PDSI Workshop
Super Computing 2006

PNNL EMSL Lustre Activities

November 2006

Evan Felix

evan.felix@pnl.gov

William R. Wiley
Environmental Molecular Sciences Laboratory



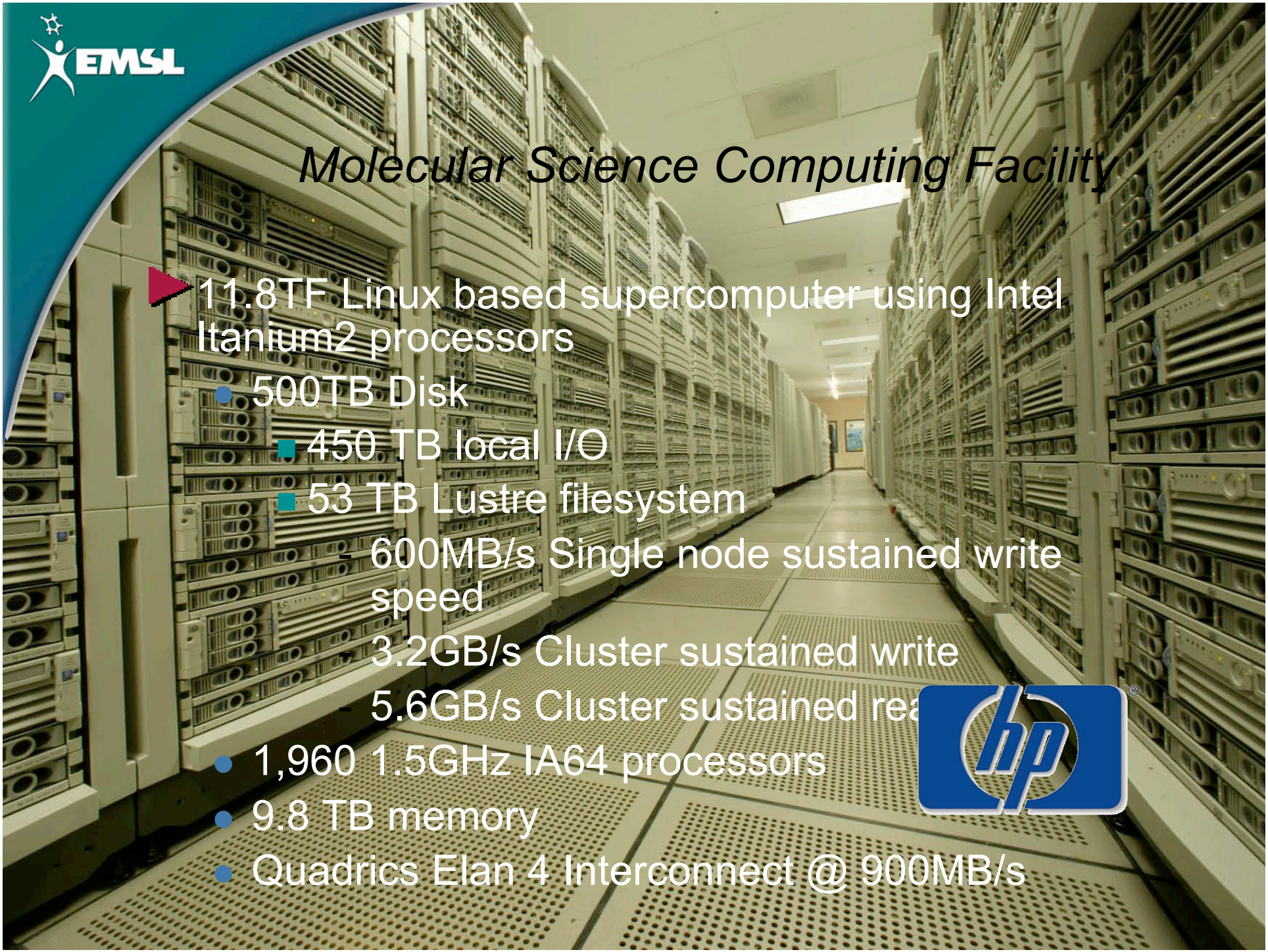



- **Who are we?**
  - A 200,000 square-foot U.S. Department of Energy national scientific user facility
  - Operated by Pacific Northwest National Laboratory in Richland, Washington
- **What we provide for users**
  - Free access to over 100 state-of-the-art research instruments
  - A peer-review proposal process
  - Expert staff to assist or collaborate
- **Why use EMSL?**
  - EMSL provides - under one roof - staff and instruments for fundamental research on physical, chemical, and biological processes.

Agenda & Stats

- MPP2(HPCS2) - Lustre Scratch File System
- NWFS v2 – Archive Lustre File System
- MPP2(HPCS2) – Lustre Scalability Test Runs

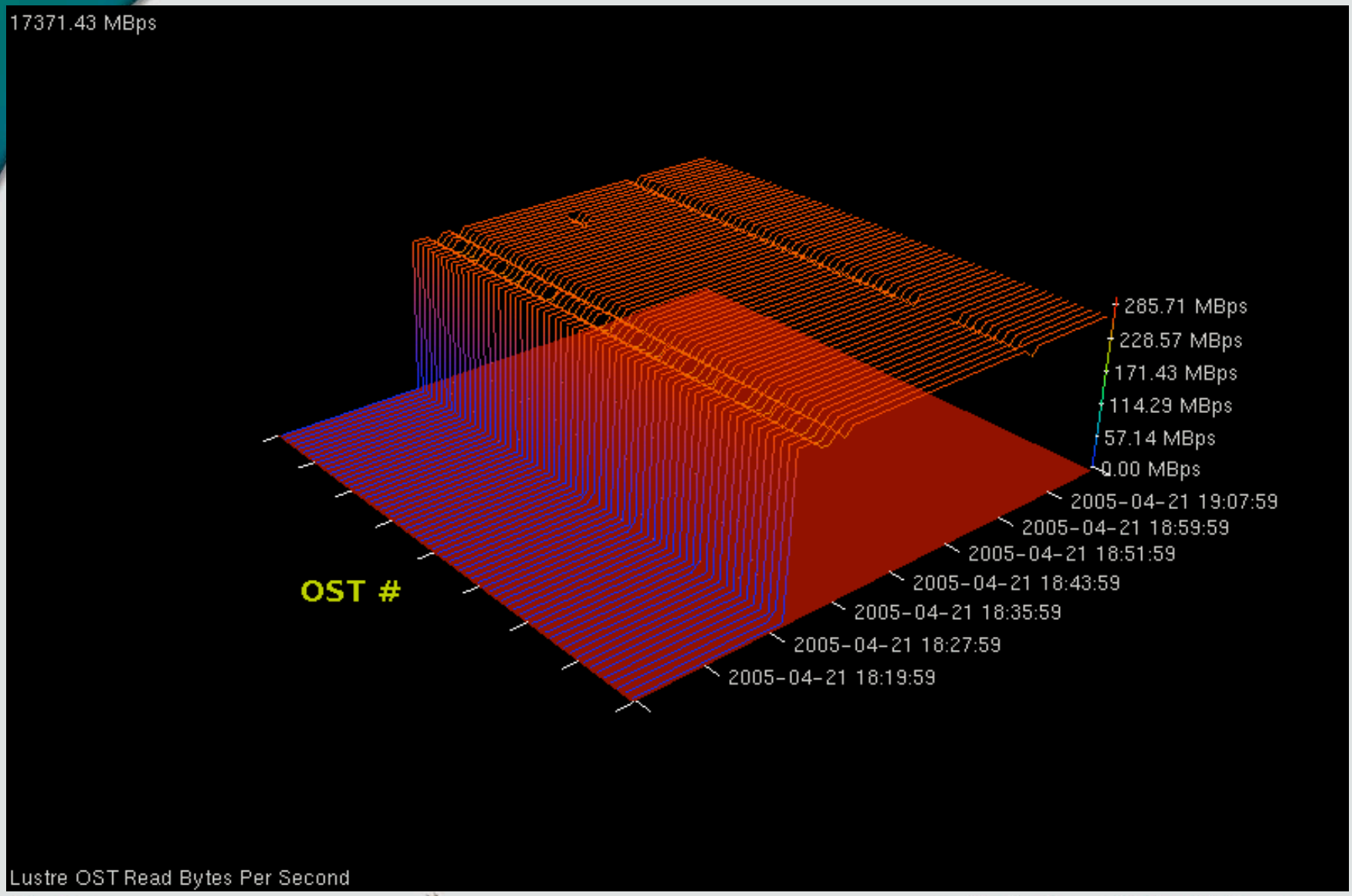
System	Kernel/Lustre Version	Packager	#OSS/# OST's	Stripe Width	Interconnect
MPP2	2.4.20+/1.4.2	HP-SFS	32/64	8	QSNet2 –elan4
NWFSv2	2.6.9/1.4.7	CFS/PNNL	34/136	1	Ethernet
MPP2 test	2.6.12/1.6beta5	PNNL	570/3990	10	QSNet2 –elan4

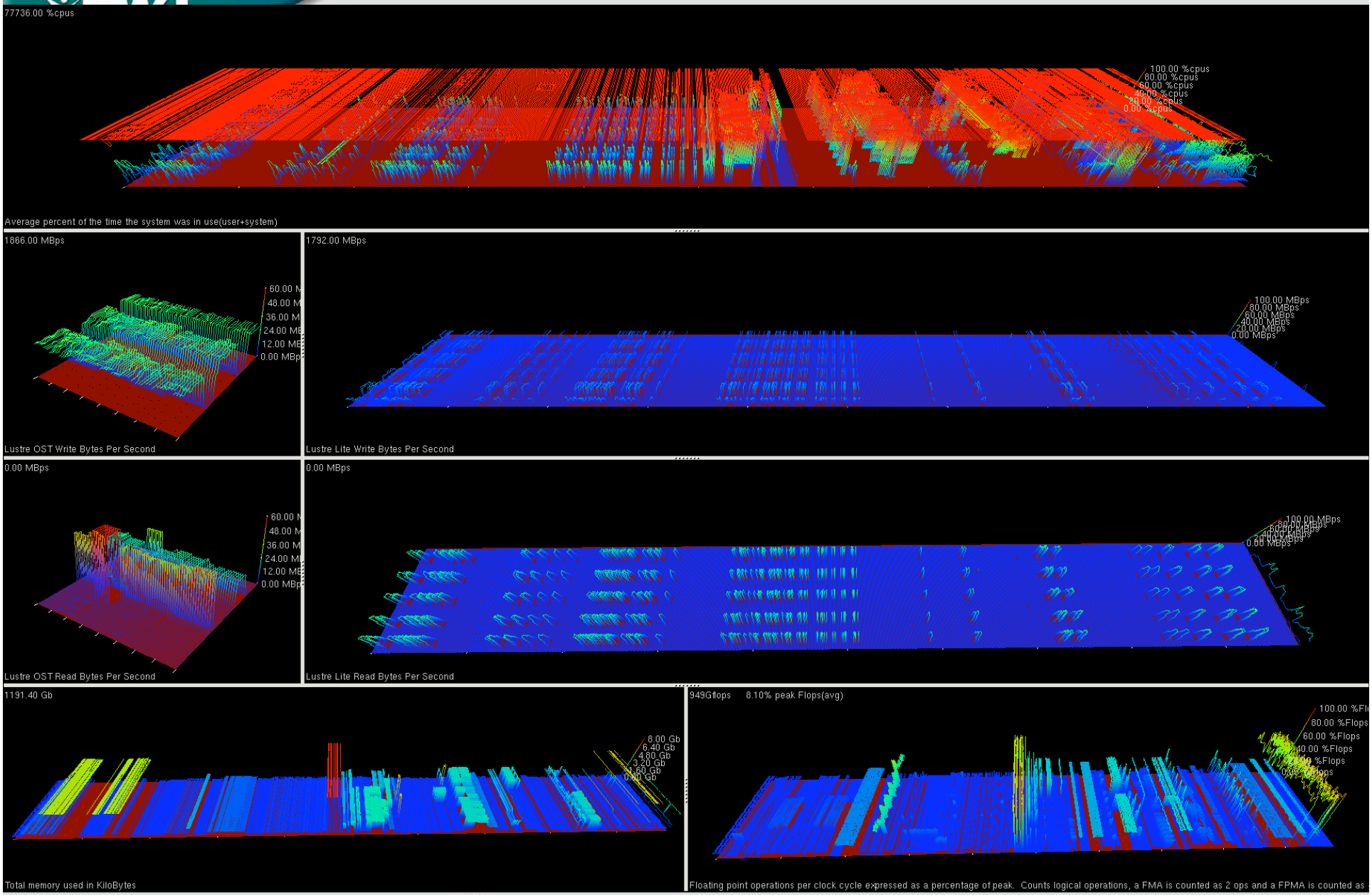

EMSL
Molecular Science Computing Facility
11.8TF Linux based supercomputer using Intel Itanium2 processors
500TB Disk
450 TB local I/O
53 TB Lustre filesystem
600MB/s Single node sustained write speed
3.2GB/s Cluster sustained write
5.6GB/s Cluster sustained rea
1,960 1.5GHz IA64 processors
9.8 TB memory
Quadrics Elan 4 Interconnect @ 900MB/s
hp

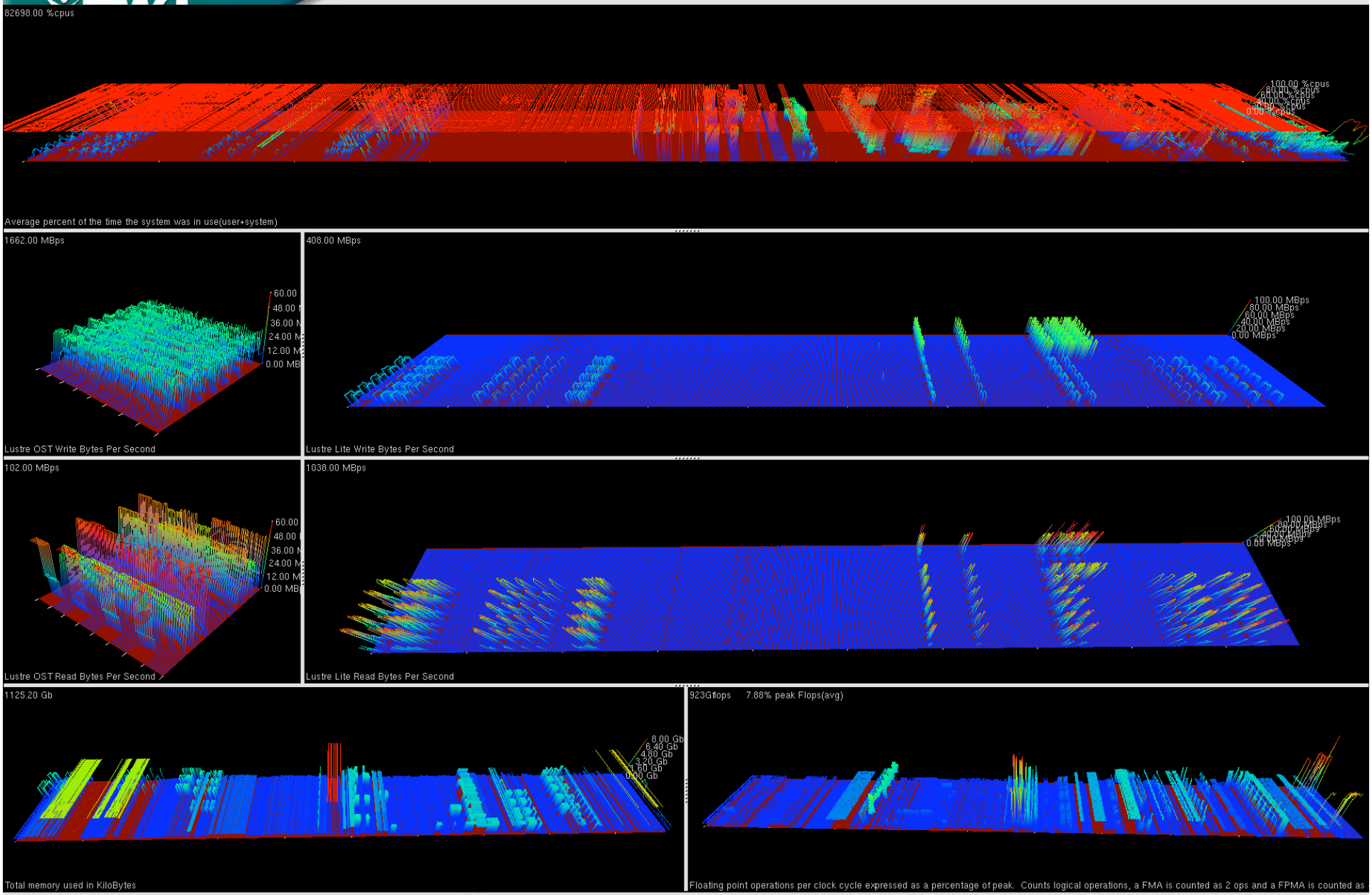
PNNL Lustre in a Nutshell

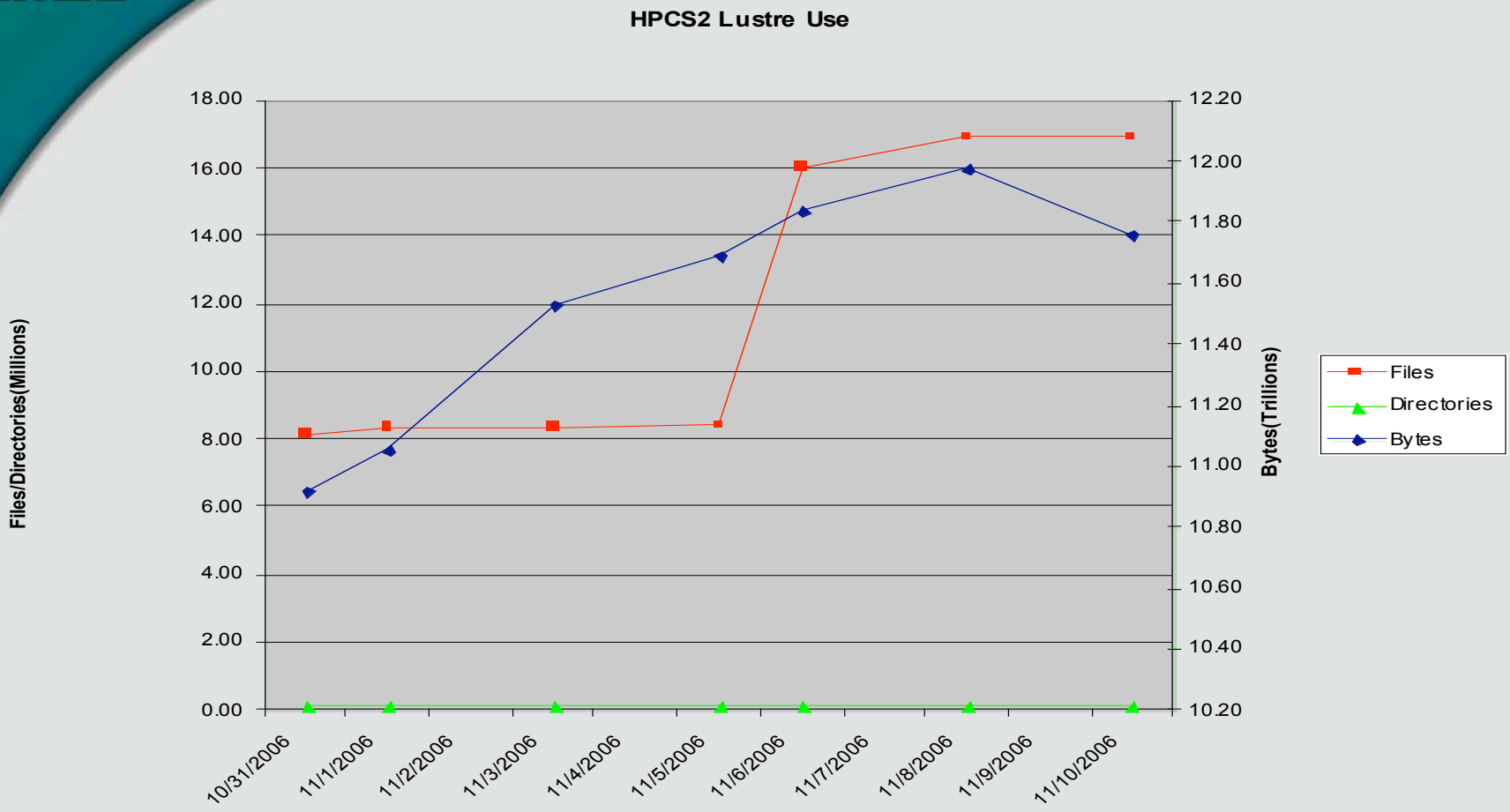
- 34 * HP rx2600 server nodes
- 32 * 2C3D HP Enterprise Virtual Arrays
- 1 * 2C2D HP Enterprise Vitual Array
- 64 * 880 GB LUN's for data
- 1 * 1TB LUN for Lustre meta data
- Dual Fibre Channel SAN

Lustre Server Read: 1024 processor

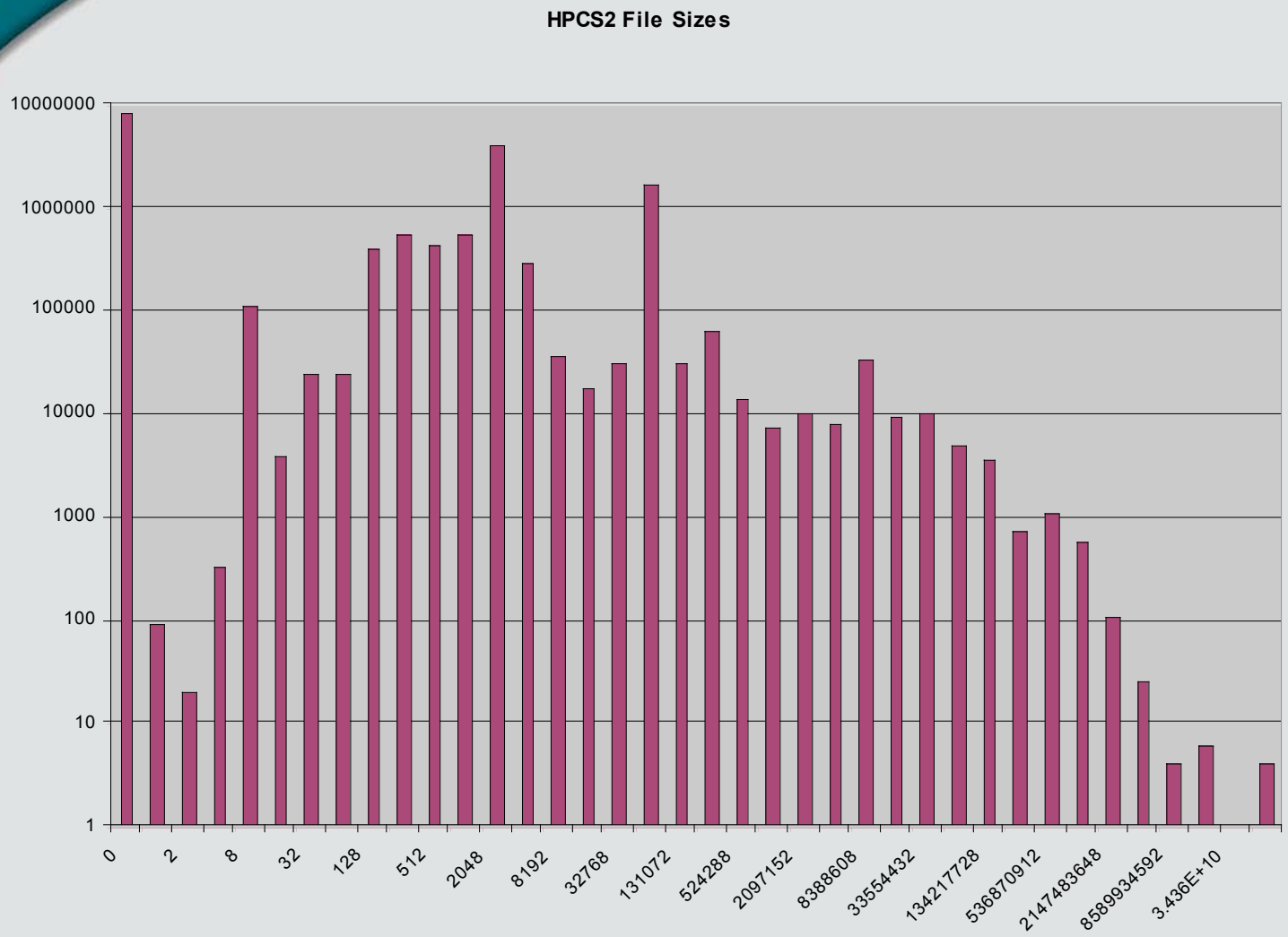


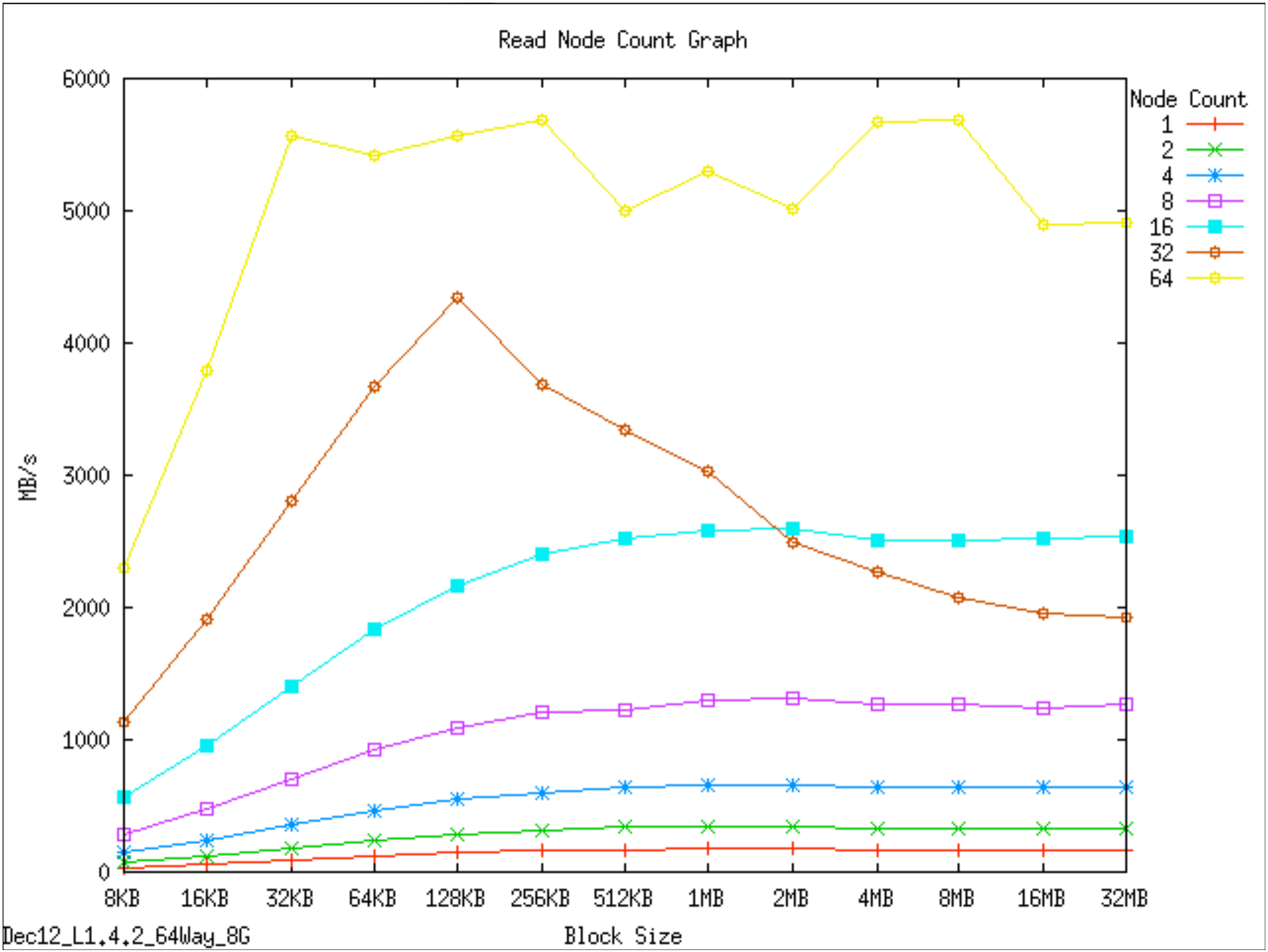

77736.00 %cpus
100.00 %cpus
80.00 %cpus
60.00 %cpus
40.00 %cpus
20.00 %cpus
0.00 %cpus
Average percent of the time the system was in use(user+system)
1866.00 MBps
60.00 M
48.00 M
36.00 M
24.00 MB
12.00 MB
0.00 MBps
Lustre OST Write Bytes Per Second
1792.00 MBps
100.00 MBps
80.00 MBps
60.00 MBps
40.00 MBps
20.00 MBps
0.00 MBps
Lustre Lite Write Bytes Per Second
0.00 MBps
60.00 M
48.00 M
36.00 M
24.00 MB
12.00 MB
0.00 MBps
Lustre OST Read Bytes Per Second
0.00 MBps
100.00 MBps
0.00 MBps
Lustre Lite Read Bytes Per Second
1191.40 Gb
8.00 Gb
6.40 Gb
4.80 Gb
3.20 Gb
1.60 Gb
0.00 Gb
Total memory used in KiloBytes
949Gflops 8.10% peak Flops(avg)
100.00 %Flops
80.00 %Flops
60.00 %Flops
40.00 %Flops
20.00 %Flops
0.00 %Flops
Floating point operations per clock cycle expressed as a percentage of peak. Counts logical operations, a FMA is counted as 2 ops and a FPMA is counted as

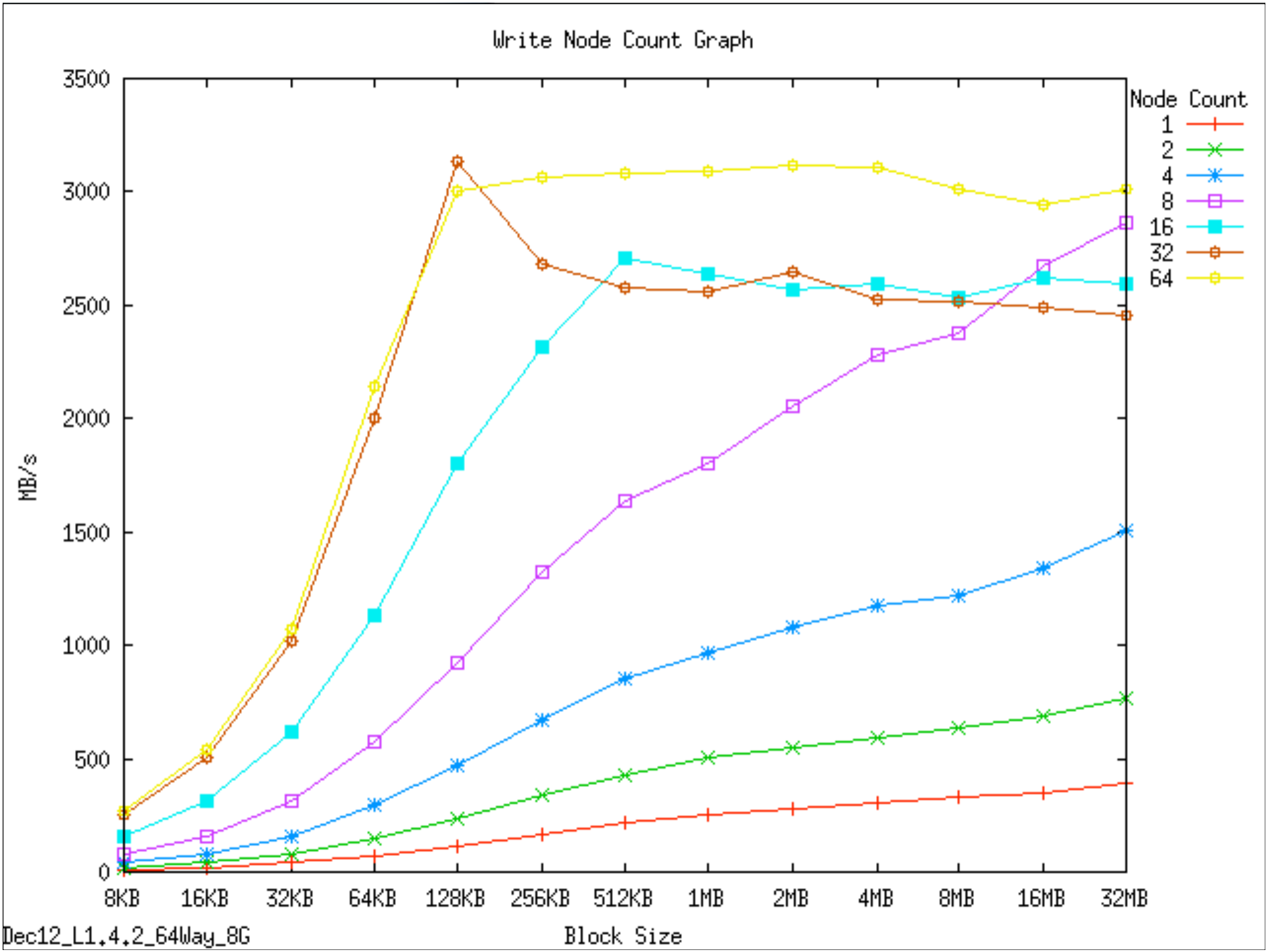

82698.00 %cpus
100.00 %cpus
Average percent of the time the system was in use(user+system)
1662.00 MBps
60.00
48.00
36.00
24.00
12.00
0.00 MB
Lustre OST Write Bytes Per Second
408.00 MBps
100.00 MBps
80.00 MBps
60.00 MBps
40.00 MBps
20.00 MBps
0.00 MBps
Lustre Lite Write Bytes Per Second
102.00 MBps
60.00
48.00
36.00
24.00
12.00
0.00 MB
Lustre OST Read Bytes Per Second
1038.00 MBps
100.00 MBps
0.00 MBps
Lustre Lite Read Bytes Per Second
1125.20 Gb
8.00 Gb
6.40 Gb
4.80 Gb
3.20 Gb
1.60 Gb
0.00 Gb
Total memory used in KiloBytes
923Gflops 7.88% peak Flops(avg)
Floating point operations per clock cycle expressed as a percentage of peak. Counts logical operations, a FMA is counted as 2 ops and a FPMA is counted as



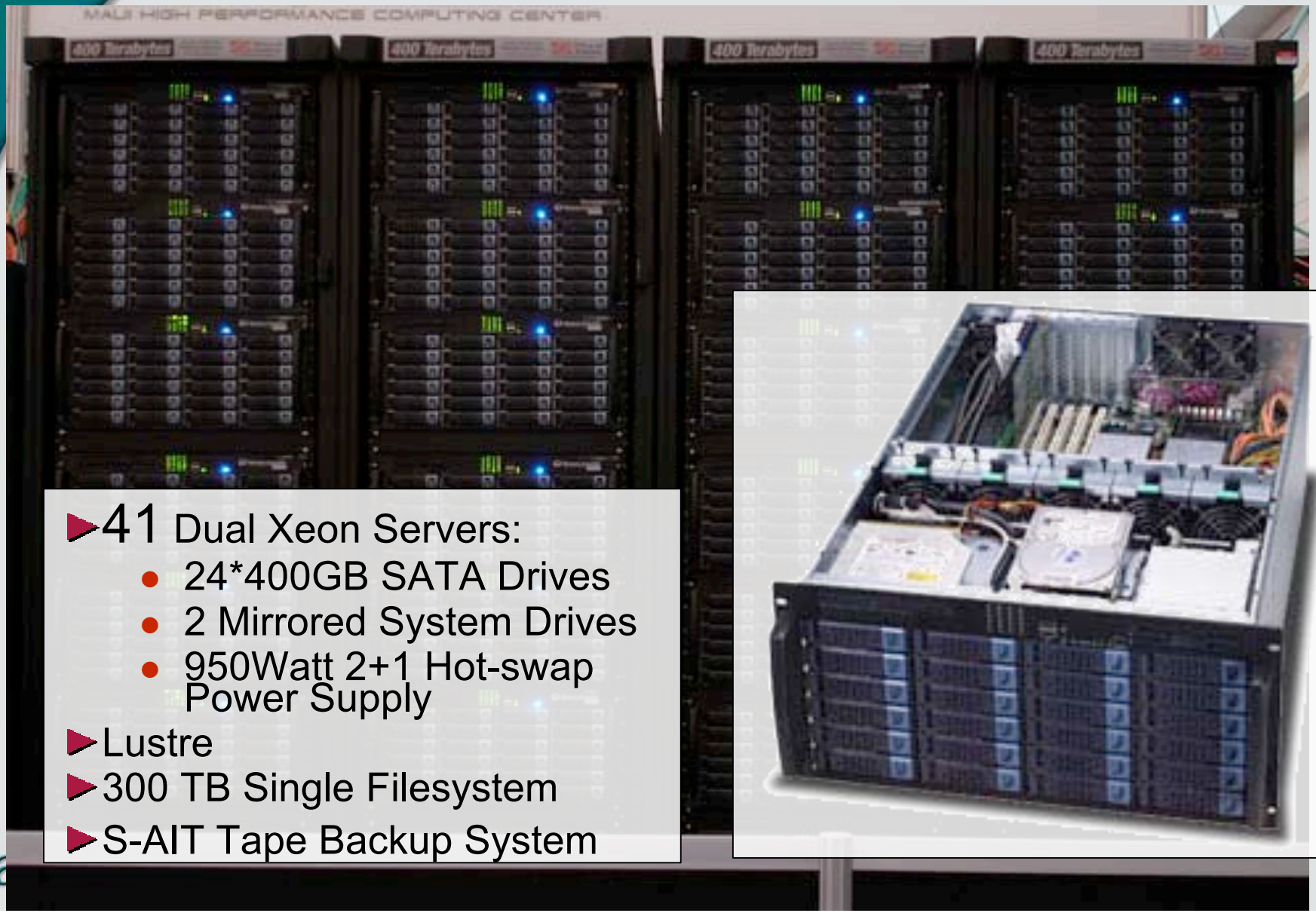
	10/31/2006	11/1/2006	11/3/2006	11/5/2006	11/6/2006	11/8/2006	11/10/2006
Files	8,119,478.00	8,361,861.00	8,369,772.00	8,378,340.00	16,038,970.00	16,933,655.00	16,928,851.00
Directories	121,378.00	117,747.00	121,489.00	122,306.00	122,405.00	122,969.00	141,107.00
Bytes	10,920,657,075,67	11,052,325,023,19	11,531,781,669,63	11,690,567,139,18	11,834,306,610,85	11,974,349,228,99	11,762,473,227,54


HPCS2 File Sizes
10000000
1000000
100000
10000
1000
100
10
1
0
2
8
32
128
512
2048
8192
32768
131072
524288
2097152
8388608
33554432
134217728
536870912
2147483648
8589934592
3.436E+10


Read Node Count Graph
6000
5000
4000
3000
2000
1000
0
MB/s
Node Count
1
2
4
8
16
32
64
8KB
16KB
32KB
64KB
128KB
256KB
512KB
1MB
2MB
4MB
8MB
16MB
32MB
Dec12_L1.4.2_64Way_8G
Block Size

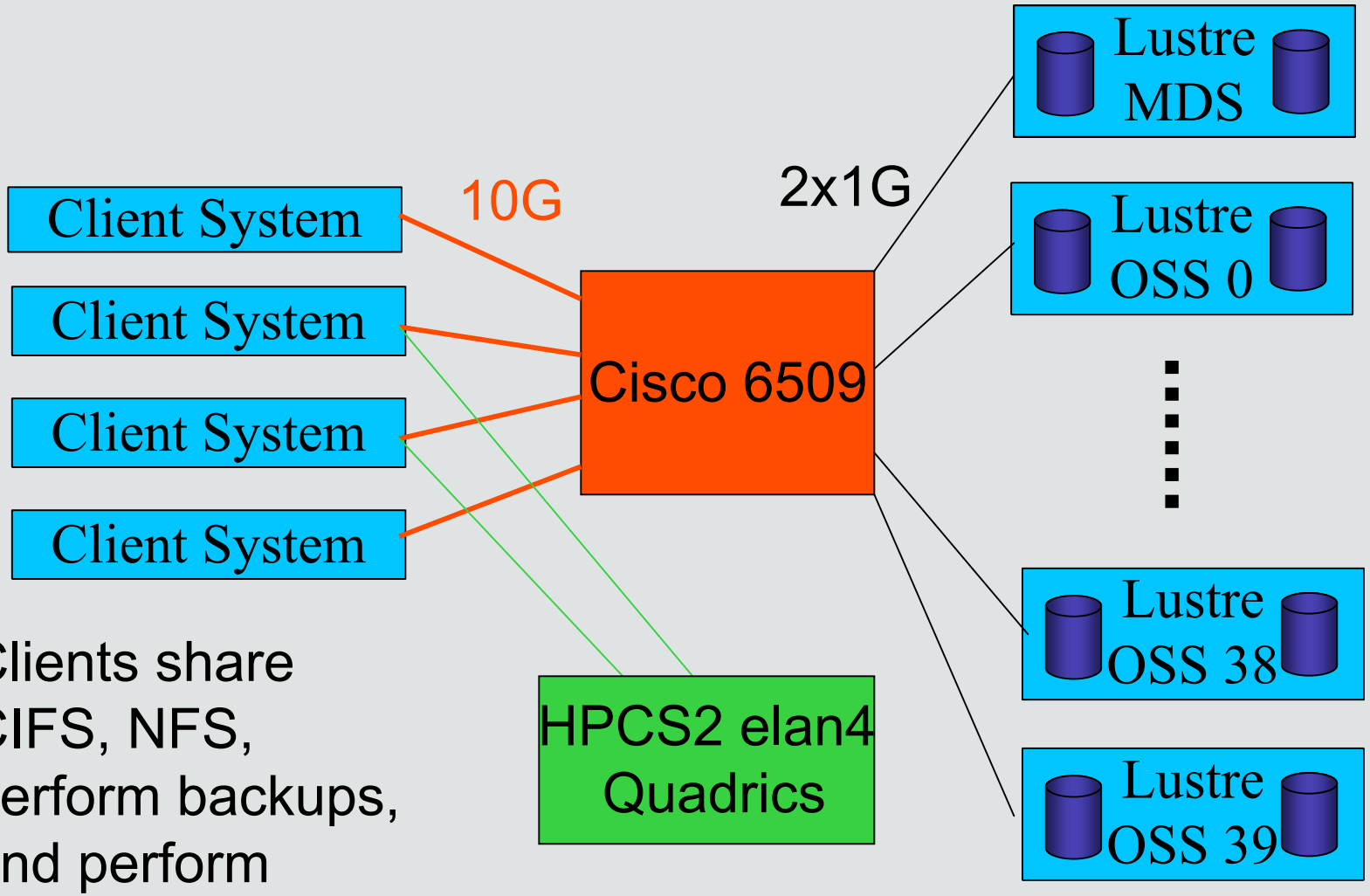

Write Node Count Graph
MB/s
3500
3000
2500
2000
1500
1000
500
0
8KB
16KB
32KB
64KB
128KB
256KB
512KB
1MB
2MB
4MB
8MB
16MB
32MB
Block Size
Node Count
1
2
4
8
16
32
64
Dec12_L1.4.2_64Way_8G

NWfs2 – Archival Storage



- 41 Dual Xeon Servers:
  - 24*400GB SATA Drives
  - 2 Mirrored System Drives
  - 950Watt 2+1 Hot-swap Power Supply
- Lustre
- 300 TB Single Filesystem
- S-AIT Tape Backup System

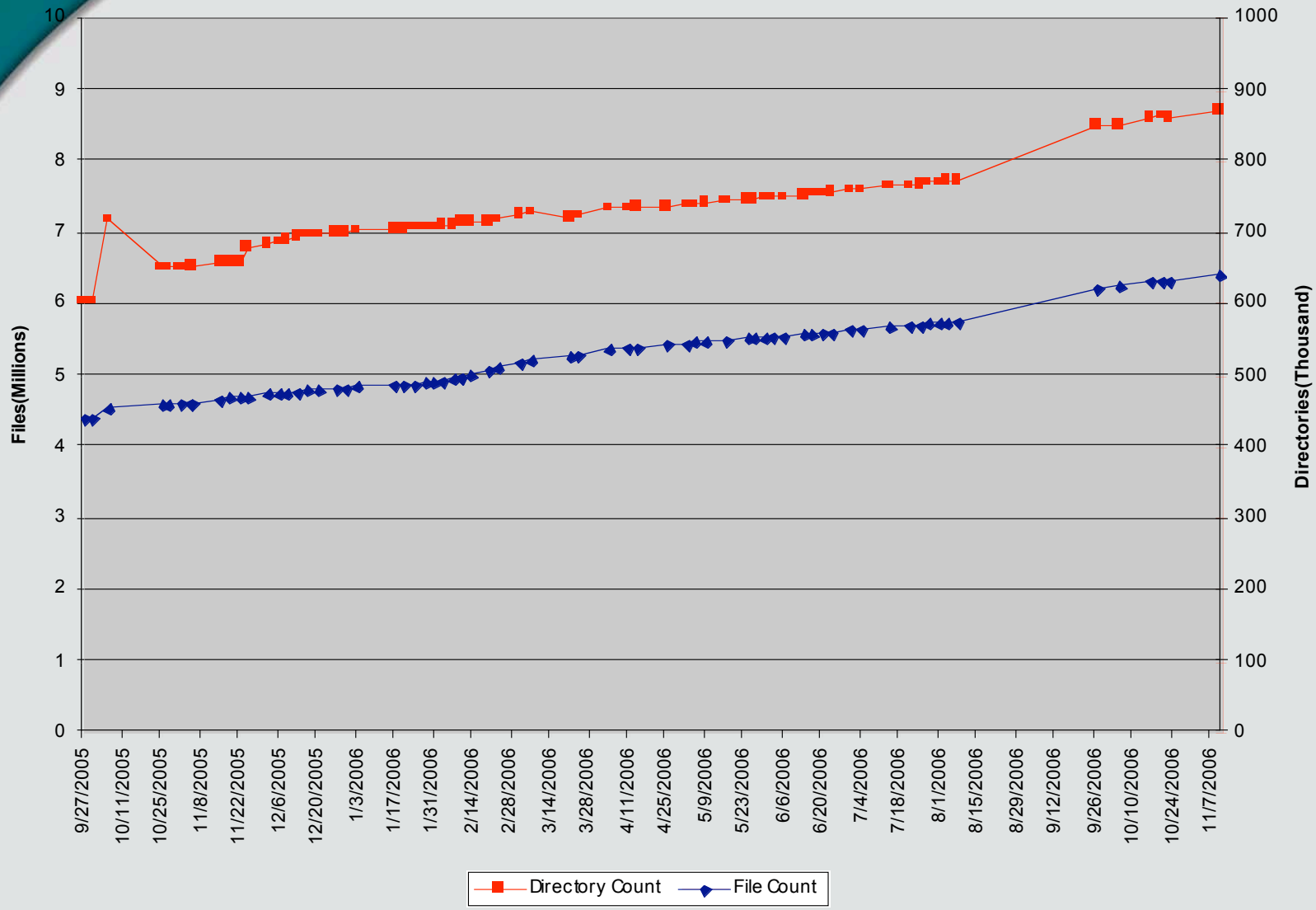
PNNL Archive Diagram

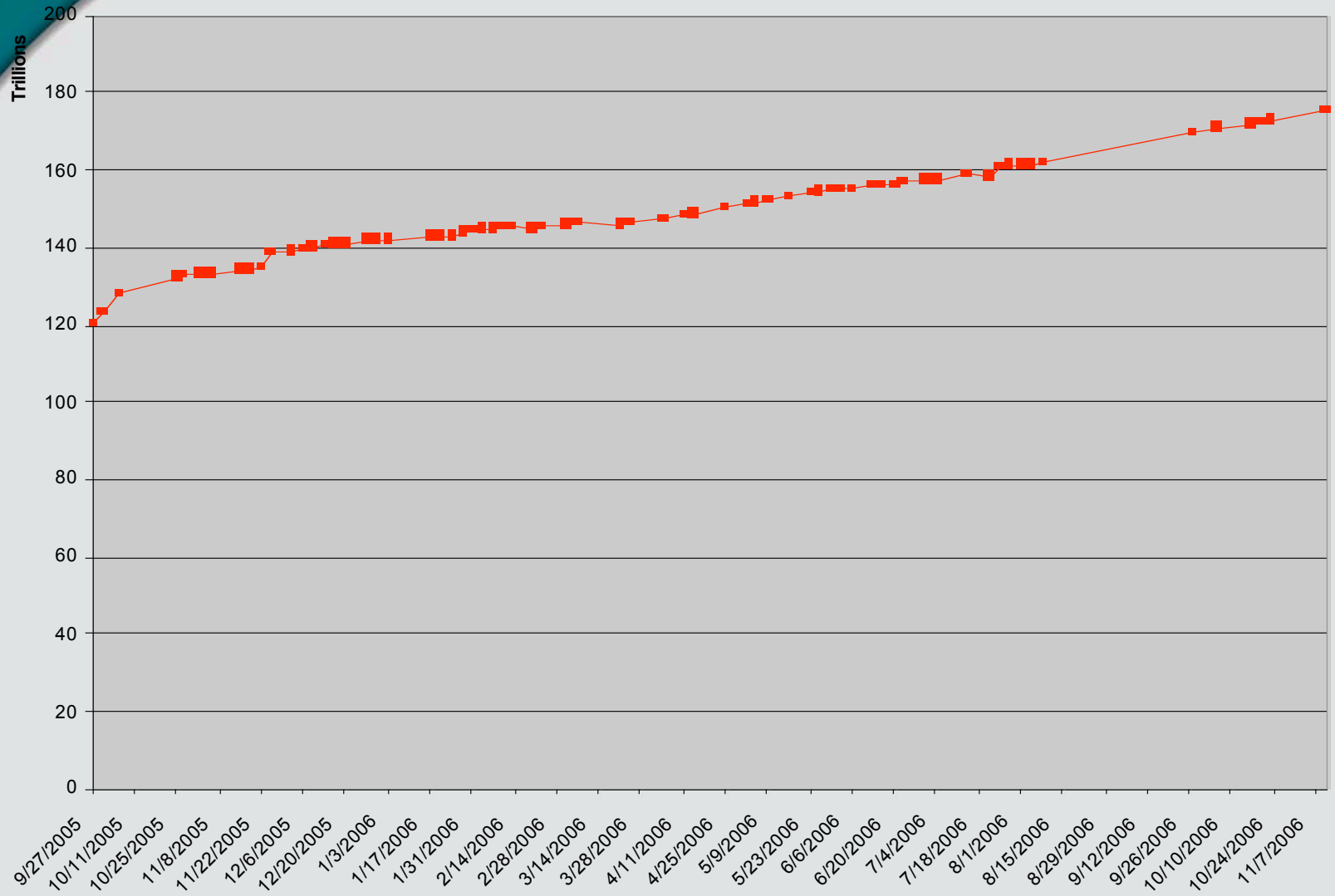


Clients share CIFS, NFS, perform backups, and perform portal routing

NWfs2 Particulars

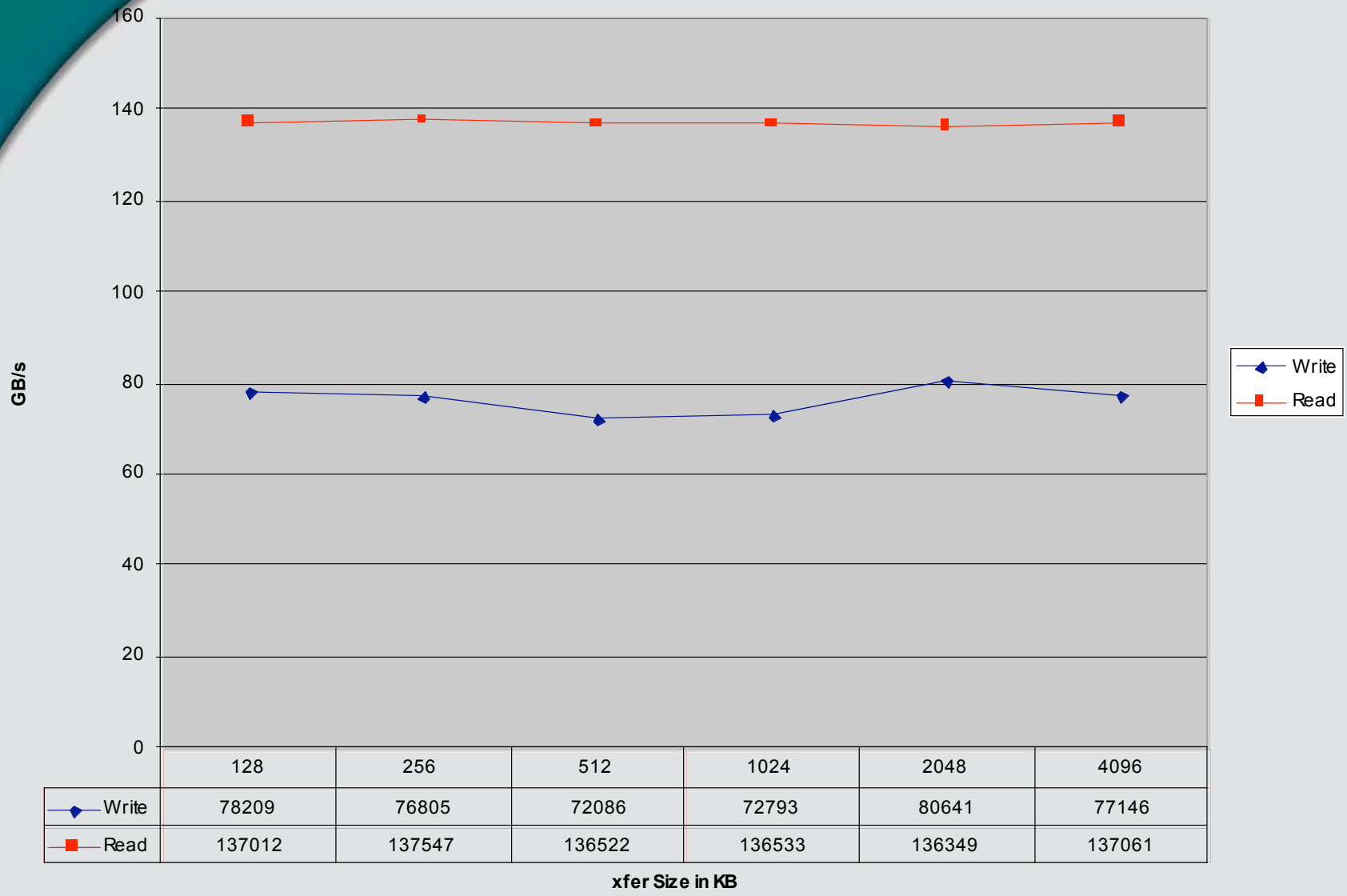
- Built for economy, capacity, reliability, expandability (Rather than performance)
  - Dual Channel-bonded Gigabit Ethernet is used as the intra-cluster interconnect
  - 10 Gigabit Ethernet connects gateway nodes to MSCF & EMSL
  - Default Lustre policy is **not** to stripe so all of a file is on a single OST
  - Tape for disaster recovery
- Lustre is the “glue”
  - Unifies locally-attached storage into a single network filesystem
  - Provides expandability of the filesystem
  - Will allow some freedom to have heterogeneous server configurations in the filesystem


NWFS v2 Lustre Storage
Files(Millions)
Directories(Thousand)
0
1
2
3
4
5
6
7
8
9
10
0
100
200
300
400
500
600
700
800
900
1000
9/27/2005
10/11/2005
10/25/2005
11/8/2005
11/22/2005
12/6/2005
12/20/2005
1/3/2006
1/17/2006
1/31/2006
2/14/2006
2/28/2006
3/14/2006
3/28/2006
4/11/2006
4/25/2006
5/9/2006
5/23/2006
6/6/2006
6/20/2006
7/4/2006
7/18/2006
8/1/2006
8/15/2006
8/29/2006
9/12/2006
9/26/2006
10/10/2006
10/24/2006
11/7/2006
Directory Count
File Count


NWFS v2 Byte Count
Trillions
200
180
160
140
120
100
80
60
40
20
0
9/27/2005
10/11/2005
10/25/2005
11/8/2005
11/22/2005
12/6/2005
12/20/2005
1/3/2006
1/17/2006
1/31/2006
2/14/2006
2/28/2006
3/14/2006
3/28/2006
4/11/2006
4/25/2006
5/9/2006
5/23/2006
6/6/2006
6/20/2006
7/4/2006
7/18/2006
8/1/2006
8/15/2006
8/29/2006
9/12/2006
9/26/2006
10/10/2006
10/24/2006
11/7/2006

MPP2(HPCS2) Lustre Scalability Test

- Use all ‘Fat’ Nodes from Mpp2 for OST targets.
- 7 – 73 gig drives in each Node
- 570 ‘Fat’ Nodes
- 3990 Total OST’s
- Lustre/Kernel Tweaks
  - Max OST defines in obd_class.h -> MAX_OBD_DEVICES=4096
  - Liblustreapi.c check for < 4096 instead of < 2049
  - Bigger Kernel Slab caches.



	128	256	512	1024	2048	4096
Write	78209	76805	72086	72793	80641	77146
Read	137012	137547	136522	136533	136349	137061

MPP2 test - Results/Issues

- Slow MGS connection rate caused slow ‘mount’ times.
- Individual Disk speeds varied from 12MB/s to 80 MB/s
- Takes some time to create/delete 399 perfectly striped files in one directory.

Future Needs/Wants

- Multi-system global file system
- File systems that handle degrading disk speeds
- Virtual I/O server research
- Object Based Disk(OSD) implementations/hardware.
- Large scale Parallel filesystem backup.
- Handle changing network congestion

Our Partners