Echo, Object Stores, Experiment Integration

Alastair Dewhurst

echo

Introduction

Q. What do I mean when I talk about Object Stores?
A. 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)

Q. Why should we care about Object Stores?
A. They have been proven to work at the scales we need for HL-LHC and they can potentially be significant simpler for a site to run.

echo

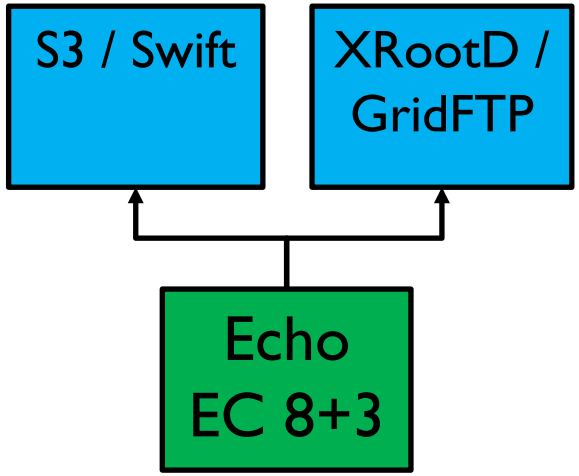
Echo - Motivation

- In October 2014 RAL decided to start work on a replacement for Castor disk.
- RAL Tier 1 supports all LHC VOs and a large number of other (mostly HEP) VOs.
  - Expect usage from non-LHC VOs to exceed LHC VOs for majority of the lifetime of this new storage.
  - Want to provide industry standard tools.
- Decided to build a Ceph object store providing S3/Swift API access.
  - Could we get LHC VOs to use S3 in time?
    - Not at that time.
  - XRootD and GridFTP plugins to librados were written.

echo

Echo - Design

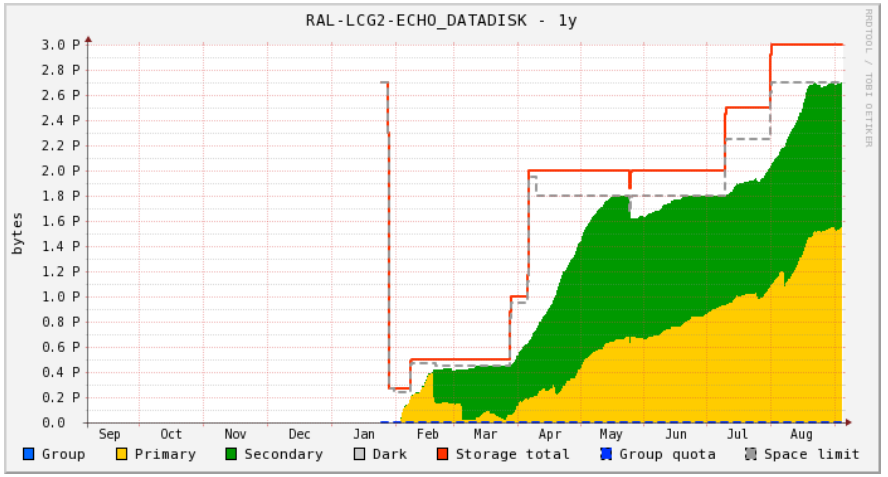
An XRootD and GridFTP plugin have been written for Ceph.
Cannot access same data via S3/Swift.



- Echo uses Erasure Coding for data resilience.
  - 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.
- Requesting any data requires the reassembly of an object.

Echo - ATLAS

Since April 2017, the RAL Tier-1 is providing some of its pledged disk resources via Echo.



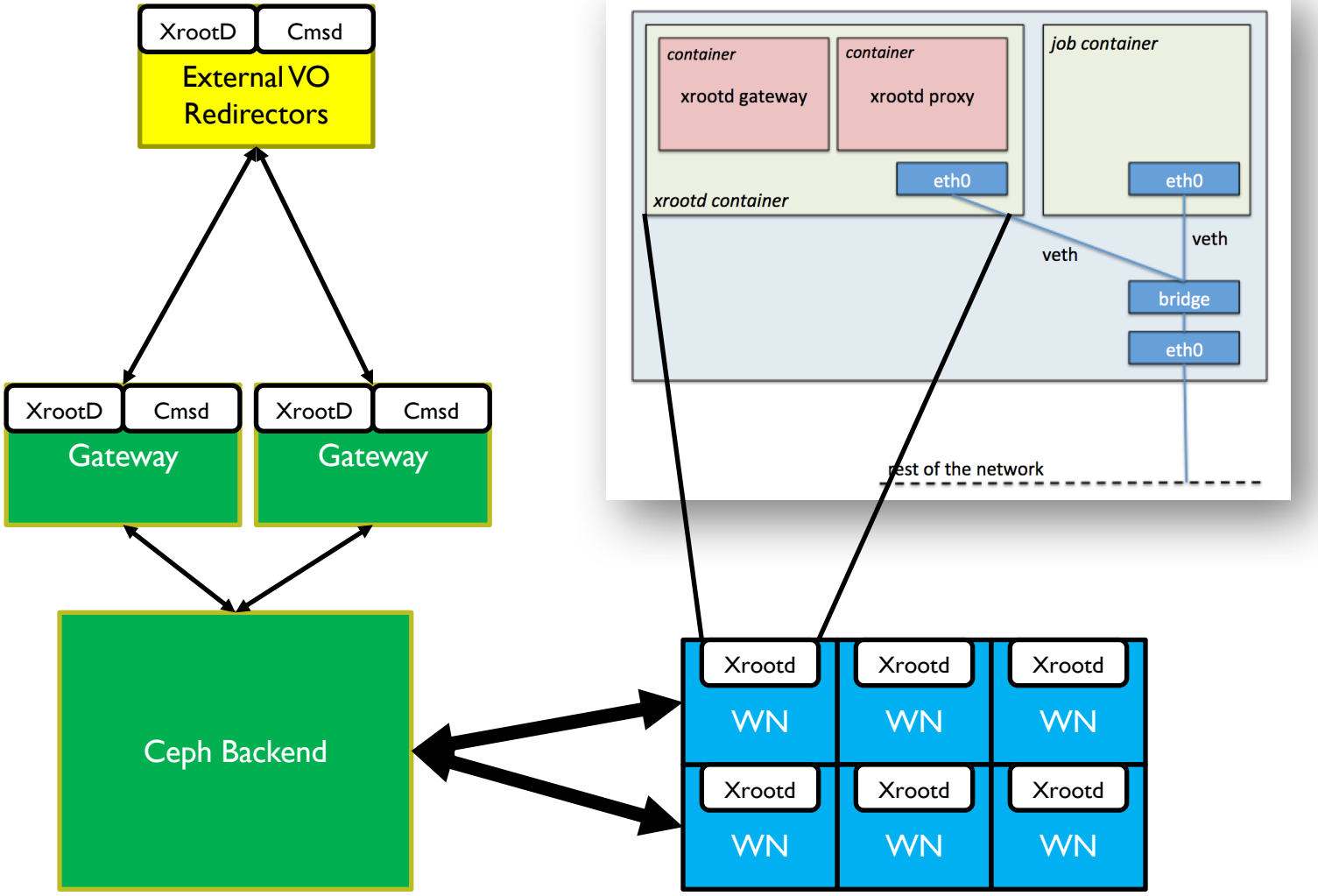
- ATLAS use GridFTP via the FTS service for WAN transfers.
- Can configure ATLAS jobs to use copy-to-scratch for data access.
- ATLAS do not require any accounting information.

Echo - CMS

- CMS also use the GridFTP for WAN transfers.
  - PhEDEx validation shows good connectivity.
- CMS use XRootD direct I/O for their jobs.
  - AAA means jobs can come in over the WAN.
- Extensive testing showed poor performance for certain types of jobs.
  - 'Lazy download' worked very well, although it appears to be out of favour with CMS.
  - Had to add a proxy cache layer to all gateways.
    - This meant adding an XRootD proxy to all WN.
- Old disk servers acting as disk cache cluster for AAA access.

echo

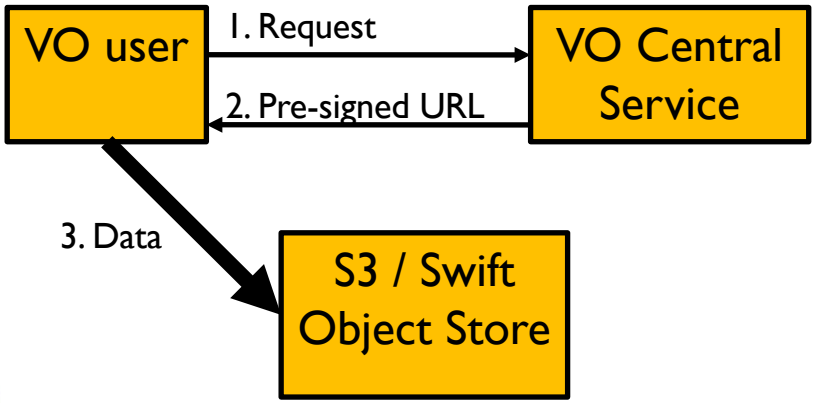
Echo - Proxy Setup



S3 / Swift

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 both 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.

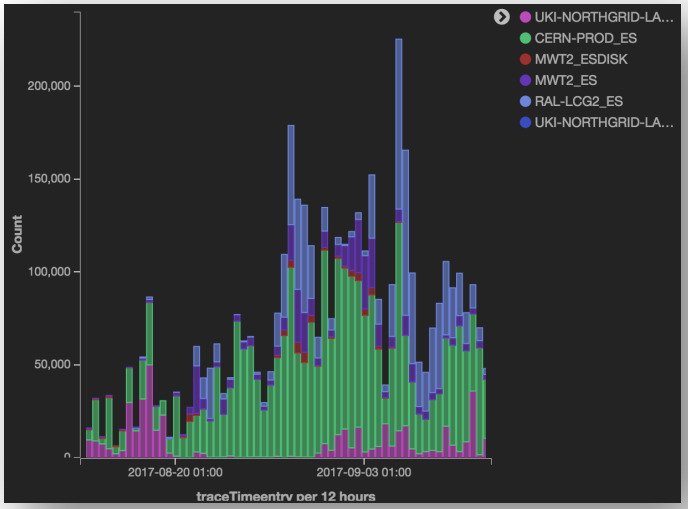
Pre-Signed URLs are a way to let users upload or download specific objects to/from buckets, but without requiring them to have the access + secret key.

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

ATLAS Event Service

- ATLAS Event Service is currently the most successful example of LHC VOs utilizing Object Stores.
- Utilizes Object Store very differently to a traditional SE.
  - Transient store for small files.
- Growing usage but not without significant teething problems.
  - ATLAS cloud working group looking at ways of validating Object Stores.

Plot shows number of successful stage-out transfers over the last month.



RAL-Azure

- RAL was provided with a grant from Microsoft.
  - One goal was to see if we could setup a complete site in the cloud.
    - We got close enough to getting it to work to call it a success!
  - Presentation to ATLAS 22nd May.
    - Slides added to this agenda as they were hidden from non-ATLAS people.
- DynaFed was used for authentication and authorization.
- An FTS instance in the cloud permitted data transfers.
  - FTS provided protocol translation for sites not supporting third party Webdav transfers.
    - Internal transfers do not normally cost for commercial clouds.

echo

Third Party Transfers

- Recent versions of dCache and DPM support third party WebDav transfers to/from S3 endpoints.
  - Not straight forward to get it working.
- Third party transfers impossible for:
  - S3 to S3.
  - StoRM to/from S3.
- Need to be able to setup service attached to Object Store.
  - GridFTP plugin for S3.
  - Service (e.g. FTS) that transfers can flow through and perform protocol translation.

echo

Secret Key Security

- Table compares the impact of an S3 secret key or a certificate with production role being compromised.
- S3 risk is lower for VO but significantly higher for sites.
  - Sites are exposed to copyright infringement claims.
- VOs will need to keep a record of the machines the keys are stored on and all names who have access.
  - Keys need to be transmitted securely and replaced annually.
- Would be better if keys never had to leave the site!

	S3 Secret Key	Certificate with production role
Potential for Data loss	One site	All sites
Potential for misuse	Trivial	Obscure
Traceability	Practically impossible	DN of user

ROOT I/O

- It is (obviously) possible to read a root file directly from S3.
- Some tools exist for performance testing of S3 endpoints.
  - I am not aware of any detailed testing.
- Started discussion with ROOT developers.
  - Intend to present at the next ROOT I/O workshop on 11th October.
- We want to put the intelligence of how best to access data from a particular storage in to client software.

echo

Conclusions

- As object store is in production and functioning as a normal Grid Storage endpoint for ATLAS.
- VO should ensure that their software can work with different types of backend storage service.
- To fully integrate Object Stores, we are going to need to deploy services on top of them. Ideally one that:
  - Stored the secret key and provided authentication / authorization.
  - Could proxy the transfer in the event that third party WebDav transfers are not supported.