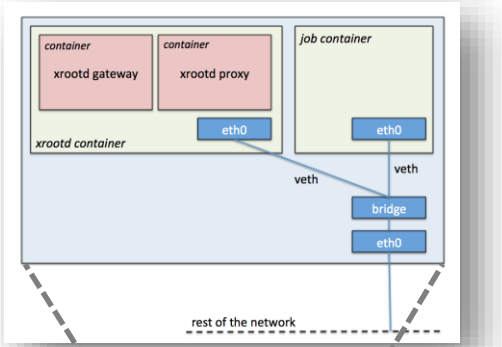
RAL QoS

Alastair Dewhurst

Ceph Object Store

- Question: Are Object stores fundamentally different from traditional Grid storage?
  - Answer: With WLCG use cases not really.
- Ceph uses algorithmic data placement:
  - No central catalogue for meta data queries.
  - Vector reads not supported.
- Very few WLCG use cases need a file system and it is mostly about educate user / fixing bugs due to invalid assumptions.

Echo



External Gateway (XRootD, GridFTP, S3)

echo

Regional Redirector

XRootD Manager

Proxy Cache

XRootD WN | XRootD WN | XRootD WN

XRootD WN | XRootD WN | XRootD WN

Ceph Backend

Erasure Coding

- Echo uses Erasure Coding rather than Replication
- 8 + 3 profile that spreads data across 11 hosts.
  - 73% RAW storage usage.
  - We can survive the loss of any 3 hosts without losing data.
  - Due to the number of hosts we have, we can operationally lose a rack.
- EC does mean that data needs to be sent to “re-assembled” one one host before being sent to the client.
  - Minimum size 8MB.

Operations

- Ceph has been extremely reliable.
  - Averaging one outage a year.
  - One data loss incident since 2017 due to a bug, which resulted in operator error.
- Vast majority of issues with XRootD plugin which are slowly being resolved.

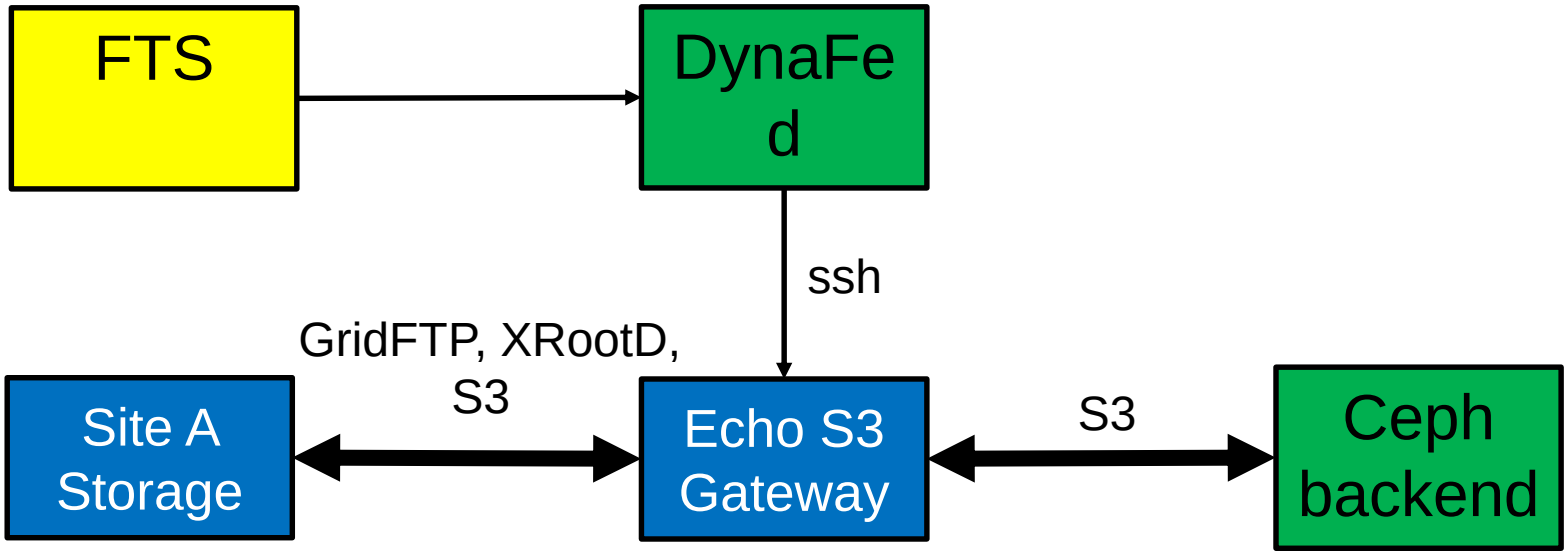
Tape

- RAL currently provides a Tape Archive via Castor.
  - Large (2PB) disk buffer externally accessible.
- In 2021 we should be migrating to CTA.
  - Small SSD buffer that will only allow transfers to internally accessible storage.
  - External experiments will need to transfer via Echo and use multi-hop FTS transfers.
- New Tape Robot installed today!
  - 6500 media slots, 20 x TS1160 drives.

SSD Caching

- RAL jobs have been observed to have lower efficiency compared to other Tier-1s.
- Evidence points to IOPs limits on the WN.
  - New WN will have SSD storage.
- Will use CTA testing to decide if SSD caches are sensible to deploy more widely.

S3 and DynaFed



- RAL are trying to use DynaFed to provide access to the S3 storage provided by Ceph.
  - Much easier to maintain than XRootD / GridFTP.
  - S3 buckets can easily have different replication policies.

Multi-VO Rucio

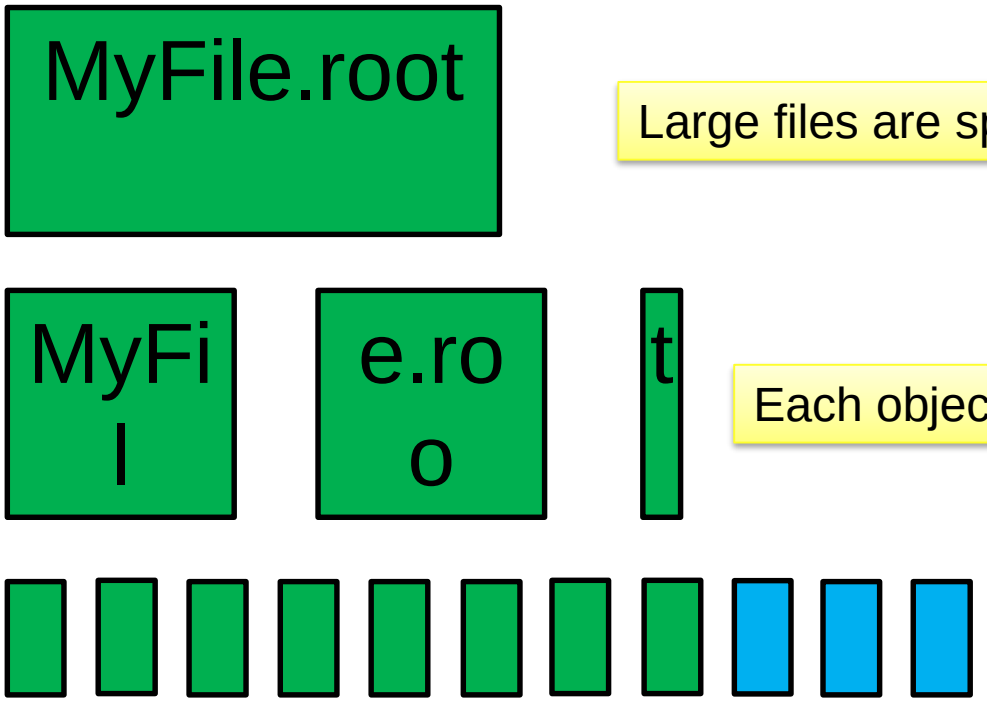
- RAL is developing and deploying a Multi-VO Rucio instance.
  - Aim for production from June 2020 onwards.
- In future, RAL will need to provide data management tools for a large number of HEP, Astronomy and Space experiments.
- Aim to encourage small VOs to use Rucio which will manage data between Echo (Grid / S3 Storage), Castor / CTA, UK sites and the rest of the world.

QoS Classes

- Start with 4 types:
  - Archival Storage.
  - High availability and reliability disk storage.
  - Disk storage.
  - High performance storage.

Backup

Erasure Coding



Large files are split into 64MB objects.

Each object is split into 8 x 8MB chunks.

3 x 8MB parity chunks are calculated.

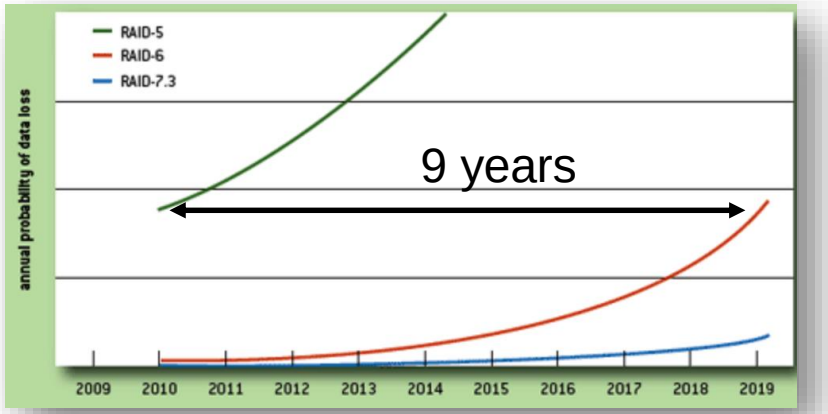
Each of the 11 chunks are written to different storage nodes.

How do we know which disks are written too?

RAID & Erasure Coding

- Hardware RAID has provided us with resilience against disk failure for decades.
  - As disks continue to grow but their performance stays relatively constant, risk of data loss increases.
- Erasure Coding (EC) can be seen as an extension to RAID and is becoming significantly more popular.
  - EC is done in software - data is spread across storage devices.
  - Requests for partial data from an object may result in storage re-assembling entire object.

Plot shows theoretical failure rate as storage grows for different RAID configurations.



http://queue.acm.org/detail.cfm?id=1670144