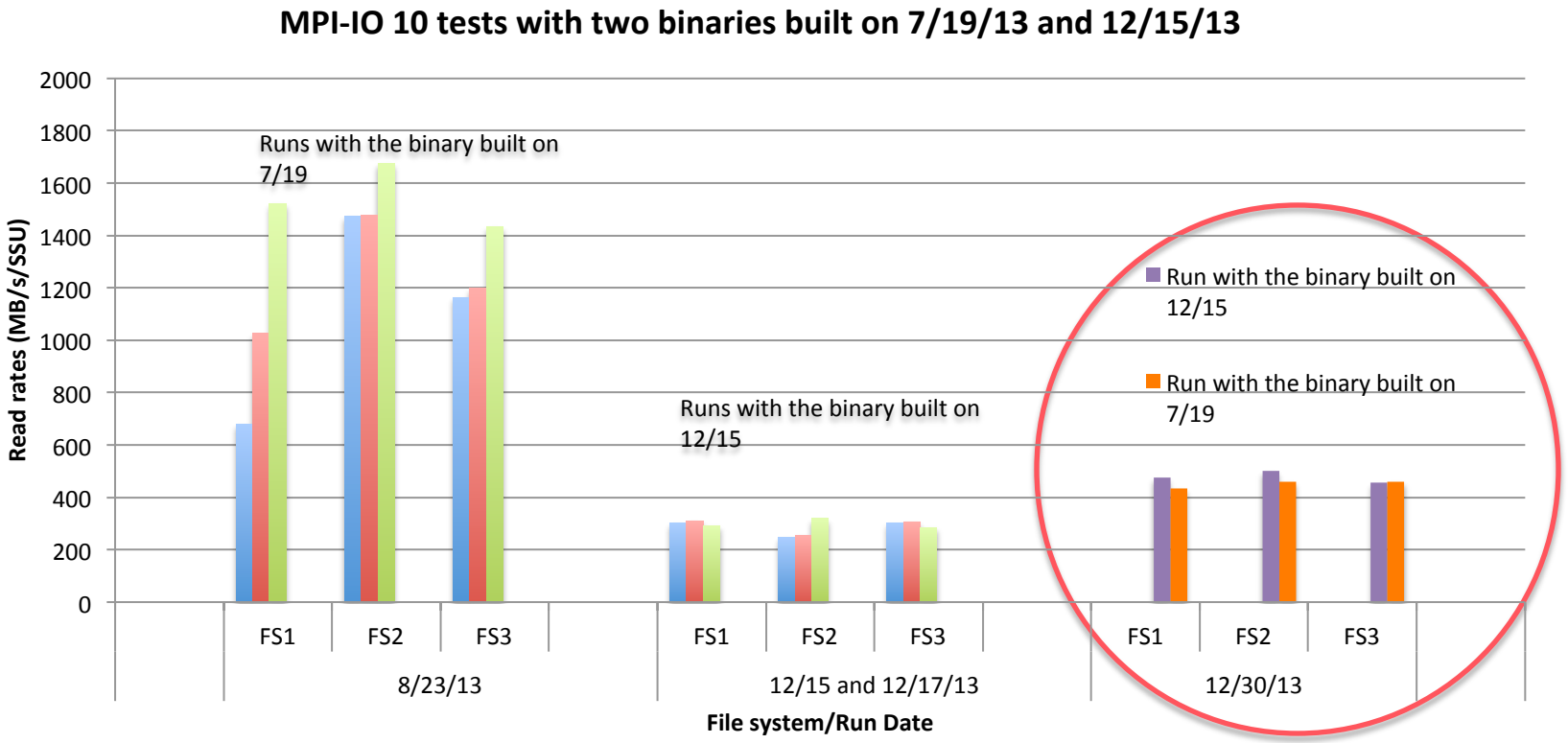


MPI-IO 10k Test

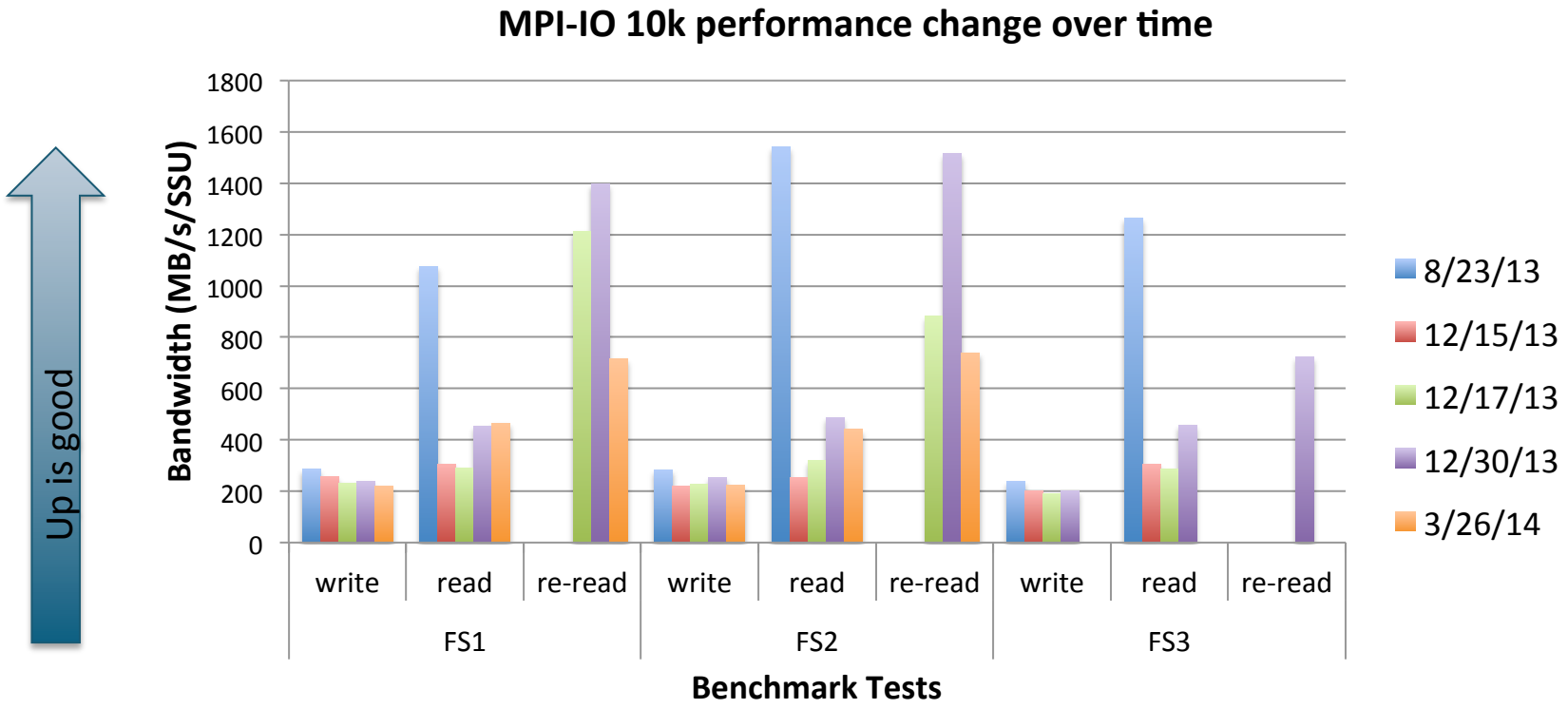
- **Worst case:** shared file, small records, large gaps, not-aligned.
  - Shared file causes file locking on writes
  - Small records causes lots of overhead per access
  - Large gaps causes large file seeks and prevents collective buffering from merging small records into large transfer. IOBUF can merge small records in some cases
  - Not-aligned causes splitting of records across OSTs and read-modify-write at physical block level
- However, it is a part of the NERSC I/O workload
  - About 50% of all I/Os on the NERSC Hopper system were unaligned, and/or small I/Os with transfer sizes that are much smaller than the Lustre block size.

Programming environment changes seemed not the cause of the MPI-IO 10k read rate slowdown.

We confirmed that none of the compilers, cray-mpich, IOBUF library changes made significant differences to the MPI-IO 10k read rate.

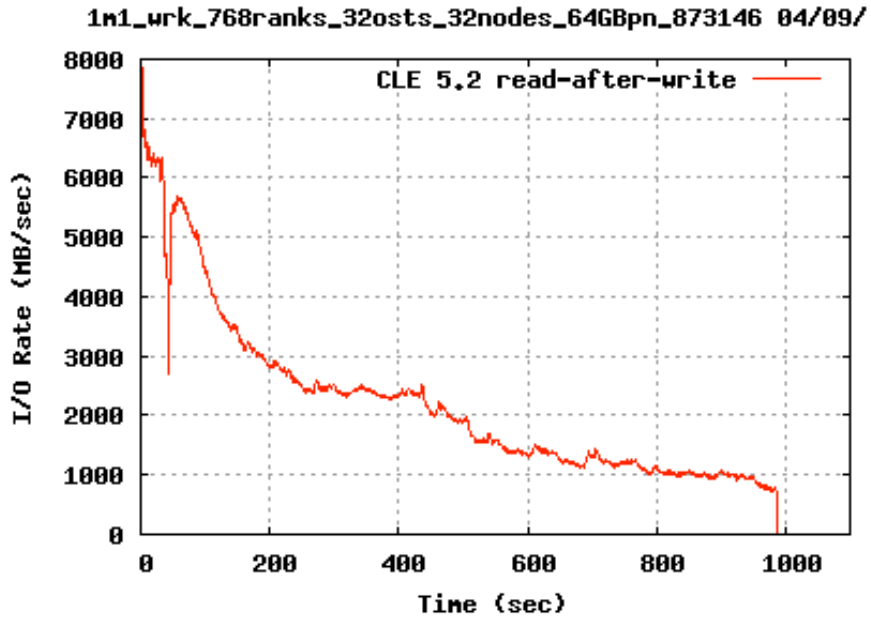


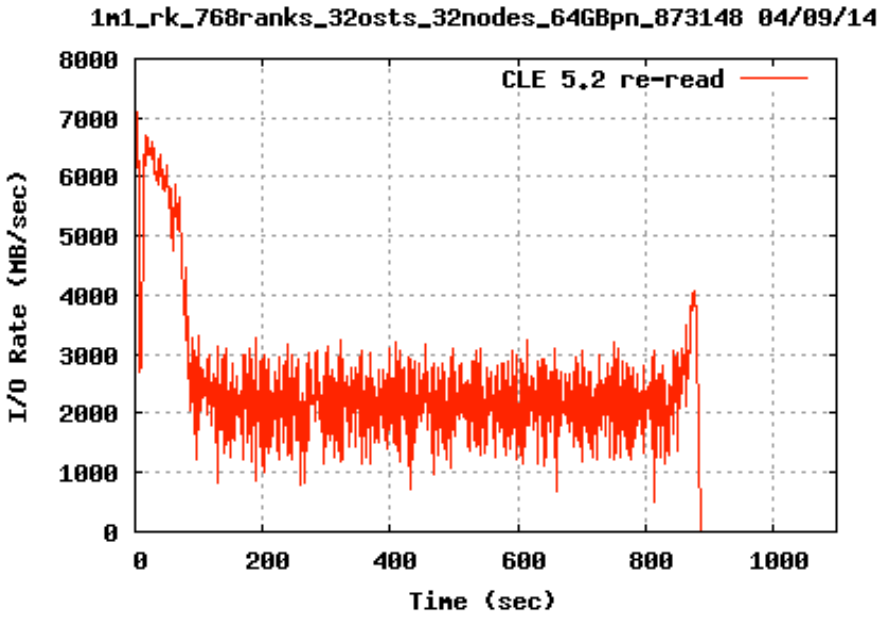
File fragmentation and physical position on the disk drives should not account for the 70% degradation



- The read rate of MPI-IO 10k has decreased significantly (up to 80%) compared to the August acceptance tests results across all three file systems.
- However, the read rate could be several times better in the read only tests than the read-after-write tests.

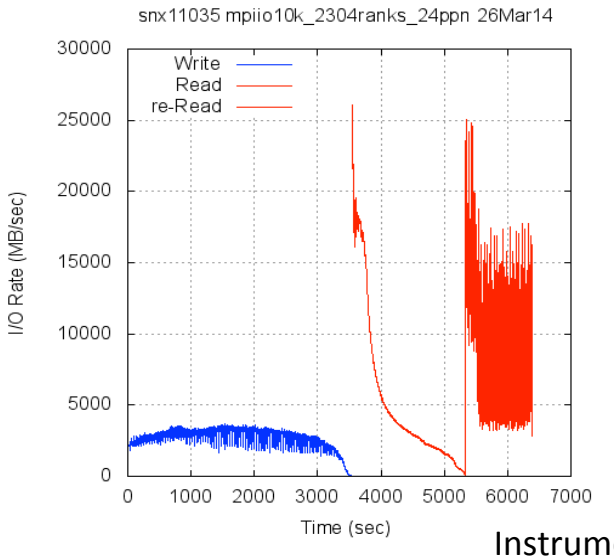
Distinctive read profile observed on internal Cray R&D XC30 system with MPI-IO 1m1 test

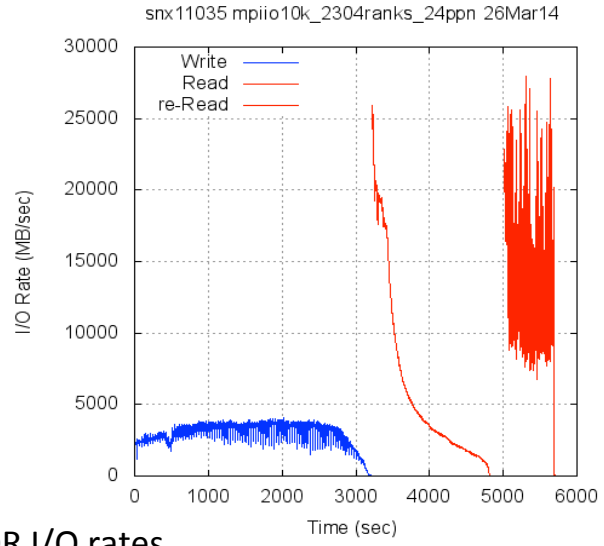




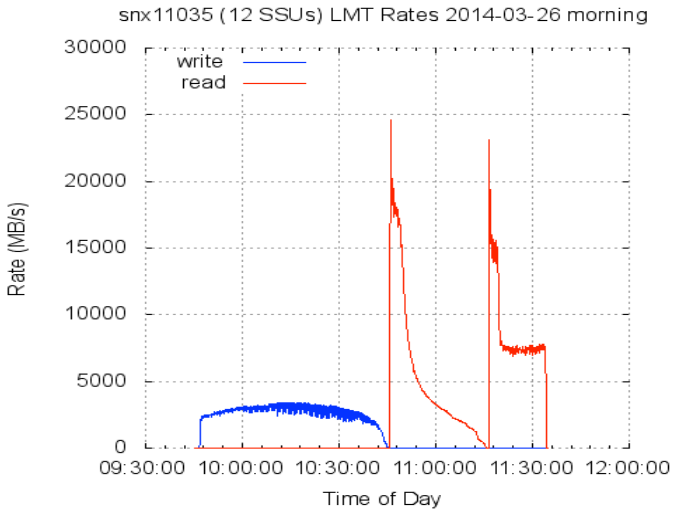
The read rate of the MPI-IO 10k read-after-write test declines steeply, while it keeps constant in the read-only test after an initial drop

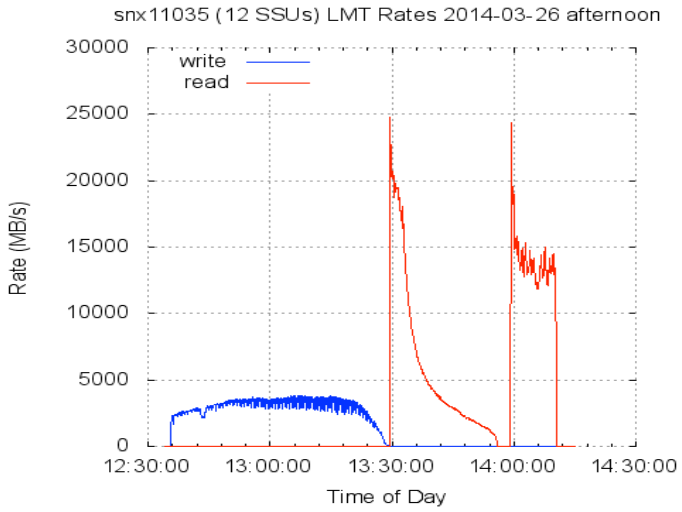
The same distinctive read profiles are observed on Edison with MPI-IO 10k tests





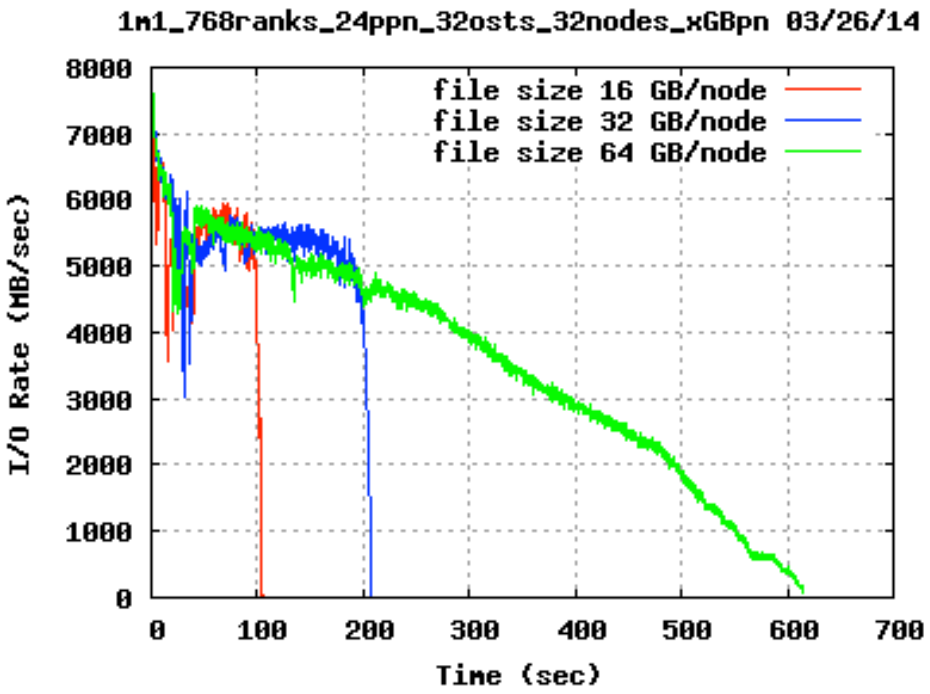
Instrumented IOR I/O rates

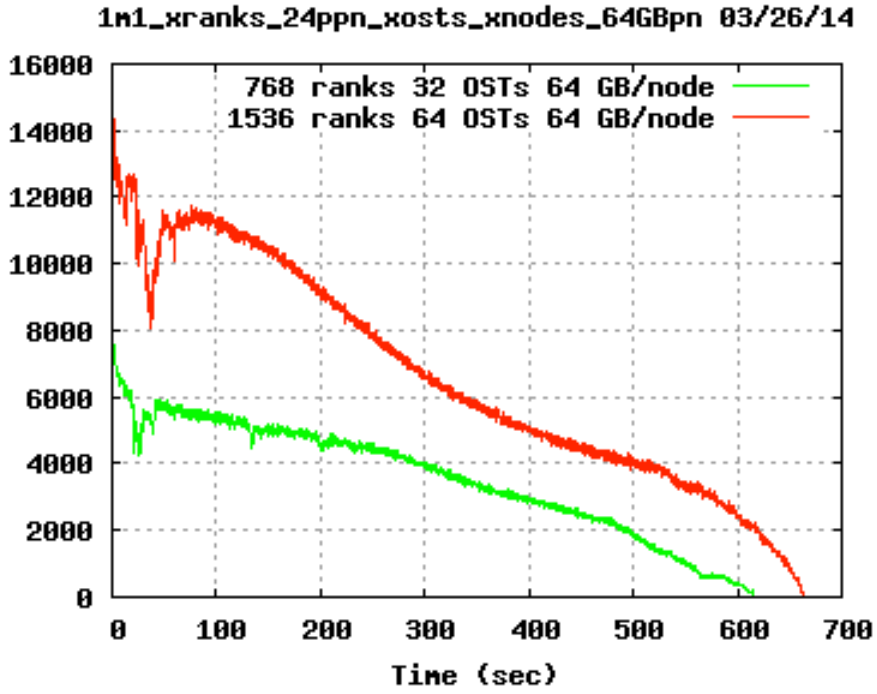




LMT data

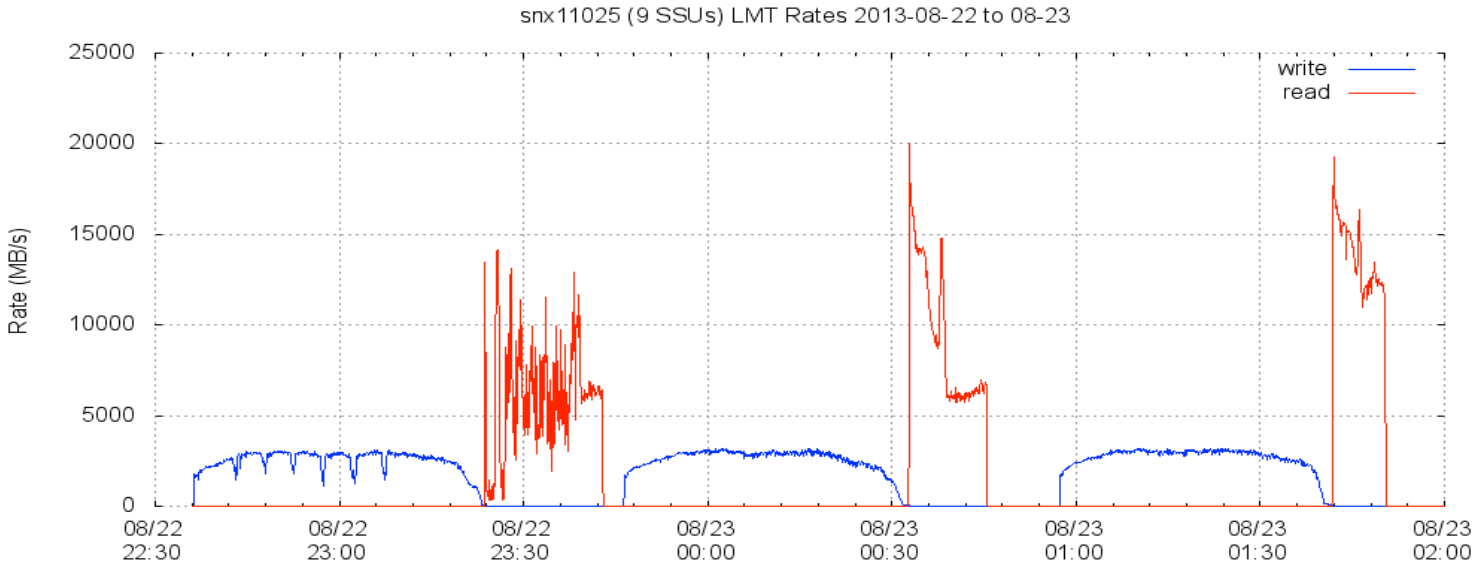
The same read pattern occurs for all read-after-write MPI-IO tests at any transfer sizes, OSTs and PE counts, and file sizes

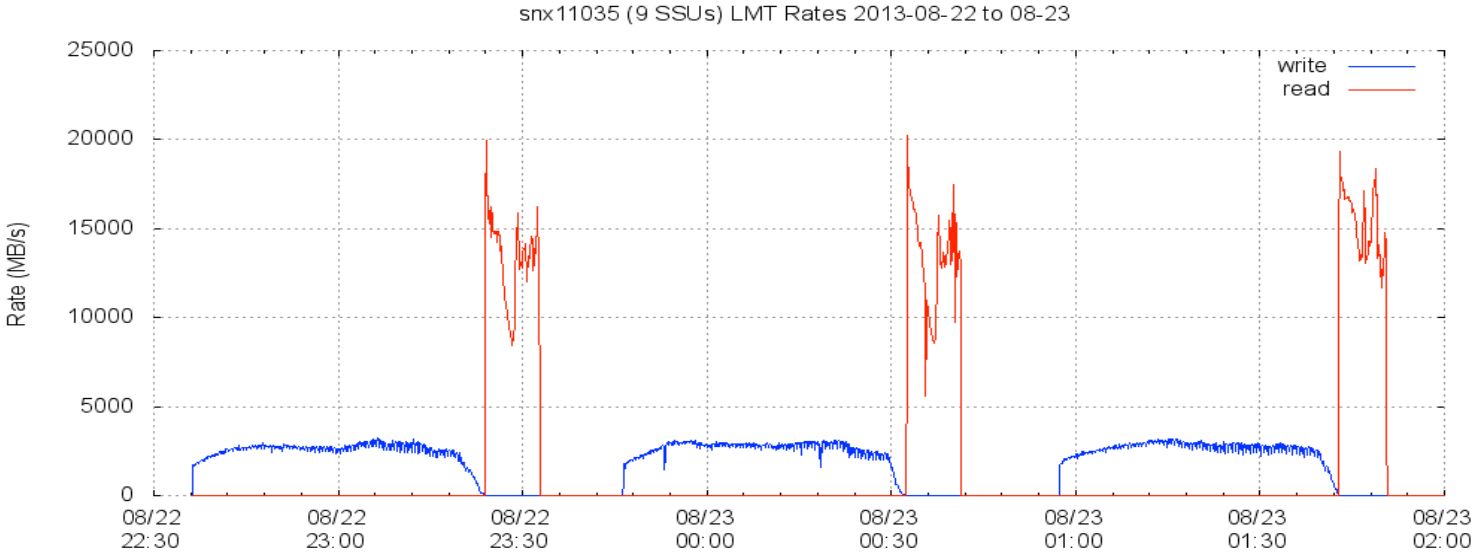




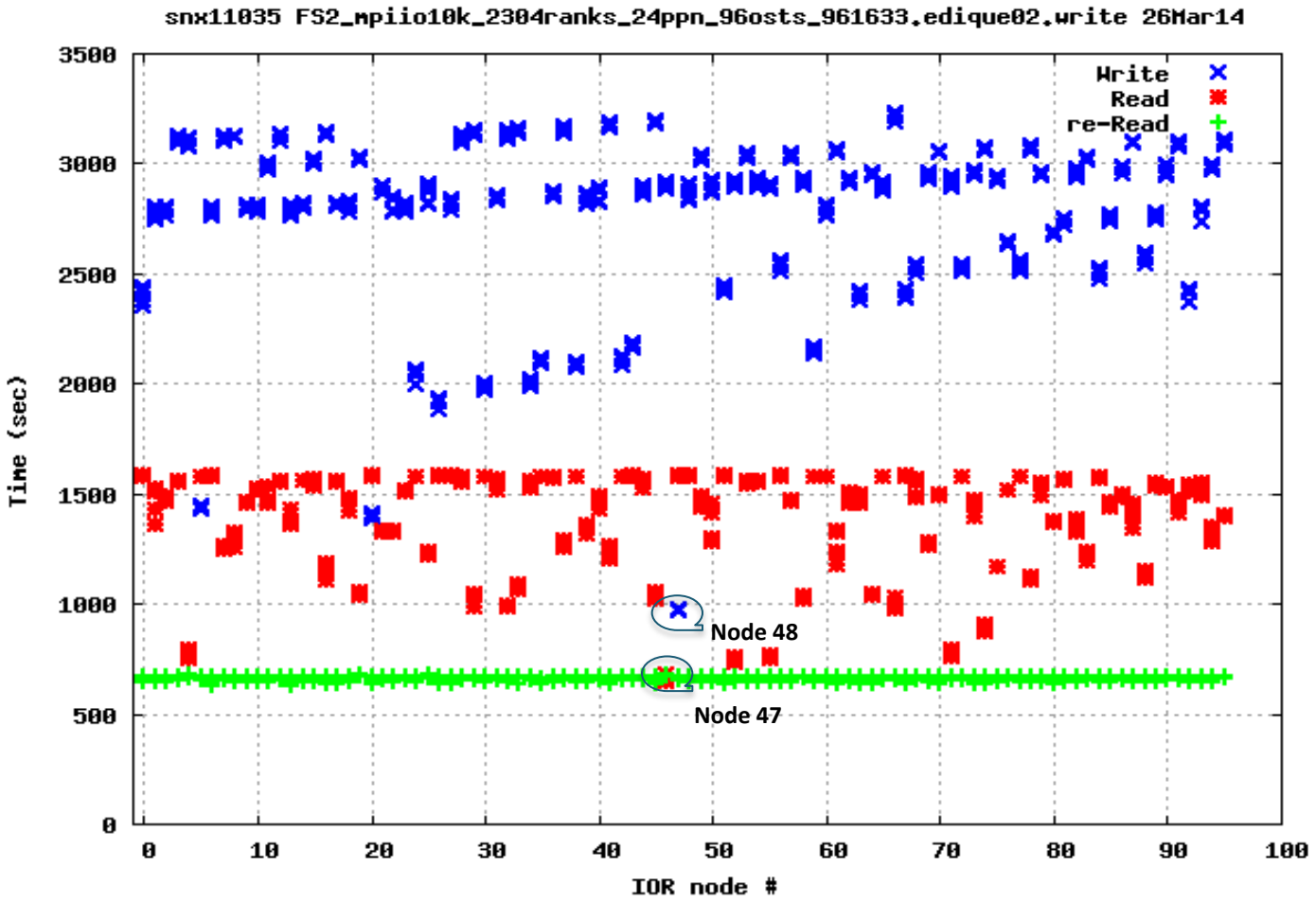
- MPI-IO 1m1 scaling tests on FS3
- MPI-IO 1m1 is equivalent to the MPI-IO 10k test when the IOBUF library is used and collective buffering is disabled.
- **Left figure**: PEs and OSTs used were kept constant, 768, and 32 OSTs (left figure). When increasing the file size, the read rate further drops down.
- **Right figure**: When using more PEs, the read rate drops more quickly

MPI-IO 10k read profiles in August 2013 were similar to the current re-read profile



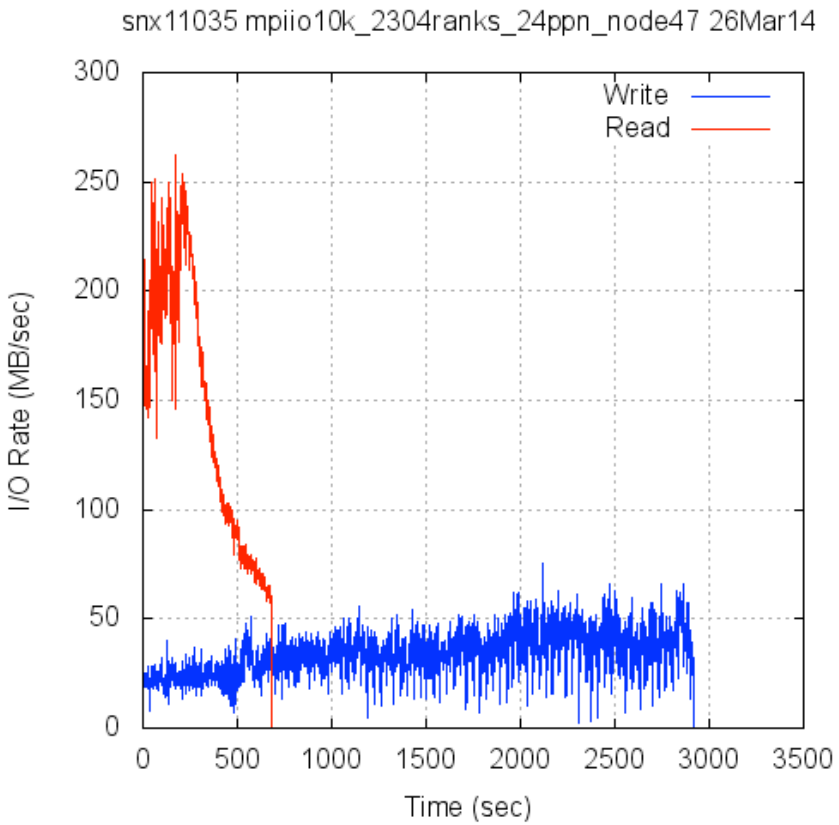


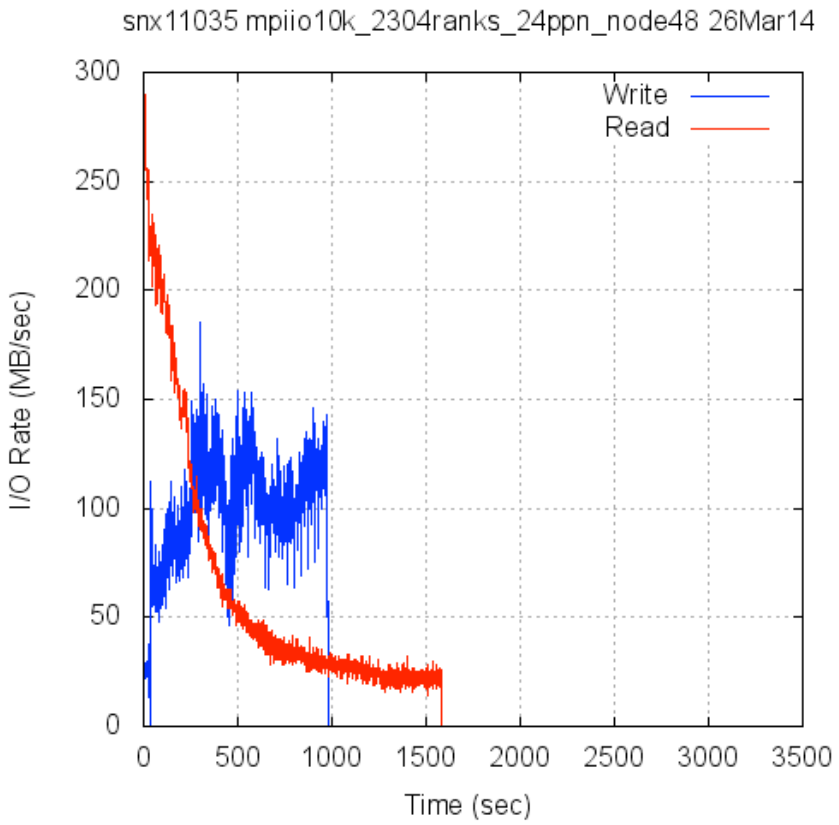
Write and Read times per Node



The I/O rates of the compute nodes differ largely in the read-after-write test, while they are very similar in the read-only test.

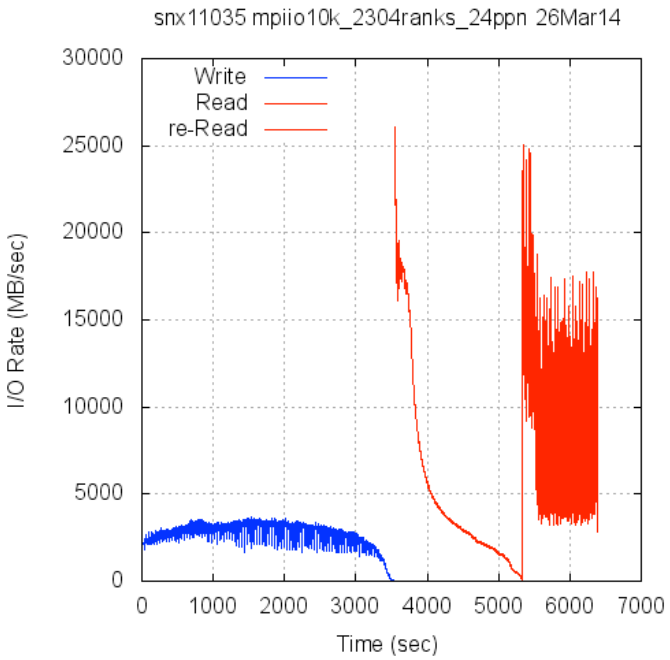
Write and read rates of the node 47 and 48



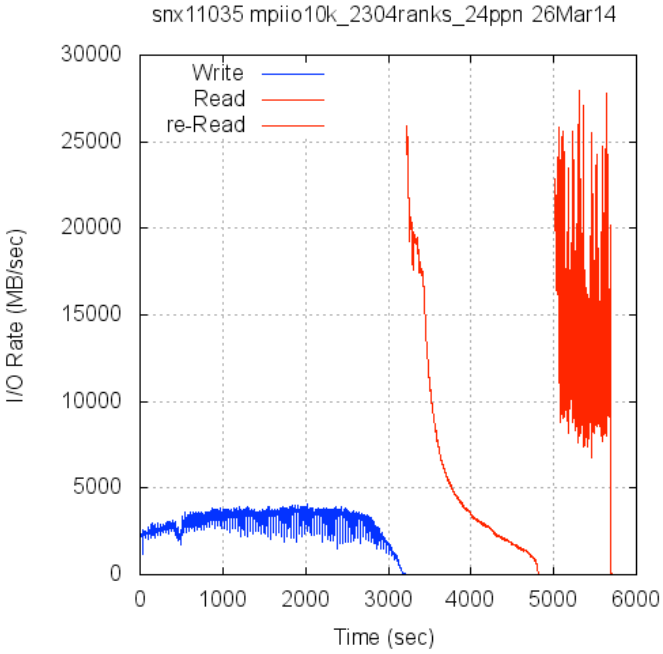


- The imbalanced I/O rates between nodes seem to account for the decreasing read rate and the long tail in the read rate curve in the read-after-write test, while the read rate is roughly constant in the read-only test.
- Why does each node perform differently with perfectly balanced I/O load?

Tests with the Sonexion parameter readcache_max_filesize

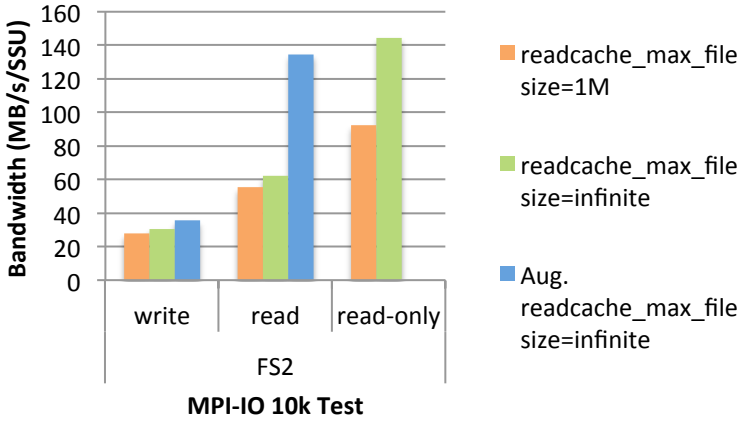


Readcache_max_filesize=1M



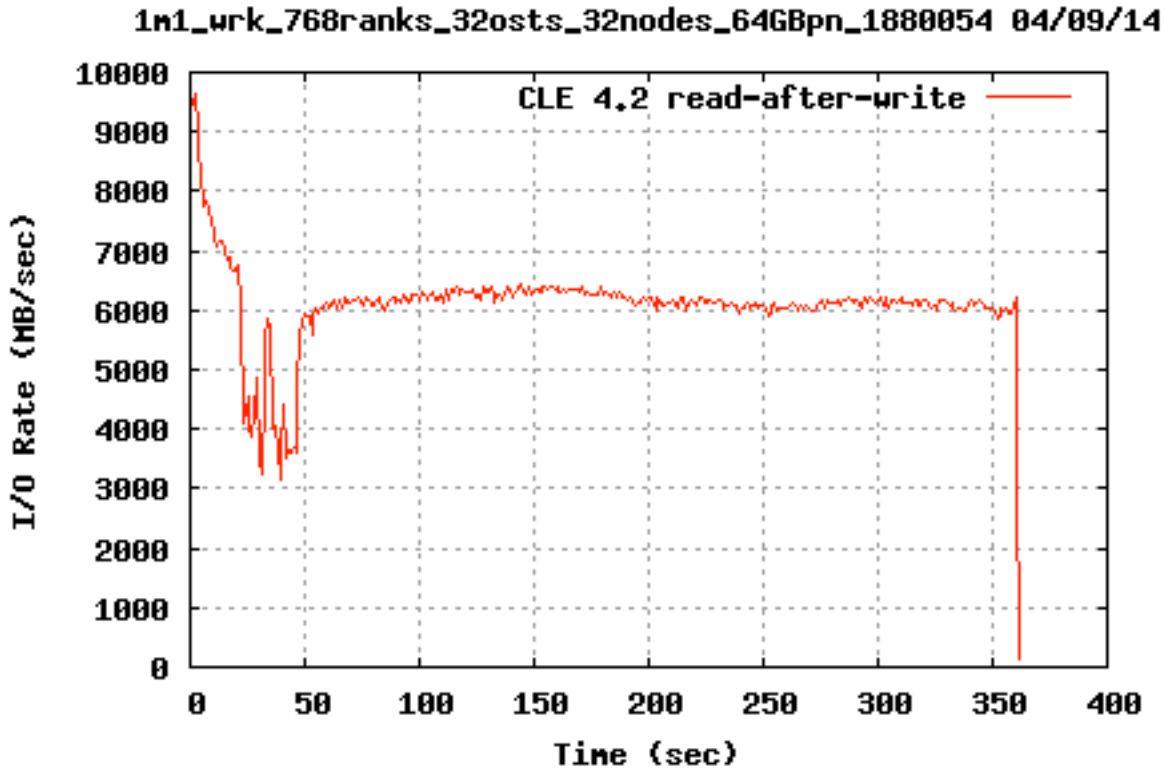
Readcache_max_filesize=infinite

- When the readcache_max_filesize=infinite (the same as in last August), read rates improved, especially in the read-only tests.
- However, the improvement was not sufficient to restore the last August read rates, and the read pattern did not change.



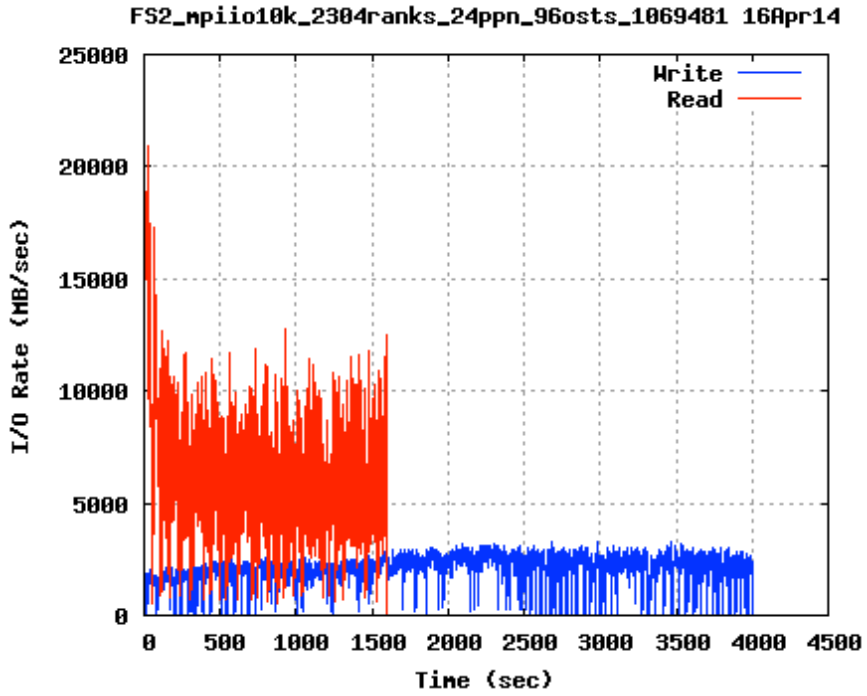
MPI-IO 10k Test

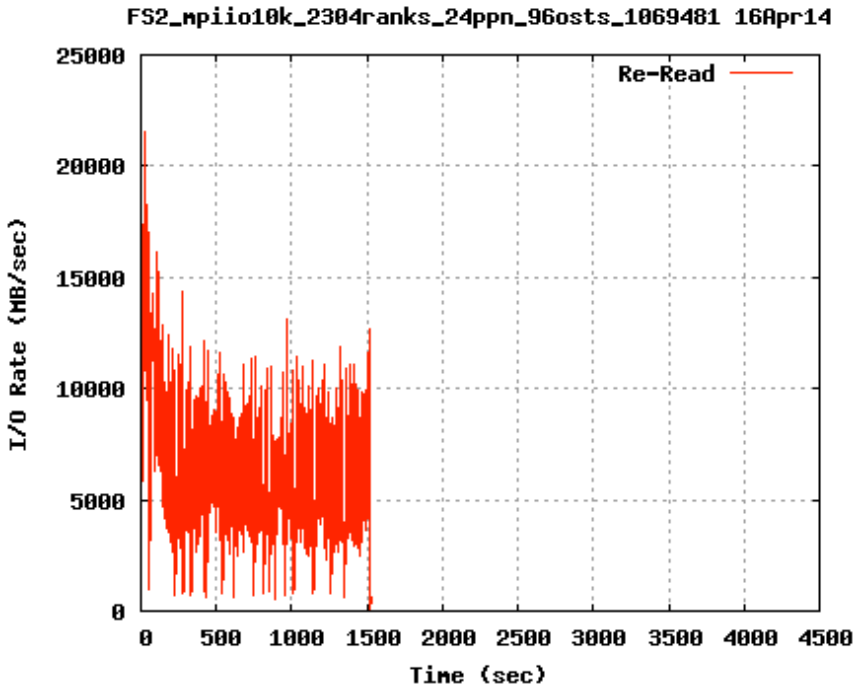
An MPI-IO 10k run with CLE 4.2 and Lustre 1.8.6 on internal Cray R&D XC30 system



- Internal Cray R&D XC30 system with a 32 OST Sonexion file system
- CLE 4.2 + Lustre client 1.8.6 shows a fairly flat performance profile.
- CLE 5.2 + Lustre client 2.4 shows the steeply declining performance profile
- Some CLE +Lustre client upgrades/patches introduced between CLE 5.0.UP03/Lustre 2.3.0 (last Aug) and CLE 5.1.UP00/Lustre 2.4.0 (last Dec).

The good read profile was observed when the Lustre caches were cleared between the write and read phases of the MPI-IO 10k test





- **15 minutes of delay was added between IOR write and read phases.**
- **The following command was run to clear compute node kernel caches:**

  echo 3 > /proc/sys/vm/drop_caches
- **The following command was run to clear Lustre caches:**

  echo 1 > /proc/fs/lustre/ldlm/drop_caches

A Lustre patch has been identified to be the cause of the MPI-IO 10k performance issue

- **We provided the Lustre logs collected on Edison to a Cray Lustre developer.**
- **The specific Lustre patch, which first introduced the problem has been identified.**

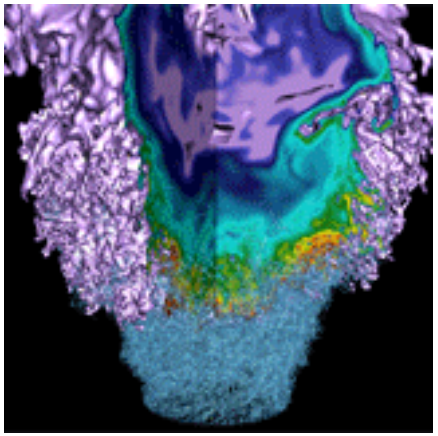
  LU-744 osc: add lru pages management - new RPC
  Add a cache management at OSC [Note: Object server client, IE, OST client. There is an OSC per OST on each client.] layer to control how much memory can be used to cache Lustre pages .
  http://review.whamcloud.com/#/c/2514/

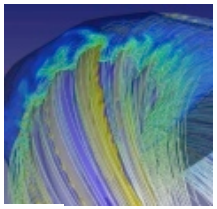
- **Unfortunately, it's both too old and too central to be removable from 2.4/2.5/etc. Further investigation to fix the problem is under the way.**

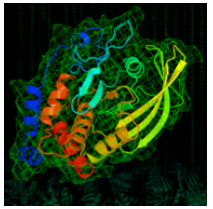
Summary of Investigation

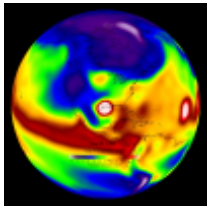
- **Disk position and fragmentation**
  - Would be the same each time a file is read
- **Compilers, Libraries, Sonexion software**
  - No problems found
- **IOBUF not a problem**
  - Replicated without IOBUF, no collective buffering
- **Sonexion readcache_max_filesize setting**
  - 1M probably hurts 10k MPI-IO test, while helps 1m1 and 1m2 MPI-IO tests
  - Infinite causes slight slowdown for Posix FpP
- **Lustre client or CLE**
  - A Lustre patch has been identified to be the cause of this performance issue

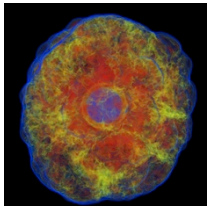
I/O Performance Monitoring in Production Environment

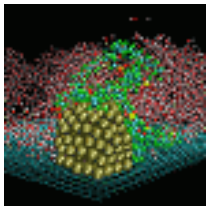












SEC and LMT

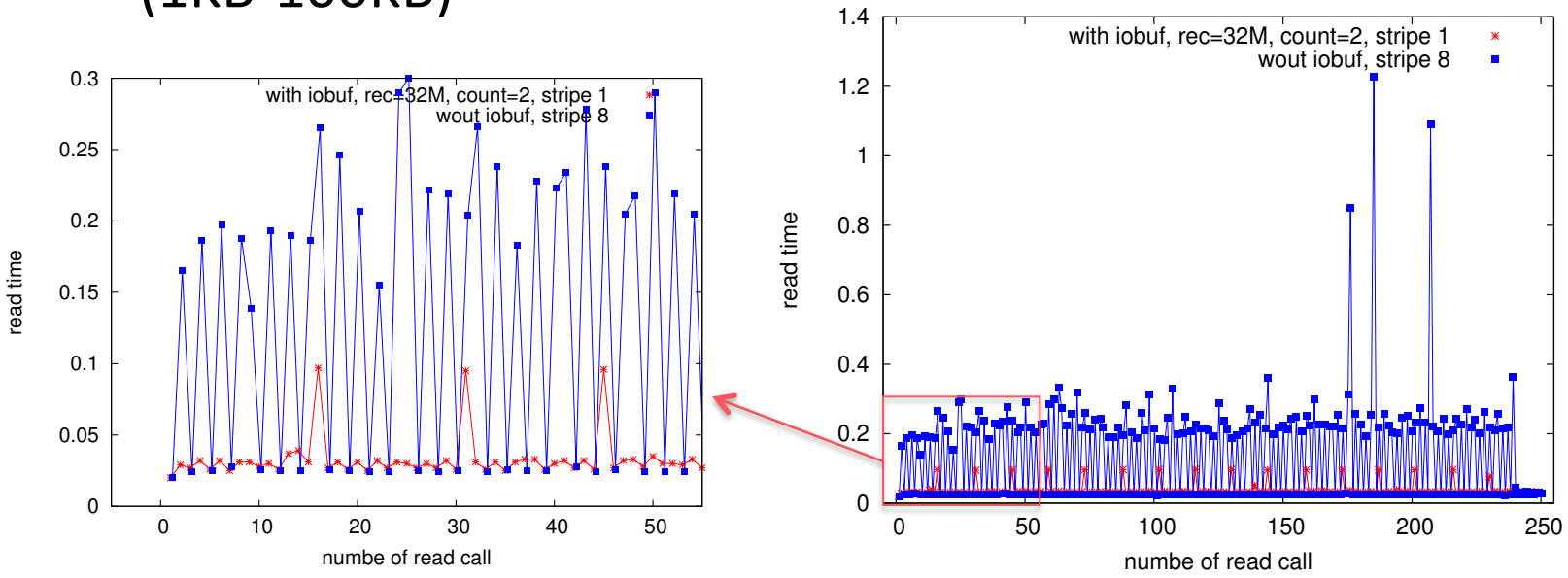
- **File system health and performance monitoring is very important on a production system.**
- **Edison uses Cray provided Simple Event Correlator (SEC) software to monitor the file system events**
  - Boot, disk in and out
  - Various failovers, eg., mds, OST, etc.
  - Slow or hung threads on OSS nodes
  - Failed to connect to database
  - Lock timed out
  - Fan enclosure error
- **However, it is difficult to tell when further investigation is needed.**
- **LMT data available, but not accessible by users.**
- **We are not using the Cray Sonexion System Manager (CSSM).**

A IOR test helped to identify a bad/slow disk drive

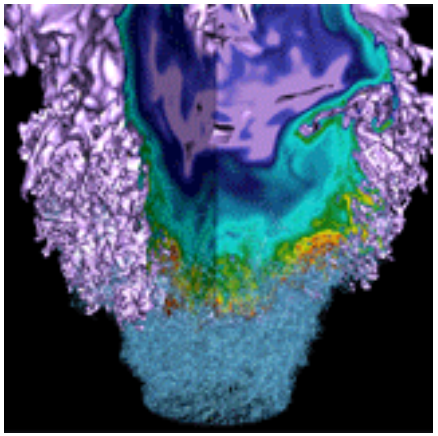
- 3/17/14, a few users reported a more than 5 times I/O slow down on FS1. We saw Lustre errors reported for one of the OSTs, and observed a high load of 450+ on a OSS node which serves that OST.
- However, it was difficult to tell that it was just a high load on the file system or it was an indication of file system problems.
- After complicated debugging (manual process), we located a bad disk drive and fixed the problem by replacing it with a spare disk drive.
- Since the Posix 1m2 IOR test takes only a few minutes to run, it was helpful to detect the slow OSTs, and also to confirm the fix during the debugging process.
- IOR Posix 1m2 is run regularly to help detect slow OSTs now.

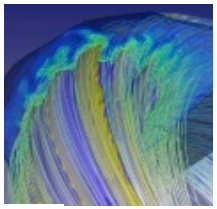
Proactively reaching out users to promote good I/O practices

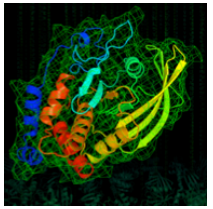
- **User case: File per process I/O with a QLUA code**
  - 11 job instances were bundled up. Each job instance uses 1024 PEs, each PE reads a 50MB file. So the job run with 11264 PEs, reading 500GB file in total.
  - Darshan data shows a small transfer size with this job (1KB-100KB)

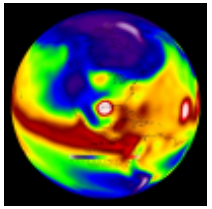


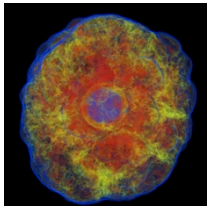
Summary and Future Work

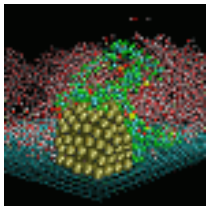












NERSC 40 YEARS at the FOREFRONT

Summary

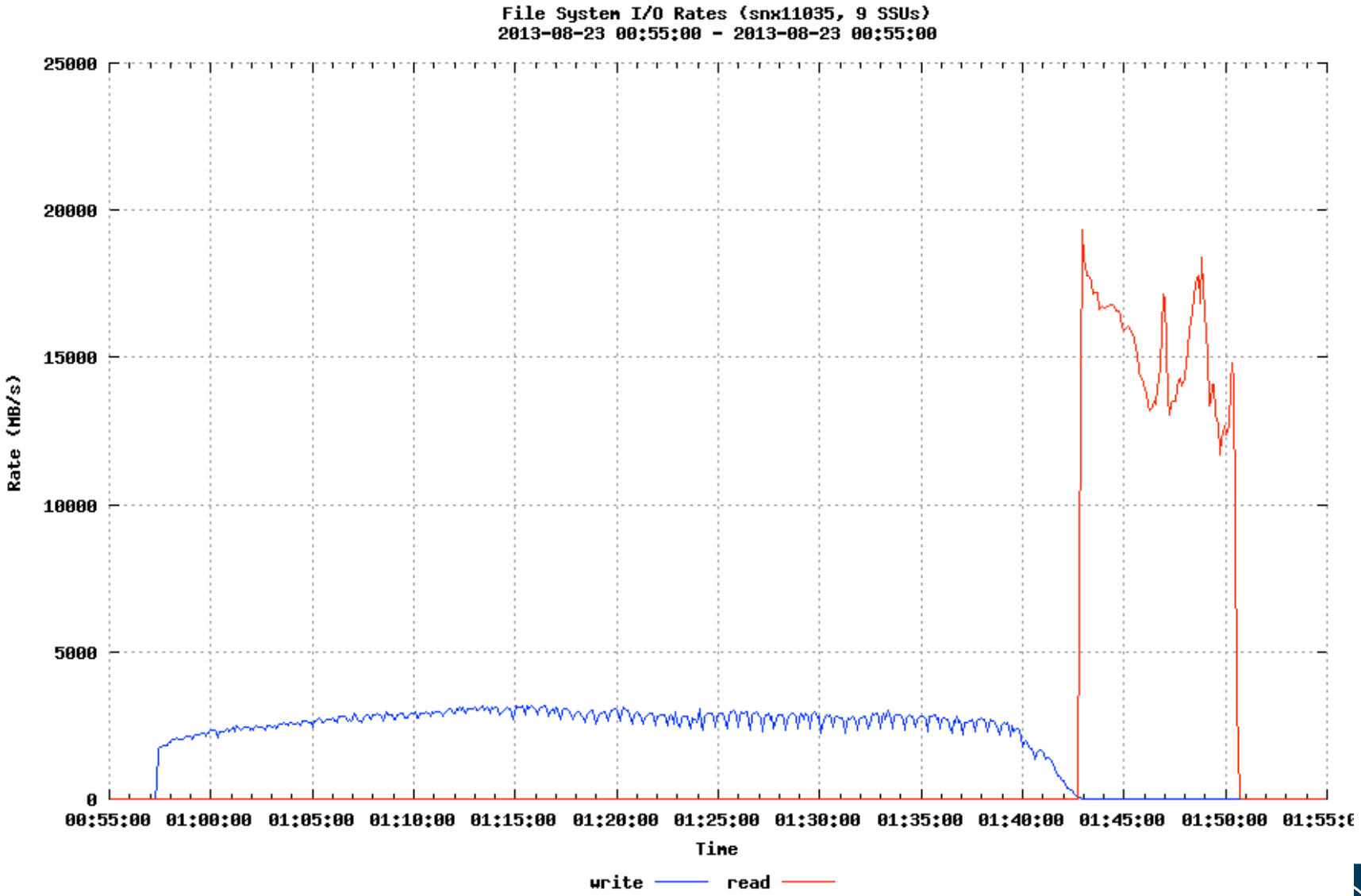
- We investigated the 70% read rate decrease with the MPI-IO 10k test on Edison after the system went through multiple software and hardware upgrades. Through an extensive series of experiments on Edison and on an internal Cray system we ruled out programing environment changes, file fragmentation and physical positions, a Sonexion caching parameter, and CLE upgrades. We were able to narrow the cause to a range of Lustre releases and eventually to a specific Lustre patch. A further investigation to fix the problem is still under way.
- The key progress we made was identifying the characteristic read profile of the MPI-IO 10k test with the instrumented IOR benchmark code, which made it possible to reproduce the dedicated performance issue of large file systems on a small Internal Cray machine, and to investigate the problem on a production environment. Catching the distinctive performance profiles using the instrumented IOR could be a general approach that helps debugging elusive IO performance issues as the performance profile is more sensitive to the changes compared to the net I/O rates.

Summary

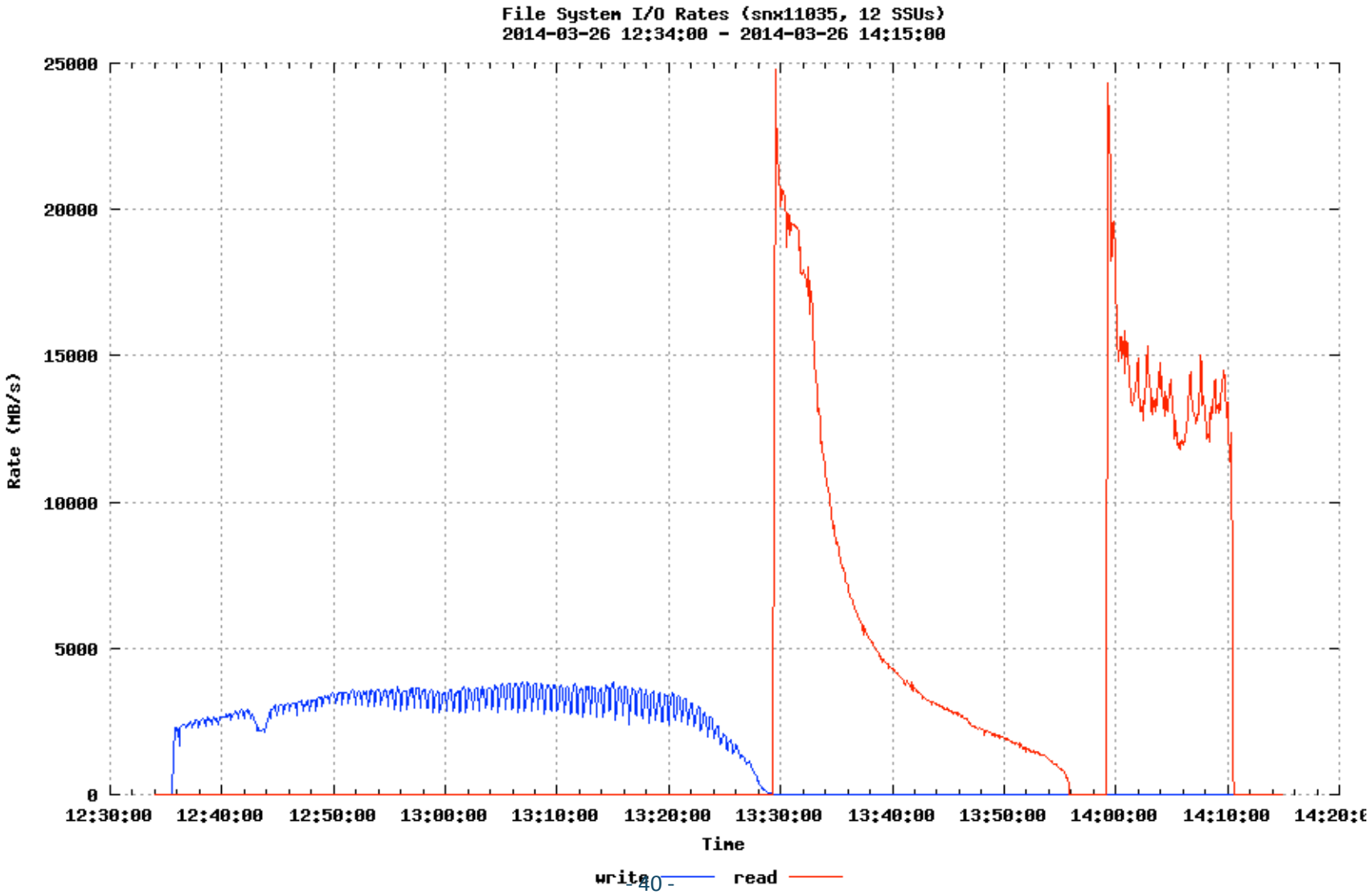
- With Sonexion 1600 storage system, the I/O bandwidth scales almost linearly up to 144 OSTs, the max number of OSTs available in a single Lustre file system. An 80-100% of the peak I/O bandwidth (4GB/s/SSU) was observed on Edison.
- I/O time variation in production environment is very disruptive to users workflows. Edison uses SEC and LMT tools to monitor the file system health and performance. In addition, the IOR tests are run regularly to help monitoring the file system performance. Promoting good I/O practices is helpful to mitigate the performance variation.
- NERSC is working on making the LMT data accessible to users; is also looking for a better benchmark options to test the file system performance with small I/Os.

Thank you.

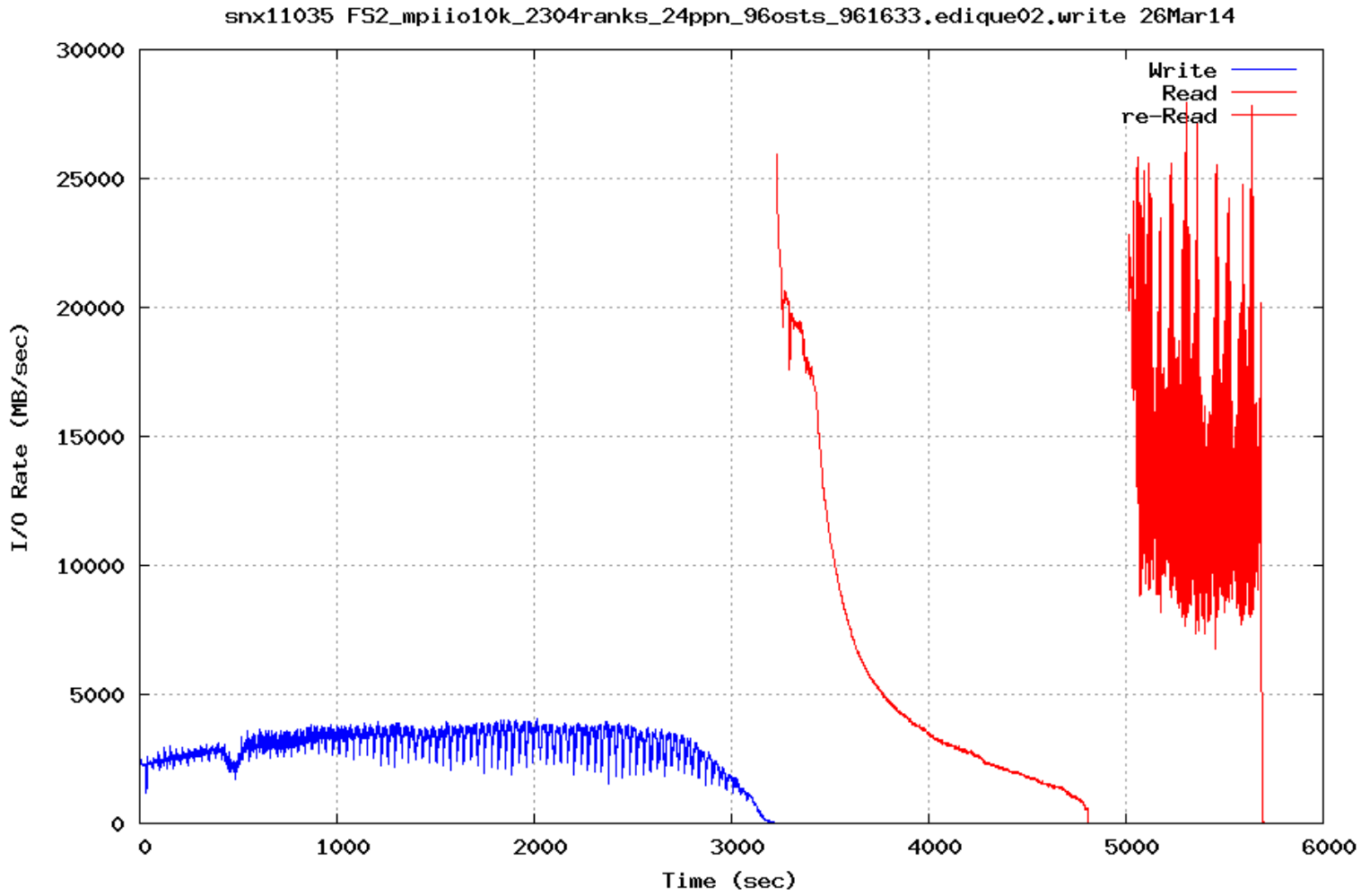
LMT data for 10k MPI-IO case from August 2013



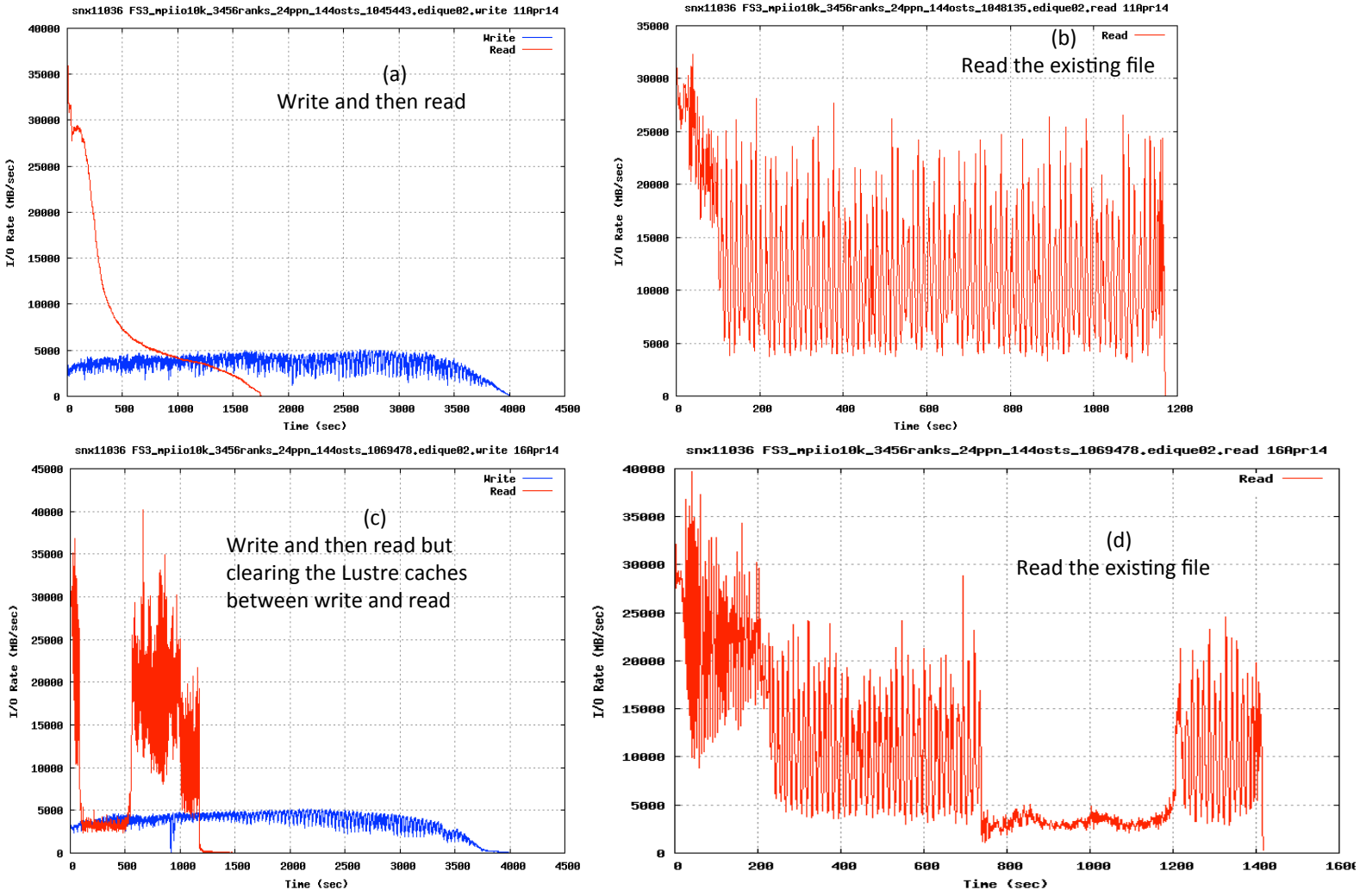
LMT data for 10k MPI-IO from March 2014



Instrumented IOR data from March 2014



Even on the production system with contentions from other users, we may still be able to tell the difference between two runs by comparing the read profiles.



These were two runs on FS3 with and without clearing Lustre Caches (non-dedicated). Although there was noised, the read profile change in (a) and (c) was obvious.