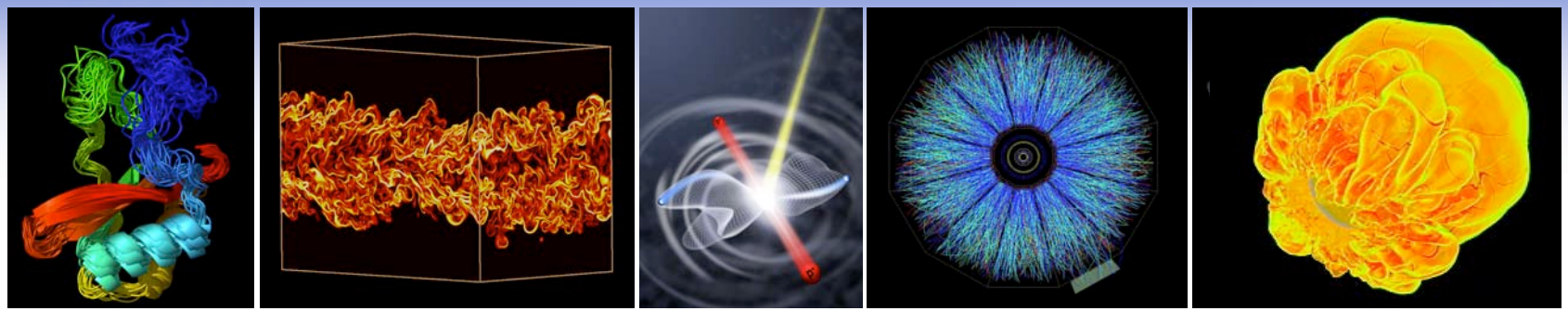


MPI-IO Performance on Franklin

Katie Antypas

Andrew Uselton, Harvey Wasserman, Shane Canon, Steve Luzmoor, Tom Davis

CUG

May 6, 2009

Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC

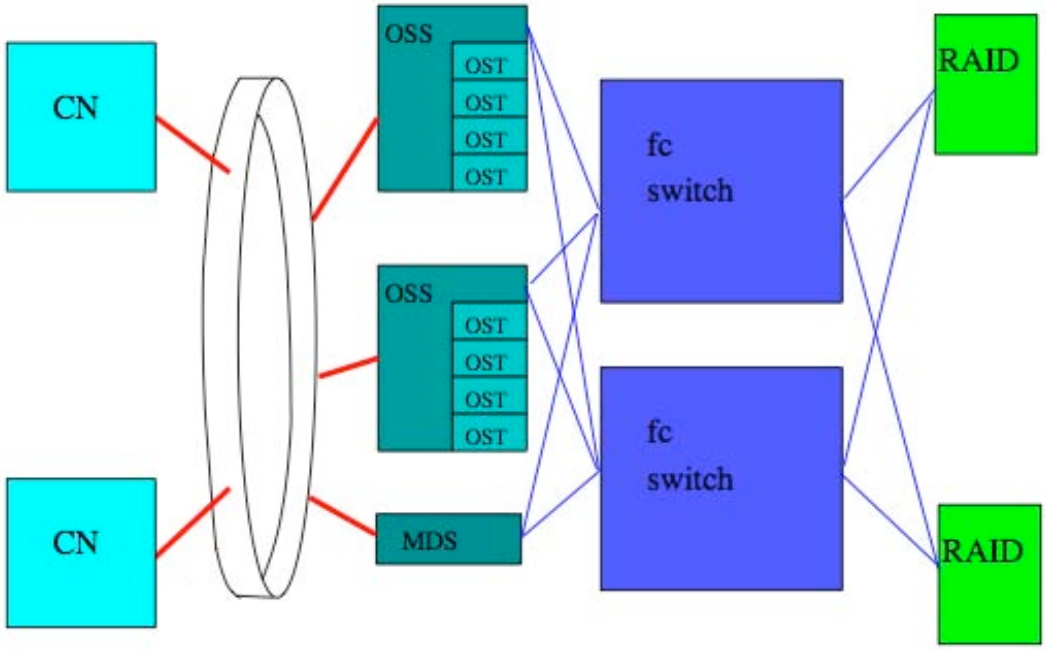
BERKELEY LAB

I/O Performance on Franklin

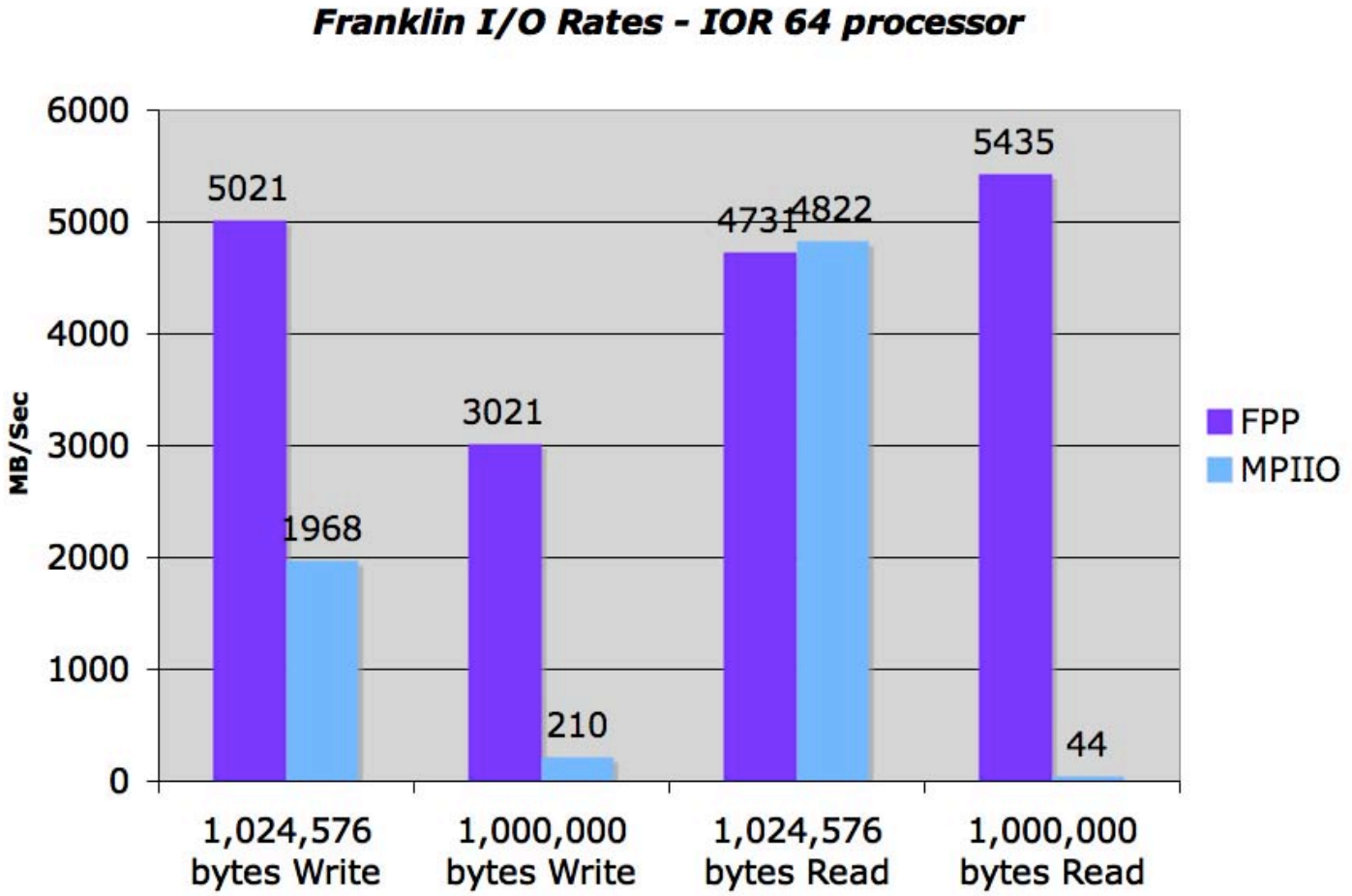
- Intended to talk about poor MPI-IO performance on Franklin
- But … the system maintenance on March 17th drastically improved MPI-IO performance and I/O performance over all
- However, none of the March 17th changes were intended to improve I/O performance -- (so we thought)
- What caused the performance improvements?
- Knowing what caused the improvement is important so we can avoid regressions and assure this improvement is available on all XTs

Franklin's I/O System

- Lustre file system -- /scratch
- 20 OSSs, 80 OSTs
- 346 TB disk
- At this time measured peak performance ~11 GB/sec



MPI-IO shared file performance compared to file-per-processor performance Feb 2009

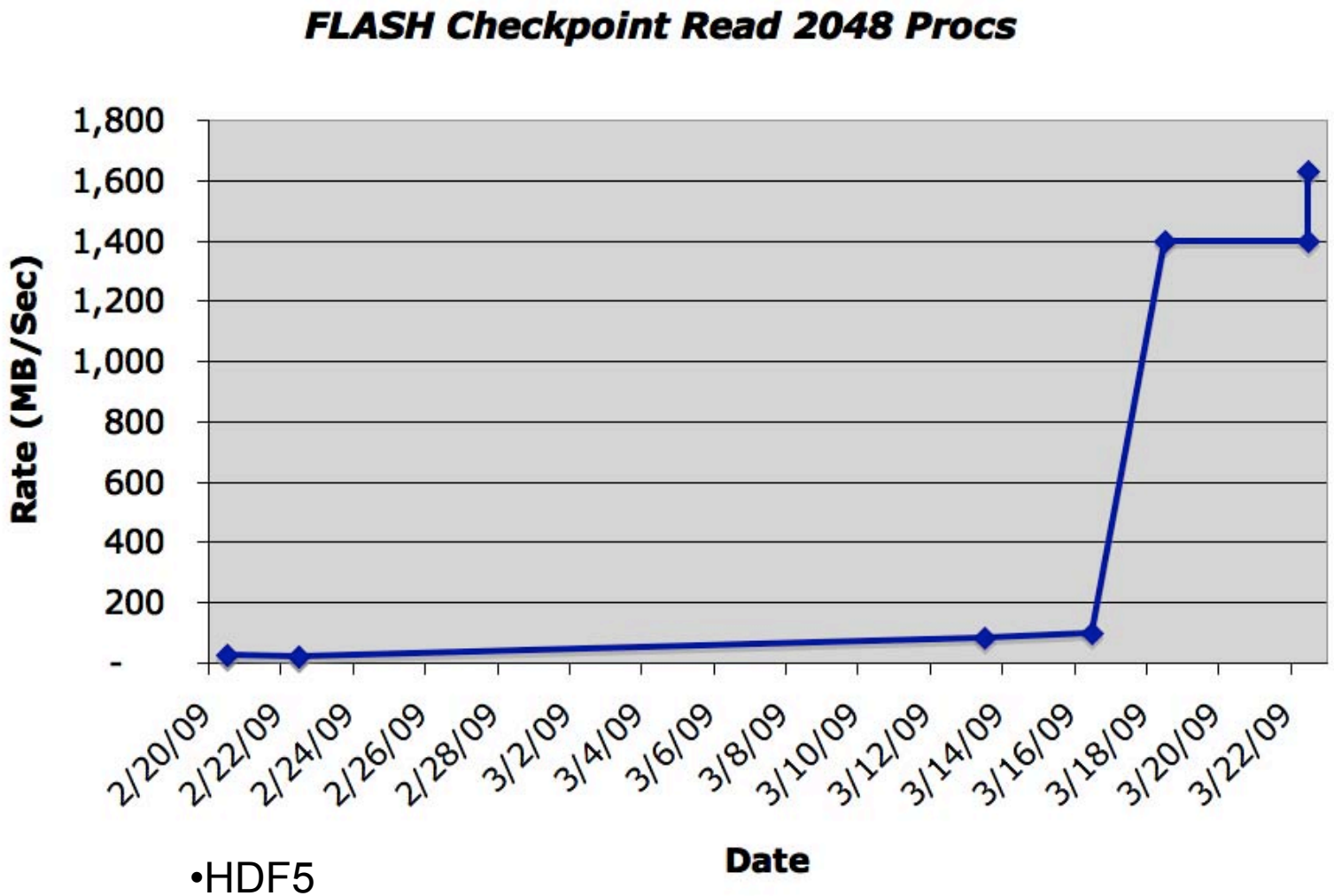


For power of ten based write and reads, MPI-IO performance achieved a very low percentage of file-per-processor performance

March 17th Maintenance

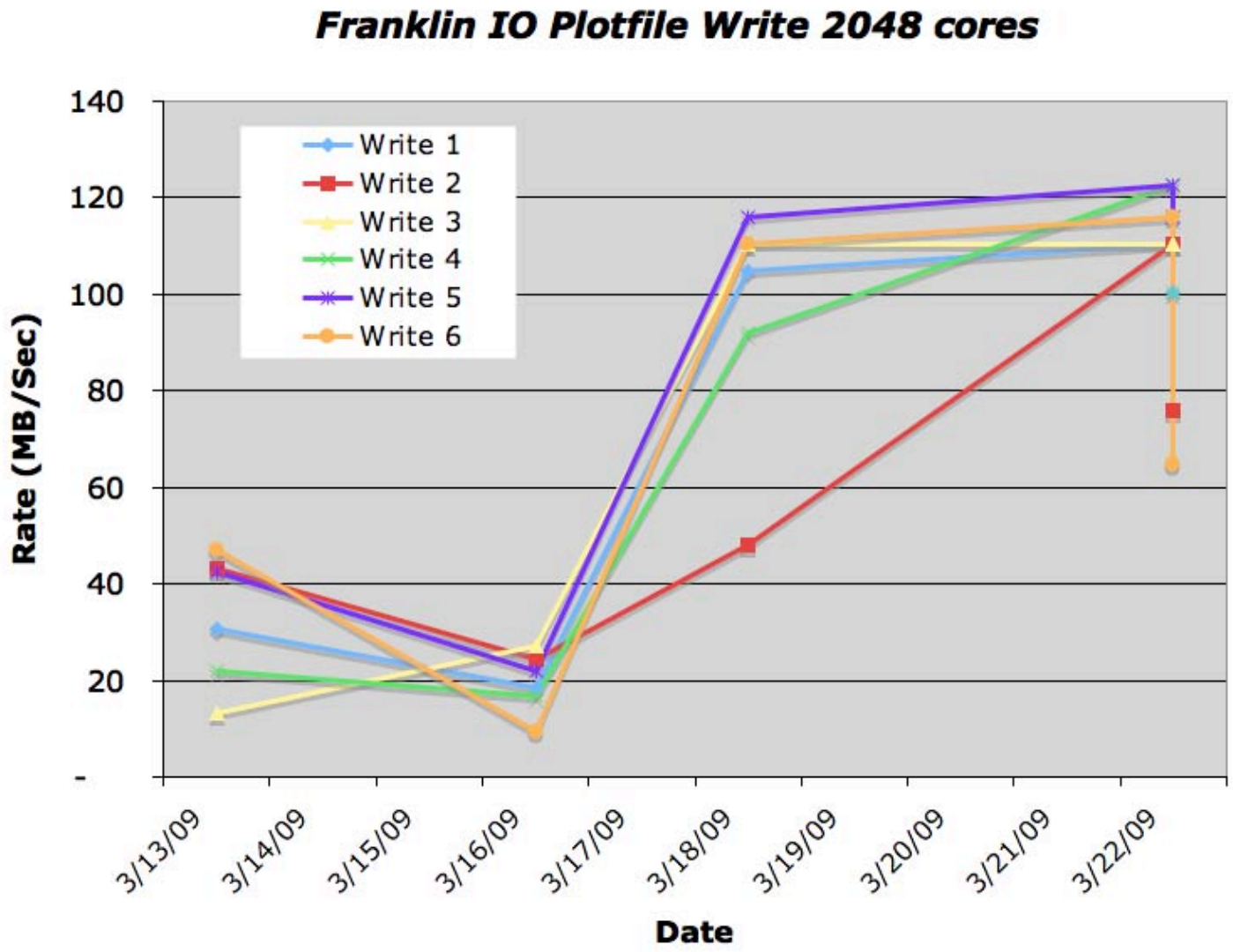
- OS Upgrade from CLE 2.1 to CLE 2.1UP01 with patch sets 01, 01A and 02
- Repair down hardware links
- Change PAM/LDAP settings
- Change network configuration from 42net to 41net
- Reduce max wall clock time for regular queues from 36 to 24 hours
- Insert new SIO modules (converted from compute nodes)

FLASH



- HDF5
- Huge performance improvement, MPI-IO Read, over 100x
- Binary was NOT recompiled
- 9GB files

FLASH

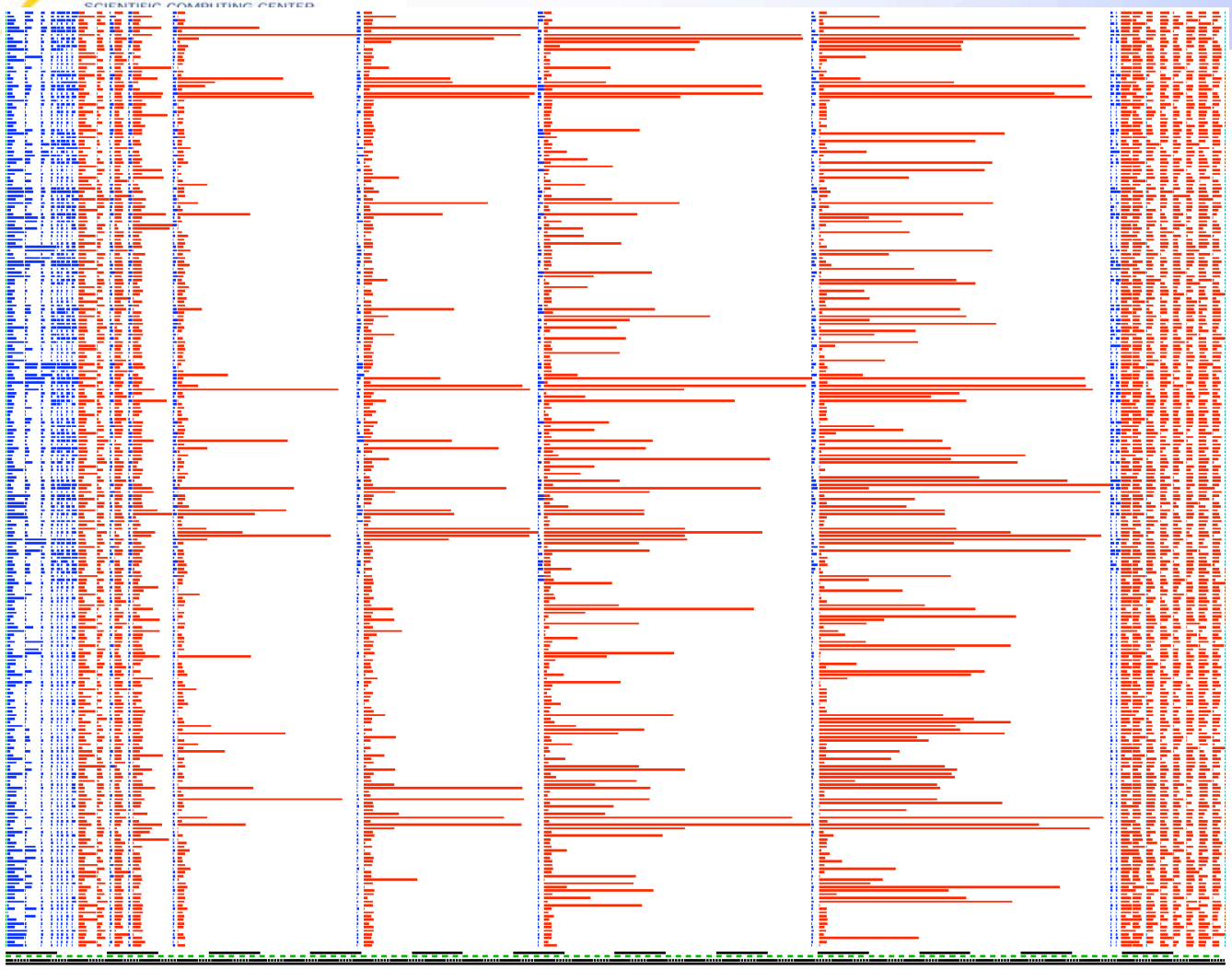


- Write performance up ~3-5x
- 2.2GB files

MADBench2 using MPI-IO Before Maintenance

- IO profile of MADBench, MPI-IO test.
- Writes -blue, Reads- Red, x axis time, y axis processors

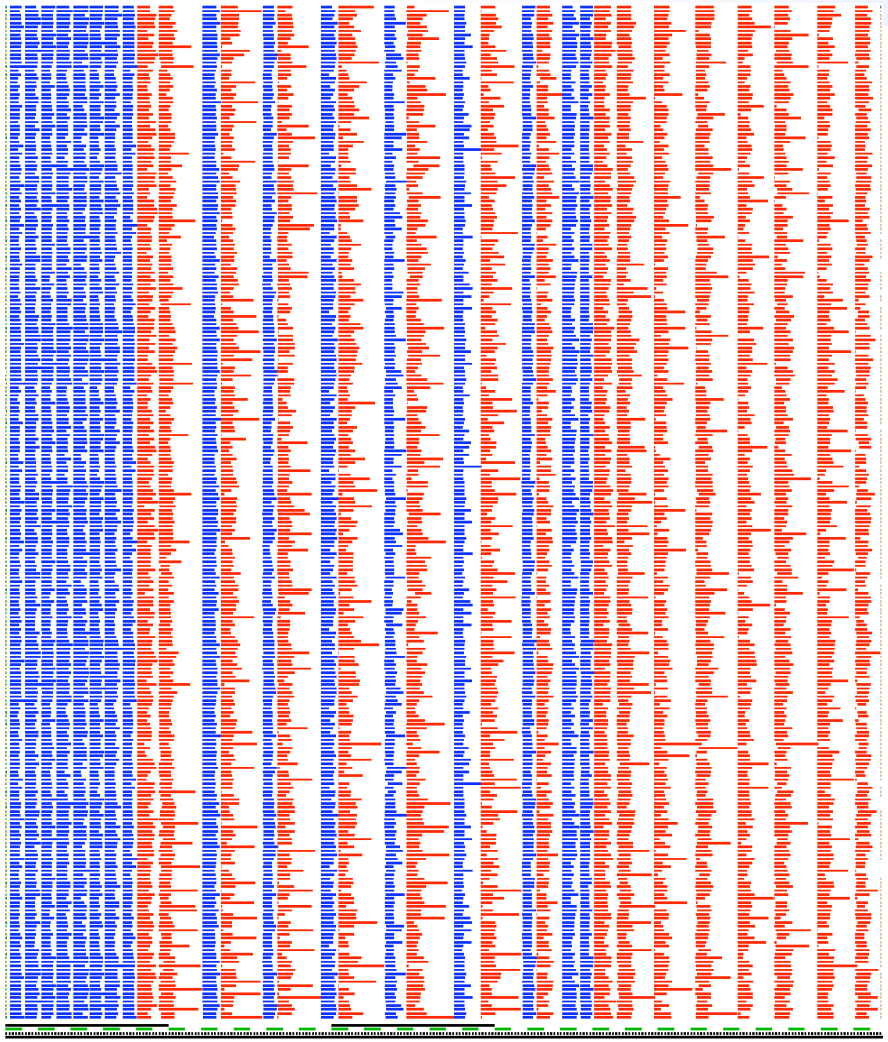
MADBench2 using MPI-IO Before and After Maintenance

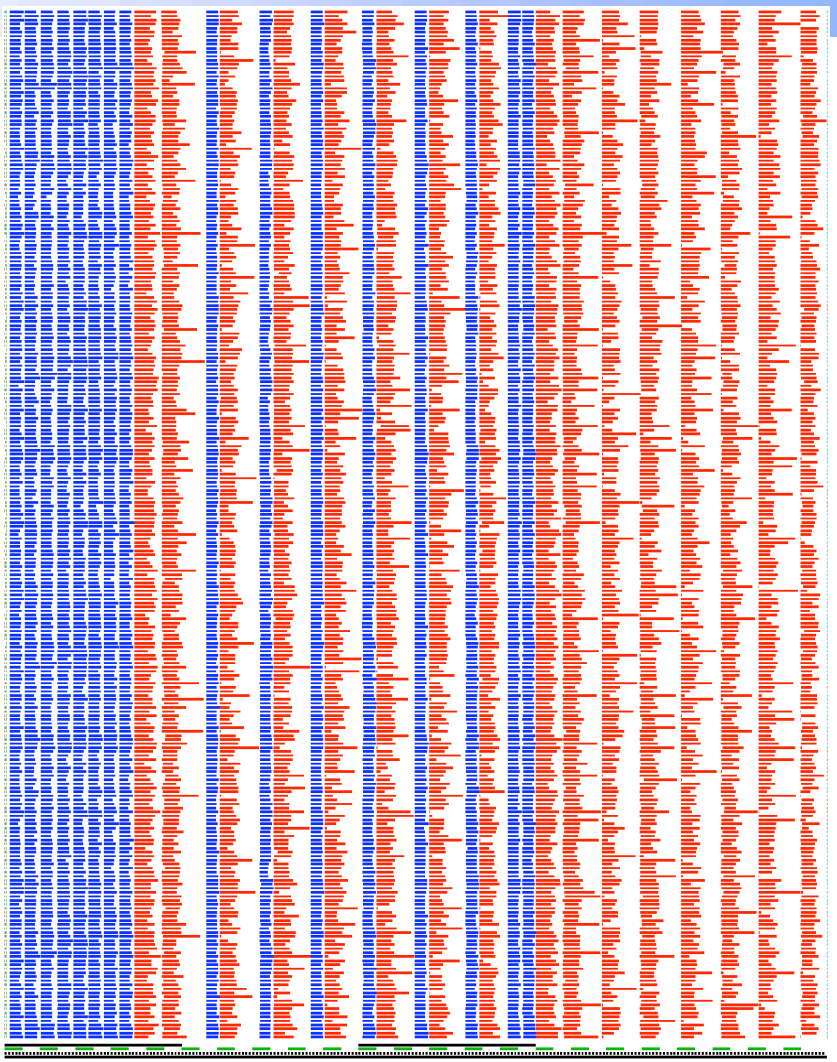


Speed up of over 4 times after the March 17th maintenance

Data from Andrew Uselton and Noel Keen

MADBench2 using POSIX Before and After Maintenance





Significant performance change?

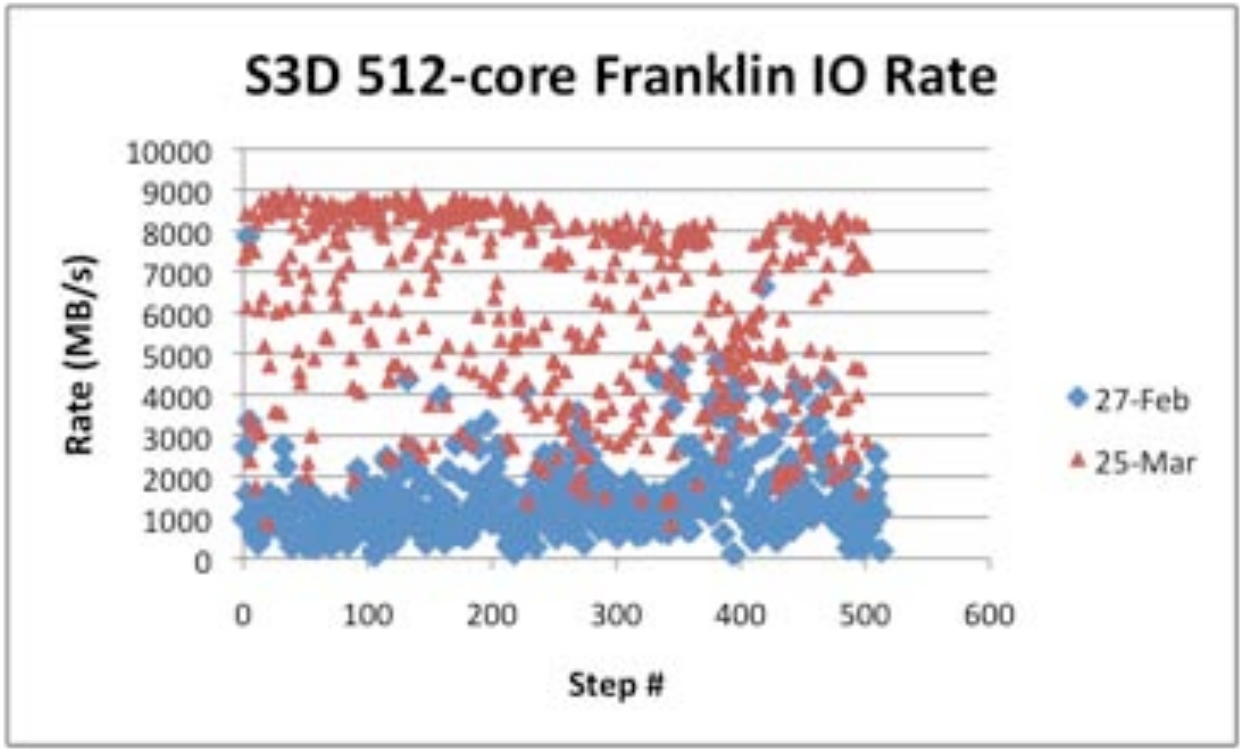
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

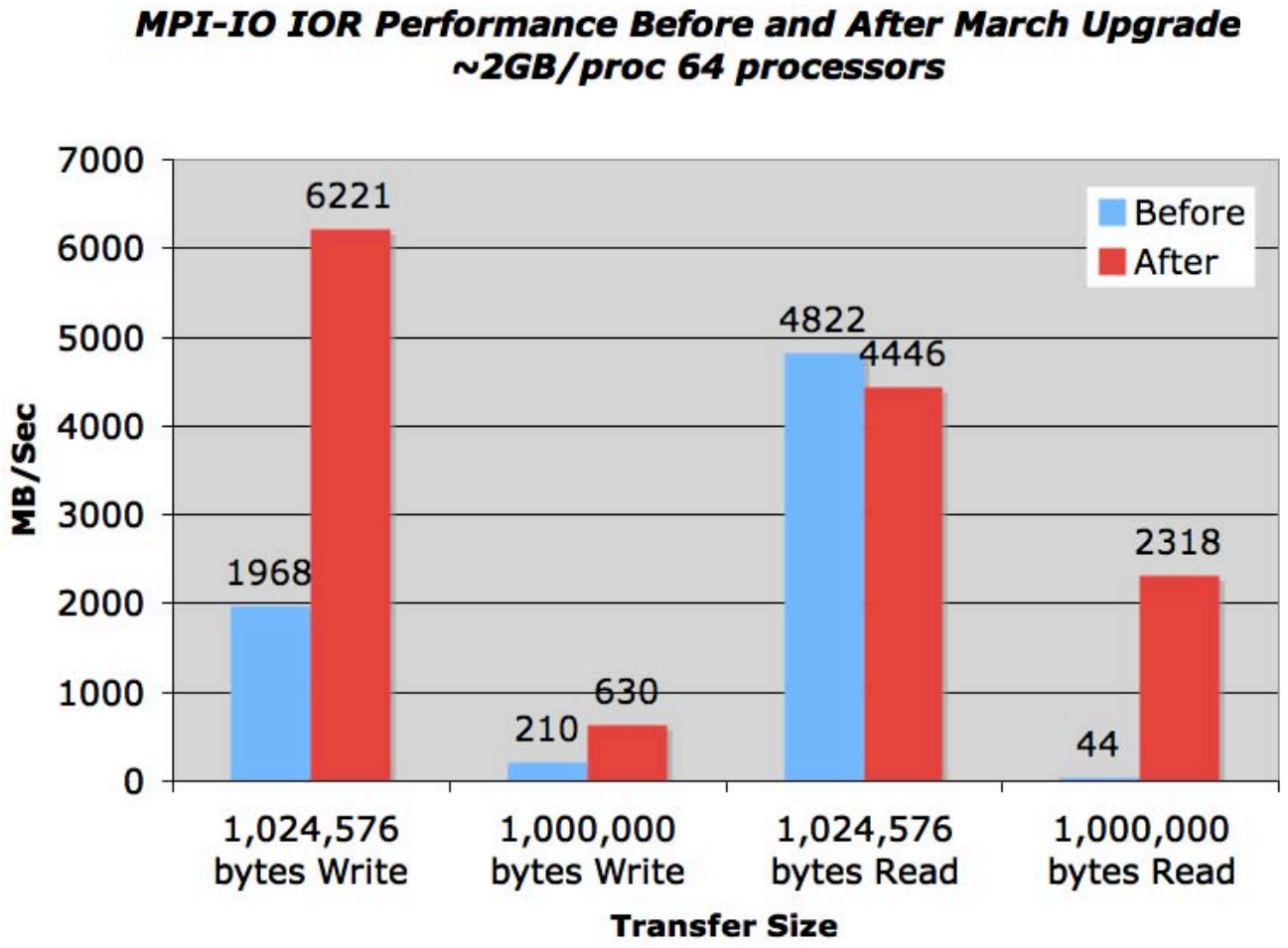
S3D I/O Write Performance

- Not an MPI-IO code, file per processor
- 200 MB/core
- Average rate before: 955 MB/sec
- Average rate after: 4972 MB/sec



Data from Harvey Wasserman

IOR MPI-IO performance before and after maintenance



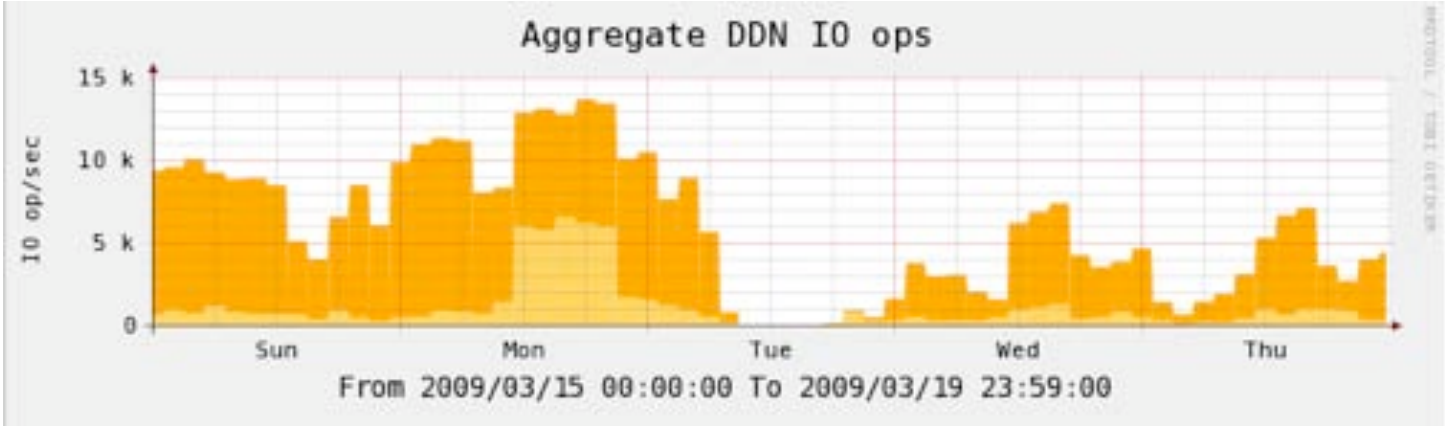
Dramatic MPI-IO performance for power-of-ten based reads

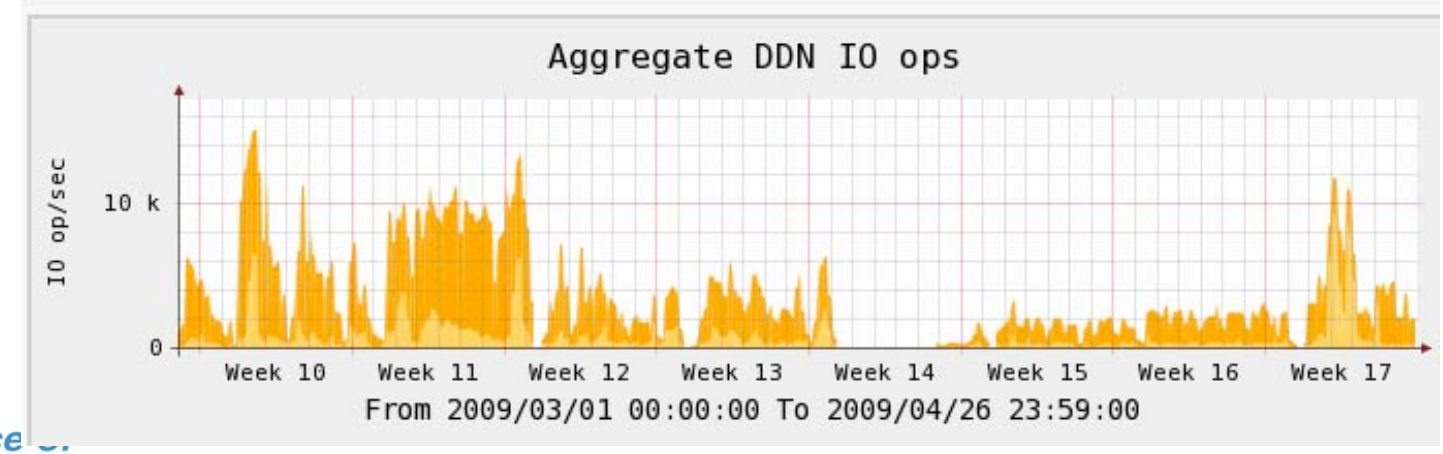
Possible Explanations

- **Repair down hardware links**
  - **If crucial links were down could this cause poor performance?**
  - **Cray staff did congestion analysis and found no differences before and after upgrade**
- **OS Upgrade from CLE 2.1 to CLE 2.1UP01 with patch sets 01, 01A and 02**
  - **Many changes in UP01**
- **Workload Changes?**

CLE2.1 UP01

- **Async journal commit mod?**
  - **Reduces frequency of writes by 8**

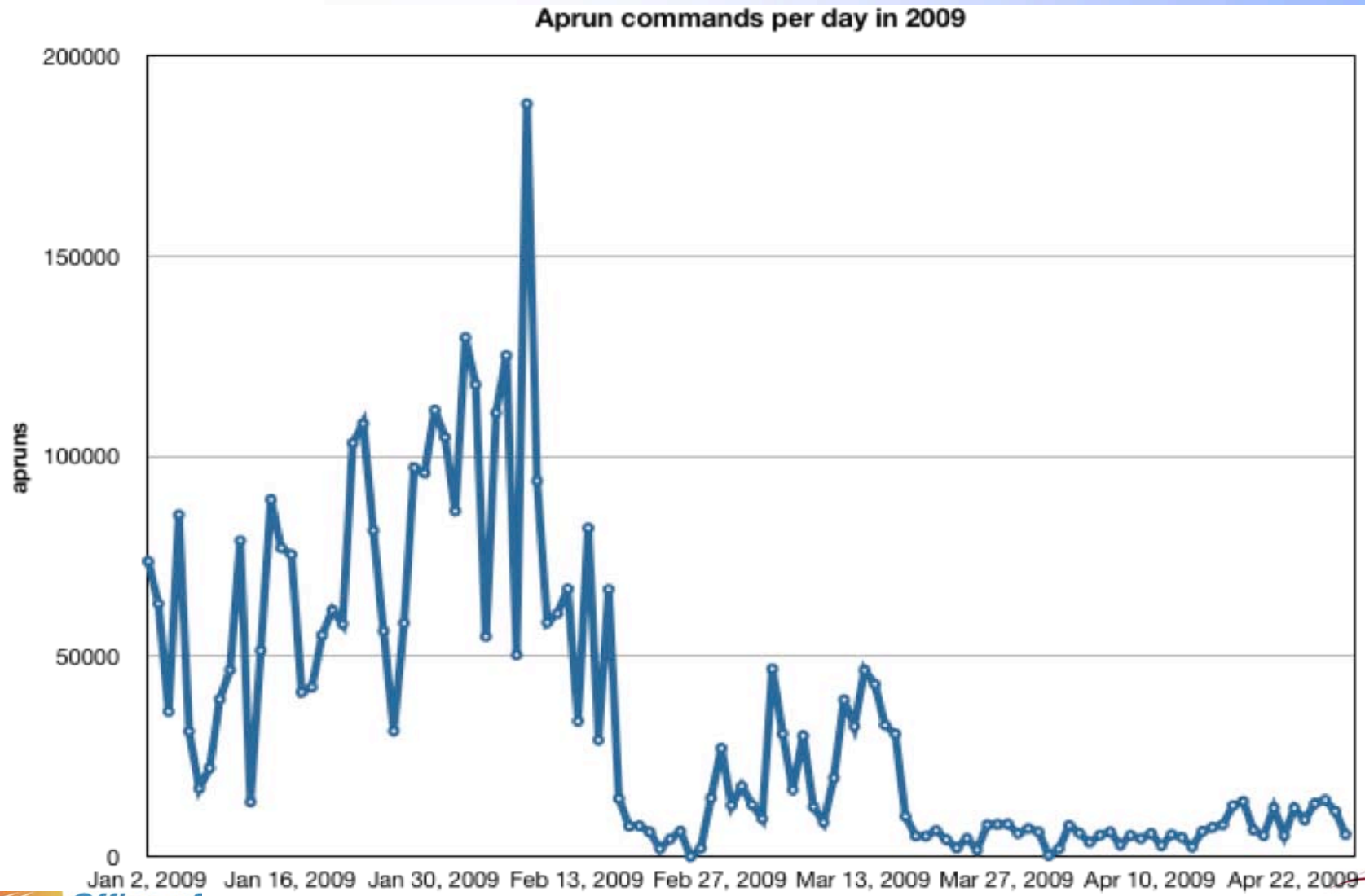




Data from Tom Davis

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Workload Changes?



Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Ending Thoughts

- Since March 17th completely reformatted I/O and moved I/O modules around
- Stability and performance of machine increased
- So can not do tests on Franklin system to recreate situation
- Priority is to keep I/O well performing, no regression
- Stresses the importance of performance monitoring
- Changes in one part of system can have unintended (but positive) consequences