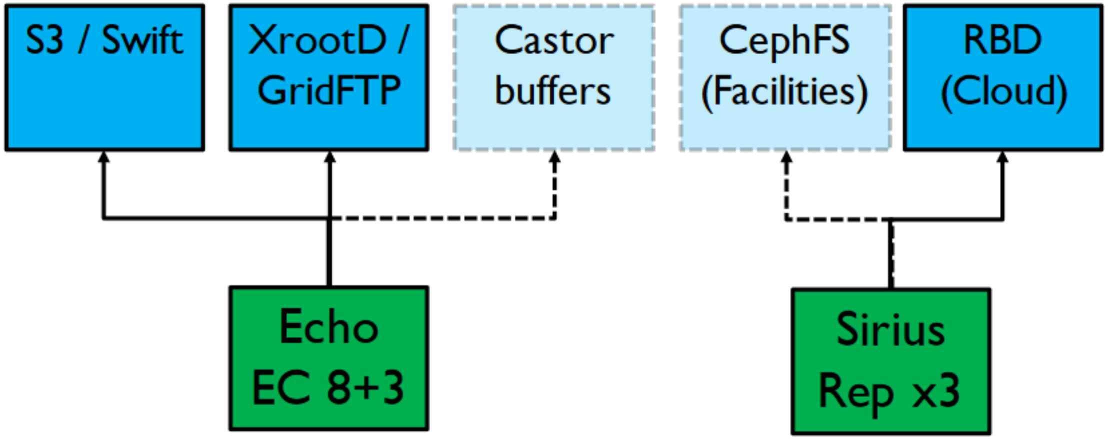
Ceph Status Report @ RAL

~~Tom Byrne~~, Bruno Canning

George Vasilakakos, Alastair Dewhurst, Ian Johnson, Alison Packer

Ceph at RAL is used for two larger projects:

- Echo
  - Disk only storage service to replace Castor for LHC VOs
    - Emphasis on TB/£ rather than IOPS/£ (thick storage nodes with EC)
  - 10 PB of space usable with EC procured in FY15-16, further 5 PB usable procured in FY16-17
- Sirius
  - Provide low-latency block storage to STFC cloud
    - Thin storage nodes with replication
  - c. 600TB raw space

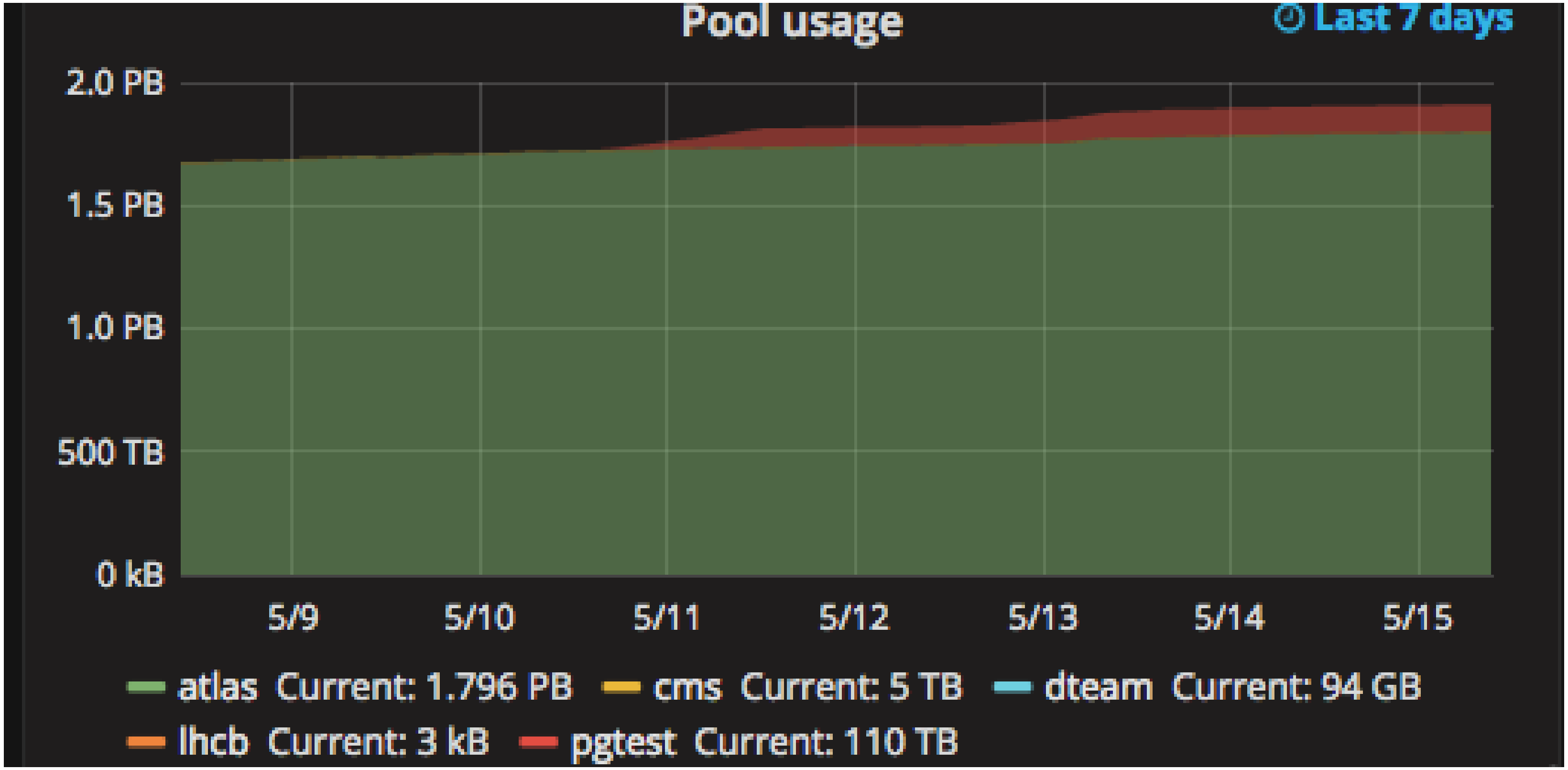


Production Status

- Echo is now accepting production data from LHC VOs
  - Currently GridFTP and XRootD supported as production I/O protocols.
  - VO data pools can be accessed via either protocol.
- 7.1 PB of WLCG pledge provided by Echo this year
  - Shared between ATLAS, CMS and LHCb
  - Already storing >1.7PB of data for ATLAS
- Lots of hard work: testing, development, on call procedures
  - The gateway development has been a large part of the work

	Castor Currently / TB	Capacity to decommission / TB	New capacity added to Castor / TB	Echo allocation / TB	Capacity provided by Echo
ALICE	480	0	0	0	0%
ATLAS	5257	1693	980	3100	41%
CMS	2287	1337	653	2500	61%
LHCb	4906	1010	545	1500	25%

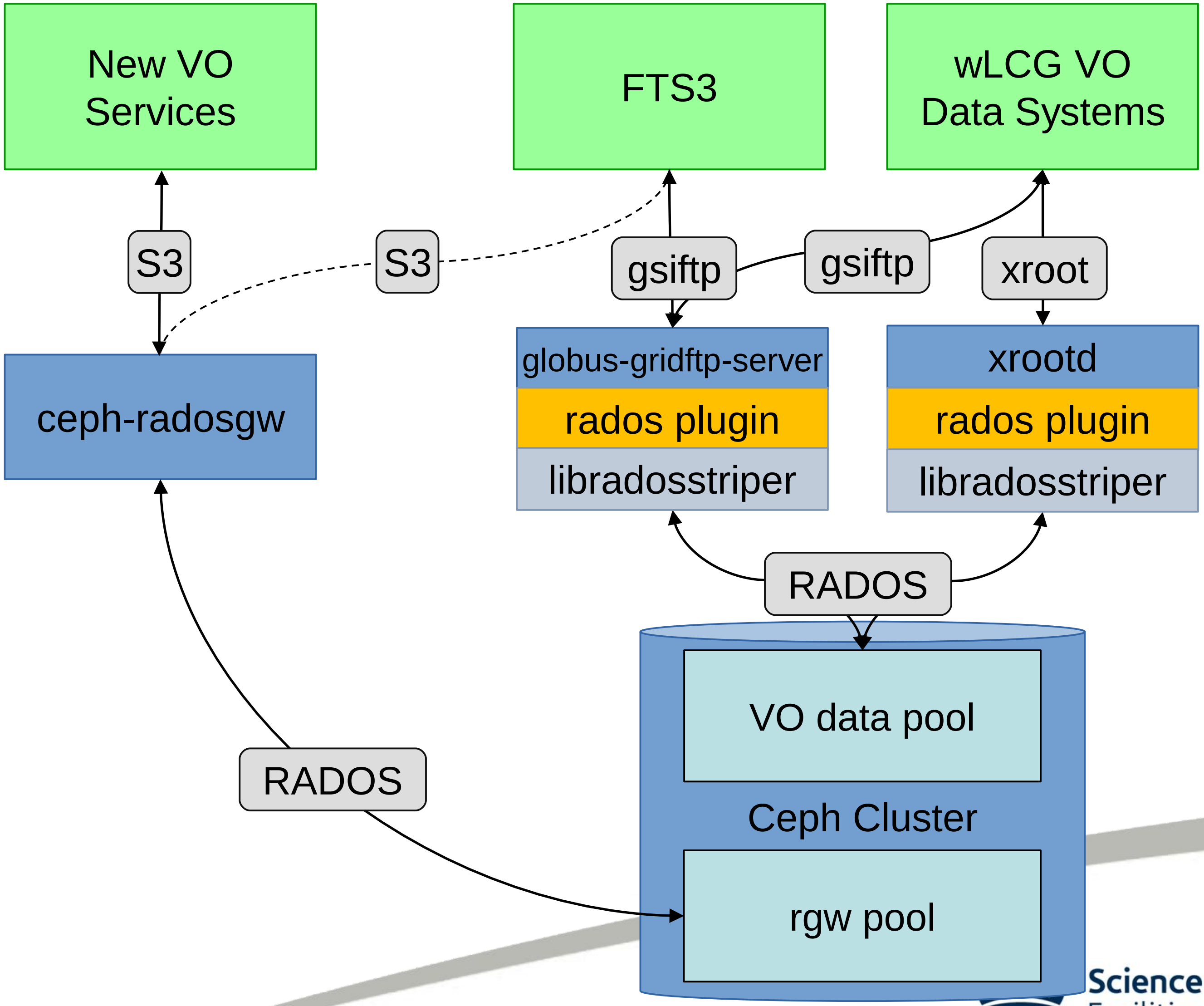
Storage Today



Gateway Specification

- Aim for a simple, lightweight solution that supports established HEP protocols: XRootD and GridFTP
  - No CephFS presenting posix interface to protocols
  - As close to a simple object store as possible
- Data needs to be accessible via both protocols
- Ready to support existing customers (HEP VOs)
- Want to support new customers with industry standard protocols: S3
  - GridPP makes available 10% of its resources to non-LHC activities

Service Architecture



Science & Technology
Facilities Council

GridFTP Plugin

- GridFTP plugin was completed at start of October 2016
  - Development has been done by Ian Johnson (STFC)
  - Work started by Sébastien Ponce (CERN) in 2015
- ATLAS are using GridFTP for most transfers to RAL
  - XRootD used for most transfers to batch farm
  - CMS Debug traffic and load tests also use GridFTP
- Recently improvements have been made to:
  - Deletion timeouts
  - Check-summing (to align with XRootD)
  - Multi-streamed transfers into EC pools

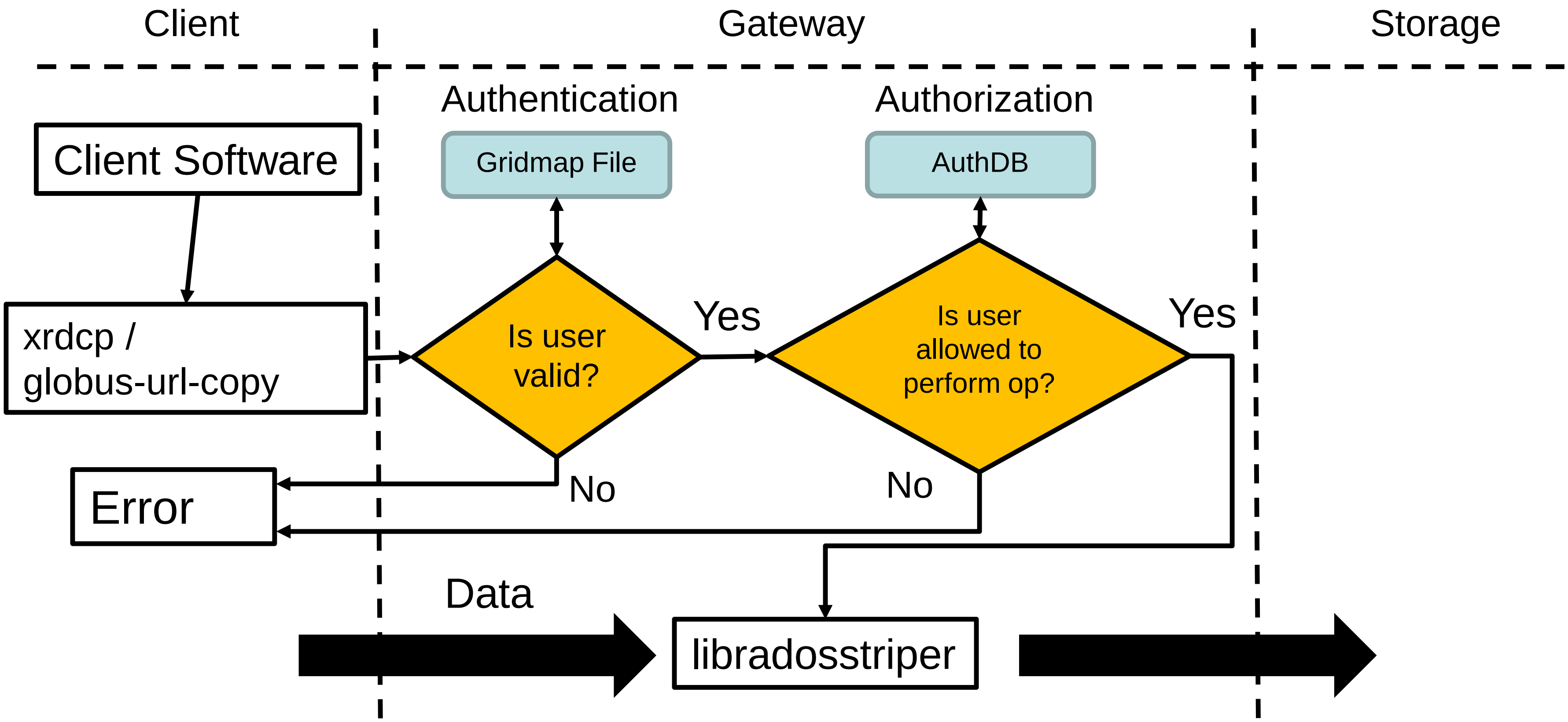
https://github.com/stfc/gridFTPCephPlugin

XRootD Plugin

- XRootD plugin was developed by CERN
  - Plugin is part of XRootD server software
  - But you have to build it yourself
- XRootD developments: Needed to enable features unsupported for objects store backends
  - Done
    - Check-summing
    - Redirection
    - Caching proxy
  - To Do
    - Over-write a file
    - Name-to-name mapping (N2N): work done, needs testing
    - Memory consumption

Science & Technology Facilities Council

XRootD/GridFTP AuthZ/AuthN

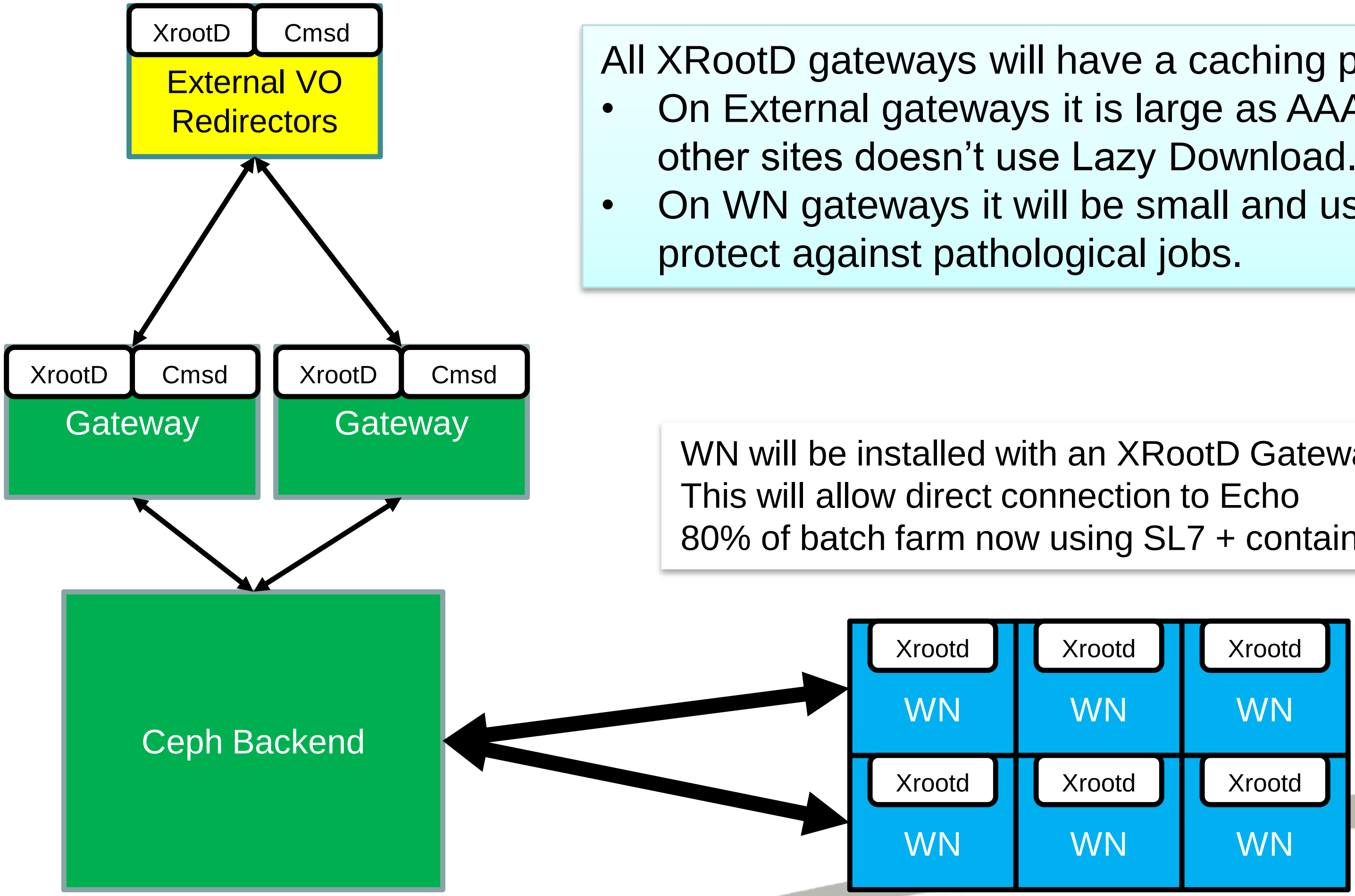


- Gridmap file used for authentication
- Authorisation is done via XRootD's authDB
  - Ian Johnson added support for this in the GridFTP plugin

Removing Gateway Bottleneck for WNs

- Having all worker nodes talk to Ceph through the gateways presents a bottleneck
- Can add gateway functionality to worker nodes:
  - Require ceph and XRootD configuration files, gridmap-file and keyring
  - Worker nodes can talk directly to Ceph object store
- Testing on a small number of WNs running an XRootD gateway in a containers.
- Container work led by Andrew Lahiff

XRootD Architecture



All XRootD gateways will have a caching proxy:

- On External gateways it is large as AAA from other sites doesn't use Lazy Download.
- On WN gateways it will be small and used to protect against pathological jobs.

WN will be installed with an XRootD Gateway
This will allow direct connection to Echo
80% of batch farm now using SL7 + containers.

Science & Technology Facilities Council

Operational issues: Stuck PG

- While rebooting a storage node for patching in February, a placement group in the atlas pool became stuck in a peering state
  - I/O hung for any object in this PG

A placement group is a section of a logical object pool that is mapped onto a set of object storage daemons

- Typical remedies for fixing peering problems were not effective (restarting/removing primary OSD, restarting set etc.)

```
pg 1.323 is remapped+peering, acting
[2147483647,1391,240,127,937,362,267,320,7,634,716]
```

- Seemed to be a communication issue between two OSDs in the set.

Stuck PG

- To restore service availability, it was decided we would manually recreate the PG
  - accepting loss of all 2300 files/160GB data in that PG
- Again, typical methods (force_create) failed due to PG failing to peer
- Manually purging the PG from the set was revealing
  - On the OSD that caused the issue, an error was seen
  - A Ceph developer suggested this was a LevelDB corruption on that OSD
  - Reformatting that OSD and manually marking the PG complete caused the PG to become active+clean, and cluster was back to a healthy state

Stuck PG: Conclusions

- A major concern with using Ceph for storage has always been recovering from these types of events
  - This showed we had the knowledge and support network to handle events like this
- The data loss occurred due to late discovery of correct remedy
  - We would have been able to recover without data loss if we had identified the problem (and problem OSD) before we manually removed the PG from the set

http://tracker.ceph.com/issues/18960

S3 / Swift

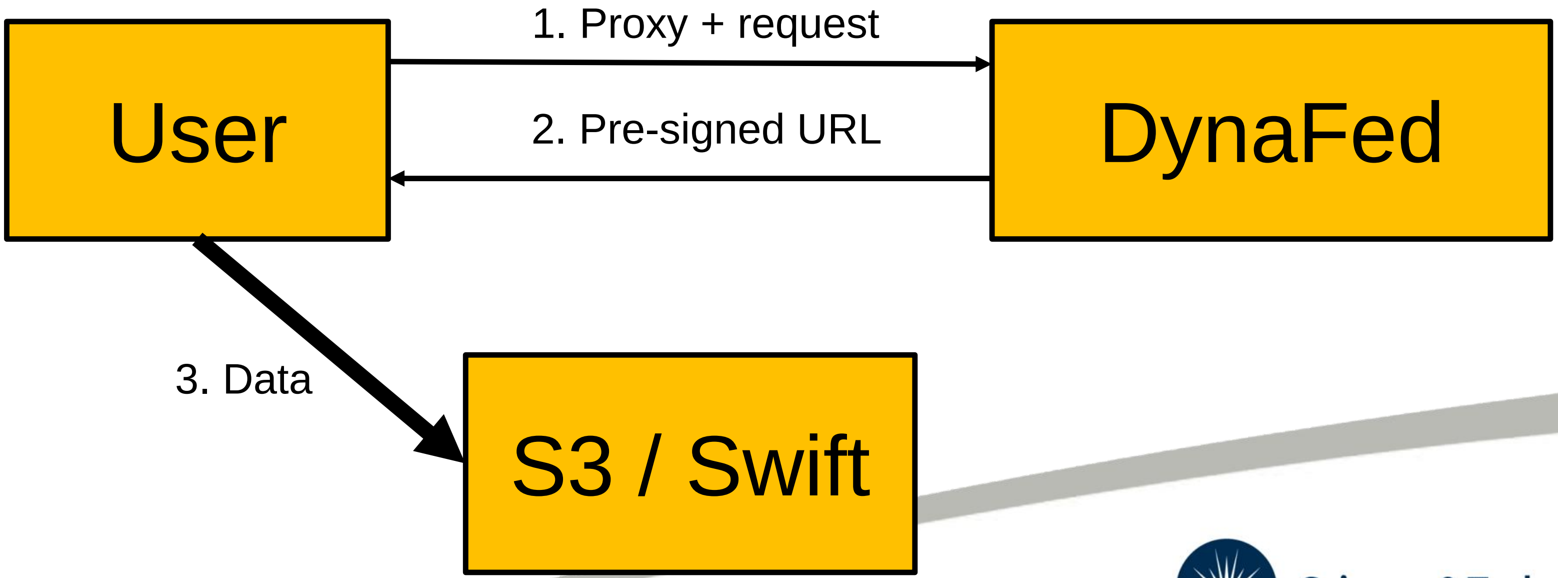
- We believe S3 / Swift are the industry standard protocols we should be supporting.

S3 / Swift API access to Echo will be the only service offered to new users wanting disk only storage at RAL.

- If users want to build their own software directly on top of S3 / Swift, that’s fine:
  - Need to sign agreement to ensure credentials are looked after properly.
- We expect most new users will want help:
  - Currently developing basic storage service product that can quickly be used to work with data.

DynaFed

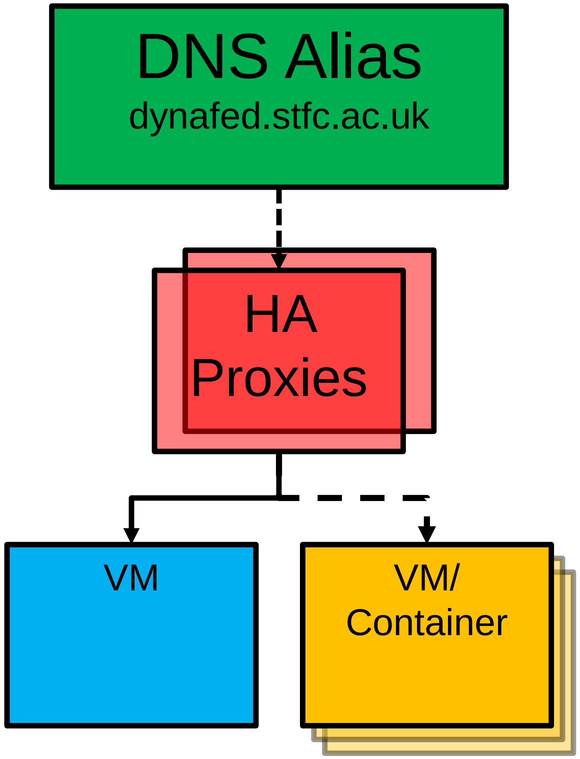
- We believe DynaFed is best tool to allow small VOs secure access.
  - S3/Swift credentials stored on DynaFed Box.
  - Users use certificate/proxy.
- Provides file system like structure.
- Good support for transfers to existing Grid storage.



Science & Technology
Facilities Council

RAL Dynafed Setup

- Service is behind HA proxy.
  - Currently just one VM but easily scalable.
- Will be in production in the next 6 months.
  - Ian J will be spending 50% of his time working on Dynafed.
- Anyone with an atlas or dteam certificate can try it out:
  - https://dynafed.stfc.ac.uk/gridpp



CLI

```
# voms-proxy-init
# davix-ls -P grid davs://dynafed.stfc.ac.uk/gridpp/dteam/disk/
# davix-put -P grid testfile davs://dynafed.stfc.ac.uk/gridpp/dteam/disk/testfile
Or
# gfal-ls davs://dynafed.stfc.ac.uk/gridpp/echo/
# gfal-copy file:///home/tier1/dewhurst/testfile davs://dynafed.stfc.ac.uk/gridpp/dteam/disk/testfile2
```

Conclusion

- Echo is in production!
- There has been a massive amount of work in getting to where we are
  - Support for GridFTP and XRootD on a Ceph object store are mature
  - Thanks to Andy Hanushevsky, Sébastien Ponce, Dan Van Der Ster and Brian Bockelman for all their help, advice and hard work.
- Looking forward: industry standard protocols are all we want to support
  - Tools are there to provide a stepping stones for VOs