Object Stores

Alastair Dewhurst

Introduction

- What do we mean by Object Store?

Backend

It is provided by a commercial provider.

Ceph is an object store.

Frontend

It uses S3 / Swift API for access.

Design Philosophy

It has multiple replicas of the data.

Object Stores

- What do I mean when I talk about Object Stores?
  - Something that stores objects (but does precious little else)!
- What is an object:
  - A piece of data
  - Metadata (extended attributes)
  - A unique identifier
- Different from a file system:
  - Get / Put semantic
  - No file descriptor
  - Flat namespace
  - Small Objects (few MBs)

Ceph is an Object Store

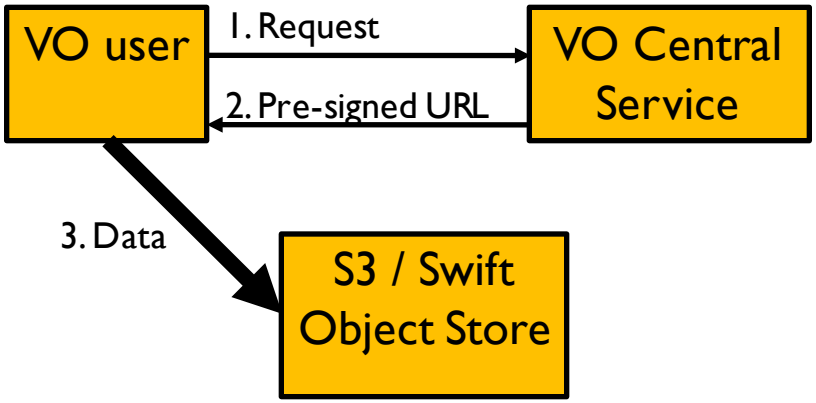
- At the core is CRUSH, which is a clever data placement algorithm:
  - No central catalogue of object placement.
  - Placement computation done by clients.
- Ceph provides block and object storage.
  - Popular as a backend for Cloud Machines (block), gaining popularity for Object Storage.
- CephFS requires the addition of Meta Data Servers and this moves away from Object Store model.
- Some commercial clouds run Ceph, others run their own software but a lot of the principles remain the same.

Commercial Clouds

- CPU has been successfully utilized on commercial clouds.
  - Less success with Storage.
- Significantly more complicated cost model.
  - Charges for moving data in/out of cloud as well as total stored.
  - Cost still an order of magnitude more than Grid resources.
- Much higher Quality of Service than Grid sites.
  - Even high priority work does not justify cost (yet).
- Different protocols to what we are used too.
  - S3 / Swift API over https.

S3 / Swift API

- S3 is the name of the Amazon's Storage Service and also the API to use it.
  - Swift is the OpenStack equivalent.
  - They use https and logically work in an almost identical manner.
- Designed for developers and encourages proper usage.
  - You can put S3FS on top of S3 and an SRM on top of a POSIX file system, but please don't.



An S3/Swift account provides a single access + secret key.

Pre-Signed URLs are a way to let users upload or download specific objects to/from buckets, but without requiring them to have the access + secret key.

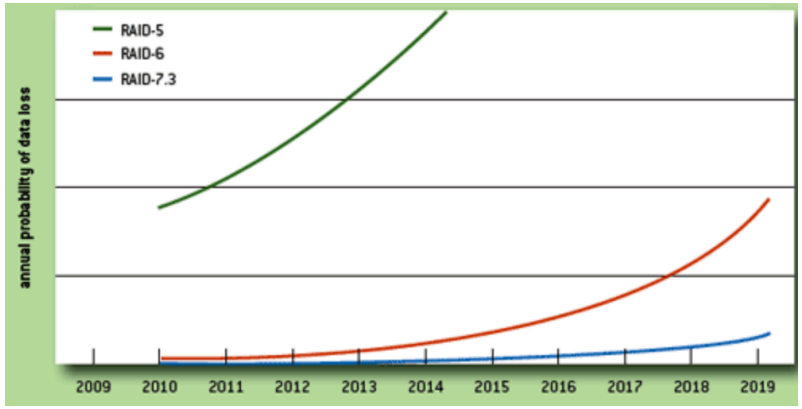
You do **not** need to interact with the Object Store in order to generate a pre-signed URL!

Replication & EC

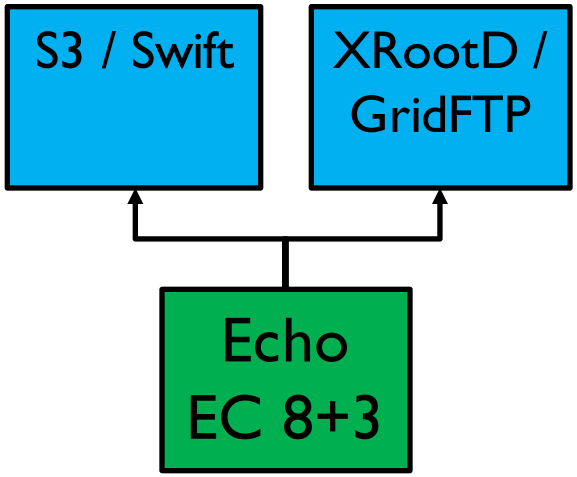
- Most commercial Object Stores replicate their data.
  - Most WLCG sites running Ceph also used replication initially.
  - A few known exceptions: Facebook photos, Spotify songs, Backblaze.
- Erasure Coding (EC) can be seen as an extension to RAID and is becoming significantly more popular recently.
  - 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.

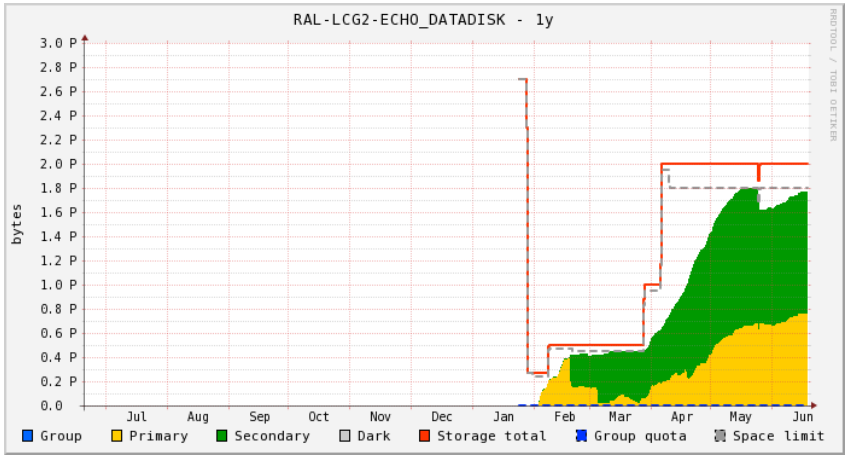
The plot shows theoretical failure rate as storage grows for different RAID configurations.

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



Echo at RAL





- RAL Tier-1 is now providing some of its pledged disk resources via a Ceph Object Store.
- XRootD and GridFTP plugins have been built.
  - Offering ONLY S3/Swift API access to non-LHC users.
- Problems / complications when not using it like an Object Store.
  - XRootD direct I/O and requests for metadata.

Conclusions

- Commercial providers have shown that Object Stores work at the scales we need for HL-LHC.
  - For cost purposes we will need to be using Erasure Coding.
- Current WLCG computing models work very well with Object Stores.
  - Most jobs contact central servers for their payload and primarily interact with storage to get/put data.
- Object Stores need to be used correctly!
  - (XRooTD) Direct I/O can be problematic.
  - Avoid metadata queries.
- Wider support of Webdav by existing storage for WAN transfers would be helpful.