S3 and Swift storage for the WLCG

Alastair Dewhurst, Dan van der Ster, Hiro Ito

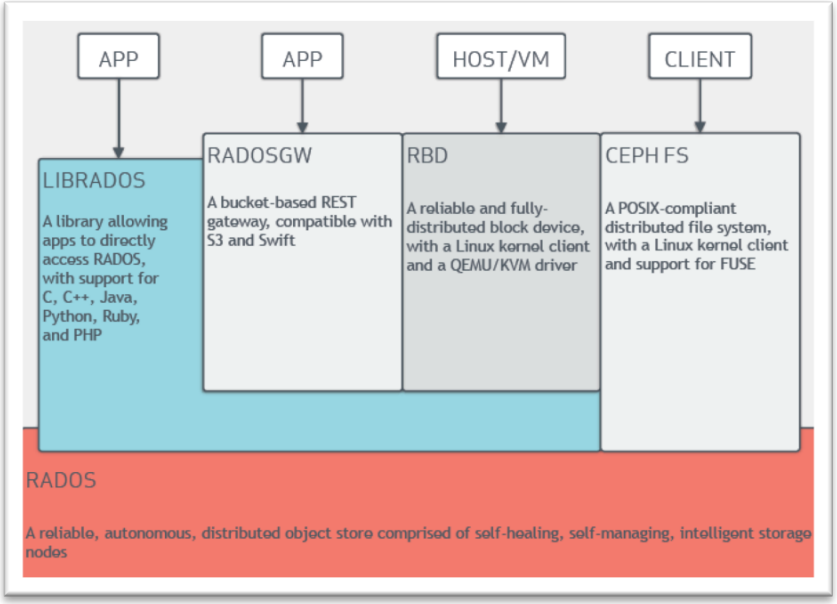
What is it?

- S3 stands for simple storage service and is Amazon's online web storage service.
  - It is an 'object store'.
- Swift is OpenStack's Storage API and is similar to S3.
- Access is via http/https: RESTful.
- Key/secret authentication. Pre-signed URLs.
- Users store objects in buckets. Each bucket is a unique namespace with hierarchy.

How does Ceph fit in?

- Ceph is an open source distributed object store.
- The RADOSGW provides S3 and Swift compatible APIs.
  - It is very straight forward to run!
- Ceph can provide RBD and POSIX like storage as well.
  - Or you can develop your own access protocol.

- All the Grid sites currently providing S3 endpoints for use by WLCG run Ceph.
- CERN, BNL and RAL run their own Ceph instances.
- Lancaster uses a commercial provider



Why is it important?

- Object stores are known to scale well.
- Lots of software can use S3 for its storage e.g.
  - CVMFS Stratum 0/1 backend storage.
  - Docker images.
  - Elastic search data.
- For certain use cases, S3 storage can be provided more easily/cheaply than a traditional SE.
  - Erasure Coding works well.
  - Need to change the belief that Storage that isn't dedicated to the WLCG isn't reliable.

Buckets

- Buckets store data:
  - They are an index of objects ... a namespace.
  - Provide some level of access control.
  - Provide usage statistics.
- The index means you can easily list all your: objects, size, checksum, last modified time etc.
- With Ceph now possible to get index-less or blind buckets.
  - Lose the namespace features but remove the metadata overhead.
- S3 has a geo-replication feature to keep buckets in sync across the WAN.

Grid tools for S3/Swift

- Davix - high performance HTTP client with S3 optimisations.
  - gfal2 - using Davix.
- Dynafed - dynamic HTTP storage federation and S3 gateway.
- FTS3 - support for HTTP (and S3 etc.)
  - Third party transfers to/from Webdav enabled sites.
  - Protocol translation (e.g. gridftp → S3) which routes transfers through FTS service.

DynaFed

Directly to S3 endpoint:

davix-ls -k --s3alternate --s3secretkey XXXXX --s3accesskey YYYYY
s3s://s3.echo.stfc.ac.uk/dynafed-test/

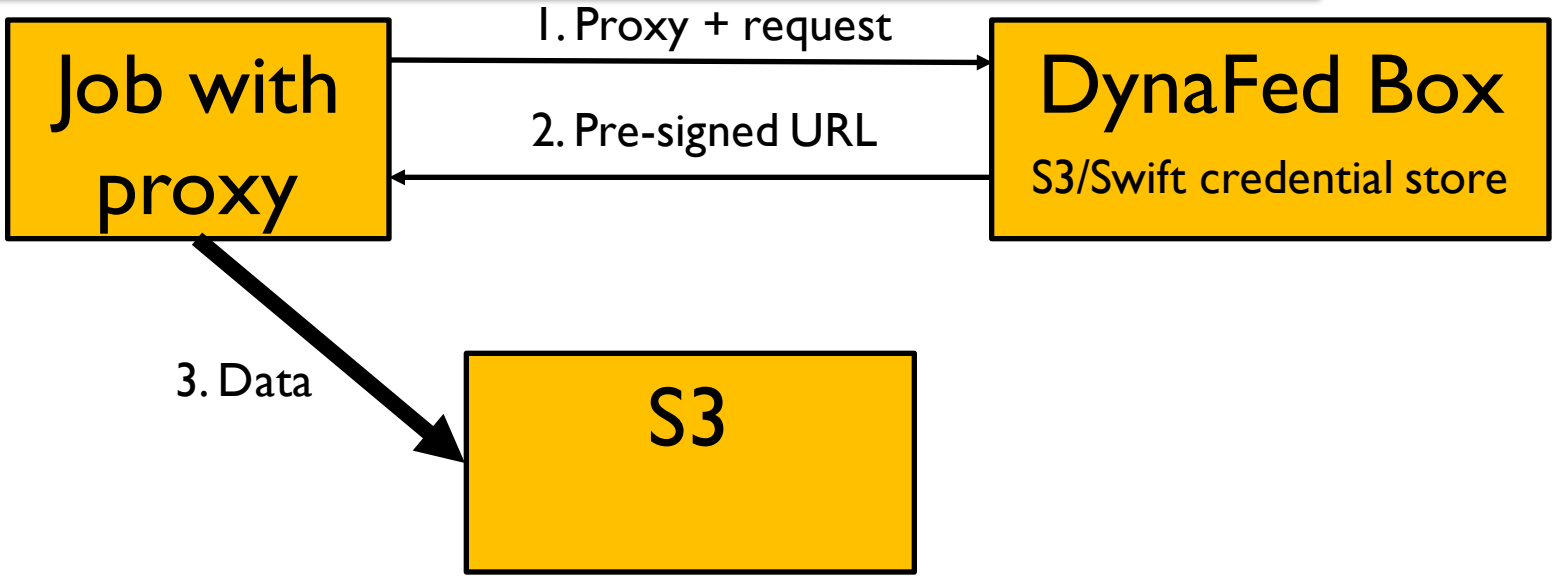
davix-put -k --s3alternate --s3secretkey XXXXX --s3accesskey YYYYY
testfile s3s://s3.echo.stfc.ac.uk/dynafed-test/testfile

With Dynafed:

voms-proxy-init

davix-ls -k https://dynafed.stfc.ac.uk/myfed/echo/

davix-put -k testfile https://dynafed.stfc.ac.uk/myfed/echo/testfile



VO Status

ATLAS

- ATLAS are looking at multiple ways of making use S3:
  - ATLAS Event Service.
  - Pilots writing log files from jobs.
  - S3 Rucio Storage Endpoints.
- AES writes output of individual (or small groups of) events to S3 endpoint.
  - For use with opportunistic (Cloud or HPC) CPU resources.
  - Uses S3 at BNL, CERN, RAL.
- ATLAS log files known to cause stress on storage which is designed for large files.
  - At RAL 20 – 30% of the transactions ~50TB space used.
  - Tested, but waiting on pilot development to implement in production.

CMS

- CMS received an AWS R&D grant with the project running from June 2015 – March 2016.
  - The goal was to run at scale, any of CMS's centrally organized workflows.
  - The focus was on running production jobs which primarily used FNAL Storage.
- S3 was required for a few workflows:
  - CMSSW supports reading from S3 via HTTP.
  - WMAgent plugin to stage out files to S3 as fallback of stage out to FNAL.
  - Tested various ROOT plugins (TWebFile, TDavixFile).

Information taken from Nicolo Magini and Burt Holzman's talks

FTS testing

- CMS setup an FTS instance at FNAL.
  - Dedicated to transfers between FNAL and Amazon S3.
  - Data from Grid Storage to S3 so streamed through FTS service.
  - 11k files (40TB) data transferred.
- RAL setup FTS development instance.
  - Trying to test third party transfers (still debugging).

ALICE and LHCb

- ALICE have no plans to change their current data management model.
  - The site would need to provide XrootD access on top of S3.
- LHCb have expressed an interest in S3 storage.
  - Limited manpower means no testing has taken place (yet).
- Https support is explicitly in the LHCb roadmap and they are engaged with the https deployment task force.
  - The more transparent S3 is the easier it will be to use it.
  - Priority is with XrootD at the moment.

Summary

- S3 style access to object stores are an attractive way for some sites to provide storage.
- S3 can fit into current WLCG storage evolution:
  - Continued deployment of http access by current sites.
  - Testing and integration with FTS.
- Object stores do not have the same limitations as traditional Grid Storage:
  - Revolutionize certain types of workflows.
  - New possibilities: Data Management → Data Curation.