Sample Lustre Performance Data

HPC Software Workshop
Open Storage Track, Regensburg 2009

Dan Ferber
Sun Microsystems

Agenda

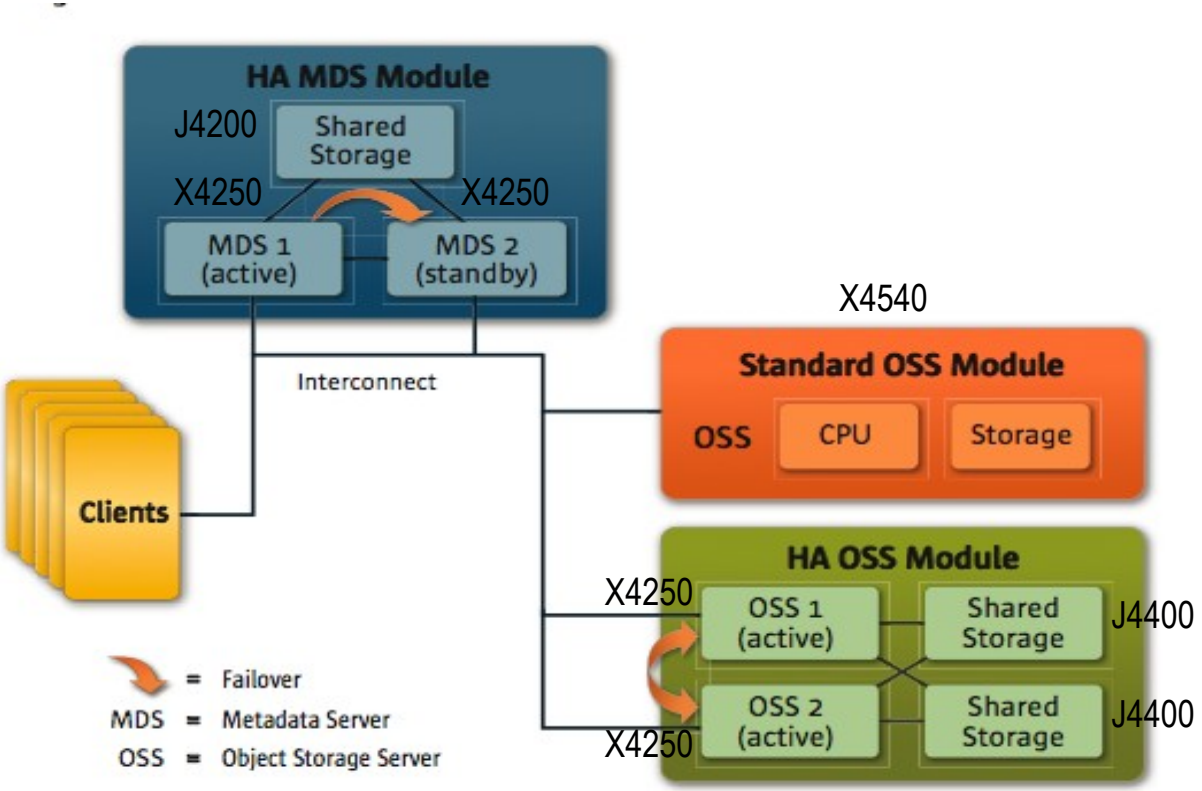
- Sample of performance data points
  - > Lustre Storage Server
  - > TACC
  - > Oak Ridge
  - > MetaData Performance
  - > Lustre and SSD
- Next talk reviews performance tips

Performance Components

- Individual disk bandwidth and IOPs performance
- Storage protocol used to access the disk e.g. SAS, SATA, FC
- Storage HBA bandwidth and IOPS and number of HBA connected
- PCI bus bandwidth and number of PCI lanes available to HBA
- IO aggregation capability of RAID layer
- CPU and memory speed and bandwidth
- Network, Clients, System and Network Workload

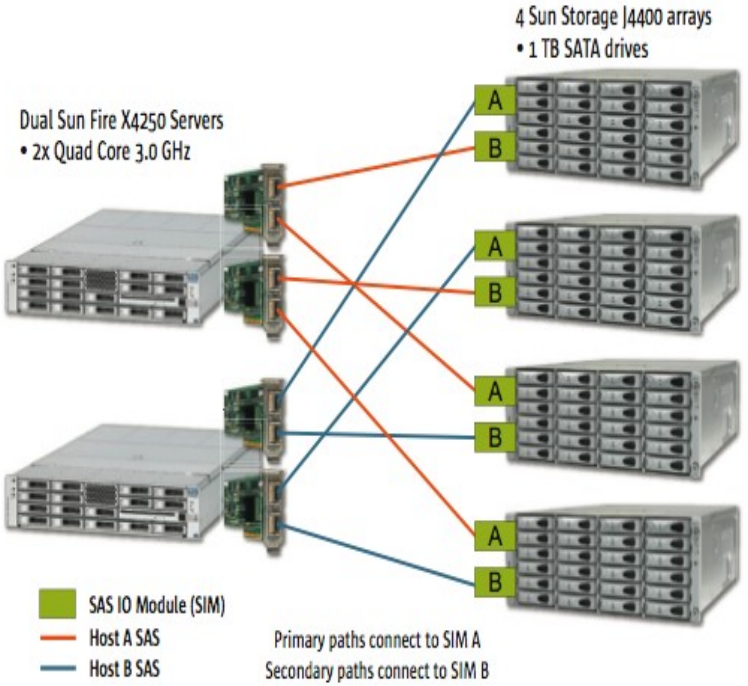
Overview – Lustre Storage Solution

- Sun Lustre Storage Server 1.0 from the Sun Blueprint
  - > See http://wikis.sun.com/display/BluePrints/Main



Deployment Modes

- HA OSS



- Standard OSS

Sun Fire X4540 Server
- 2x Quad Core 2.3 GHz
- 32 GB memory
- I/O cards (IB or 10 GbE)
- 48 internal 1 TB 7200 RPM SATA drives

Snowbird 1.0 Performance

- Each standard OSS module sustains 970 MB/sec

The testing was performed with 9 nodes that were configured to run the IOZone benchmark. A file size of 16 GB was used to minimize the impact of client side caching. Tests were run with 256 KB, 512 KB and 1 MB block sizes. Performance differences were discovered to be negligible between block sizes; for brevity only 1 MB block size results are captured in this document.

Peak performance with a 9 client compute cluster was observed with 63 threads (7 threads per client) attaining a sustained initial write speed of approximately 974 MB/sec.

These benchmark results show that approximately 970 MB/sec can be sustained for a single Sun Lustre Storage System Standard OSS module on initial writes. To achieve higher aggregate performance, multiple Standard OSS modules can be deployed within a single cluster.

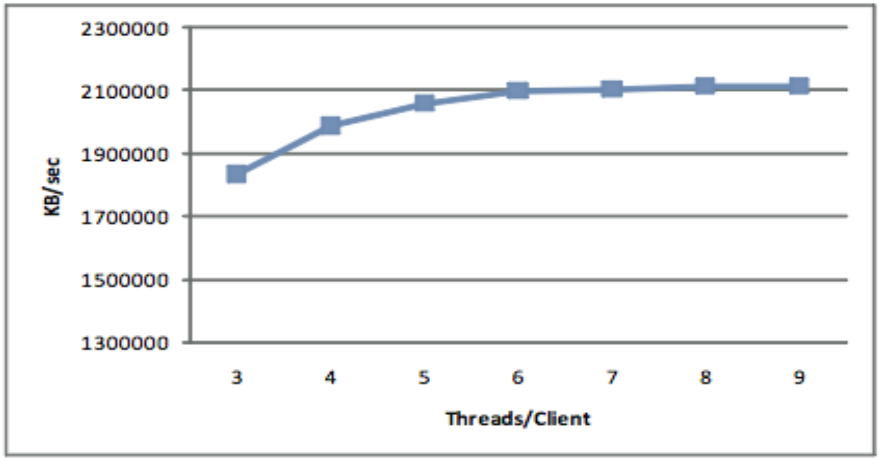
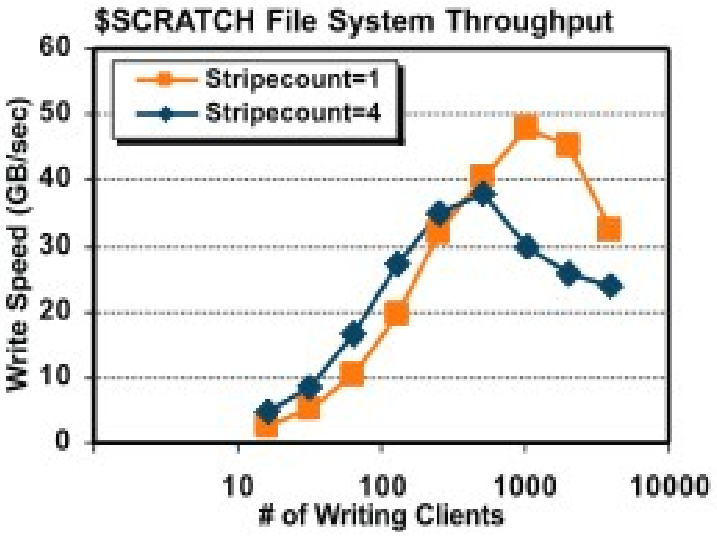


Figure 9. HA OSS: Initial write performance—12 clients.

- Each HA OSS module sustains 2.1 GB/sec

Lustre at TACC Performance

- TACC ranger system – has observed 46 GB/sec throughput
- They use 50 Sun Fire X4500 servers as OSS
- A single app achieved 35 GB/sec throughput



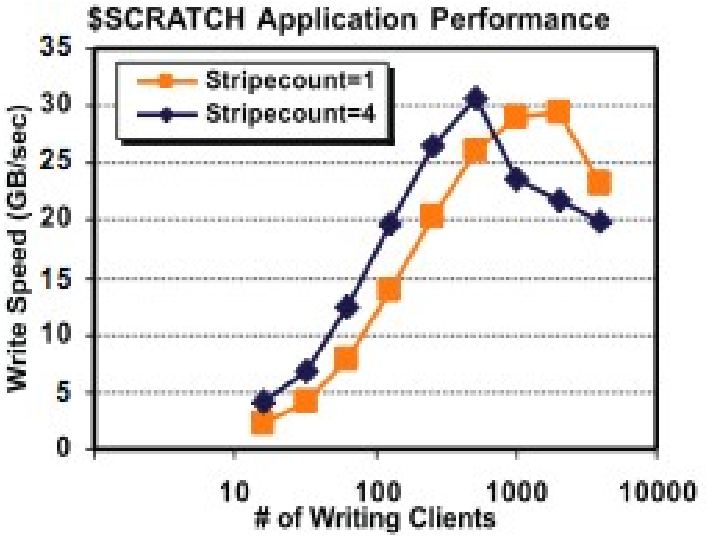
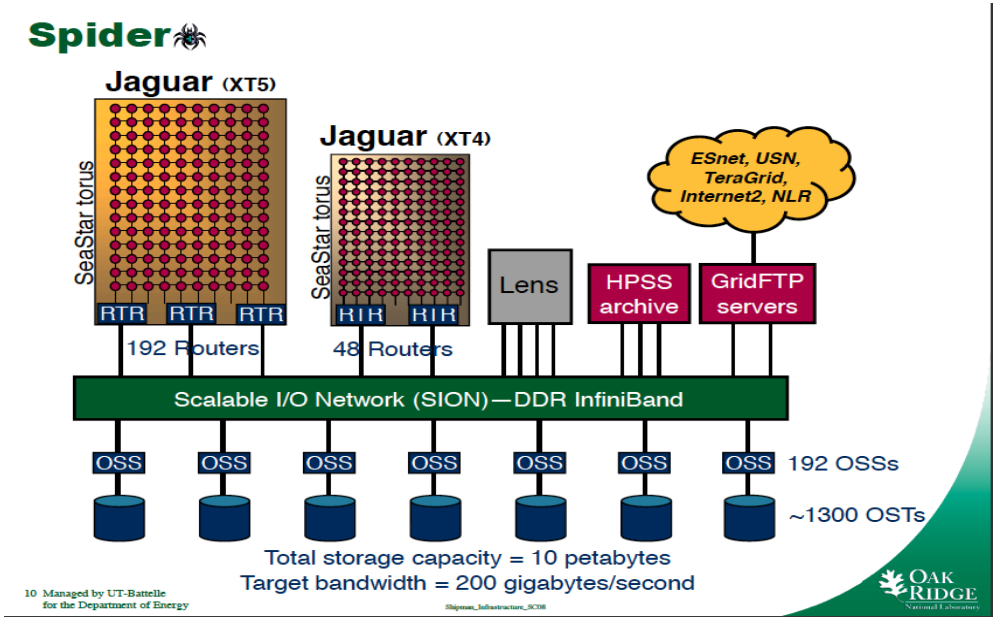


Figure 12. Lustre file system performance at TACC.

Lustre at Oak Ridge Performance

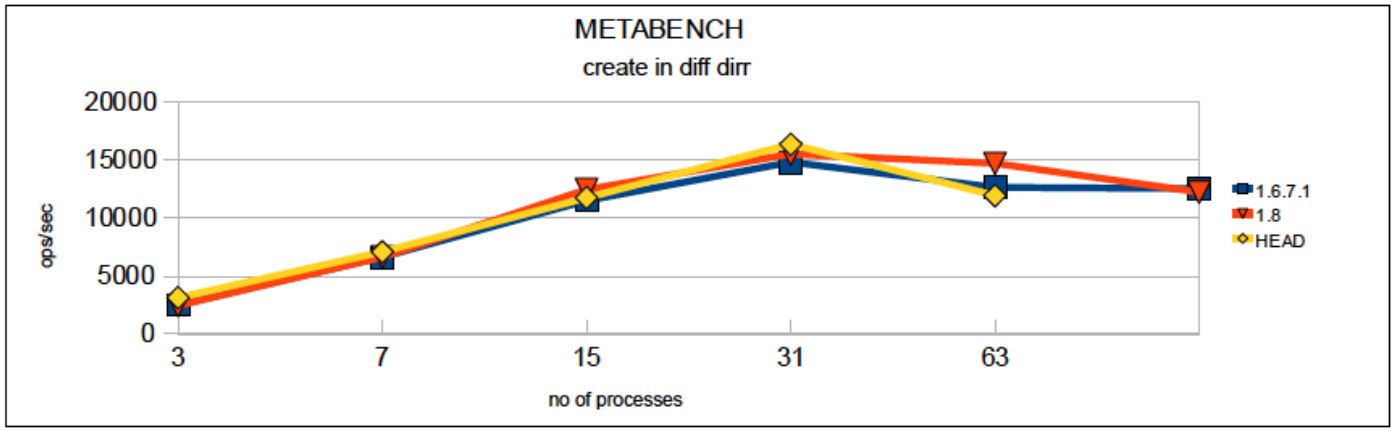
- "Spider" Lustre file server for their compute facility
- 10.7 Petabytes of storage
- 26,000 clients on IB, 192 servers, 13,000 disks
- 240 GB/sec throughput

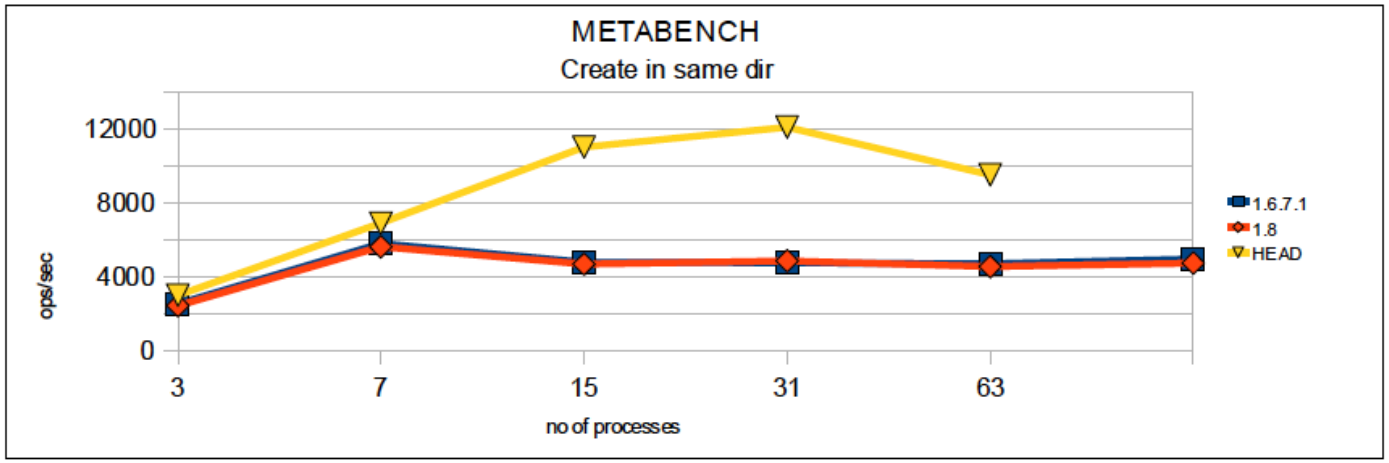


Current Data Point – MetaData

- See http://wiki.lustre.org/index.php/File:MD-perf-comparison.pdf
- Results also show stat, delete, and directory create

graphs- multi-clients



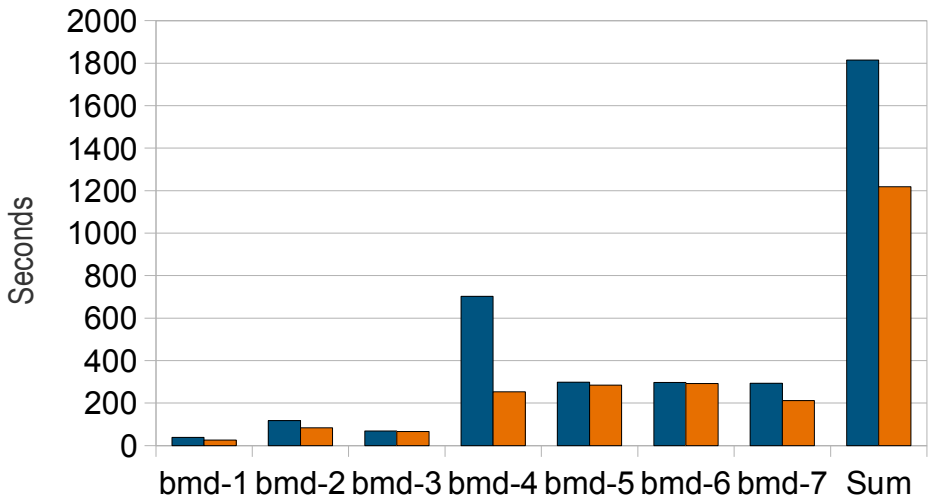


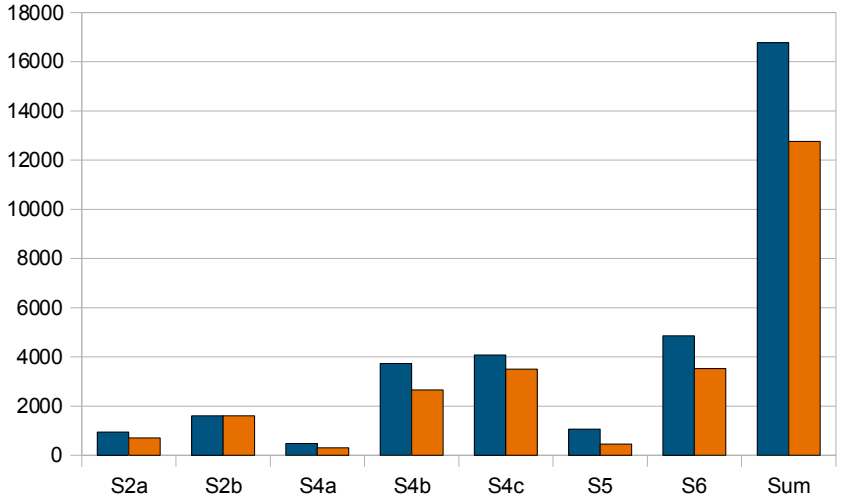
ANSYS & Abaqus

I/O Intensive Benchmarks Faster with Lustre and SSDs

Hardware configuration

- Sun Fire X4450 server
- Four 2.93 GHz quad-core Intel Xeon Processor X7350 CPUs
- Four 15,000 RPM 500 GB SAS drives
- Three 32 GB SSDs





Disk Drives Flash based SSD

Overall Test Time Savings – 32% for ANSYS; 24% for Abaqus

ANSYS & Abaqus

I/O Intensive Benchmarks
Faster with Lustre and SSDs

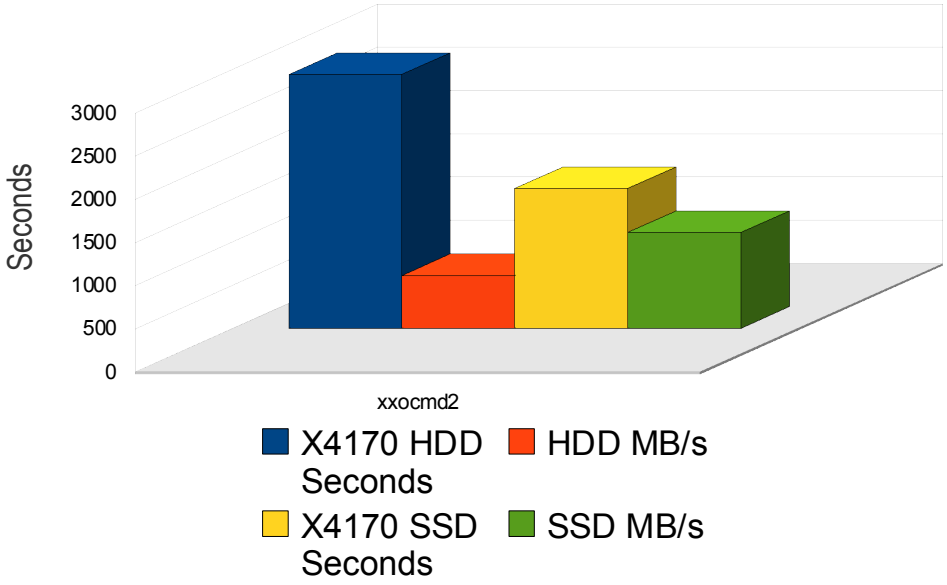
Table 1. ABAQUS Benchmark standard test suite: HDDs vs. SSDs

Test, cores	Time(sec) x4450 HDD	Time (sec) x4450 SSD	Time Ratio HDD:SSD	Improvement	Sockets	Cores
s2a-1,1	2787	2464	1.13	12.00%	1	1
s2a-2,2	1659	1298	1.28	22.00%	2	2
s2a-4,4	949	709	1.34	25.00%	4	4
s2b-1,1	3074	3111	0.99	-1.00%	1	1
s2b-2,2	1684	1753	0.96	-4.00%	2	2
s2b-4,4	1608	1606	1	0.00%	4	4
s4a-1,1	679	613	1.11	10.00%	1	1
s4a-2,2	628	419	1.5	33.00%	2	2
s4a-4,4	480	303	1.58	37.00%	4	4
s4b-1,1	11698	8115	1.44	31.00%	1	1
s4b-2,2	6162	4520	1.36	27.00%	2	2
s4b-4,4	3734	2655	1.41	29.00%	4	4
s4c-1,1	6608	6743	0.98	-2.00%	1	1
s4c-2,2	5499	4571	1.2	17.00%	2	2
s4c-4,4	4073	3509	1.16	14.00%	4	4
s5-1,1	1708	1051	1.63	38.00%	1	1
s5-2,2	1345	675	1.99	50.00%	2	2
s5-4,4	1069	456	2.34	57.00%	4	4
s6-1,1	9040	7175	1.26	21.00%	1	1
s6-2,2	6128	4741	1.29	23.00%	2	2
s6-4,4	4864	3520	1.38	28.00%	4	4

NASTRAN Testing Results

Tremendous Performance Boost with Lustre and SSDs

Sun Fire X4170/HDD & X4170/SSD



SunFire X4170

- 2 x Intel Nehalem 5570 2.93 GHz, Quad Core, 24GB RAM 1333 MHz
- Six SAS disk drives vs. six SSDs

MSC/Nastran "Vendor 2008" benchmark test suite"

Time Savings of 45%; Bandwidth Improvements of 82%

NASTRAN Testing Results

Tremendous Performance Boost with Lustre and SSDs

Seconds

Test–number of cores	Sun Fire x2270 server Time (sec) x2270 HDD	Sun Fire x2270 server Time (sec) x2270 SSD	Time Ratio HDD:SSD	Improvment	Number Cores
vlosst1-1	127	126	1.007936508	0.79%	1
xxocmd2-1	895	884	1.012443439	1.23%	1
xxocmd2-2	614	583	1.053173242	5.05%	2
xxocmd2-4	631	404	1.561881188	35.97%	4
xxocmd2-8	1554	711	2.185654008	54.25%	8
xlotdf1-1	2000	1939	1.031459515	3.05%	1
xlotdf1-2	1240	1189	1.042893188	4.11%	2
xlotdf1-4	833	751	1.10918775	9.84%	4
xlotdf1-8	1562	712	2.193820225	54.42%	8
sol400_1-1	2479	2402	1.032056619	3.11%	1
sol400_s-1	2450	2262	1.08311229	7.67%	1
getrag-1	843	817	1.031823745	3.08%	1

Lustre Performance Work Continues

- SMP scalability
- Size on MetaData
- Clustered MetaData
- Sun Lustre Storage Server Enhancements
  - > Including SSD benchmarks
- Customer and partner collaboration

References

- sun.com/lustre
- wikis.sun.com/display/BluePrints/Main
- wiki.lustre.org/index.php/Lustre_User_Group
- wikis.sun.com/display/BluePrints/Main

- Thank You