Upgrading and Expanding Lustre Storage for use with the WLCG

IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 062004 doi:10.1088/1742-6596/898/6/062004

Upgrading and Expanding Lustre Storage for use with the WLCG

D P Traynor, T S Froy, C J Walker

Queen Mary University of London, Mile End Road, E1 4NS, UK

E-mail: d.traynor@qmul.ac.uk

Abstract. The Queen Mary University of London Grid site's Lustre file system has recently undergone a major upgrade from version 1.8 to the most recent 2.8 release, and a capacity increase to over 3 PB. Lustre is an open source, POSIX compatible, clustered file system presented to the Grid using the StoRM Storage Resource Manager. The motivation and benefits of upgrading including hardware and software choices, are discussed. The testing, performance improvements and data migration procedure are outlined as is the source code modifications needed for StoRM compatibility. Benchmarks and real world performance are presented and future plans discussed.

1. Introduction

The Queen Mary WLCG tier two site has successfully operated a reliable, high performance, efficient, budget oriented storage solution, utilising Lustre [1] StoRM [2] and xrootd [3], since 2010 [4] [5].

Lustre is a open-source (GPL), POSIX compliant, parallel file system used in over half of the worlds Top 500 supercomputers. Lustre is made up of three components: One or more Meta Data Servers (MDS) connected to one or more Meta Data Targets (MDT), which store the namespace metadata such as filenames, directories and access permissions; One or more Object Storage Servers (OSS) connected to one or more Object Storage Targets (OST) which store the actual files; and clients that access the data over the network using POSIX filesystem mounts. The network is typically either Ethernet or Infiniband.

StoRM (STorage Resource Manager, SRM) is a scalable and file system independent storage manager service. It supports standard access and transfer protocols like HTTP(S), WebDAV and GridFTP. It it is designed to work on top of any POSIX filesystem with Access Control Lists (ACL) support such as Lustre.

Over the past 5 years the Lustre storage file system at Queen Mary has undergone expansion from 300TB to 1.5PB, an upgrade of Lustre from version 1.6 to 1.8, a network upgrade from multiple 1Gb to 10Gb Ethernet, and migration of the MDS and MDT to new hardware. Most recently a major upgrade of all aspects of the Lustre file system has been undertaken. This upgrade has involved new hardware, a complete reinstallation and upgrade the OS and Lustre software on every storage server (MDS/OSS) and a migration of data from the old Lustre filesystem to the new.

2. Motivation for Upgrade

In 2015 it was decided that a major software and hardware upgrade was required. This was driven by several reasons: The need to upgrade the Operating System (OS) from Scientific Linux (SL)5 to a supported OS such as SL6 or CentOS7; Use a supported Lustre version compatible with SL6 or CentOS7; To take advantage of new software developments providing improved performance and reliability; Migrate to a new MDS/MDT with hardware in warranty; Double the storage capacity to over 3PB and allow for a doubling again before 2020.

Consideration was given to use of other open source file systems such as Ceph and GlusterFS. However, it was decided early on that local knowledge and experience with Lustre; its maturity, reliability and performance; clear long term development and support from Intel and others; and POSIX support made Lustre the obvious choice. While it is possible to buy a commercially supported solution but this was beyond the budget available. Therefore the specification, installation, configuration and operation of hardware and software had to be done by the site team.

3. Hardware Choice and Setup

In order to reduce costs the existing Lustre OSS/OSTs, made up of 70 Dell R510s, with 12 two or three TB hard disks in RAID6, were reused, providing 1.5 PB of usable storage. An additional 20 Dell R730XDs, with 16 six TB disks in RAID6 were also purchased, providing 1.5 PB of usable storage, matching the size of the existing Lustre file system. The Dell 730XDs have two Intel E5-2609 V3 processors and 64GB of RAM. Lustre is a light user of CPU resources on the OSS/OST and the E5-2609 processor is one of the cheapest CPUs available. Cost saving were also made by not utilising failover OSS/OSTs hardware which helped reduce costs by 40%.

However, the new MDS/MDT was set up in a resilient, automatic failover configuration utilising two Dell R630s connected to a MD3400 disk array. The Dell R630s have two Intel E5-2637 V3 processors and 256 GB RAM. The disk array has 12 600GB 15K SAS disks in RAID10. Only one MDS/MDT is used in the cluster and the hardware has been specified as high as affordable. The automatic failover was configured using Corosync, Cman, Fence-agents and the Red Hat resource group manager (rgmanager) packages. Lustre itself has protection against the MDT being mounted by more than one MDS at a time.

All servers (storage, compute and service nodes) are connected to one of seven top of rack Dell S4810 network switches with a single 10Gb SFP+ Ethernet connection (the max attainable server to server bandwidth), which in turn are connected with four 40Gb QSFP+ connections to a distributed core switch made up of two Dell Z9000s in a Virtual Lan Trunk (VLT) configuration (figure 1).

As a result of design choices and several years of evolution in hardware the network connections from storage and compute servers are mixed in the top of rack switches. This has the advantage of balancing power and network IO in a rack [4] but at the expense of a more complicated hardware layout.

4. Software Setup

The Lustre software was installed on a standard SL6 OS configured server. A patch has been applied to Lustre due to a bug, LU1482 [1], causing incorrect interaction between Access Control Lists (ACLs) and the extended attribute permissions. This is required by StoRM as attributes are used to store checksums of every file which, after every GridFTP transfer, are compared between source and destination. This bug is be fixed in the future 2.9 release of Lustre.

The Lustre manual [1] describes in detail how to setup and configure a Lustre system. The MDT is formatted and mounted on the MDS using the following commands. On the MDS add the `acl` option when mounting the MDT to ensure ACL and extended attributes support. For simplicity we install the Lustre Management Server (MGS) on the MDS.

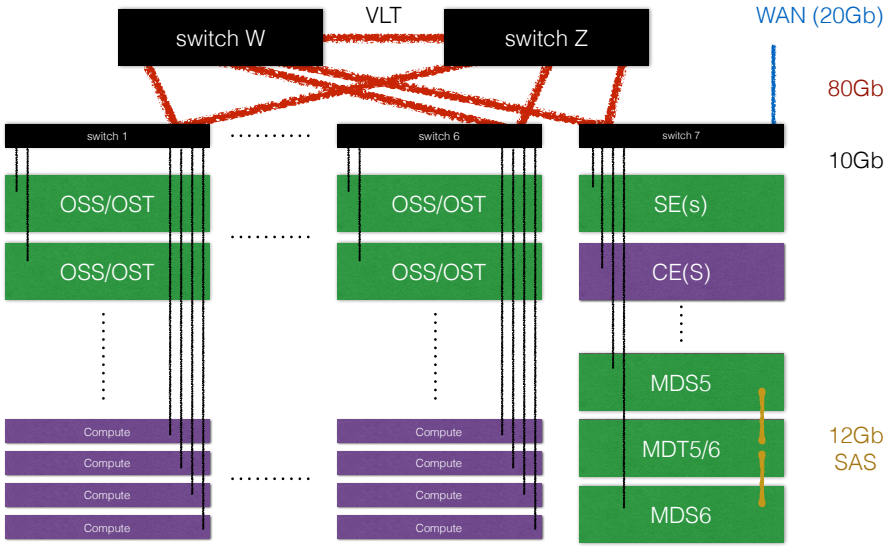

Figure 1. Queen Mary GridPP Cluster Layout including the StoRM Storage Element (SE) and Cream Compute element (CE)

```
[~] mkfs.lustre -fsname=lustre_1 --mgs --mdt --servicenode=10.0.0.5
    --servicenode=10.0.0.6  --index=0 /dev/mapper/mpathb
[~] cat /etc/fstab
...
/dev/mapper/mpathb  /mnt/mdt   lustre   rw,noauto,acl,errors=remount-ro,user_xattr  0 0
```

On the OSS/OST you need to specify each of the MDSs when you configure a Lustre OST. Once each file system has been mounted it becomes visible to Lustre.

```
[~] mkfs.lustre -fsname=lustre_1 --mgsnode=mds05@tcp0 --mgsnode=mds06@tcp0 --ost
    --index=0 /dev/sdb
[root@sn100 ~] cat /etc/fstab
...
/dev/sdb                /mnt/sdb                lustre  defaults        0 0
```

Lustre Clients need to know about both MDS/MGS nodes when mounting lustre in order to be able to fail over. Lustre is mounted as standard POSIX file system, of type lustre, on clients.

```
[~] cat /etc/fstab
...
mds05@tcp0:mds06@tcp0:/lustre_1 /mnt/lustre_1    lustre  flock,user_xattr,_netdev 0 0
```

The file system mounted on a client appears as a local file system.

```
[~] df -h
Filesystem                      Size  Used Avail Use\% Mounted on
mds05@tcp0:mds06@tcp0:/lustre_1  2.9P  2.1P  710T  75\% /mnt/lustre_1
```

StoRM is used for remote data management for all Virtual Organisations (VOs) supported by the site and supports SRM, HTTP(S) and GridFTP. Most data is transferred via GridFTP.

Dedicated WAN data transfer tests were conducted and three GridFTP nodes were found to be needed to provide the capacity to fully utilise the 20Gb/s WAN link. A standalone, readonly, installation of XRootD is deployed and is remotely usable by all site supported VOs using standard Grid authentication.

5. Performance Tuning

A number of optimisations were made to improve the performance of the OSSs. To test these optimisations the IOzone [6] benchmarking program was used. IOzone is used to perform a variety of read and write tests. It is able to operate on a single server or on multiple clients at the same time.

It is useful to have an estimate of possible performance before undertaking benchmarking. The typical maximum sustained throughput of a single disk is quoted at approximately 200MB/s. For a 16 disk RAID6 array the maximum sustained throughput for a single server is expected to be 2.8GB/s (excluding the two parity disks). For a Lustre system made up of 20 Dell R730XDs, with 16 disks in each, this should scale to 56GB/s. However, each server is only connected with a 10Gb/s ethernet connection. Therefore the maximum sustained throughput obtainable, taking into account the topology outlined in figure 1, is 25GB/s.

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`). As well as the standard sequential read and write tests results were obtained for stride reads, and mixed workloads, which does reading and writing of a file with accesses being made to random locations within the file. The values were chosen to match the expected workload (i.e. the reading of large, GigaByte sized) and to match the 1024k buffer size used in Lustre network transfers.

Using the BgFS Tips and Recommendations for Storage Server Tuning [7] for the Linux IO Scheduler and Virtual memory setting as reference we applied different sets of optimisations to the storage server.

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 or (3), used in conjunction with optimisation 1, optimises the Linux file system cacheing which is used by Lustre to help improve performance.

```
echo 5(1) > /proc/sys/vm/dirty_background_ratio
echo 10(75) > /proc/sys/vm/dirty_ratio
echo 262144 > /proc/sys/vm/min_free_kbytes
echo 50 > /proc/sys/vm/vfs_cache_pressure
```

RAID misalignment can cause performance degradation. To reduce RAID alignment complications the partition was made directly on to the storage device (e.g. /dev/sdb) taking into account the RAID configuration (block size, stripe size and width). Lustre uses the Ext4 file system although it is possible to use ZFS instead.

The results for a single server with different optimisations are shown in figure 2. The results clearly show the benefits of applying optimisations to the OS to improve file system performance. As optimisation 1+3 show the highest throughput this has been applied to the Lustre file system.

The single server tests were carried out for each of the 20 R730XD servers as a cross check of performance and as a check for hardware issues. All servers were found to produce similar performance. A cross check for of the single server benchmark test, for optimisation 1 only, limiting the storage servers to only 2GB RAM, to remove caching effects, was performed and results were found to be consistent with the results presented here.

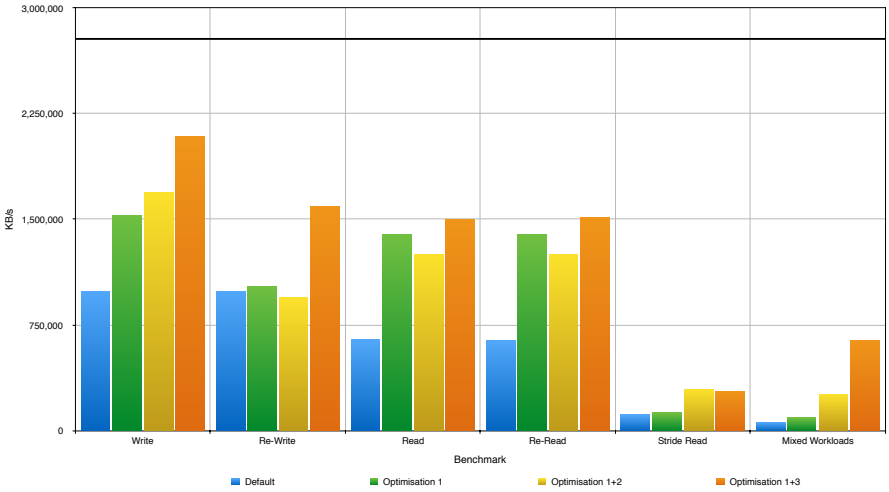

Figure 2. Effect of tuning on single server storage performance

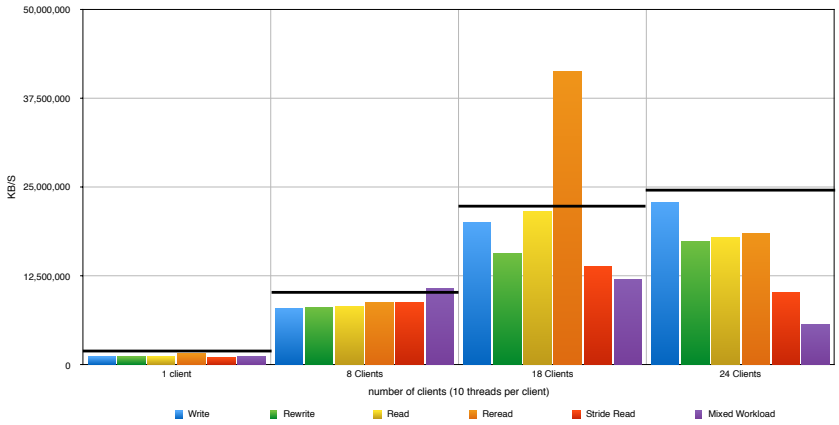

Figure 3. 1.5PB Lustre benchmark system performance

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. Lustre is set up such that individual files remain on a single OSS (i.e. there is no striping of files across OSSs). The well known Lustre clients tunes were included by default [1].

```
echo 256 > /proc/fs/lustre/osc/*/max_pages_per_rpc
echo 1024 > /proc/fs/lustre/osc/*/max_dirty_mb
```

For Lustre benchmarking using multiple clients IOzone is run with the `-+m filename` option to specify the client nodes (`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`). Figure 3 shows the benchmark results for different number of clients. Each client has a 10Gb/s network connection so this sets the upper limit of the storage performance until we have more than 20 clients (black solid line). As the number of clients increase the performance first increases and then falls off for all but the initial write test. The maximum performance of the storage is seen with 18 clients. The anomalous reread result for 18 clients is reproducible and may to be due to client side cacheing effects. With 24 clients the mixed workload performance is below that for 8 clients. The reason for the fall off in performance for large number of active clients is probably due to contention for resource when seeking data on the file system, this would be less important for the initial writes tests.

If we assume that a typical data analysis job uses 5MB/s and there is a maximum of 4000 job slots, then a throughput of the complete Lustre system of 20GB/s is required. The performance of the full Lustre file system, including 20 R730XDs and 70 R510s, is expected to be at least double that of the benchmarked system. If the real world workload is dominated by read type workflows, as is expected, then the full Lustre system should be able to provide the 20GB/s performance required.

A number of network optimisations were deployed in production based on recommendations found on the faster data web site [8], for both data transfers within the cluster and for those over the WAN by StoRM. These were not applied during the benchmarking.

6. Migration of Data

it was not possible to mount both Lustre 1.8 and 2.8 on the same client, therefore migration of data had to be done via rsync between two clients mounting the different Lustre file systems. Setting up an rsync daemon on the clients was found to be an order of magnitude quicker than using rsync over ssh for transferring data between the two clients. Hard links, ACLs and extended attributes are preserved by using the `-HAX` option when transferring data. Over the course of six weeks up to a dozen clients were utilised for the initial transfer of 1.5PB of data between the old and new Lustre file systems. After the initial transfer the old and new systems were kept in sync with repeated rsync runs, remembering to use the `-delete` option to remove files that no longer existed on the live Lustre system. MD5 checksums were compared for a random selection of files. The final transfer from the old to new Lustre took about a day, during which the file system was unavailable to external users. Then all clients were updated to the new Lustre version.

7. Real World Experience

The present Queen Mary Grid cluster has 3PB of Lustre storage and 4000 job slots in over 200 Lustre client compute nodes. Compute nodes fill the bottom 12U of every rack, where the air is cooler, and storage above them in the next 24U.

Real world performance over half a year, March to September 2016, is shown in figure 4. Figure 4(a) Shows that in 6 months ATLAS transferred 2.4 PB into QMUL and (b) transferred 2.3 PB from QMUL to the rest of the world. The Weekly transfer rates in to QMUL by ATLAS

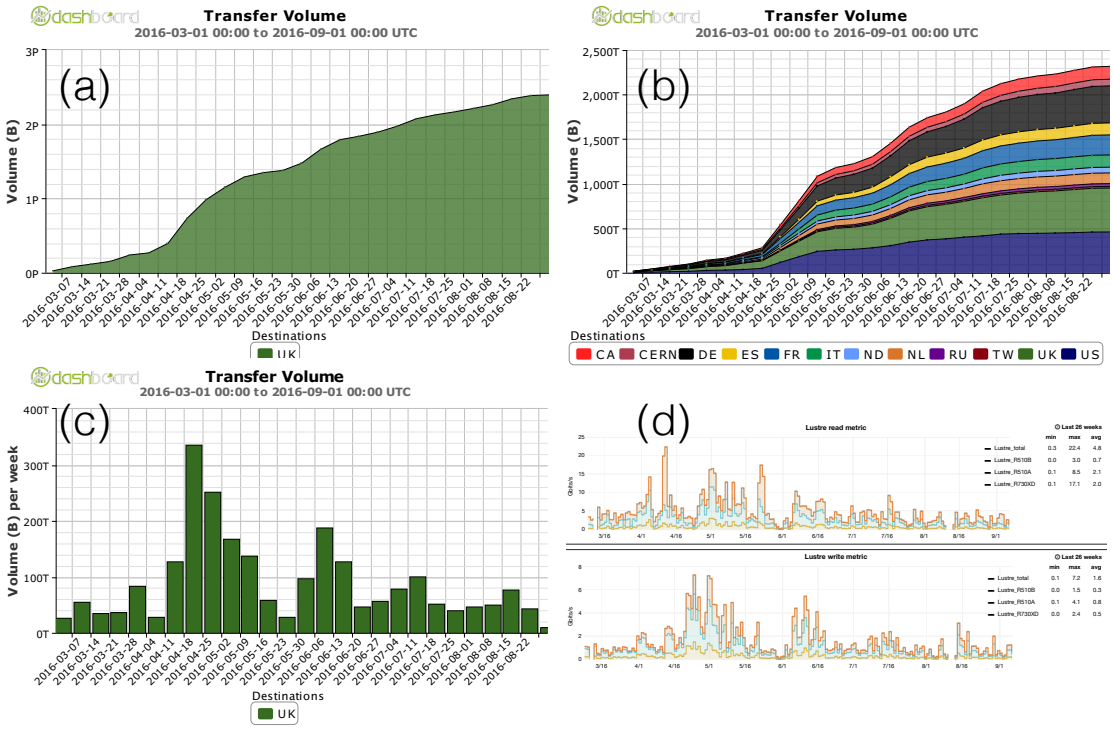

Figure 4. Six months data for (a) Cumulative data transferred into Queen Mary through StoRM. (b) Cumulative data transferred out of Queen Mary though StoRM. (c) Weekly totals of data transferred into Queen Mary through StoRM. (d) Actual Lustre read and write activity.

are shown in figure 4(c), with up to 340 TB transferred in one week. figure 4(d) show the Lustre activity over 6 months, an average read speed = 4.8 Gb/s and average write speed = 1.6 Gb/s is observed. During this period the Queen Mary Grid site was responsible for processing about 2.5% of all ATLAS data globally and about 20% of data processed in the UK [9].

It is possible for all job slots to be running analysis workloads, requiring access to data stored in Lustre, during these periods no slow down in data analysis was observed. However, in one case a local user simultaneously ran more than 1500 jobs each accessing a very large number of small files, in this case Bioinformatics data, on Lustre and a slow down in performance was observed. Once the user was limited to no more than 500 jobs no further issues were seen. It is expected that accessing small files on the Lustre filesystem is not efficient [1] and should be avoided or limited where possible. A future enhancement to Lustre is planned that will enable small files to be stored on the MDS which should improve small file performance.

8. Future Plans

- Double the Storage of the cluster to 6PB in 2018.
- Consider an upgrade to Lustre 2.9 which will have bug LU1482 fixed and also provide additional functionality such as user and group ID mapping which would allow the storage

to be used in different clusters. However Lustre 2.9 is SL/Centos7 only.

- Upgrade OSS servers to SL/CentOS 7 from SL6.
- 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.

9. Conclusions

Details of a successful major upgrade of the Lustre storage at the Queen Mary WLCG tier 2 Grid site have been presented with various performance measurements. During the upgrade and migration process the site was able to operate continuously with only one day of scheduled down time. The site continues to provide the largest UK tier two contribution to ATLAS storage requirements. There have been no observed performance or availability issues since the new Lustre file system went into production. There is expected to be no impediments to future evolution of the storage to CENTO 7, Lustre 2.9, or the use of ZFS.

10. Reference

[1] Lustre
http://www.seagate.com/files/www-content/solutions-content/cloud-systems-and-solutions/high-performance-computing/_shared/docs/clusterstor-inside-the-lustre-file-system-ti.pdf
http://www.intel.com/content/www/us/en/lustre/architecting-a-high-performance-lustre-storage-solution.html
Lustre Documentation: http://lustre.org/documentation/
Lustre BUG LU-1482: https://jira.hpdd.intel.com/browse/LU-1482
Lustre small files: http://www.opensfs.org/wp-content/uploads/2014/04/D1_S10_LustreFeatureDetails_Pershin.pdf
[2] StoRM: http://italiangrid.github.io/storm/index.html
[3] XrootD: http://xrootd.org/
[4] Scalable Petascale Storage for HEP using Lustre: Journal of Physics: C.J. Walker D.P. Traynor and A.J. Martin. *Journal of Physics: Conference Series* **396** (2012) 042063
[5] 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
[6] IOzone: http://www.iozone.org/
[7] BeeGFS Tips and Recommendations for Storage Server Tuning: http://www.beegfs.com/wiki/StorageServerTuning
[8] ESnet Fasterdata Knowledge Base: http://fasterdata.es.net/
[9] ATLAS dashboard: http://dashb-atlas-job.cern.ch/dashboard/request.py/dailysummary