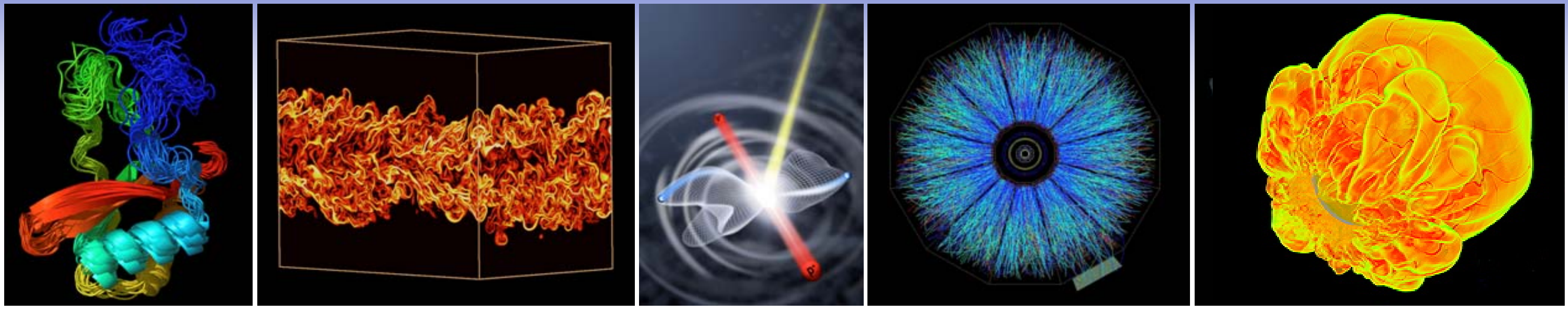


Monitoring Lustre I/O on the Franklin Cray XT4

Andrew Uselton

NERSC

acuselton@lbl.gov

I/O on Franklin

- Early Franklin file system performance was not where we wanted it
  - Interactive responsiveness at the command line
  - Bandwidth in parallel I/O
- Things have gotten better
  - LDAP changes improved responsiveness
  - We are better able to understand bandwidth issues
- We are looking for ways make things even better
  - Monitoring

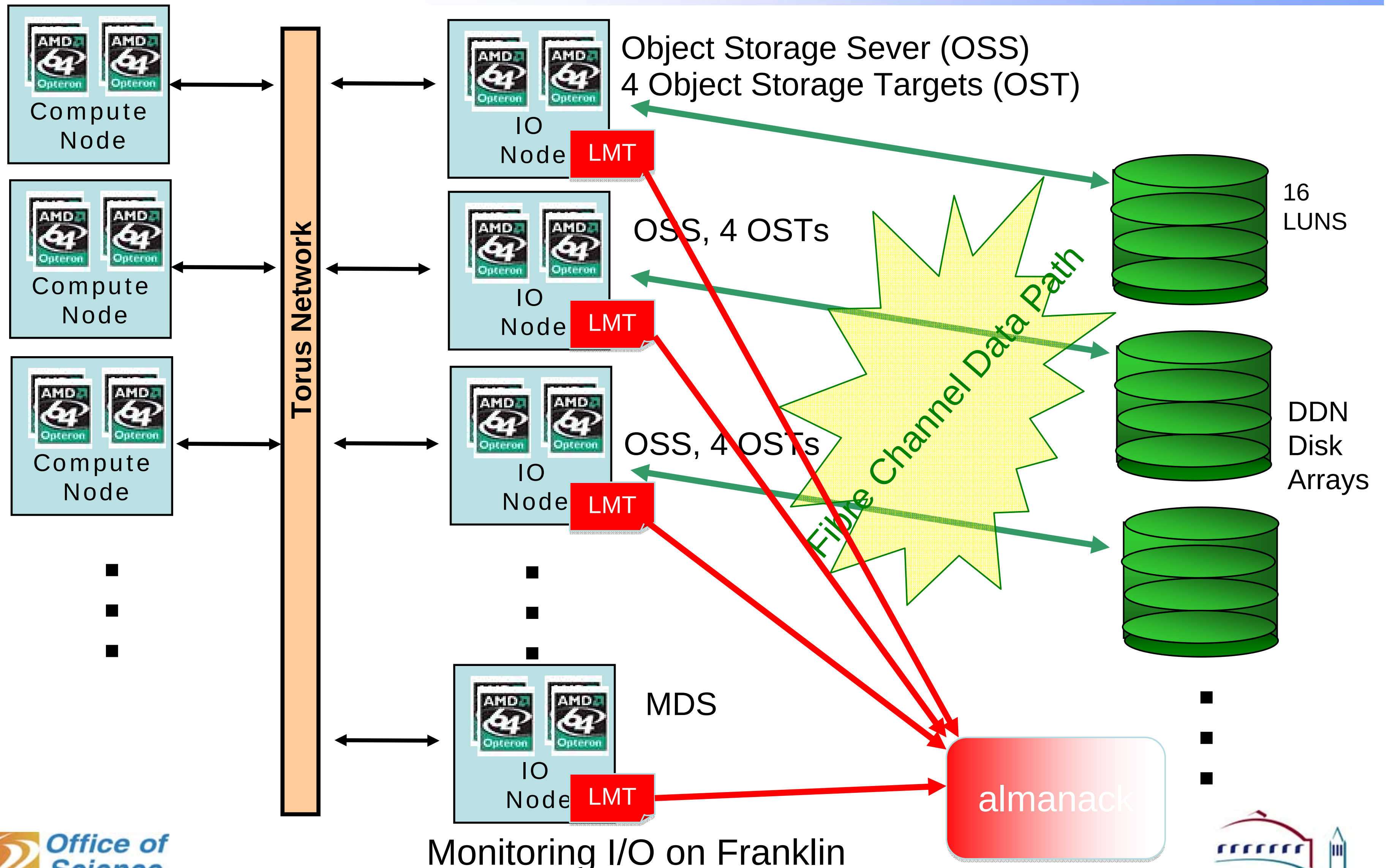
The Uses of Monitoring

- After-the-fact incident investigation
- Checking the health of the system
- Noticing and exploring anomalous behavior
- Documenting use statistics
- System analysis and tuning

Overview

- The Franklin I/O Pipeline
- The Lustre Monitoring Tool (LMT)
- Accessing LMT data
- Analyzing I/O performance

The Franklin I/O Pipeline



Monitoring I/O on Franklin

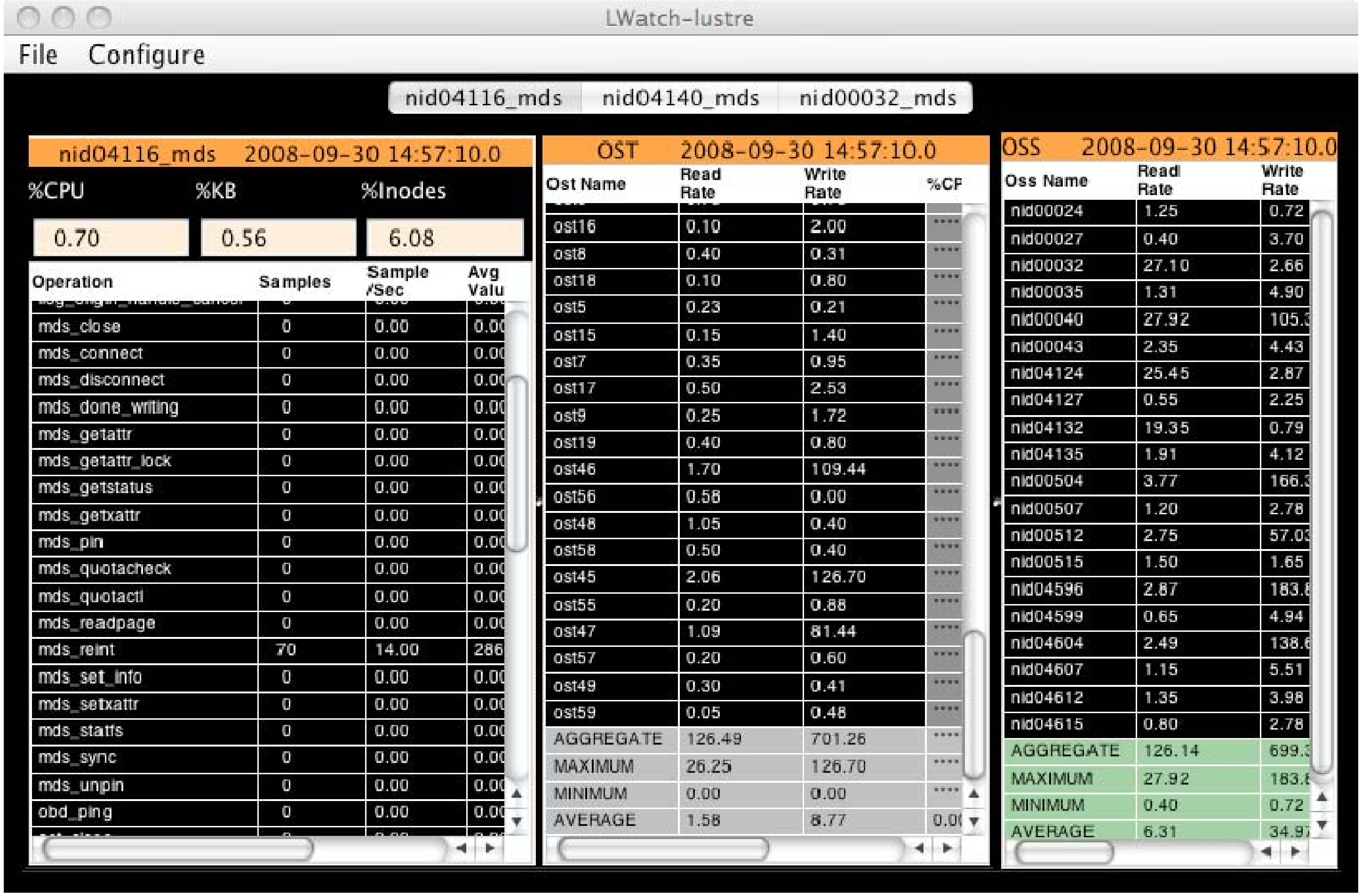
The Lustre Monitoring Tool

The Lustre Monitoring Tool (LMT) is a set of plug-in modules for Cerebro, which is a monitoring package (resembling Ganglia). LMT and Cerebro were developed at LLNL.

LMT gathers data from /proc on Lustre servers and ships it to a database.

Additional tools query the data from there.

The LMT Dashboard



nid04116_mds 2008-09-30 14:57:10.0

%CPU	%KB	%Inodes
0.70	0.56	6.08

Operation	Samples	Sample /Sec	Avg Valu
mds_close	0	0.00	0.0
mds_connect	0	0.00	0.0
mds_disconnect	0	0.00	0.0
mds_done_writing	0	0.00	0.0
mds_getattr	0	0.00	0.0
mds_getattr_lock	0	0.00	0.0
mds_getstatus	0	0.00	0.0
mds_getxattr	0	0.00	0.0
mds_pin	0	0.00	0.0
mds_quotacheck	0	0.00	0.0
mds_quotactl	0	0.00	0.0
mds_readpage	0	0.00	0.0
mds_reint	70	14.00	286
mds_set_info	0	0.00	0.0
mds_setxattr	0	0.00	0.0
mds_statfs	0	0.00	0.0
mds_sync	0	0.00	0.0
mds_unpin	0	0.00	0.0
obd_ping	0	0.00	0.0

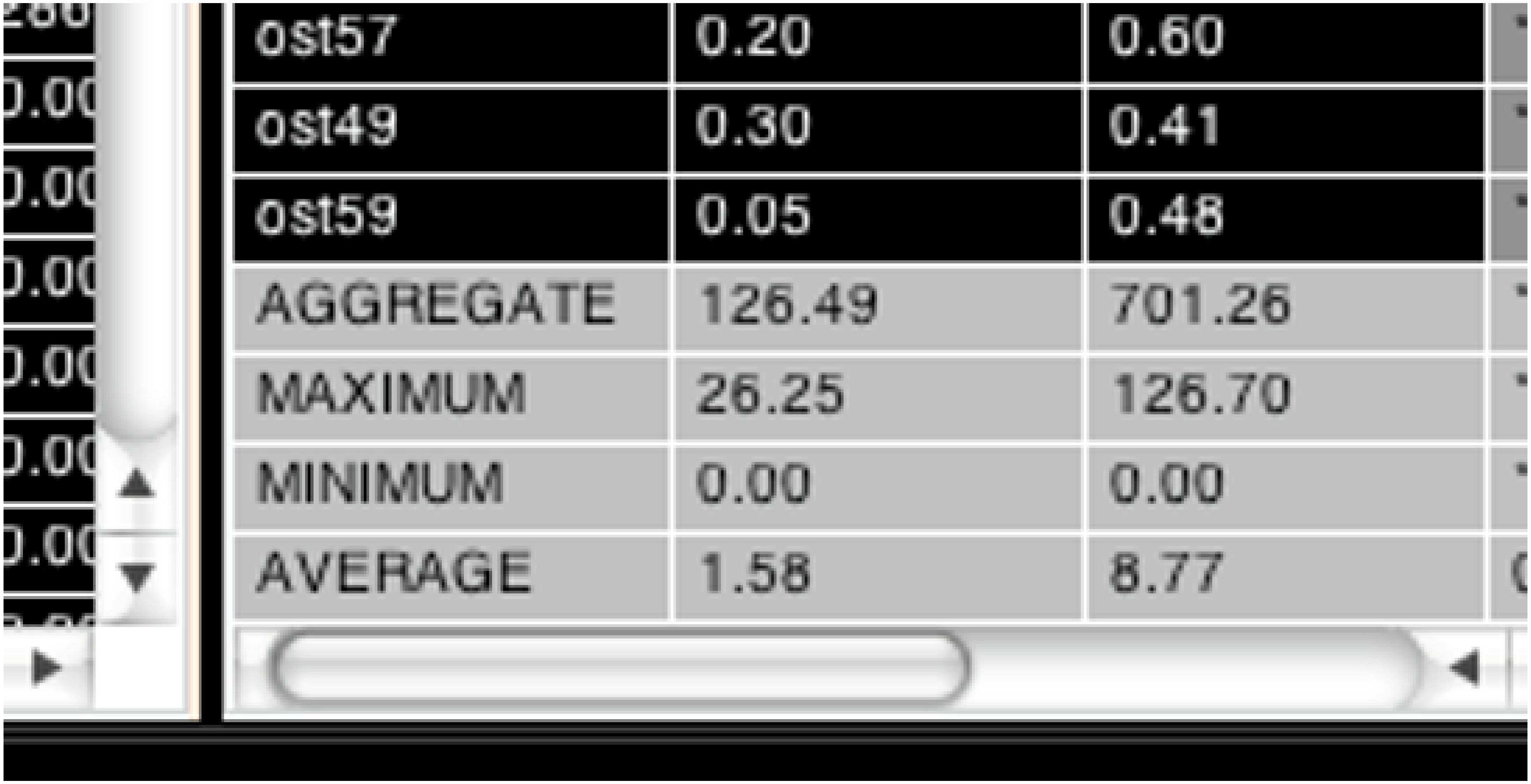
OST 2008-09-30 14:57:10.0

Ost Name	Read Rate	Write Rate	%CP
ost16	0.10	2.00	****
ost8	0.40	0.31	****
ost18	0.10	0.80	****
ost5	0.23	0.21	****
ost15	0.15	1.40	****
ost7	0.35	0.95	****
ost17	0.50	2.53	****
ost9	0.25	1.72	****
ost19	0.40	0.80	****
ost46	1.70	109.44	****
ost56	0.58	0.00	****
ost48	1.05	0.40	****
ost58	0.50	0.40	****
ost45	2.06	126.70	****
ost55	0.20	0.88	****
ost47	1.09	81.44	****
ost57	0.20	0.60	****
ost49	0.30	0.41	****
ost59	0.05	0.48	****
AGGREGATE	126.49	701.26	****
MAXIMUM	26.25	126.70	****
MINIMUM	0.00	0.00	****
AVERAGE	1.58	8.77	0.0

OSS 2008-09-30 14:57:10.0

Oss Name	Read Rate	Write Rate
nid00024	1.25	0.72
nid00027	0.40	3.70
nid00032	27.10	2.66
nid00035	1.31	4.90
nid00040	27.92	105.3
nid00043	2.35	4.43
nid04124	25.45	2.87
nid04127	0.55	2.25
nid04132	19.35	0.79
nid04135	1.91	4.12
nid00504	3.77	166.3
nid00507	1.20	2.78
nid00512	2.75	57.03
nid00515	1.50	1.65
nid04596	2.87	183.8
nid04599	0.65	4.94
nid04604	2.49	138.6
nid04607	1.15	5.51
nid04612	1.35	3.98
nid04615	0.80	2.78
AGGREGATE	126.14	699.3
MAXIMUM	27.92	183.8
MINIMUM	0.40	0.72
AVERAGE	6.31	34.97

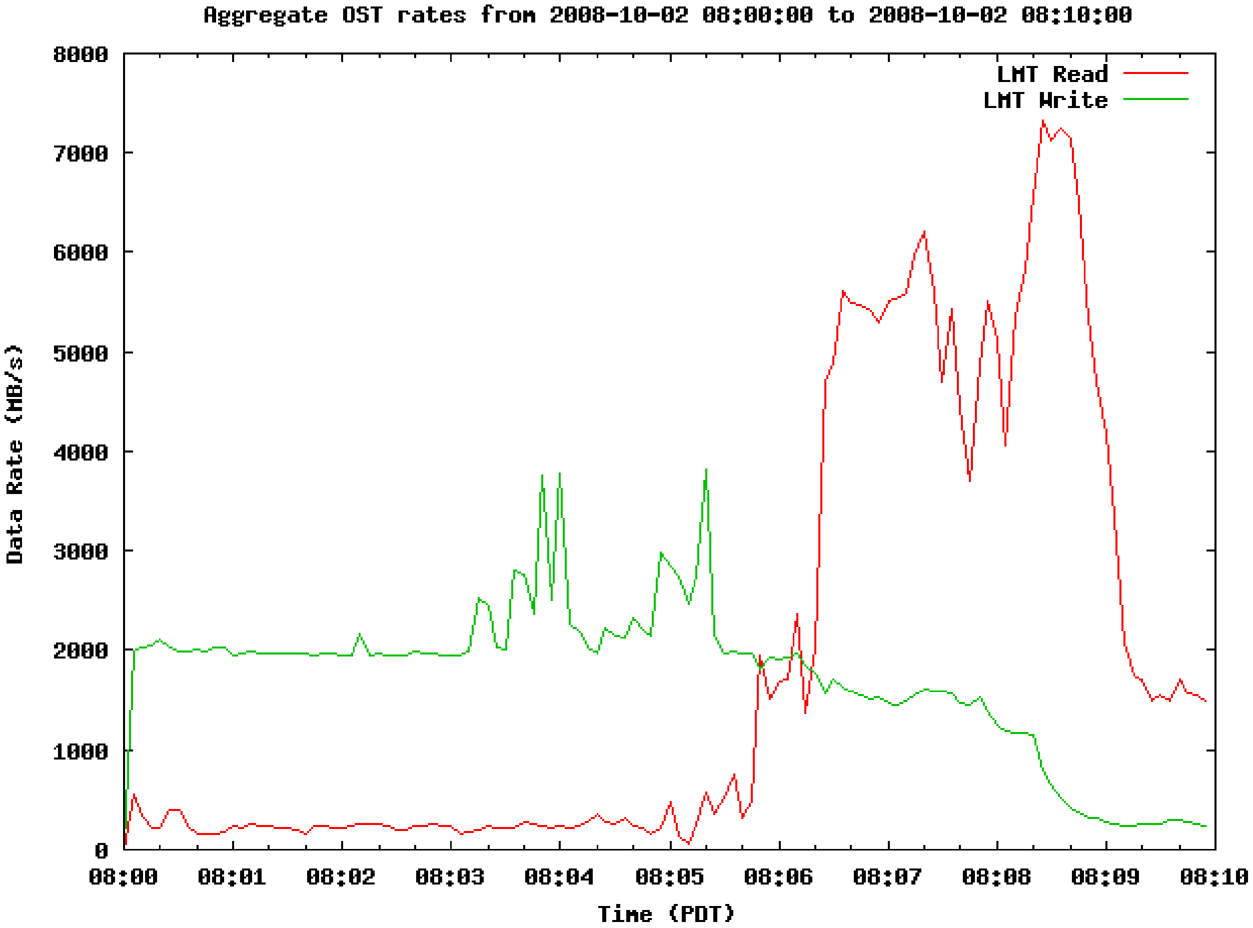
Detail from the Dashboard



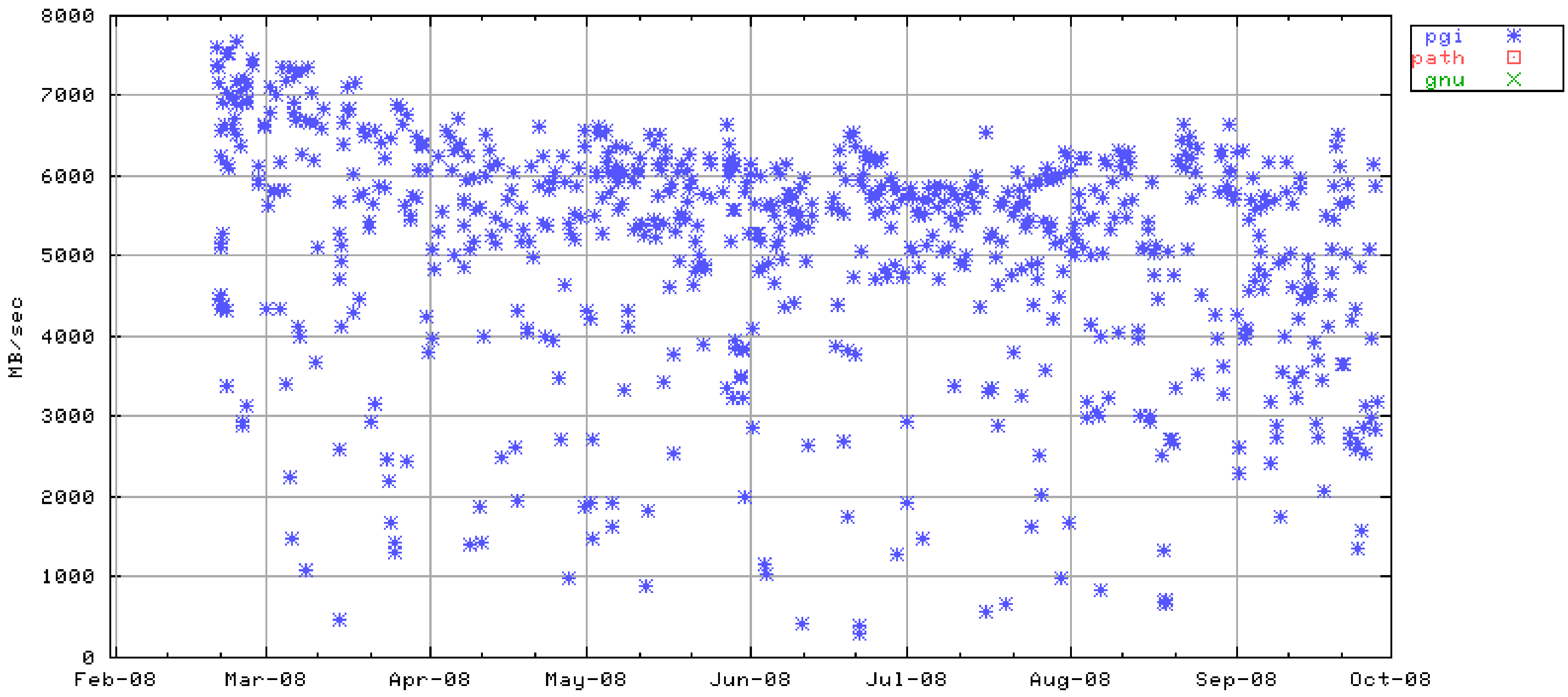
ost57	0.20	0.60
ost49	0.30	0.41
ost59	0.05	0.48
AGGREGATE	126.49	701.26
MAXIMUM	26.25	126.70
MINIMUM	0.00	0.00
AVERAGE	1.58	8.77

LMT from the Command Line

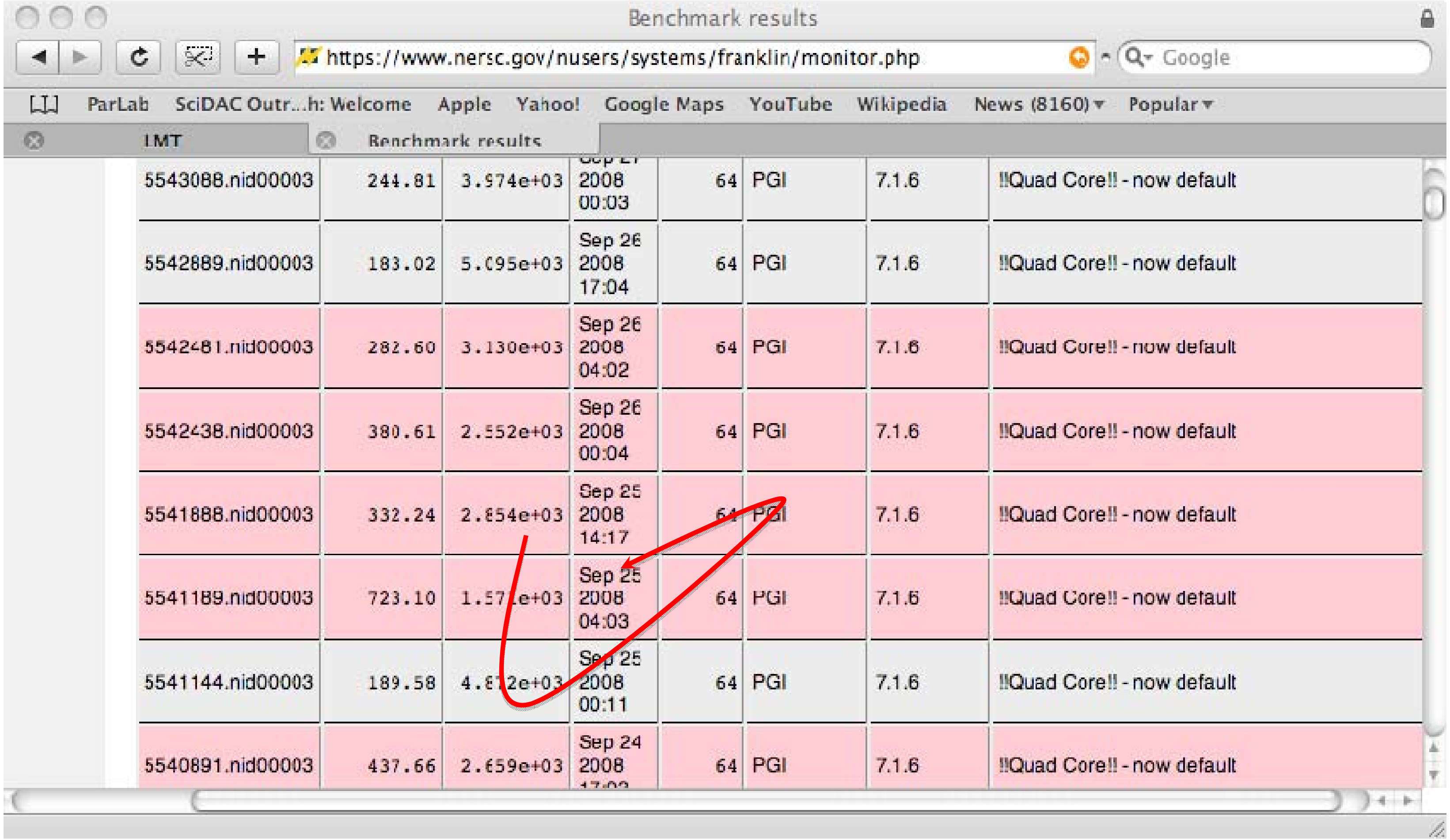
```
bash-3.2$ /usr/share/lmt/bin/lmt.pl -s "2008-10-02 08:00:00" -e "2008-10-02 08:10:00"
```



Recent IOR Test Results (reads)

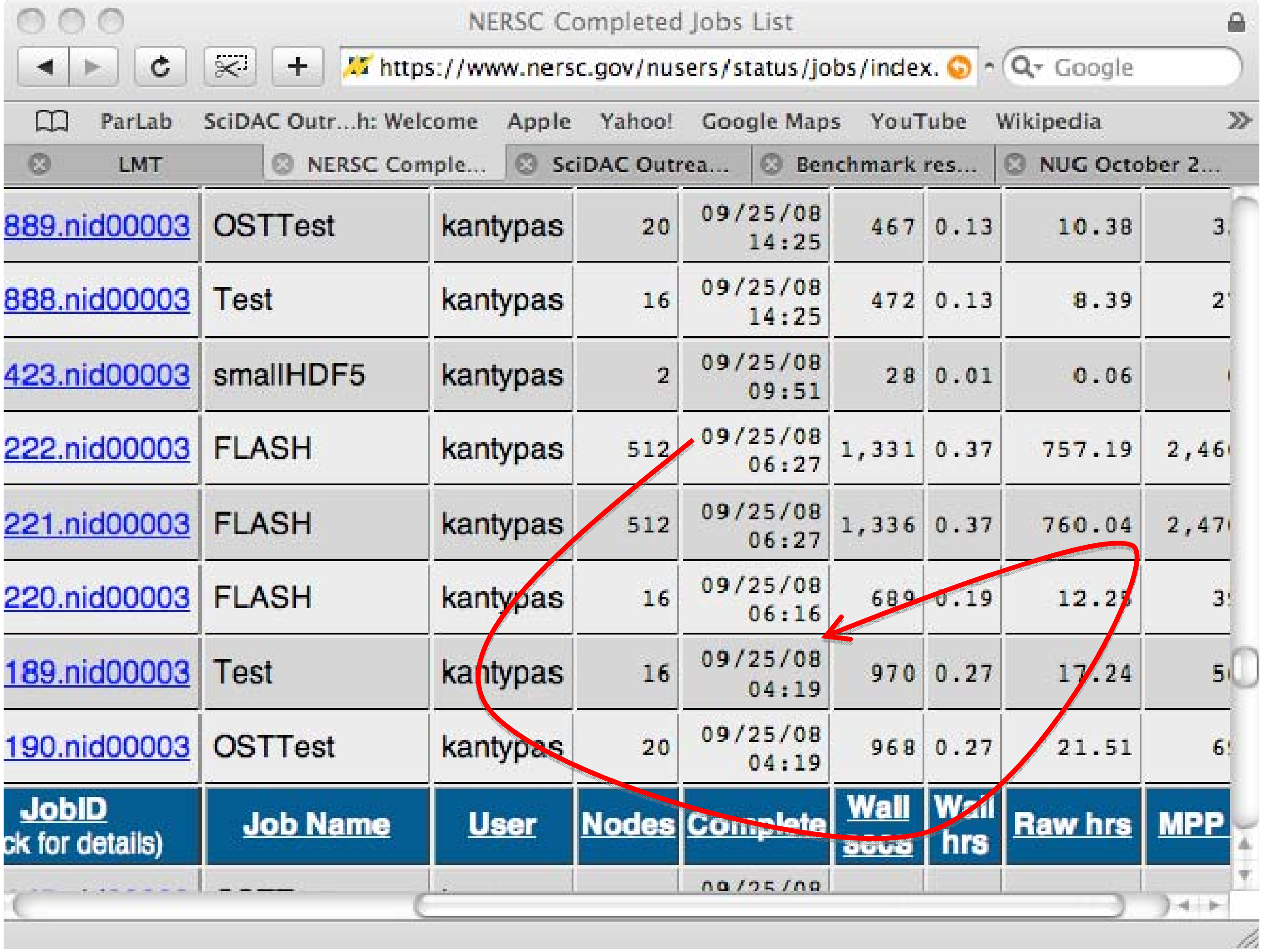


A Particularly Slow IOR



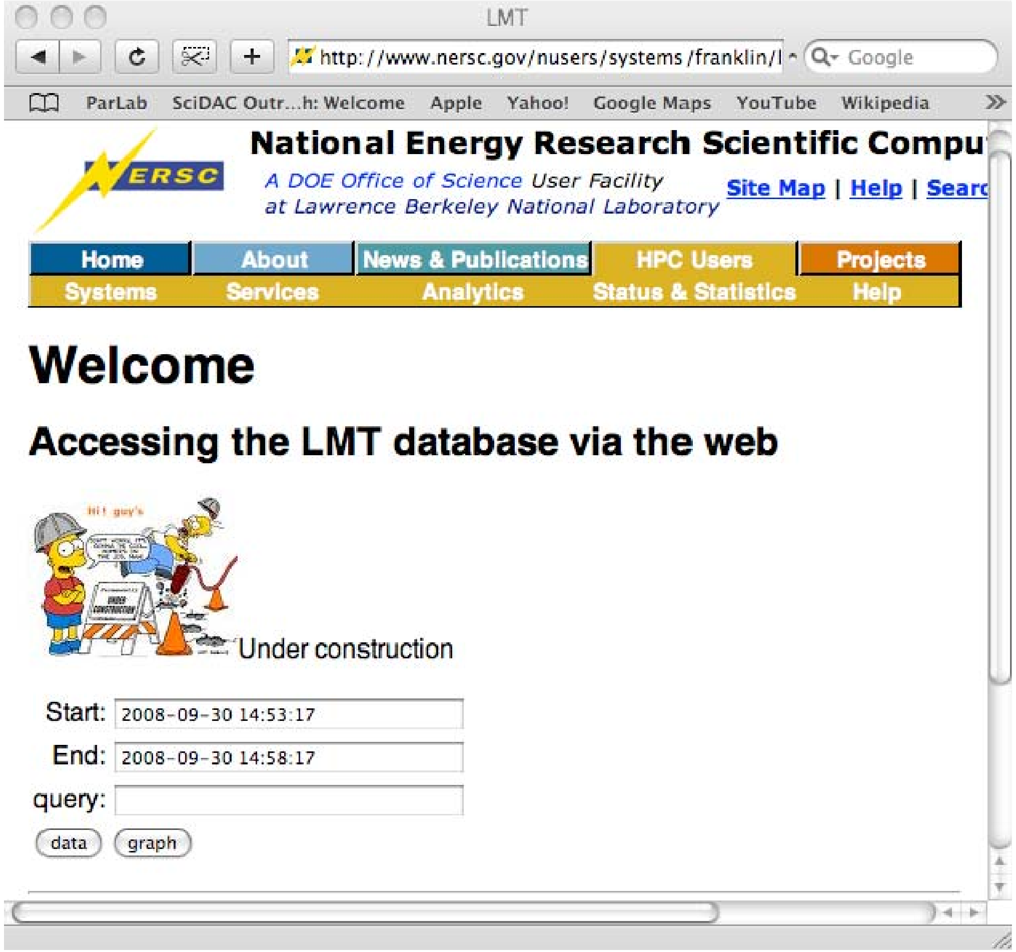
5543088.nid00003	244.81	3.974e+03	Sep 27 2008 00:03	64	PGI	7.1.6	!!Quad Core!! - now default
5542889.nid00003	183.02	5.095e+03	Sep 26 2008 17:04	64	PGI	7.1.6	!!Quad Core!! - now default
5542481.nid00003	282.60	3.130e+03	Sep 26 2008 04:02	64	PGI	7.1.6	!!Quad Core!! - now default
5542438.nid00003	380.61	2.552e+03	Sep 26 2008 00:04	64	PGI	7.1.6	!!Quad Core!! - now default
5541888.nid00003	332.24	2.854e+03	Sep 25 2008 14:17	64	PGI	7.1.6	!!Quad Core!! - now default
5541189.nid00003	723.10	1.571e+03	Sep 25 2008 04:03	64	PGI	7.1.6	!!Quad Core!! - now default
5541144.nid00003	189.58	4.872e+03	Sep 25 2008 00:11	64	PGI	7.1.6	!!Quad Core!! - now default
5540891.nid00003	437.66	2.659e+03	Sep 24 2008 17:02	64	PGI	7.1.6	!!Quad Core!! - now default

What Happened During My Job?



JobID ck for details)	Job Name	User	Nodes	Complete	Wall secs	Wall hrs	Raw hrs	MPP
889.nid00003	OSTTest	kantypas	20	09/25/08 14:25	467	0.13	10.38	3.
888.nid00003	Test	kantypas	16	09/25/08 14:25	472	0.13	8.39	2
423.nid00003	smallHDF5	kantypas	2	09/25/08 09:51	28	0.01	0.06	
222.nid00003	FLASH	kantypas	512	09/25/08 06:27	1,331	0.37	757.19	2,46
221.nid00003	FLASH	kantypas	512	09/25/08 06:27	1,336	0.37	760.04	2,47
220.nid00003	FLASH	kantypas	16	09/25/08 06:16	689	0.19	12.25	3
189.nid00003	Test	kantypas	16	09/25/08 04:19	970	0.27	17.24	5
190.nid00003	OSTTest	kantypas	20	09/25/08 04:19	968	0.27	21.51	6

Accessing LMT Data on the Web



LMT
http://www.nersc.gov/nusers/systems/franklin/l
Google
ParLab SciDAC Outr…h: Welcome Apple Yahoo! Google Maps YouTube Wikipedia
Home About News & Publications HPC Users Projects
Systems Services Analytics Status & Statistics Help
data
graph

National Energy Research Scientific Compu

A DOE Office of Science User Facility at Lawrence Berkeley National Laboratory

Site Map | Help | Searc

Welcome

Accessing the LMT database via the web

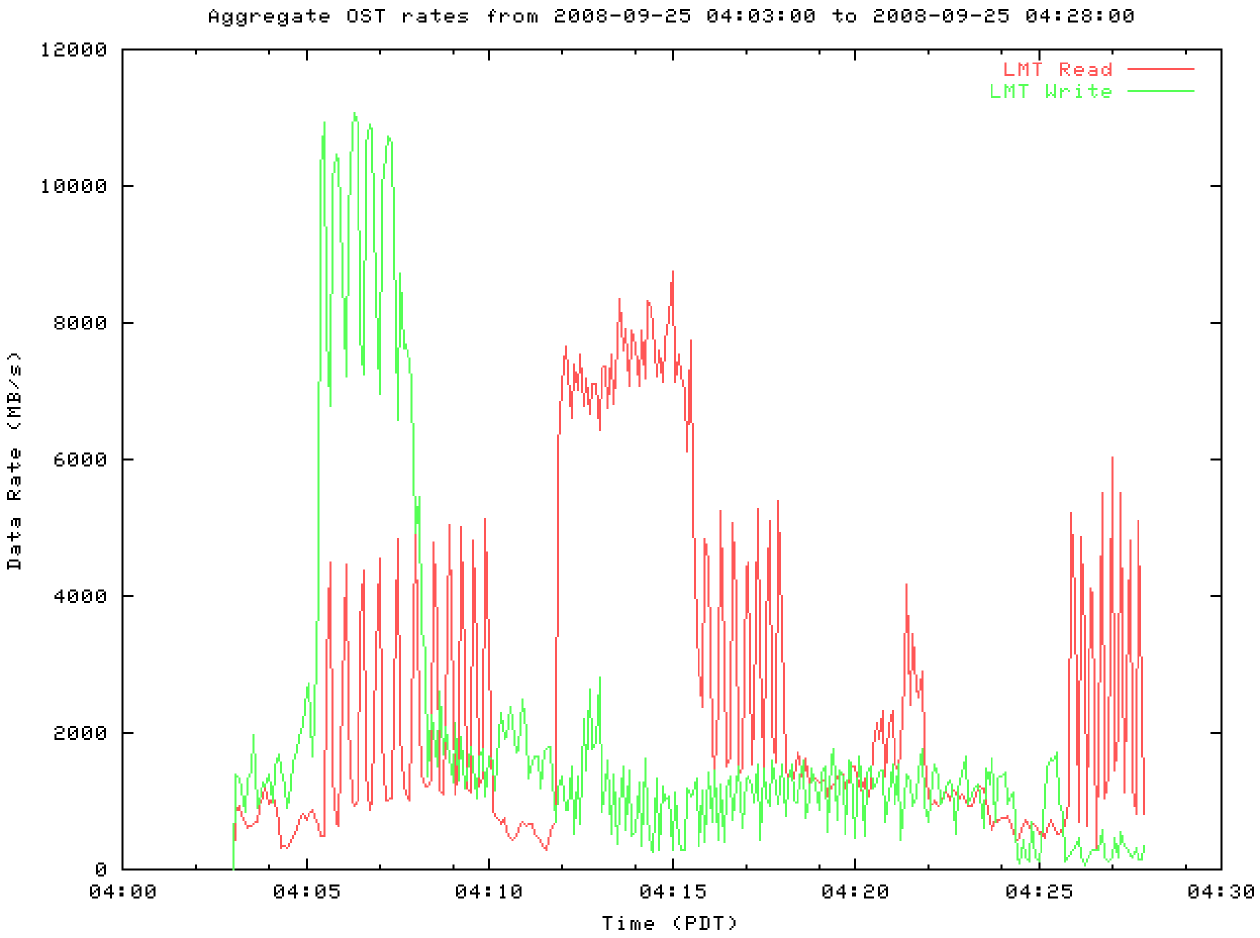
Under construction

Start: 2008-09-30 14:53:17

End: 2008-09-30 14:58:17

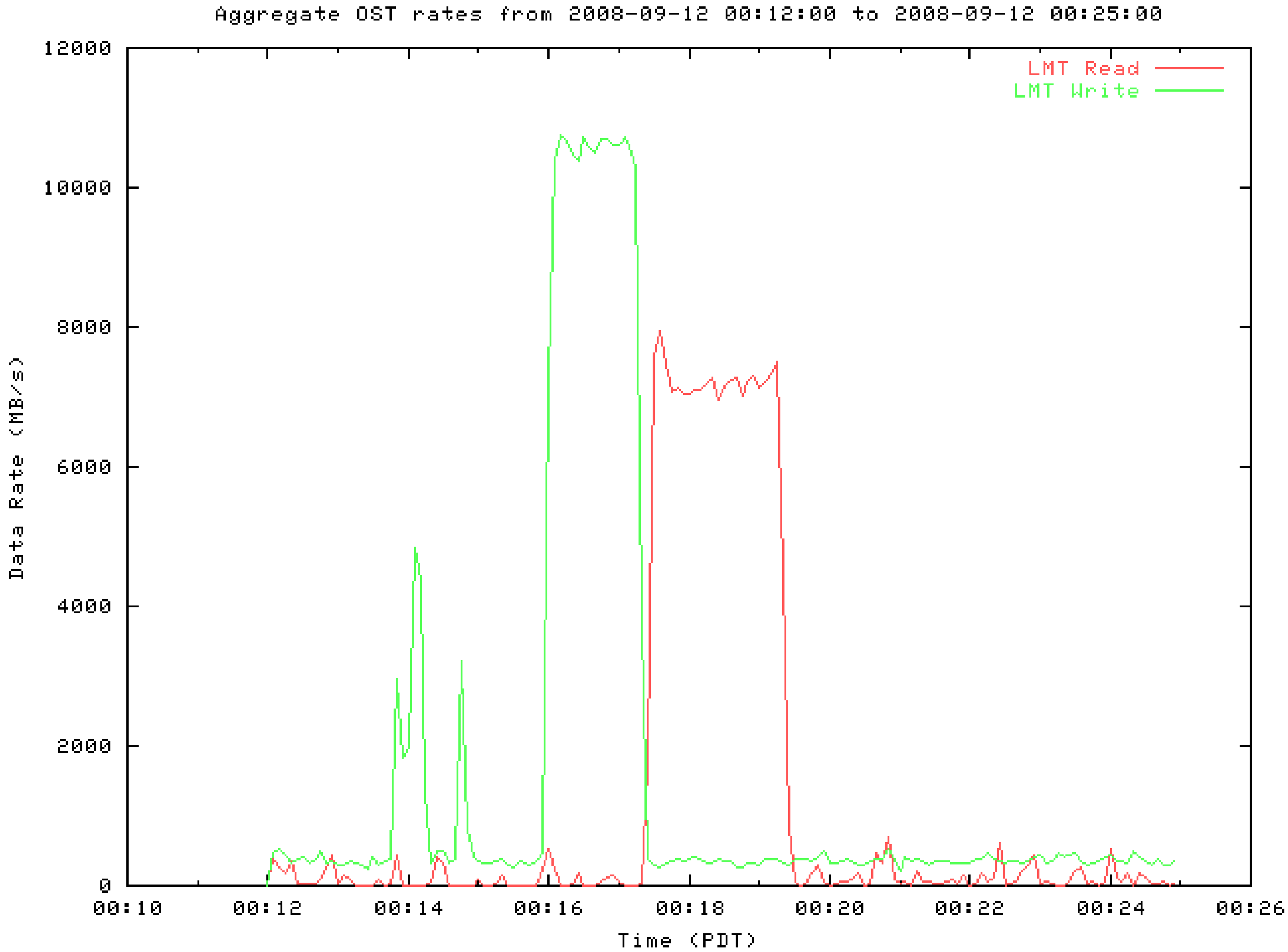
query:

The View from LMT

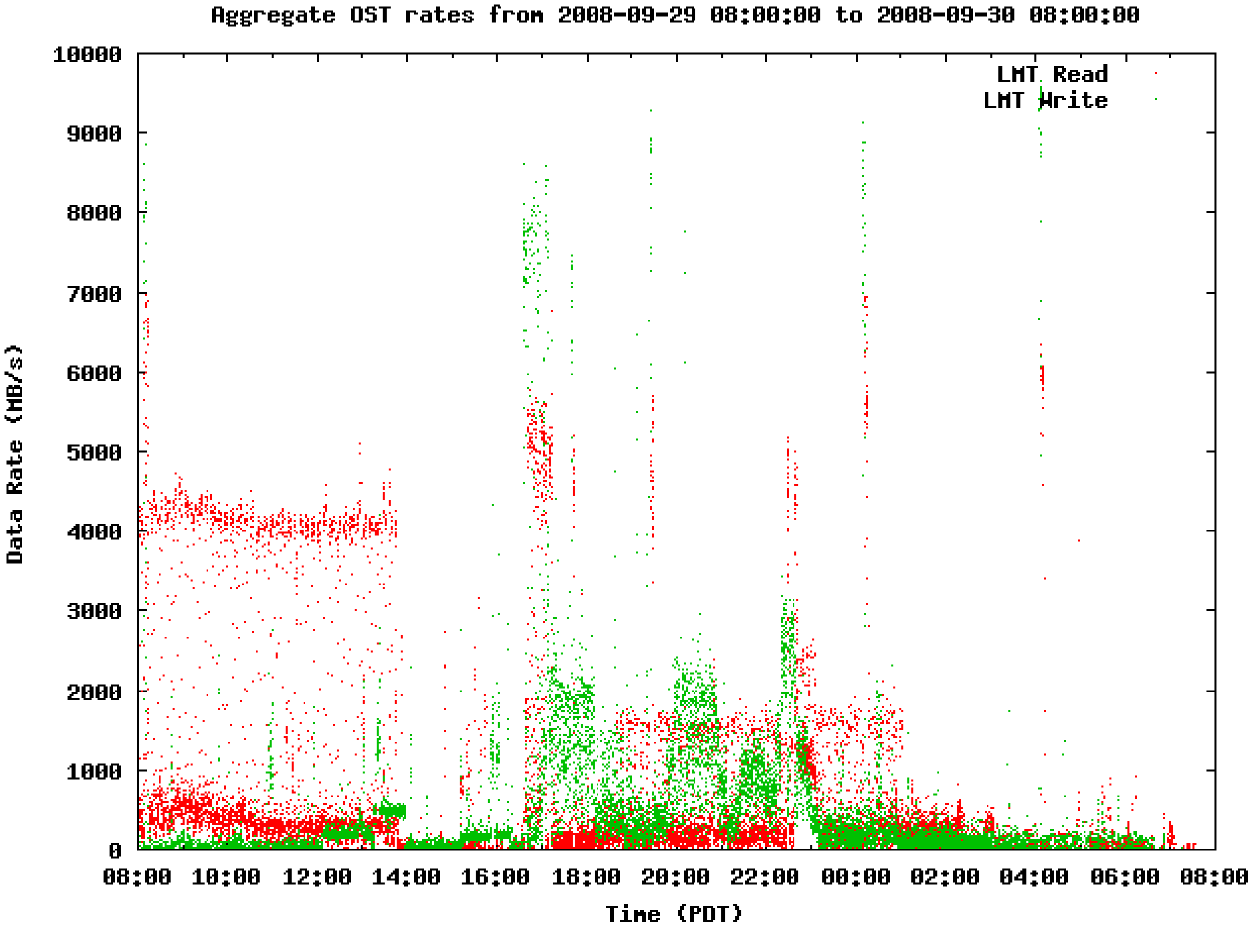


Monitoring I/O on Franklin

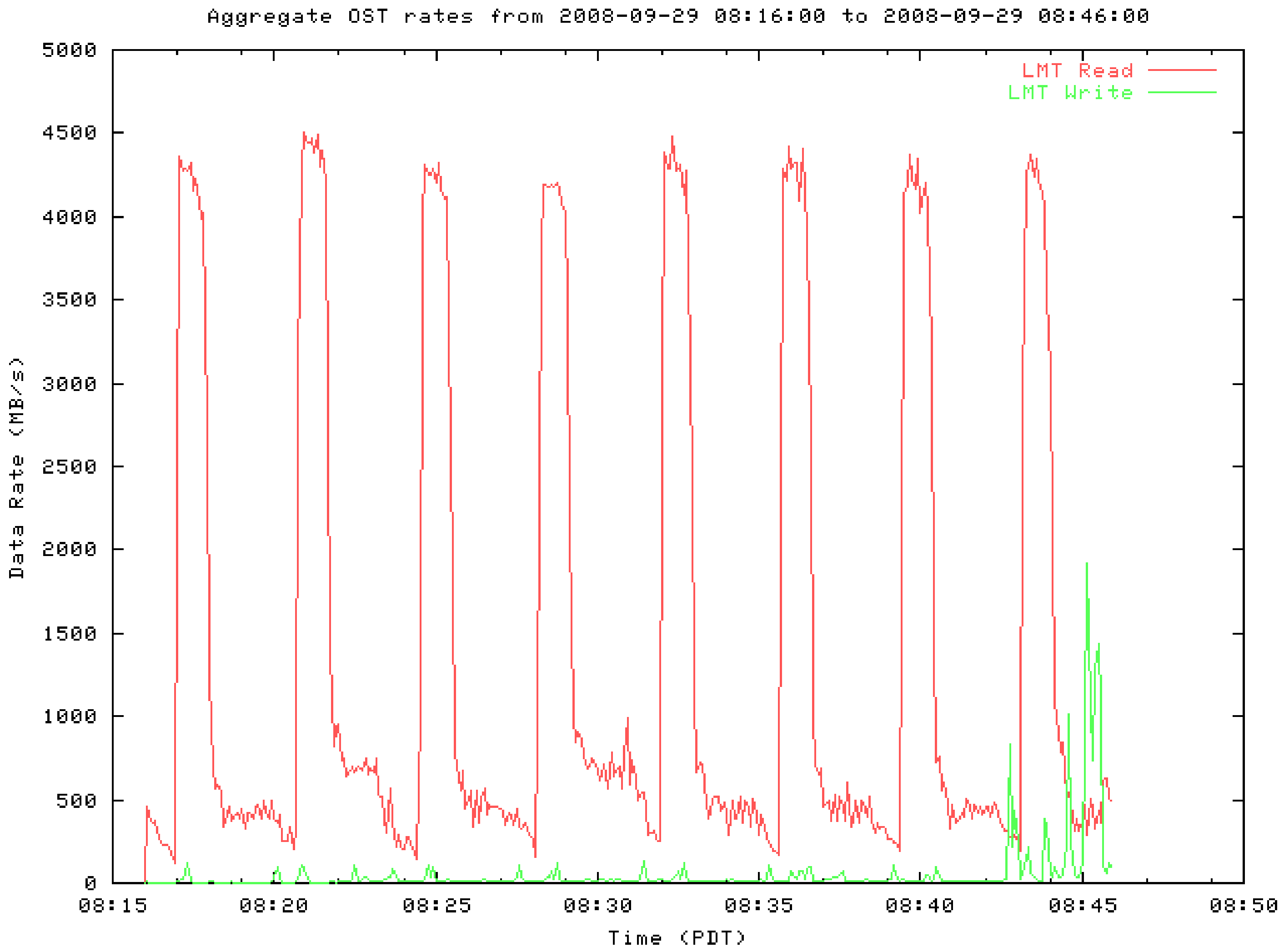
A “Clean” IOR Run



24 hours of I/O



Zooming in on the Reads



Useful Links

- **Franklin LMT web interface (may change)**

  http://www.nersc.gov/nusers/systems/franklin/lmt/
- **Franklin benchmarking page**

  https://www.nersc.gov/nusers/systems/franklin/monitor.php
- **LMT source code**

  http://sourceforge.net/projects/lmt/
- **Cerebro source code**

  http://sourceforge.net/projects/cerebro/