Research Storage Platforms

Matt Rásó-Barnett
Research Computing Platforms

Platforms Overview

Lustre

Large-scale, distributed parallel filesystem

Lustre HSM

Robinhood Policy Engine coordinates tiering of data between Lustre and Tape system

QStar Archive Manager

FUSE filesystem combining disk cache with tape archive

ZFS - NexentaStor

2x two-node ZFS clusters in different sites, asynchronous replication between each other

Spectra Logic T950

Two T950 libraries each containing 10PB LTO7

Platforms Overview

Lustre

Large-scale, distributed parallel filesystem

Research Data Store - (RDS)

Our new HPC parallel filesystem(s) - single global namespace

Currently composed of 5x 1.4PB Lustre filesystems

- **7PB storage in total**

Accessible both remotely via ssh:

- scp/rsync/sftp to `rds.uis.cam.ac.uk`

or through our HPC clusters - CSD3 and Darwin/Wilkes

```
login.hpc.cam.ac.uk
login-gpu.hpc.cam.ac.uk
login-knl.hpc.cam.ac.uk
```

Platforms Overview

Lustre

Large-scale, distributed parallel filesystem

Lustre HSM

Robinhood Policy Engine coordinates tiering of data between Lustre and Tape system

QStar Archive Manager

FUSE filesystem combining disk cache with tape archive

Research Cold Store - (RCS)

Hierarchical Storage system in front of large-scale Tape archival system, comprising:

- 240TB Lustre filesystem - the front-end ‘interface’ and cache, utilising Lustre HSM
- QStar + Tape libraries - the back-end receiving archival data and moving to tape

Spectra Logic T950

Two T950 libraries each containing 10PB LTO7

Platforms Overview

Research File Share - (RFS) and Institutional File Share (IFS)

Each share is a ZFS dataset with regular snapshot schedule and replicated to secondary NexentaStor cluster in remote DC

NexentaStor servers are bound to UIS Blue AD, providing users and group information

Shares are accessible via SMB/CIFS at

```
rfs.uis.private.cam.ac.uk
```

ZFS - NexentaStor

Two 2-node ZFS clusters in different sites, async replication between each other

QStar Archive Manager

FUSE filesystem combining disk cache with tape archive

Spectra Logic T950

Two T950 libraries each containing 10PB LTO7

Lustre

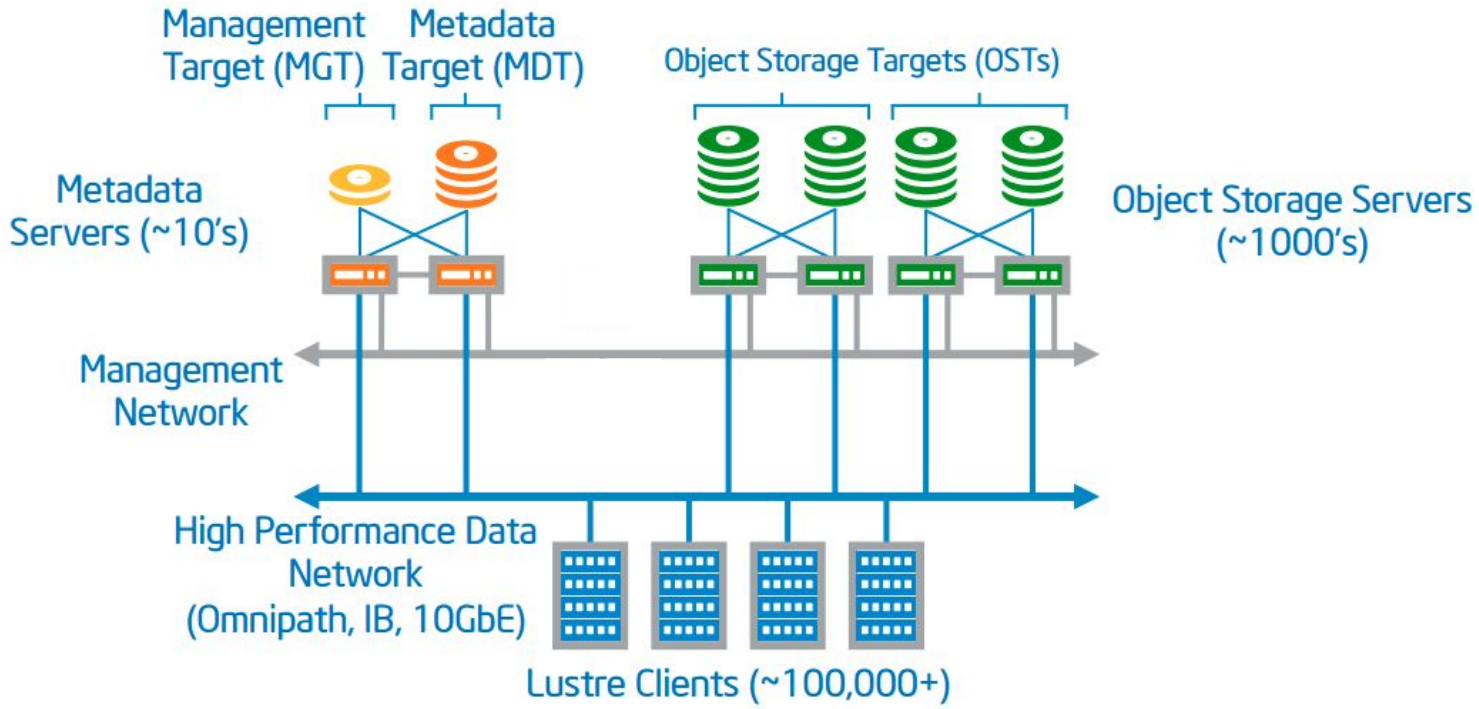
What is Lustre?

- Scale-out, parallel distributed Filesystem
  - Tens of Thousands of clients
  - Large capacities (some US institutes have 70PB+ filesystems)
  - High bandwidth (>TB/s, scales close to linearly for seq IO with number of storage servers)
  - POSIX semantics
  - Open source under GPL2 - runs as Linux kernel modules
  - Very popular in HPC community - used by vast majority of Top500 supercomputers

- Some of its features:

File striping across storage targets (multi-TB file sizes)	RDMA support
Multiple metadata servers	I/O Routing between network technologies (eth, IB, OPA)
Multiple backend-storage formats (ldiskfs, ZFS)	Storage pools
HSM integration	High availability

Lustre Architecture



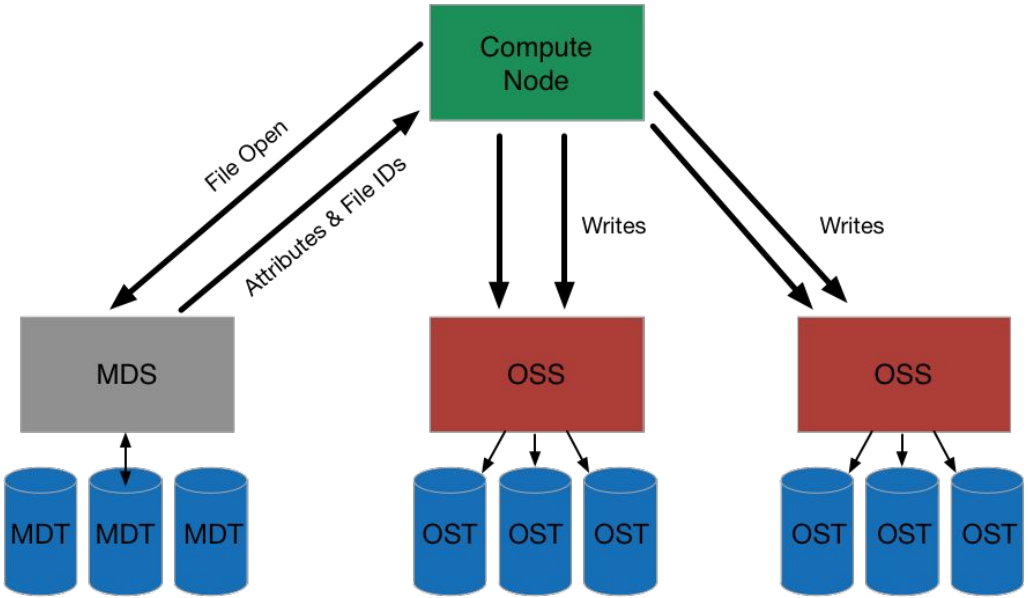
Lustre IO Path

When a client accesses a file, it performs a lookup on the MDS

The MDS server responds to the client with information about how the file is striped (which OSTs are used, stripe size of file, etc.)

Client directly contacts appropriate OST to read/ write data

After the initial lookup of the file layout, the MDS is not involved in file IO operations since all block allocation is managed internally by the OST



Lustre - Typical Filesystem Hardware

Type	Quantity	Specs
MDS Server	2	Dell R630 - Dual E5-2667v3 3.2GHz 8 Core 128GB RAM FDR IB, 10GB Ethernet
MDT Storage	1	Dell PowerVault MD3420 20x 300GB SAS 15K HDDs Dual RAID controller with 8GB cache
OSS Server	6	Dell R630 - Dual E5-2623v3 3.0GHz 4 Core 64GB RAM FDR IB, 10GB Ethernet
OSS Storage	6	Dell PowerVault MD3460 60x 6TB NL-SAS HDDs Dual RAID controller with 8GB cache
Network	Infiniband	Mellanox SX6036 FDR IB switches 36x 56Gb FDR
	Ethernet	Mellanox SN2410 Spectrum switches 48x 10GbE + 8x 100GbE

Dell MD3420
MDT - 20 disk Raid 10 with 4x SSD cache device

Dell MD3460
Each chassis supports 6x OSTs
Each OST: 10-disk Raid-6, ~58TB per OST
~350TB per chassis

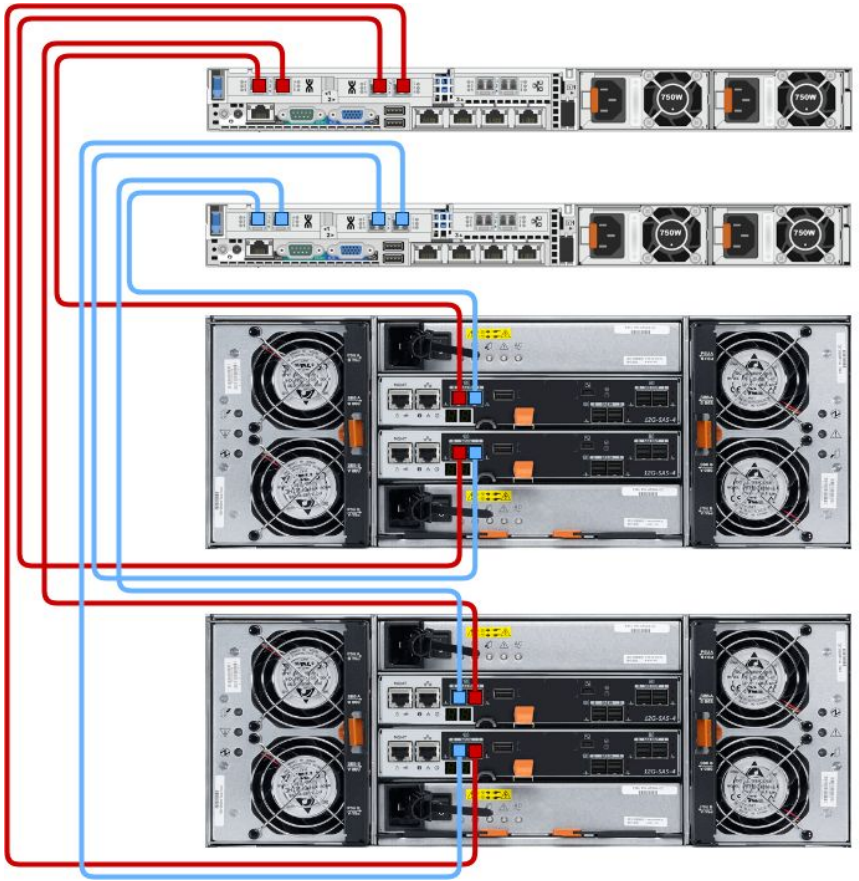
Lustre Redundancy

Each Lustre server part of 2-node active-active HA cluster - every storage target can be served by one of two nodes

Failover will cause IO to the affected target block until failover is complete, after which IO will resume transparently

Each server and storage array has fully redundant IO paths

Each 10-disk Raid-6 is spread over the 5 shelves of the array, accommodating an entire tray failure (this has actually happened to us once!)

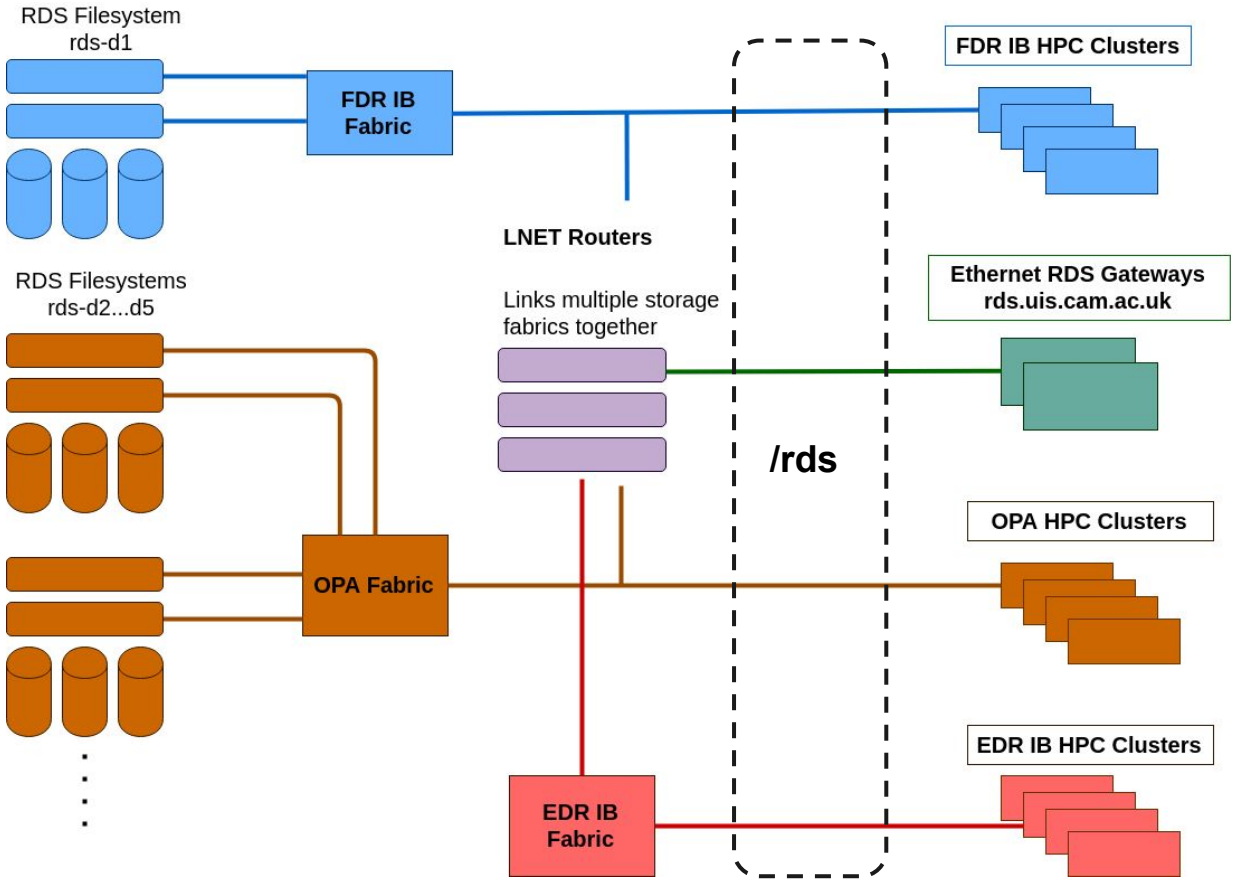


RDS

Multiple filesystems abstracted into one global namespace

Multiple network technologies across various access points

LNET routing bridges networks to ensure access from all user gateways



Tape

Why Tape?

Need for ever-growing amounts of storage

Expensive to just keep expanding Lustre/ZFS to handle this

Often research outputs must be stored for long periods beyond active usage

Tape offers unparalleled storage density for cost

However tape latency is high - best suited for archival data

Tape libraries are not easy to work with directly, need good abstraction (tape filesystem, backup utility)

Research Cold Store (RCS) aims to address this use-case

Tape Library

We have two Spectralogic T950 Tape libraries, each with two expansion chassis attached

One located in WCDC, one in Soulsby

Each currently has just over 1500 LTO7 tapes, with uncompressed ~6TB capacity, making approx 10PB of uncompressed capacity in each location

Each Library has 12 IBM Ultrium-TD7 drives allowing large number of simultaneous read/write operations to be performed in parallel

Linked via inter-DC fiber-channel network

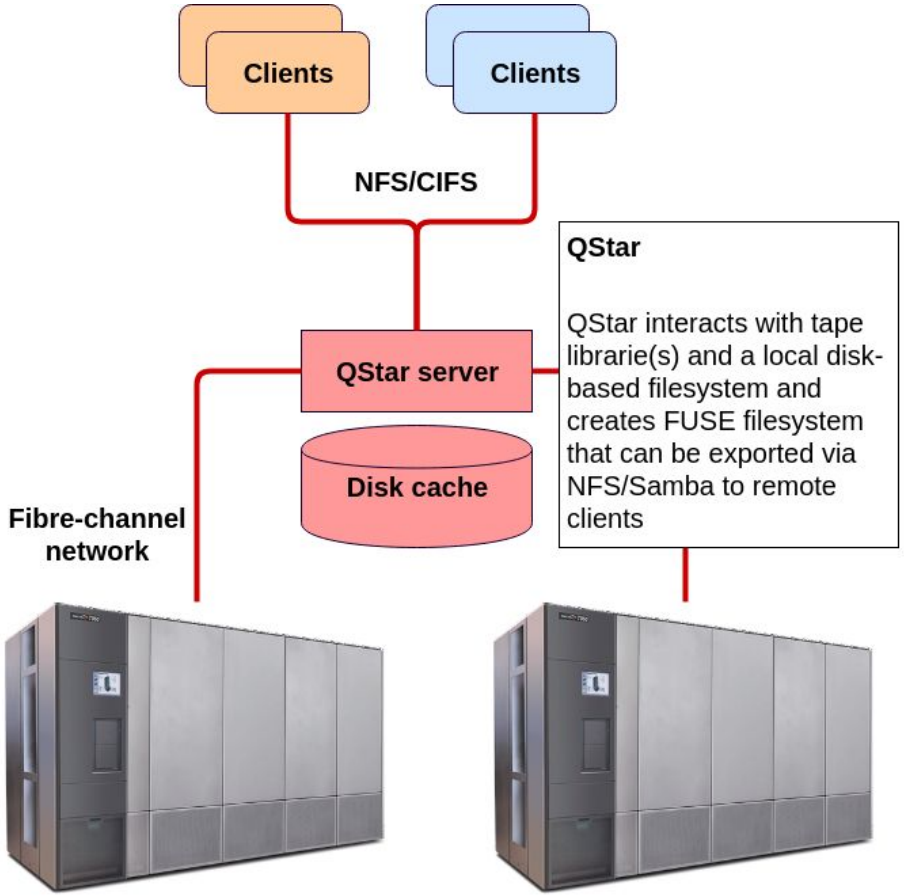
QStar ‘Active Archive’

QStar Archive Manager simplifies how we interact with our tape libraries

QStar creates a FUSE filesystem combining a local disk-based filesystem with a tape filesystem (LTFS, QStar’s TDO…), providing POSIX semantics.

The QStar FUSE filesystem is itself a type of HSM, where data is initially written to the disk cache and asynchronously replicated to tape(s) via configurable policies

Reading data from the filesystem that has been evicted from cache will block while the tape robot replicates the data back into the cache



RCS

For Research Cold Store, we put an additional Lustre ‘cache’ filesystem in front of the QStar tape-system

Lustre fits our needs as more capable interactive filesystem, with user/group(and recently project)-quotas, more robust POSIX semantics, and has it’s own HSM capability we can take advantage of

Users interact directly only with Lustre - Lustre HSM works in the background staging data between it and QStar

Lustre

240TB ‘cache’ filesystem

Tier 1

Robinhood

Filesystem Analysis and HSM Policy Manager

QStar Archive Manager

350TB disk cache

Tier 1.5

Spectralogic T950

10PB tape - Data replicated to two tapes in different libraries

Tier 2

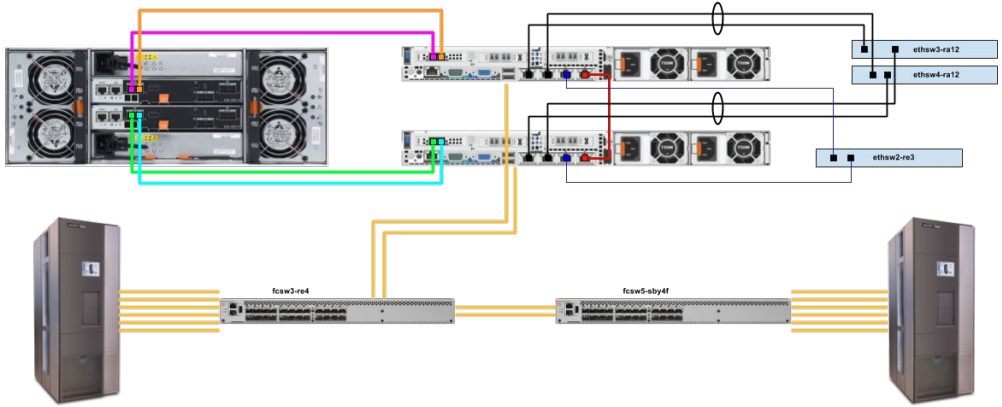
Tape System Hardware

Type	Quantity	Specs
QStar Server	**2**	**Dell R630 Dual E5-2667v3 3.2GHz 8C 128GB RAM Bonded 2x 10GB Ethernet**
Cache Storage	**1**	**Dell PowerVault MD3460 60x 6TB NL-SAS HDDs 6x 10-disk RAID-6 arrays composed into single striped-LVM volume**
Network	**Fibre Channel**	**Brocade 6505 FC Switches 24x 16Gb**
	Ethernet	**Mellanox SN2410 switches 48x 10GbE + 8x 100GbE**
Tape	**2**	**Spectralogic T950 libraries 10 PB each library 12x IBM Ultrium TD7 Drives**

Approximately 350TB of disk-cache managed by QStar

QStar runs on RHEL 7.3, can be made HA via standard pacemaker+corosync toolchain

We use nfs-ganesha NFS server implementation which provides factor-2x improved performance for QStar FUSE backend

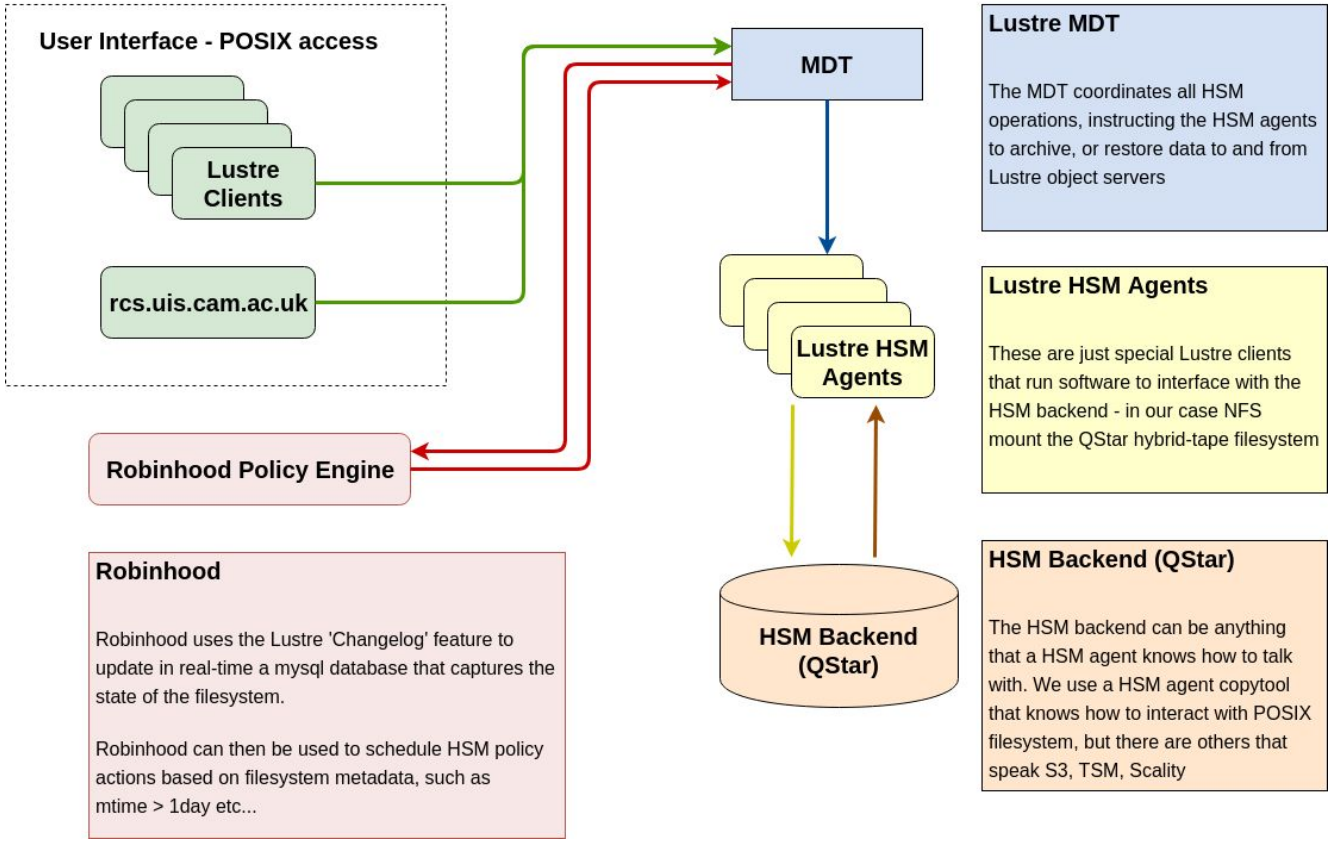


Lustre HSM

Lustre HSM functionality is transparent to the user

Data replicated to HSM backend periodically based on configured policy - approximately daily

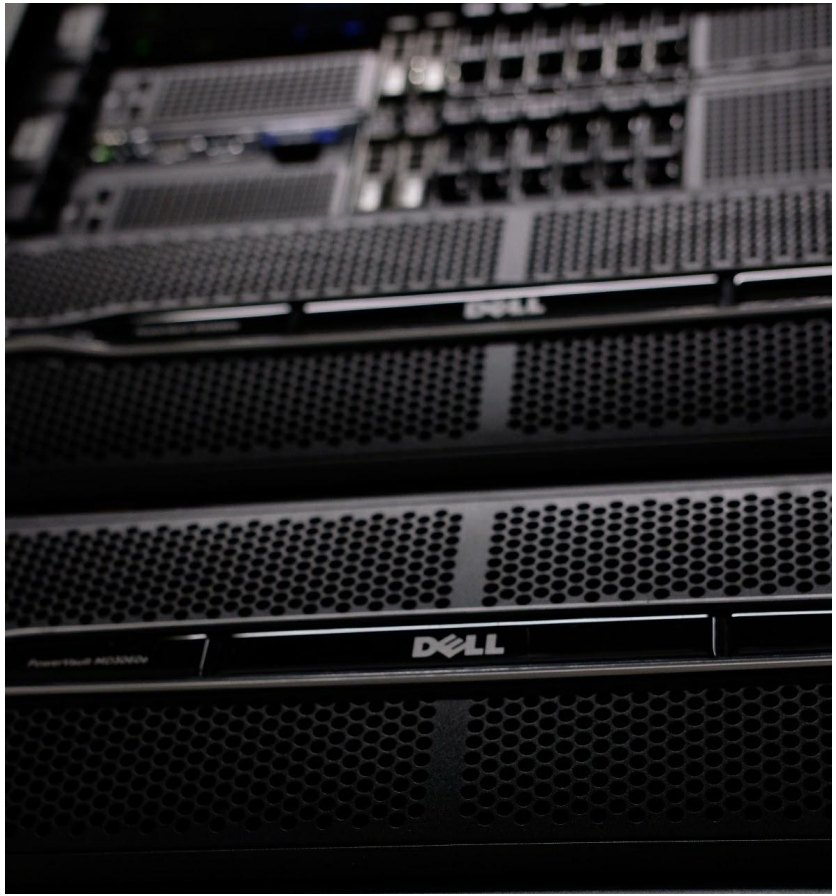
Data released from Lustre ‘cache’ when cache begins to fill up, configurable by policy (so eg: larger, least recently used files released first etc…)



How to use RCS

- Due to Lustre's architecture, filesystem namespace remains the same whether file's data is on tape or on disk
- File that is replicated to tape and removed from Lustre's disk will show as having zero size
- Accessing a file replicated to tape **will block while data staged back to lustre** - your process will resume once this has completed
- Due to this, RCS is heavily suited towards **LARGE** files, (eg: pre-packaged tar-archives). Requesting large numbers of small files will incur a heavy latency penalty as the tape system seeks around different tapes
- We are working on a major documentation update to provide more fine-grained user-interface, so can query file's state before requesting it be replicated back to disk, and async-replication etc...

ZFS



ZFS Overview

Advanced copy-on-write filesystem and logical volume manager

Originally developed at Sun Microsystems, first introduced in 2005, now OpenZFS open-source project available on wide-range of *nix-based OS's

Many features:

Block-level checksums - provably consistent tree, self-healing	Transparent compression - (HPCS home directories see compression ratio of 1.83x with LZ4)
Extremely lightweight efficient snapshots	Transaction groups - high level of data integrity
Online filesystem replication via zfs send/receive	Raid-Z

ZFS Overview

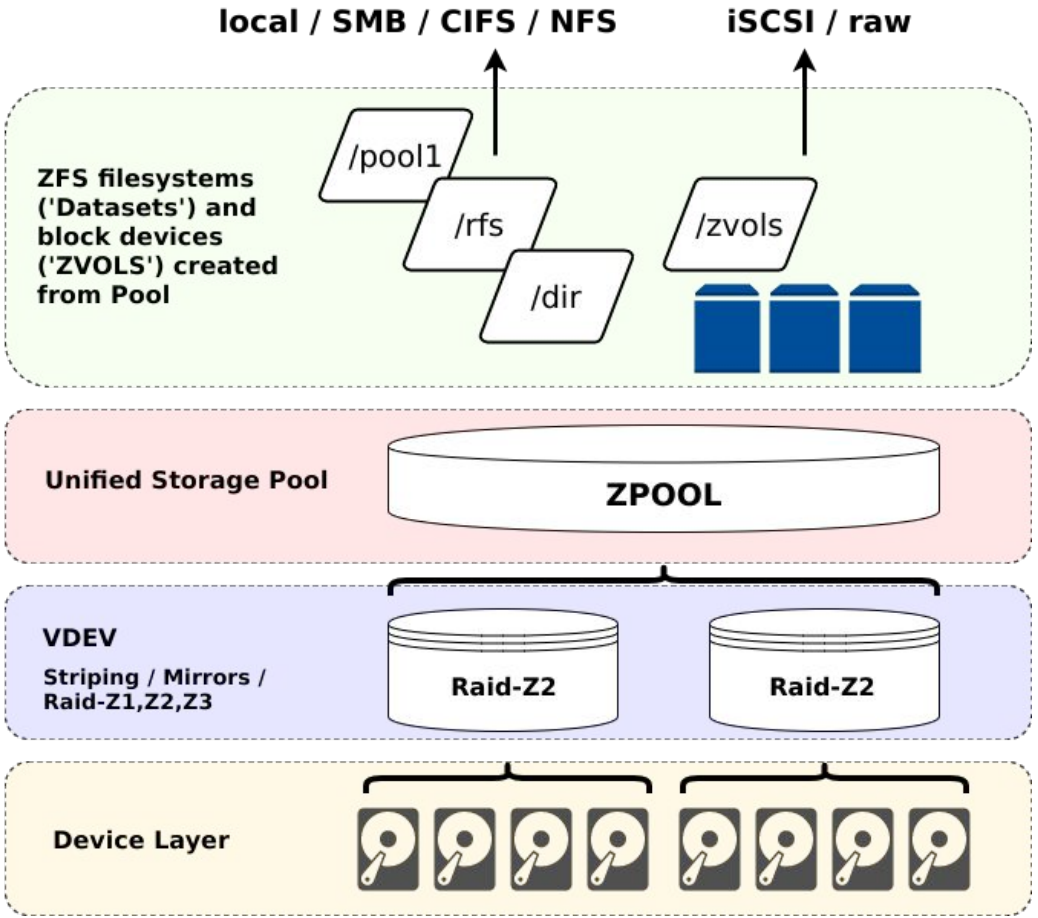
ZFS manages the whole IO path from filesystem to disk

Raw disks are combined into VDEVs which have various analogues to Raid levels

VDEVs are aggregated into a zpool, which stripes writes across all VDEVs

ZFS filesystems (‘datasets’) can then be created on the zpool, in a hierarchical manner.

ZVOLs are zfs-backed block devices - can be exported to other machines



NexentaStor

Commercial ZFS appliance based on Illumos

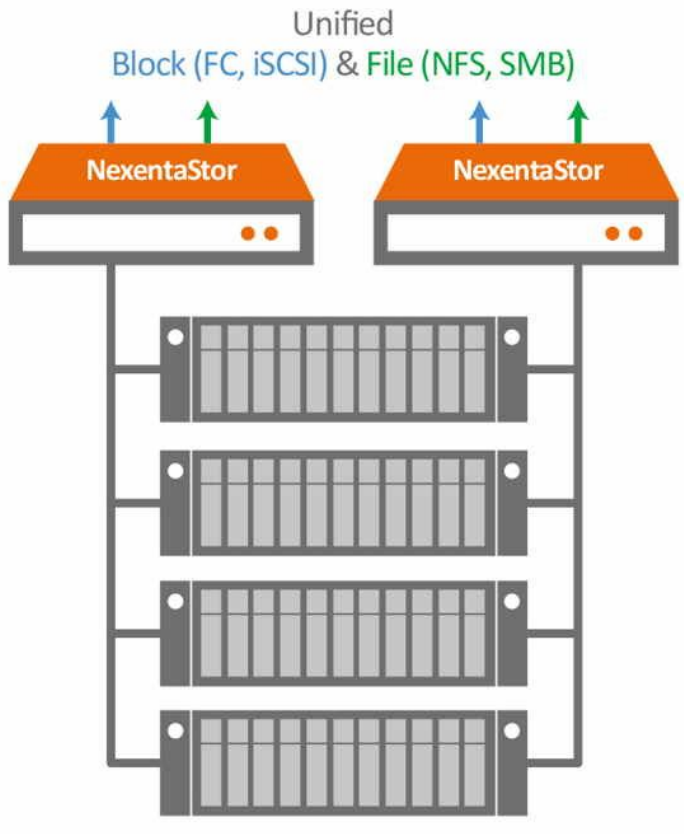
Offers simplified management and enterprise support

There is GUI/CLI configuration, as well as API and Openstack Cinder plugin for automated changes

Offer a HA-cluster product which we use

Automated jobs for snapshots schedules and replication for remote synchronisation between clusters

Active-Directory integration



ZFS Hardware Configuration

Per cluster:

Type	Quantity	Specs
Servers	**2**	**Dell R730** **Dual E5-2637v3 3.5GHz 4 Core** **256GB RAM** **2x 120GB SSD**
Storage	**5**	**Dell PowerVault MD3060e** **58x 6TB NL-SAS HDDs** **Across all 5 arrays:** **8x write-intensive 200G SSD for SLOG** **2x read-intensive 400G SSD for L2ARC** **590x Disks for 10-disk RAIDZ2 VDEVs** **1x Pool containing 14x RAIDZ2 VDEVs** **1x Pool containing 15x RAIDZ2 VDEVs**
Network	**Ethernet**	**2x 10G bonded for data network** **2x 10G bonded for replication network** **Replication network links sites via dedicated fiber**

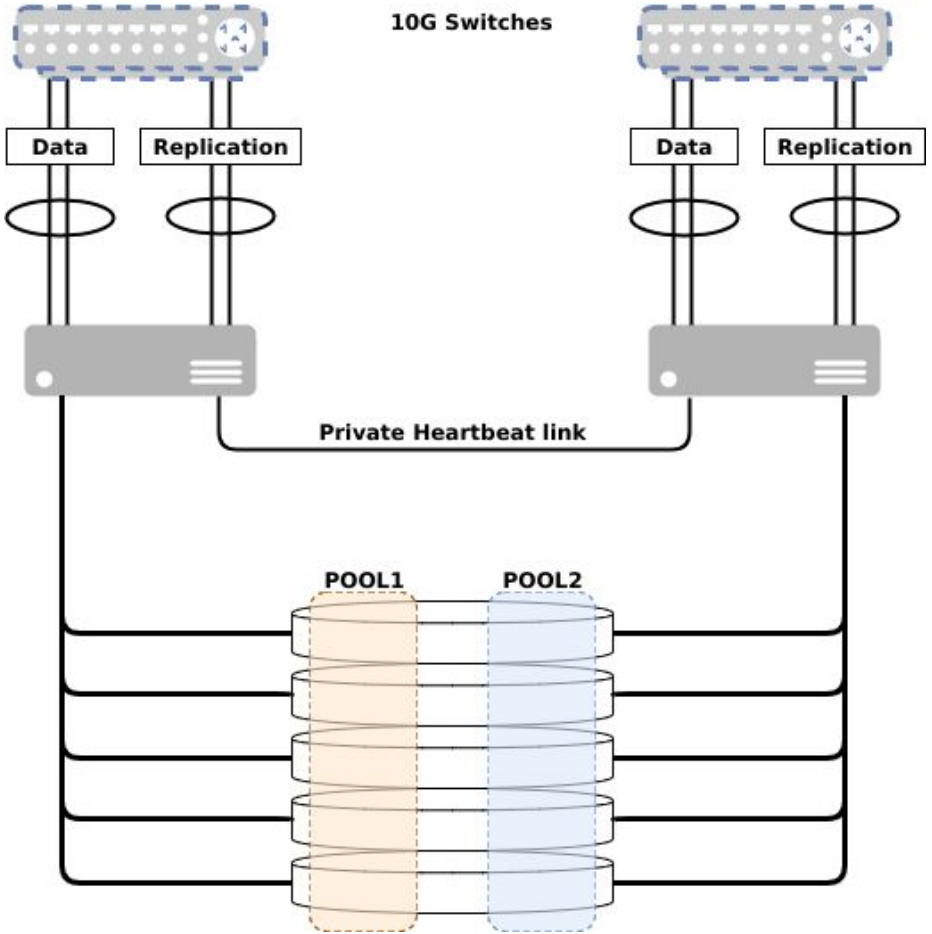
ZFS Resiliency

Each two-node NexentaStor cluster servers two pools, in active-active configuration

Each pool consists of either 14 or 15 10-disk RAIDZ2 VDEVs

Raid groups arranged such that we can tolerate either a whole tray failure in any JBOD, or a whole single JBOD failure

ZFS provides provable end-to-end data integrity - block checksums that can be validated through the entire hierarchy protecting against silent data corruption



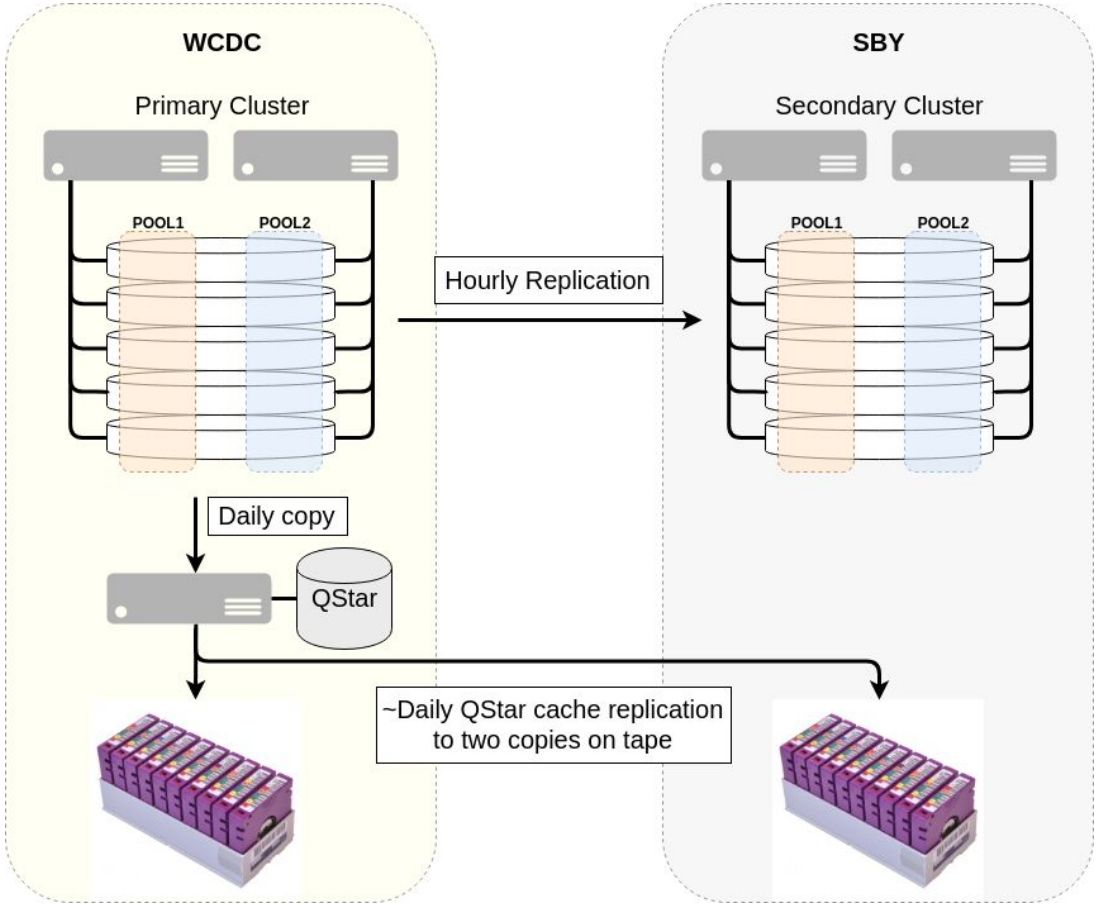
RFS Storage Architecture

With both services you are effectively purchasing ZFS datasets on one of the pools

Hourly, daily, weekly snapshots

Hourly dataset replication (along with all snapshots) to remote cluster

Optional daily copy of dataset to QStar tape archive (which then asynchronously pushes data to 2x tapes one in each location)

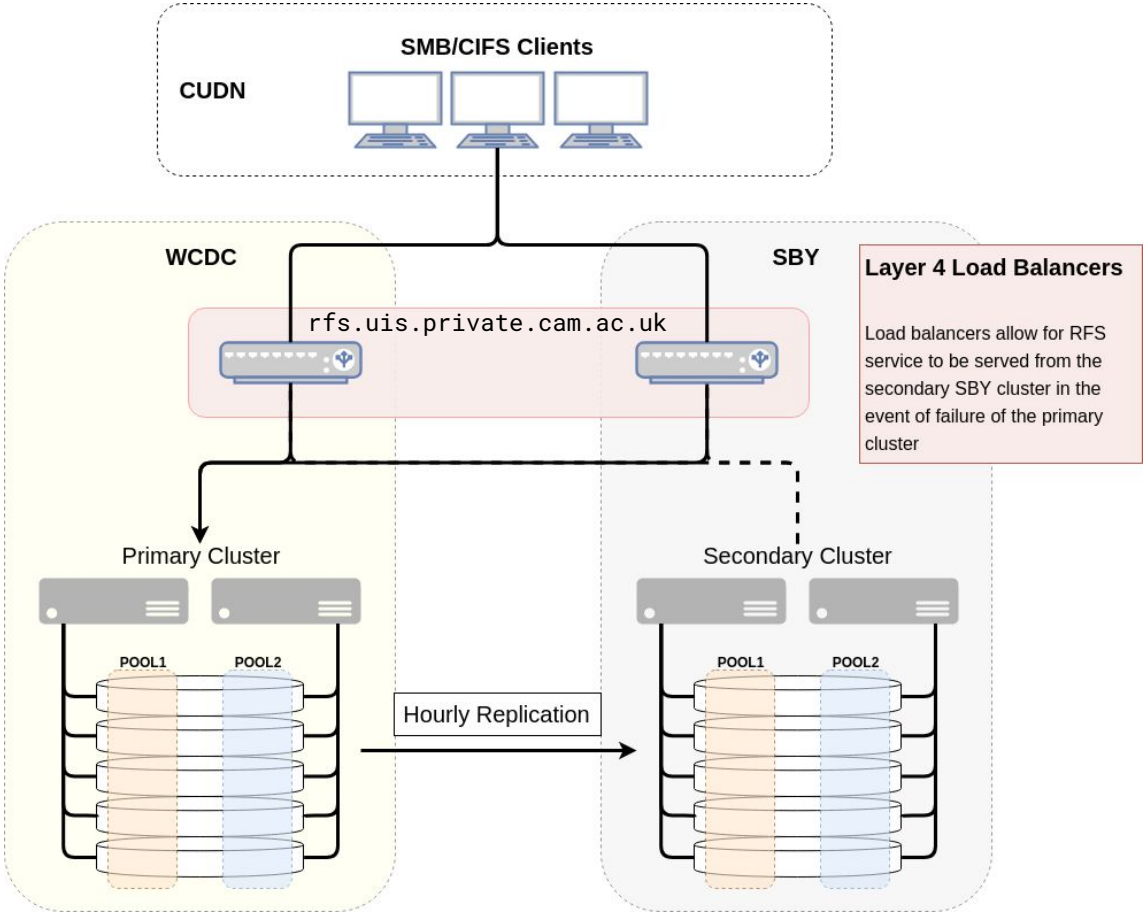


RFS Access

RFS access via SMB/CIFS protocol through rfs.uis.private.cam.ac.uk

SMB service sits behind layer-4 load-balancers which provides ability to serve a dataset from its replica on the secondary cluster

This provides a level of service continuity in the event of a failure of the primary cluster



RFS Administration

All managed through Storage self-service gateway - `https://selfservice.uis.cam.ac.uk/`

LZ4 Compression as standard

Standard snapshot schedule:
- Hourly snapshots for 24 hours
- Daily snapshots for 7 days
- Weekly snapshots for 30 days

Snapshot storage is counted towards storage allocation

Currently working towards exposing this through self-service gateway and allowing a degree of configurability

Summary

Platform	Research Data Store	Research Cold Store	Research File Share
Technologies	Lustre	Lustre HSM QStar Tape	ZFS / NexentaStor QStar Tape
Features	Highest performance High capacity Low cost per TB Single copy on disk Accessible remotely via SSH Mounted on HPC clusters	High latency, archival Suits large, infrequently accessed data Highest capacity Lowest cost per TB Two copies on tape Accessible remotely via SSH Mounted on HPC clusters	High performance Strongest data integrity Snapshots Replicated offsite High availability Accessible over CUDN via SMB/CIFS