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 existing networking equipment. Reuse existing storage, (1.5 PB in 12 Dell R510 servers); Don’t use failover for bulk storage.

3) Minimise do-ityourself 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 support RPM, GridFTP, Webad, Xrootd

6) Optimise for data analysis workloads (reading files)

Can’t afford off-the-shelf commercial system. Lustre/OSS etc... software support required.

Fastest – Low cost maintenance.

Amphitheatre: Our and Rost in own. does neither very well.

Mini Copper Winner: Monitor Carys only. 20k, 64, 4x.

Echo madvise > /sys/kernel/mm/

Echo madvise > /sys/kernel/mm/

### Solution

#### Existing Storage (OSI/OST) 1.5PB

- Dell M4400 / 7x Dell R510
- RAID 6

#### New Storage (OSI/OST) 1.5PB

- Dell 13овое с 16TB SAS disks, в RAID 6

#### Meta Data Server (MDM/MDT) HA

- Dell M4400 / 7x Dell R510

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

OS = Scientific Linux 6.7 Lustre version ~ 2.8

StorRM for SGB/GrifFTP/weekday (3xGrifFTP 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

- Optimisation 1
  - To test a single server client runs with 16 threads each transferring a file size of 24GB in chunks of 100KB
  - Sizes ~ 5x 1.5x 30GB/s ➔ 50GB/s ➔ 150GB/s

- Optimisation 2
  - For Lustre benchmarking up to 15clients, each client runs 10 threads transferring a file size of 24GB in chunks of 100KB
  - Sizes ~ 5x 1.5x 30GB/s ➔ 50GB/s ➔ 150GB/s

- Optimisation 3

OR

- Optimisation 1-2

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

A near complete 1.5 PB lustre file system with 20 Dell 130XD servers was created with up to 24 client nodes dedicated to the benchmark tests

### Real World Performance of full 3 PB system

In 6 months ATLAS transferred 2.4 PB to QMUL

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

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