Building a large scale object store for the RAL Tier 1

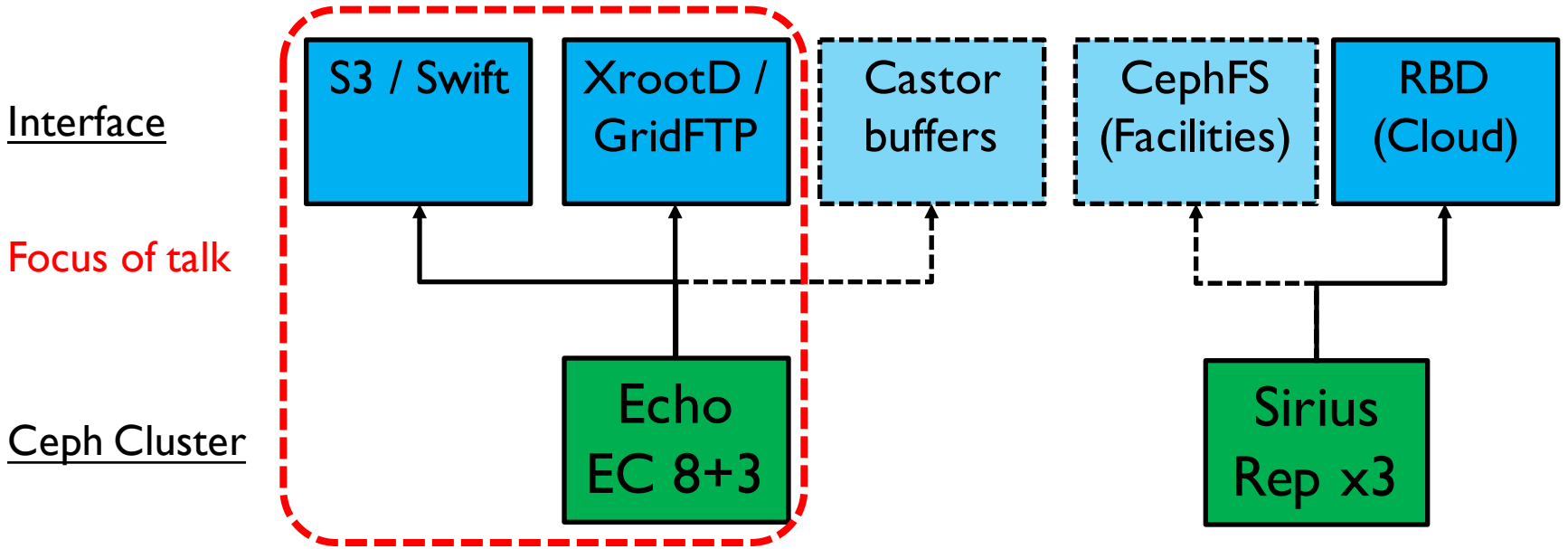
James Adams, Bruno Canning, Alastair Dewhurst,
Ian Johnson, Alison Packer, George Vasilakakos

Email: <firstname>.<lastname>@stfc.ac.uk

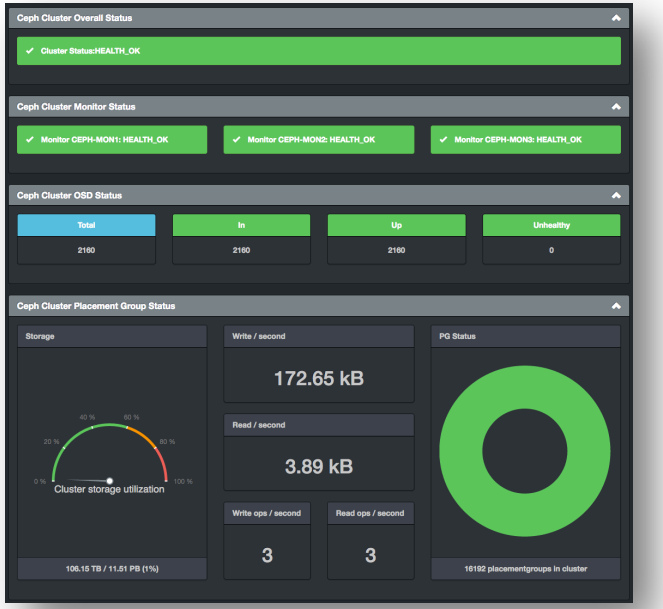
echo

Motivation

- CERN moved their disk storage from Castor to EOS several years ago.
  - RAL only remaining site using Castor for Disk which is reaching limits.
- Future storage must be simpler to run and have a similar hardware cost per TB.
  - RAL is running other services on Ceph.
- Service must appeal to wider audience than just HEP.



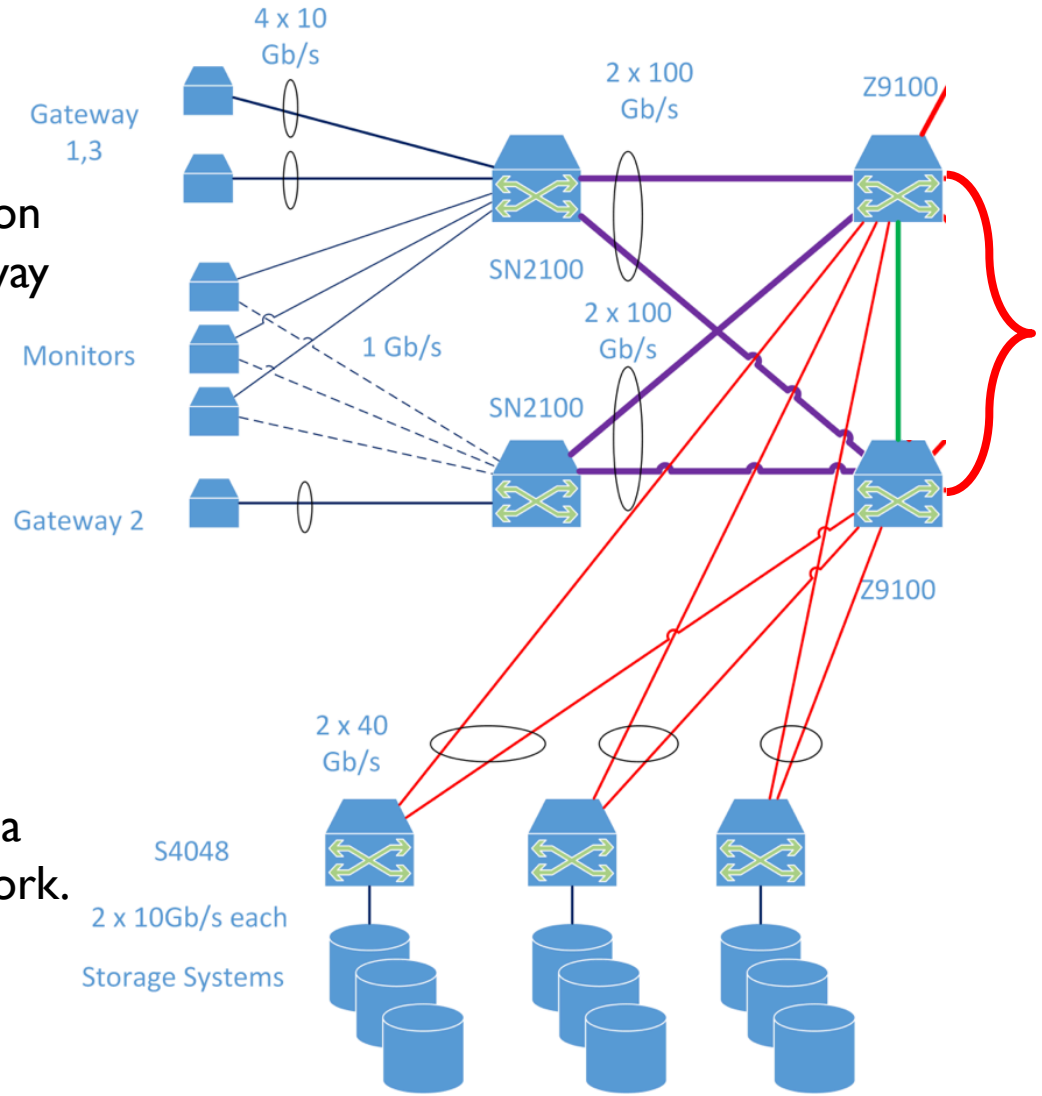
Echo cluster



- Working cluster since July 2016.
  - 3 physical Monitors + hot spare.
  - 60 x 216TB storage nodes.
  - 3 gateways machines.
  - Running Jewel release on SL7.
- 9.4PB usable storage (8+3 Erasure Coding).
  - Intend to provide ATLAS and CMS 2.8PB each as part of pledged resources in April 2017.
- Configuration management of all Ceph clusters via Quattor.

Echo Network

Designed with expansion in the number of gateway machines in mind.

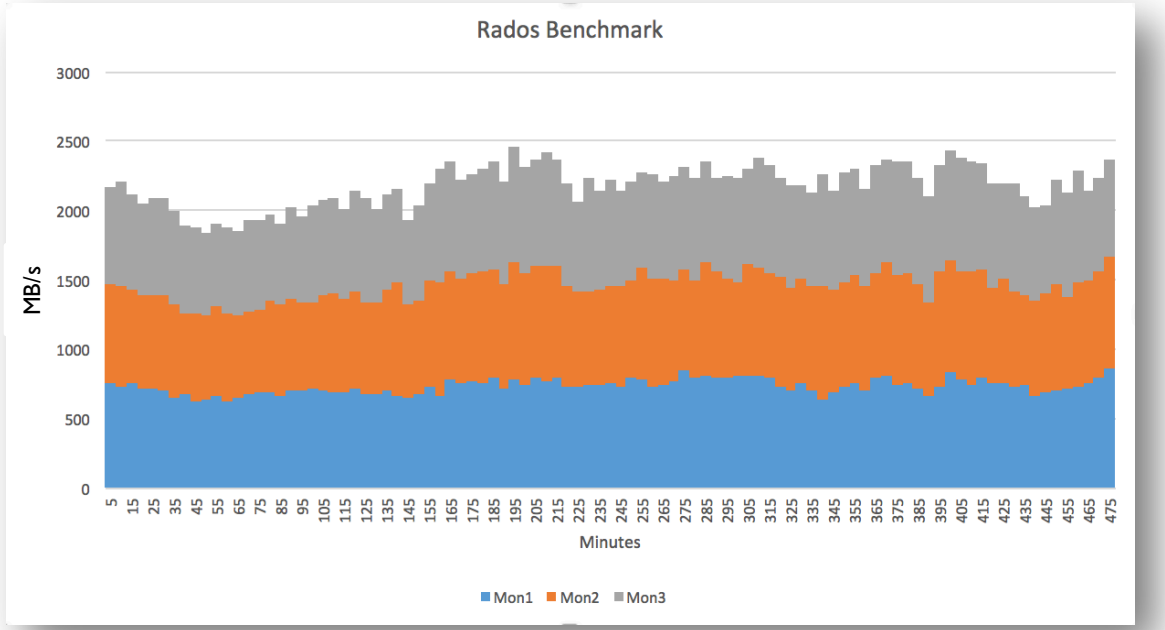


To the rest of the Tier 1 and internet

Each Storage Node has a public and cluster network.

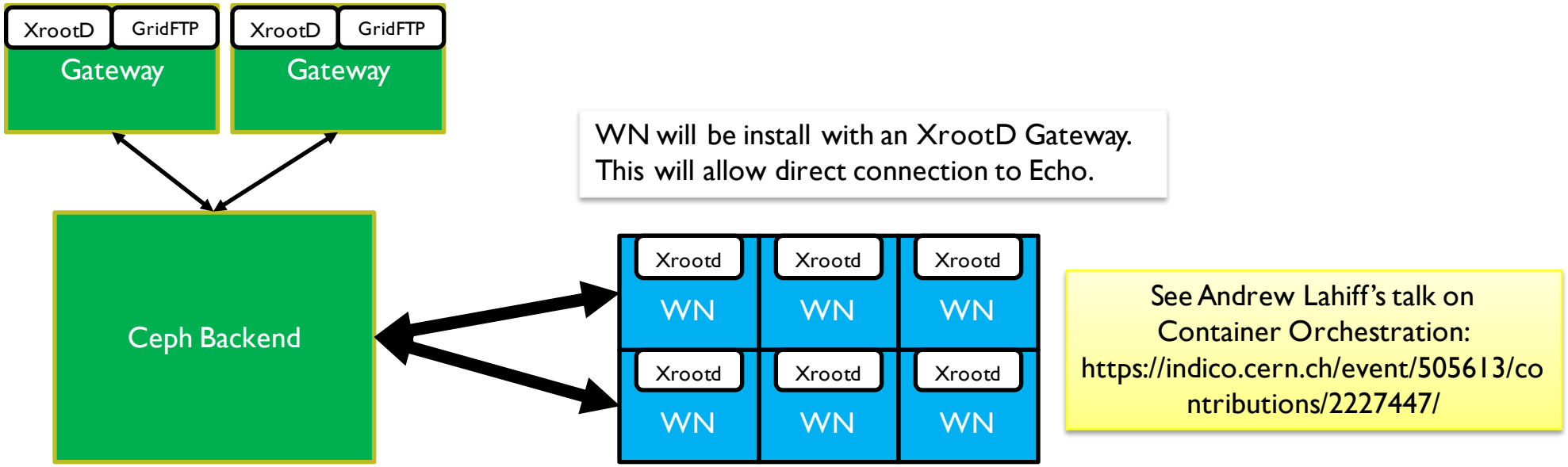
Early Benchmarking

- Ran Ceph 'rados bench' from 3 machines with 10Gb/s links.
- Each machine concurrently writing 64 x 4MB objects.
- Average latency (time to write file) 0.35s.
- Bottleneck appears to be with benchmarking machines (not the cluster).

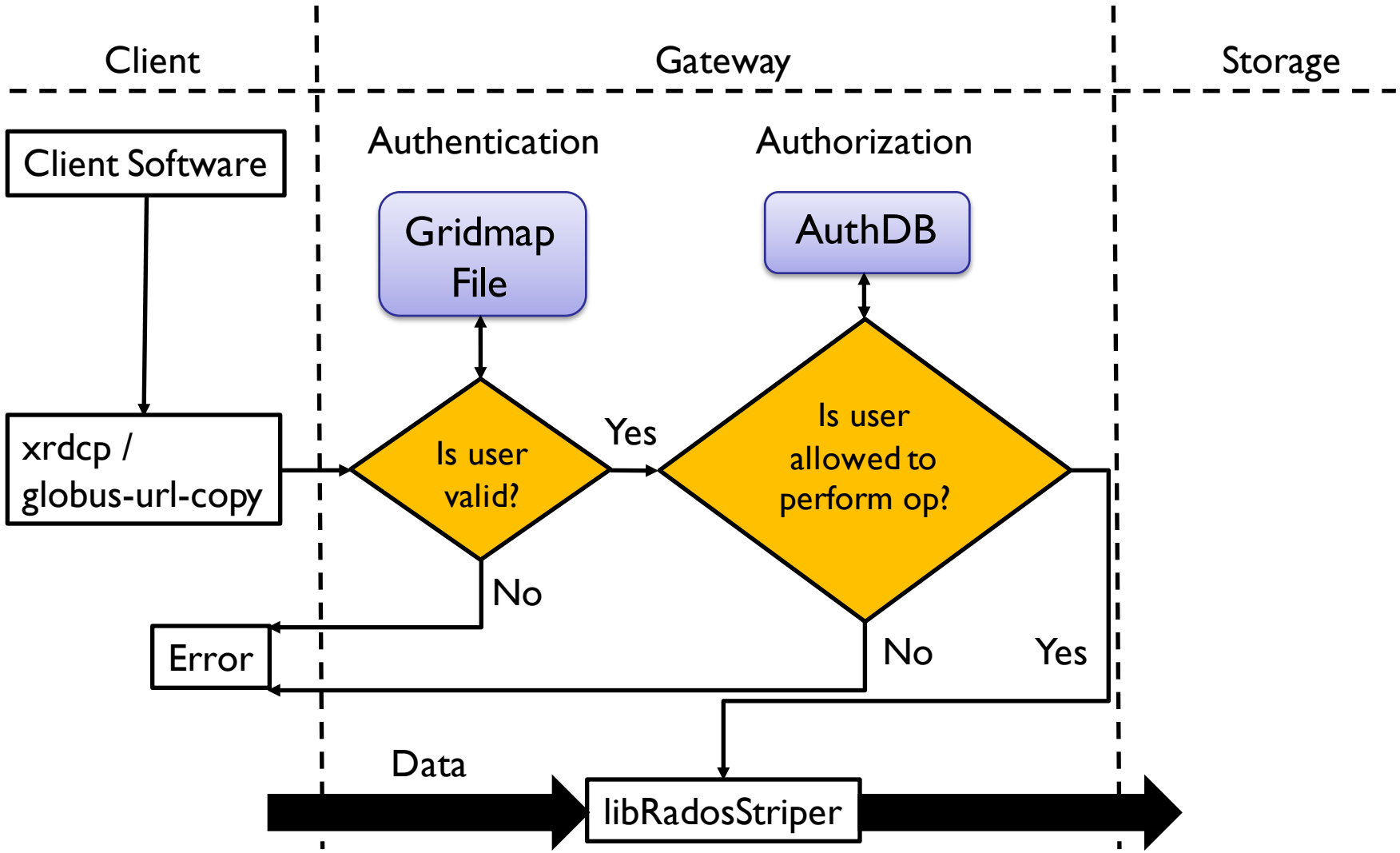


GridFTP + XrootD plugins

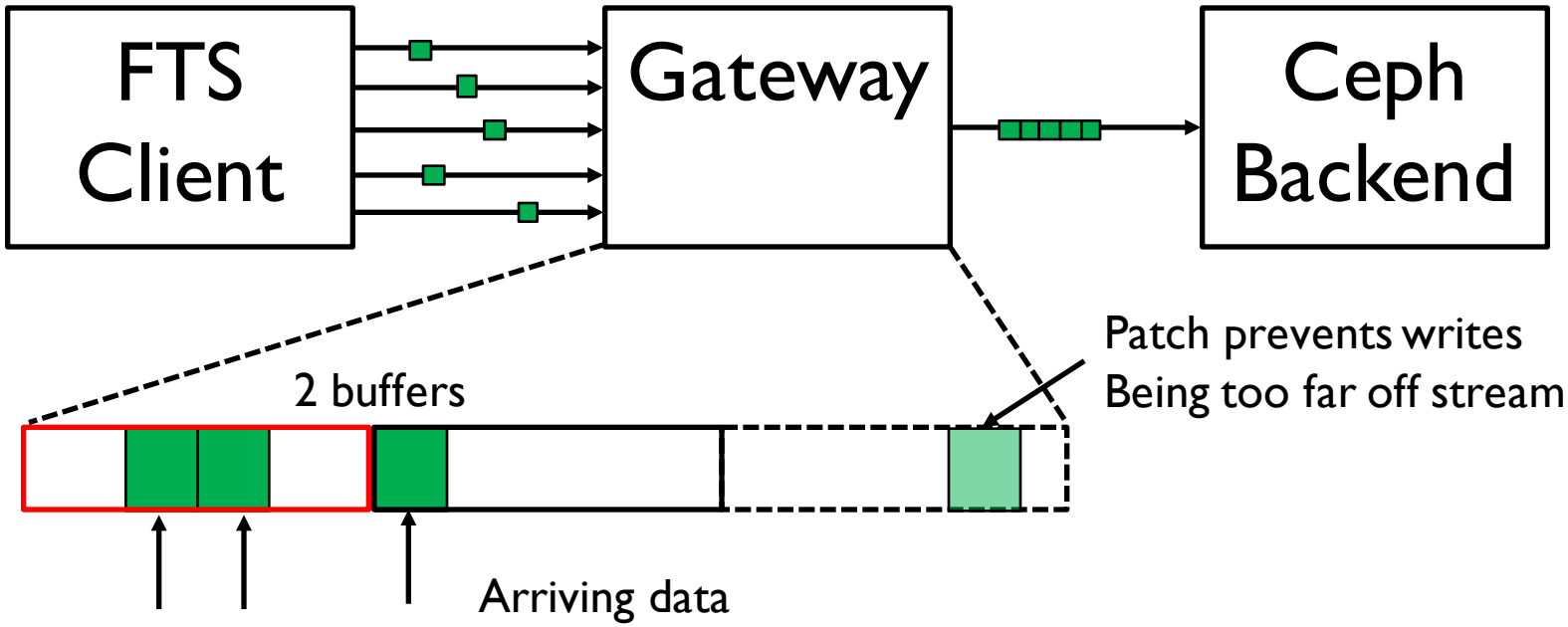
- For the LHC VOs we need working GridFTP and XrootD access.
  - CERN had written plugin for XrootD.
  - RAL wrote GridFTP plugin.
- No SRM – Accounting via .json files.



Plugin architecture



GridFTP plugin design

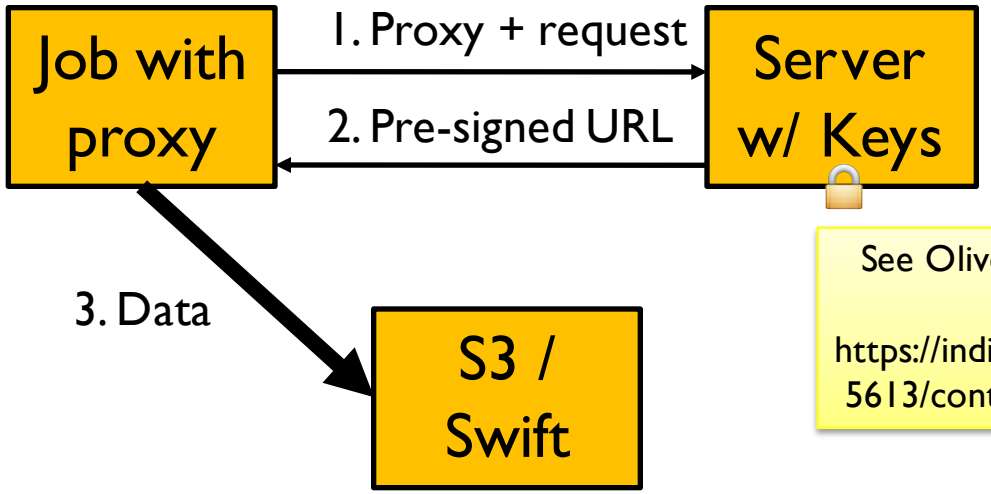


- FTS transfers use GridFTP with multiple streams.
- GridFTP plugin re-assembles data to optimise writing to Ceph.
- Ongoing work to prevent certain write streams getting too far ahead.
  - Large buffer size currently ensures data is not lost.

S3 / Swift setup

- We believe S3 / Swift are the industry standard protocols we should be supporting in the long term.
  - S3 / Swift Gateway is being provided for all users.
- Need to ensure credentials are looked after properly.
  - Some VOs or internal users we can trust.
  - Make use of pre-signed (or temporary) URLs for others.

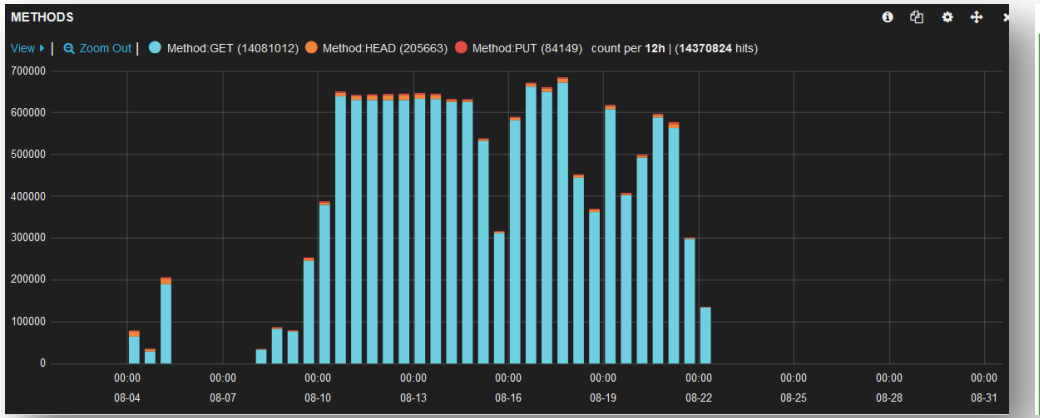
Pre-Signed URLs are a way to let users upload or download specific objects to/from buckets, but without requiring them to have the username/password.



See Oliver Keeble's talk on DynaFed: https://indico.cern.ch/event/505613/contributions/2230903/

Usage

- S3 / Swift used by or tested with:
  - ATLAS Event Service.
  - Storing Docker containers.
  - Backend for CVMFS Stratum 1.
  - Dynafed.
- GridFTP / XrootD used by:
  - ATLAS RSE: RAL-LCG2-ECHO successful FTS transfers.
  - Working on getting ATLAS Hammer Cloud jobs to run.
  - Working on adding Echo to PhEDEx framework.
  - Successfully tested CMS workflows.



Source	Destination	VO	Submitted	Active	Staging	S.Active	Finished	Failed	Cancel	Rate (last 1h)	VO Thr.
srm://se3.itep.ru	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	3	-	-	100.00 %	-
srm://srm-cms.jinr-t1.ru	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	21	3	-	87.50 %	-
srm://ingrid-se02.cism.ucl.ac.be	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	4	-	-	100.00 %	-
srm://cmssrm.hep.wisc.edu	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	3	-	-	100.00 %	-
srm://srmcms.pic.es	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	40	5	-	88.89 %	-
srm://pcncp22.ncp.edu.pk	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	4	3	-	57.14 %	-
srm://t2-srm-02.lnl.infn.it	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	1	-	-	100.00 %	-
srm://srm.ciemat.es	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	3	-	-	100.00 %	-
gsiftp://lcgse01.phy.bris.a	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	16	-	-	100.00 %	-
srm://srm.ihepa.ufl.edu	gsiftp://gridftp.echo.stfc.a	cms	-	-	-	-	1	1	-	50.00 %	-

Summary & Acknowledgements

- Future RAL Tier 1 storage heavily Ceph based.
- RAL have developed a GridFTP plugin for Ceph, that works with the XrootD plugin.
- We intend to provide ATLAS and CMS some of their pledged storage on Echo in 2017.
- Have been exploring ways to use S3 / Swift and will continue to push for their adoption.

- We would like to thank Sebastien Ponce and Brian Bockelman for their assistance in developing the GridFTP plugin.

echo