Upgrading and Expanding Lustre Storage for use with the WLCG

Daniel Traynor, Terry Froy, Chris Walker

School of Physics and Astronomy, Queen Mary University of London

We put together a 3 PB budget Lustre file system with the performance to support 4000 analysis jobs and the capability to double in size.

Requirements

1) Typical data analysis job uses 5MB/s, there are 4000 job slots, then the throughput of the complete Lustre system is required to be capable *of 20GB/s*

2) Keep costs down by; Reusing exiting networking equipment. Reuse existing storage, (1.5 PB in 72 Dell R510 servers); Don't use failover for bulk storage.

3) Minimise downtime by migrating data between live systems. Will need to match size of existing storage (1.5 PB) with new hardware.

4) Expandable to at least 6PB and maintainable by local system admin

5) Works with the WLCG grid, supports SRM, GridFTP, Webdav, Xrootd

6) Optimise for data analysis workloads (reading files)

Can't afford off the shelf commercial system, Lustre/GPFS etc... or software support contract.

Ferrari - High cost maintenance

Don't yet trust CEPH, Not yet widely used in HPC/HTC/WLCG. POSIX interface not well tested/tuned.

Amphicar - Car and Boat in one, does neither very well

Go for tried and tested, open source (GPL) Lustre files system. Widely used, community support. High performance.

Mini Copper Winner Monte Carlo rally 1964, 65, 66, 67.

Solution

Existing Storage (OSS/OST) 1.5PB

72 Dell 510with 12*2/3TB NLSAS Disks. in RAID 6

+

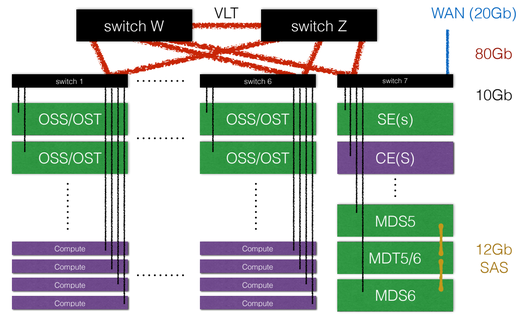
New Storage (OSS/OST) 1.5PB

20 Dell 730XDs with 16*6TB NLSAS Disks. in RAID 6

+

Meta Data Server (MDS/MDT) HA

Dell MD3400 + 2*Dell R630



Layout mix of storage and compute in a rack to balance power and network bandwidth. Every server connected with 10Gb/s Ethernet. WAN link at 20Gb/s

OS = Scientific Linux 6.7

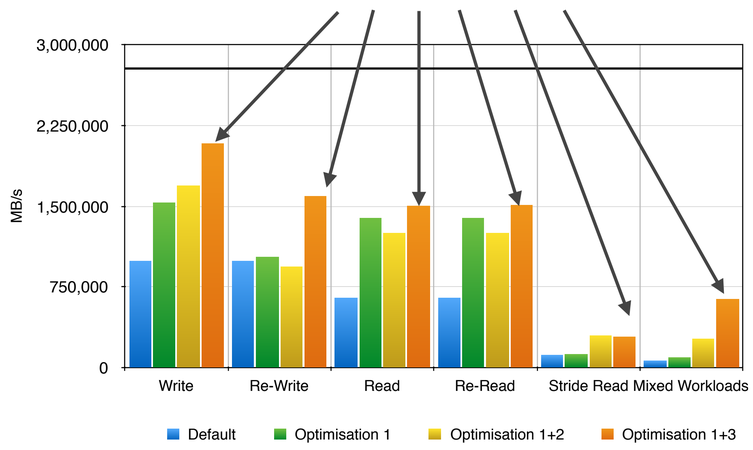
Lustre version = 2.8

StoRM for SRM/GridFTP/webdav (3xGridFTP nodes to fill 20Gb/s WAN)

Standalone XrootD (read only) server

Benchmark Performance

Double (or better) performance of a single storage server with a few optimisations



To test a single server IOzone was run with 12 threads each transferring a file size of 24GB in chunks of 1024kB

iozone -e -+u -t 12 -r 1024k -s 24g -i0 -i1 -i 5 -i 8

For Lustre benchmarking up to 24 IOzone clients, each client runs 10 threads transferring a file size of 24GB in chunks of 1024kB

(iozone -+m iozone_client_list_file -+h [IP of master IOzone node] -e -+u -t 10 -r 1024k -s 24g -i0 -i1 -i 5 -i 8).

Optimisation 1

```
echo deadline > /sys/block/sdb/queue/scheduler
echo 4096 > /sys/block/sdb/queue/nr_requests
echo 4096 > /sys/block/sdb/queue/read_ahead_kb
echo madvise > /sys/kernel/mm/
redhat_transparent_hugepage/enabled
echo madvise > /sys/kernel/mm/
redhat_transparent_hugepage/defrag
```

+

Optimisation 2

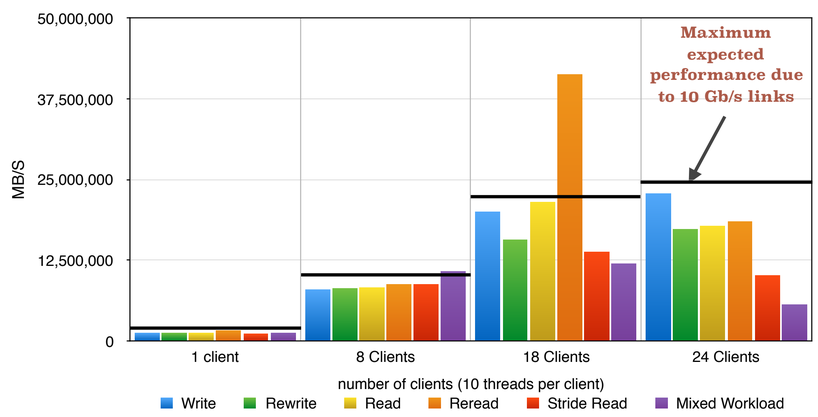
```
echo 5 > /proc/sys/vm/dirty_background_ratio
echo 10 > /proc/sys/vm/dirty_ratio
echo 262144 > /proc/sys/vm/min_free_kbytes
echo 50 > /proc/sys/vm/vfs_cache_pressure
```

OR

Optimisation 3

```
echo 1 > /proc/sys/vm/dirty_background_ratio
echo 75 > /proc/sys/vm/dirty_ratio
echo 262144 > /proc/sys/vm/min_free_kbytes
echo 50 > /proc/sys/vm/vfs_cache_pressure
```

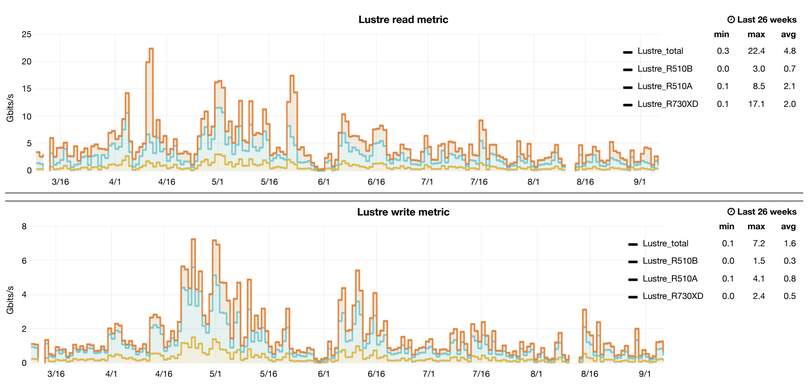
Results show that for 1.5 PB system, read and write speed greater than 15GB/s. Confident that full system (3PB) will perform grater than required 20 GB/s



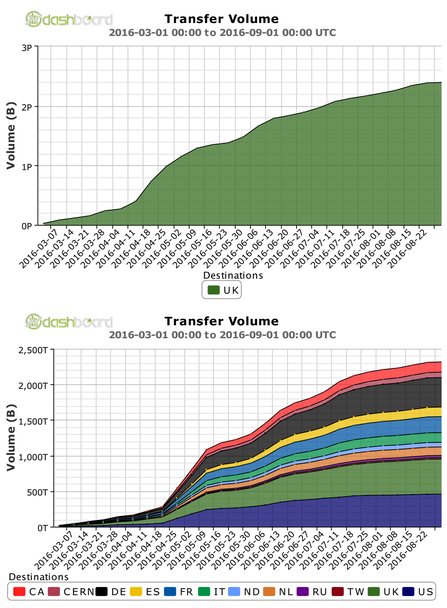
A near complete 1.5 PB lustre file system with 20 Dell 730XD servers was created with up to 24 client nodes dedicated to the benchmark tests

Real World Performance of full 3 PB system

6 month average read speed = 4.8 Gb/s

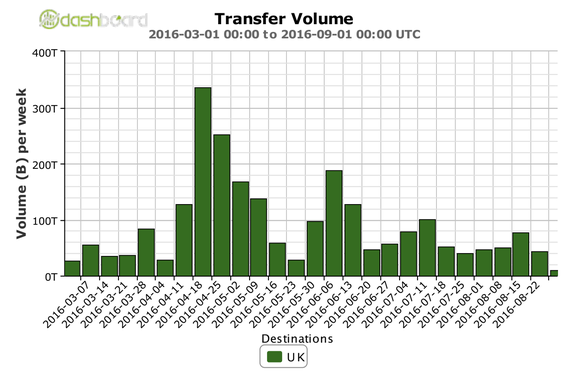


6 month average write speed = 1.6 Gb/s



In 6 months ATLAS transferred 2.4 PB to QMUL

In 6 month ATLAS transferred 2.3 PB from QMUL to the rest of the world



Weekly transfer rates in to QMUL by ATLAS, up to 340 TB in one week.

Future Plans:

Double the Storage of the cluster to 6PB in 2018. Upgrade OSS servers to SL/CentOS 7 and Lustre 2.9 which provides additional functionality such as user and group ID mapping which would allow the storage to be used in different clusters. Examine the use of ZFS in place of hardware raid which might help mitigate very long raid rebuild times after replacement of a failed hard drive.

Links:

StoRM: http://italiangrid.github.io/storm/index.html

CHEP2012: Scalable Petascale Storage for HEP using Lustre: Journal of Physics: C.J. Walker D.P. Traynor and A.J. Martin. Conference Series 396 (2012) 042063

CHEP2015: Optimising network transfers to and from Queen Mary University of London, a large WLCG tier-2 grid site: C J Walker, D P Traynor, D T Rand, T S Froy and S L Lloyd. Journal of Physics: Conference Series 513 (2014) 062048

IOzone: http://www.iozone.org/

BeeGFS Tips and Recommendations for Storage Server Tuning: http://www.beegfs.com/wiki/StorageServerTuning

ESnet Fasterdata Knowledge Base: http://fasterdata.es.net/

Contacts:

Daniel Traynor: d.traynor@qmul.ac.uk

Terry Froy: t.froy@qmul.ac.uk

Christopher J. Walker: c.j.walker@qmul.ac.uk,

School of Physics and Astronomy, Queen Mary University of London, Mile End Road, London, E1 4NS

GridPP UK Computing for Particle Physics

Queen Mary University of London

Science & Technology Facilities Council