CernVM-FS at the Tier-1

Catalin Condurache

GridPP42, Abingdon, 24 April 2019

Outline

- CernVM-FS ?
- CernVM-FS infrastructure @RAL
- EGI CernVM-FS service
- Users
- Recent developments and plans

Outline

- CernVM-FS ?
- CernVM-FS infrastructure @RAL
- EGI CernVM-FS service
- Users
- Recent developments and plans

CernVM File System ?

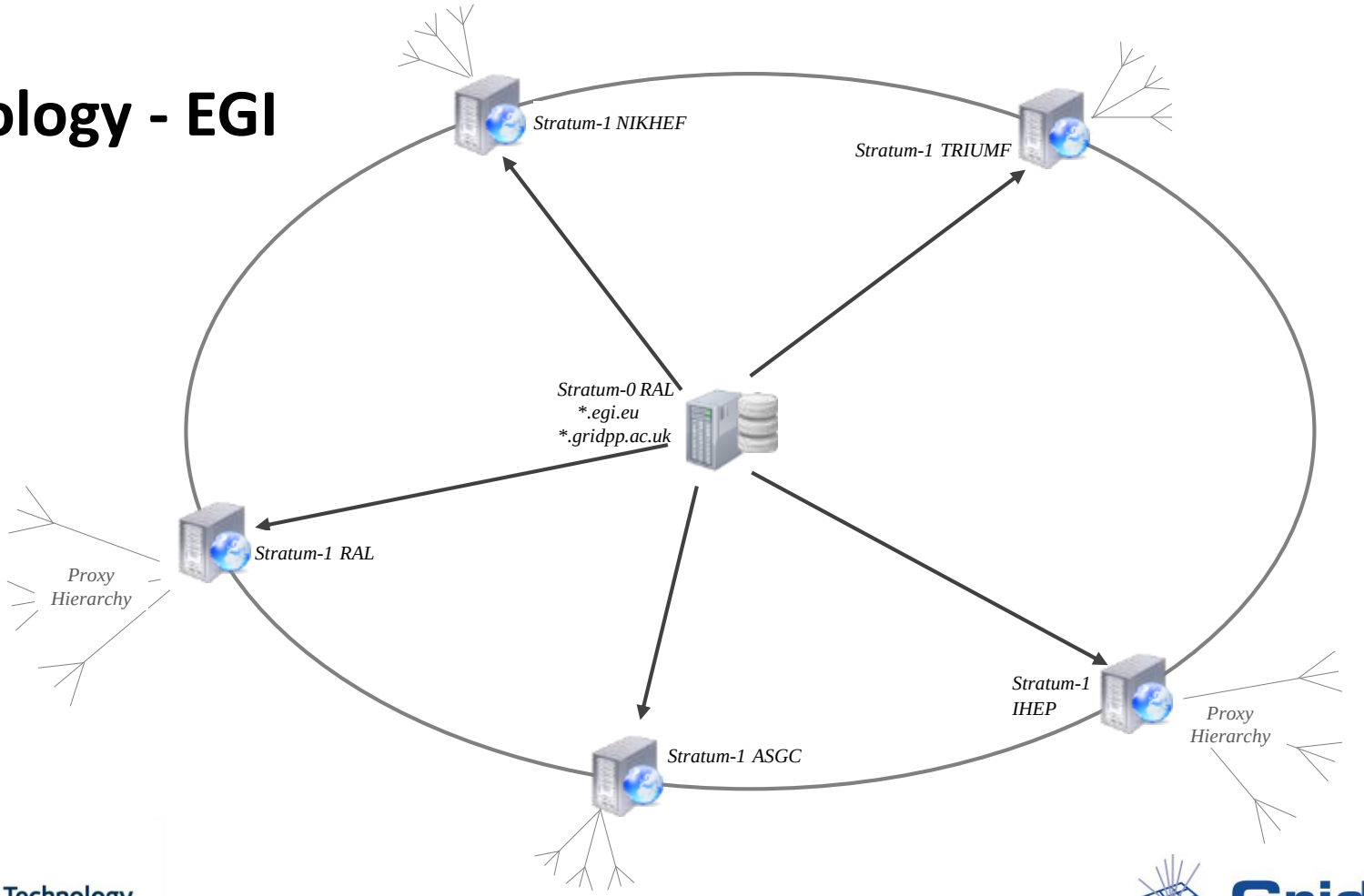
- Read-only filesystem that provides a low-maintenance, reliable software distribution service for HEP and non-HEP communities
- Built using standard technologies (http, sqlite, fuse, squid)
- Data is aggressively cached, de-duplicated across different SW releases and transported via HTTP and proxy servers for the fastest possible retrieval of files
- Digitally signed repositories ensure data integrity

Outline

- CernVM-FS ?
- CernVM-FS infrastructure @RAL
- EGI CernVM-FS service
- Users
- Recent developments and plans

CernVM-FS Infrastructure @RAL

Topology - EGI



CernVM-FS Infrastructure @RAL

Stratum-0 service (EGI, STFC)

- Maintains and publishes the current state of the repositories
- 32GB RAM, 12TB disk, 2x E5-2407 @2.20GHz
- cvmfs-server v2.5.2
- 35 repositories (*egi.eu* and *gridpp.ac.uk*) – 2.7 TB – 24.4 million files
- Repository stats
  - largest number of files: *mice* (9.2×10^6), *facilities* (2.3×10^6), *t2k* (1.9×10^6)
  - largest total file size: *chipster* (644 GB), *mice* (346 GB), *t2k* (151 GB)
  - largest revision number: *pheno* (735), *t2k* (256), *auger* (178)

CernVM-FS Infrastructure @RAL

Stratum-1 service (WLCG, EGI, STFC)

- Part of the worldwide network of servers (RAL, NIKHEF, TRIUMF, ASGC, IHEP) replicating the *egi.eu* repositories
- RAL - 2-node HA cluster (cvmfs-server v2.5.2)
  - each node – 64 GB RAM, 55 TB storage, 2xE5-2620 @2.4GHz
  - dual stack IPv4/IPv6 since September 2017
  - it replicates 92 repositories – total of 33 TB of replica
  - *egi.eu*, *gridpp.ac.uk* and *nikhef.nl* domains
  - also many *cern.ch*, *opensciencegrid.org*, *desy.de*, *africa-grid.org*, *ihep.ac.cn* and *in2p3.fr* repositories

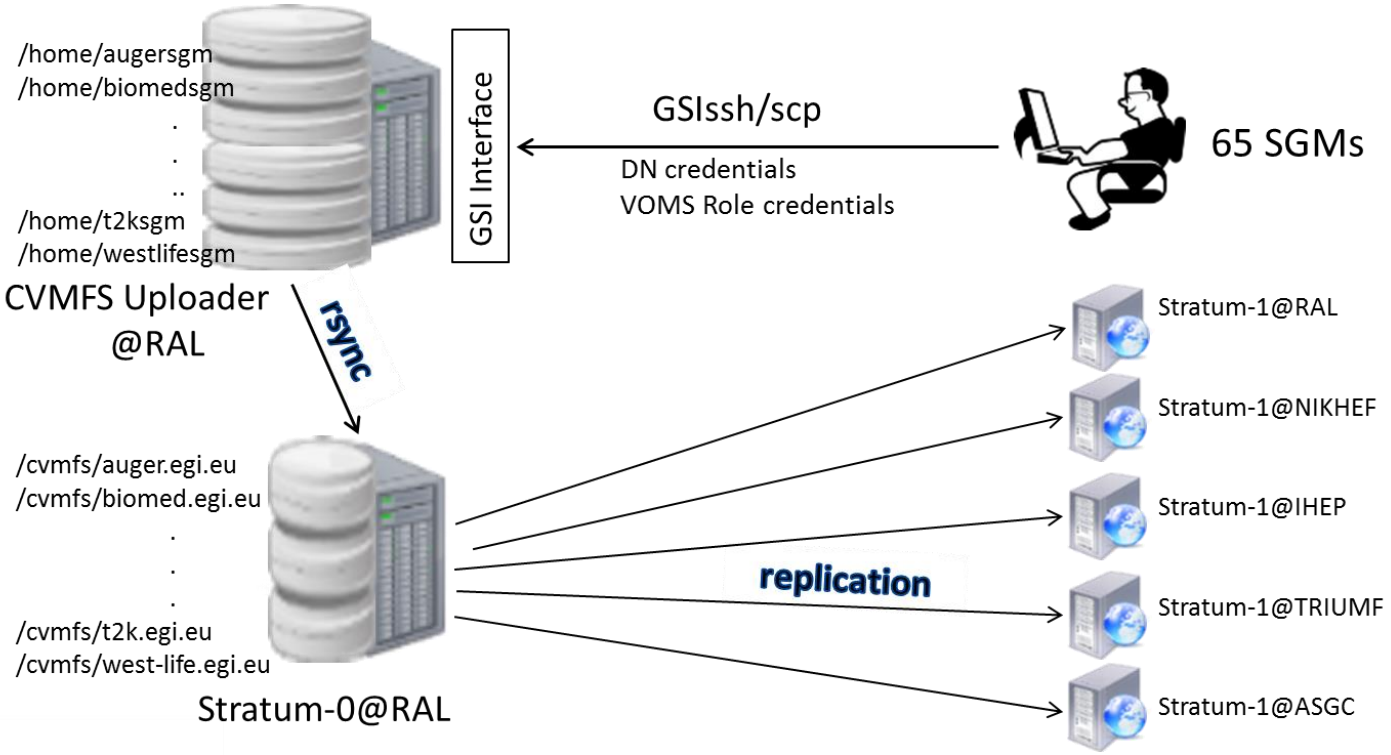
CernVM-FS Infrastructure @RAL

Uploader service (EGI, STFC)

- In-house implementation that provides upload area for *egi.eu* (and *gridpp.ac.uk*) repositories
- Currently 2.0 TB – repo master copies
- GSI-OpenSSH interface (gsissh, gsiscp, gsisftp)
  - similar to standard OpenSSH tools with added ability to perform X.509 proxy credential authentication and delegation
  - DN based access, also VOMS Role possible
- *rsync* mechanism between Stratum-0 and Uploader

CernVM-FS Infrastructure @RAL

Repository uploading mechanism



CernVM-FS Infrastructure @RAL

Squid service (UK-wide)

- Battery of 5 frontier-squid nodes under *cvmfs-squid.gridpp.rl.ac.uk* alias
- Currently frontier-squid v3.5.28 – plans for frontier-squid-4
- Dual stack IPv4/6
- Serves requests from **.ac.uk* hosts
- Also used by ATLAS and CMS to access the Frontier launchpads
  - *atlas-squid* and *cms-squid* aliases

Outline

- CernVM-FS ?
- CernVM-FS infrastructure @RAL
- EGI CernVM-FS service
- Users
- Recent developments and plans

EGI CernVM-FS Service

Two EGI Operational Procedures

- Process of enabling the replication of CernVM-FS spaces across OSG and EGI CernVM-FS infrastructures - *https://wiki.egi.eu/wiki/PROC20*
- Process of creating a repository within the EGI CernVM-FS infrastructure for an EGI VO – *https://wiki.egi.eu/wiki/PROC22*

Operations Level Agreement for Stratum-0

- Between STFC and EGI.eu
- Provisioning, daily running and availability of service
- Service advertised through the EGI Service Catalogue

The EGI Staged Rollout

- RAL is an early Adopter for cvmfs client and server, also frontier-squid
- Once reports are accepted, upload to UMD-4 repo

EGI CernVM-FS Service

Two EGI Operational Procedures

- Process of enabling the replication of CernVM-FS spaces across OSG and EGI CernVM-FS infrastructures - *https://wiki.egi.eu/wiki/PROC20*
- Process of creating a repository within the EGI CernVM-FS infrastructure for an EGI VO – *https://wiki.egi.eu/wiki/PROC22*

Operations Level Agreement for Stratum-0

- Between STFC and EGI.eu
- Provisioning, daily running and availability of service
- Service advertised through the EGI Service Catalogue

The EGI Staged Rollout

- RAL is an early Adopter for cvmfs client and server, also frontier-squid
- Once reports are accepted, upload to UMD-4 repo

EGI CernVM-FS Service

Two EGI Operational Procedures

- Process of enabling the replication of CernVM-FS spaces across OSG and EGI CernVM-FS infrastructures - *https://wiki.egi.eu/wiki/PROC20*
- Process of creating a repository within the EGI CernVM-FS infrastructure for an EGI VO – *https://wiki.egi.eu/wiki/PROC22*

Operations Level Agreement for Stratum-0

- Between STFC and EGI.eu
- Provisioning, daily running and availability of service
- Service advertised through the EGI Service Catalogue

The EGI Staged Rollout

- RAL is an early Adopter for cvmfs client and server, also frontier-squid
- Once reports are accepted, upload to UMD-4 repo

Outline

- CernVM-FS ?
- CernVM-FS infrastructure @RAL
- EGI CernVM-FS service
- Users
- Recent developments and plans

Who Are the Users?

- Broad range of HEP and non-HEP communities
- High Energy Physics
  - *hyperk, mice, t2k, snoplus*
- Medical Sciences
  - *biomed, neugrid*
- Physical Sciences
  - *cernatschool, pheno*
- Space and Earth Sciences
  - *auger, extras-fp7*
- Biological Sciences
  - *chipster, enmr*

The Users – What Are They Doing?
Grid Environment

- snoplus.snolab.ca VO
  - uses CernVM-FS for MC production (also ganga.cern.ch)
- cernatschool.org VO
  - educational purpose, young users get used with grid computing
  - software unit tests maintained in the repository
- dirac.egi.eu
  - repository maintained by the DIRAC interware developers
  - contains the DIRAC clients, environment settings for various DIRAC services (France Grilles, GridPP, DIRAC4EGI)
  - repository is therefore accessed by any user submitting to a DIRAC service

The Users – What Are They Doing?
Grid Environment

- auger VO
  - simulations for the Pierre Auger Observatory at sites using the same software environment provisioned by the repository
- pheno VO
  - maintain HEP software – Herwig, HEJ
  - daily automated job that distributes software to CVMFS
- other VOs
  - software provided by their repositories at each site ensures similar production environment

The Users – What Are They Doing?
Cloud Environment

- chipster
  - the repository distributes several genomes and their application indexes to ‘chipster’ servers
  - without the repo the VMs would need to be updated regularly and become too large
  - four VOs run ‘chipster’ in EGI cloud (test, pilot level)
- enmr.eu VO
  - use DIRAC4EGI to access VM for GROMACS service
  - repository mounted on VM
- other VOs
  - mount their repo on the VM and run specific tasks (sometime CPU intensive)

Outline

- CernVM-FS ?
- CernVM-FS infrastructure @RAL
- EGI CernVM-FS service
- Users
- Recent developments and plans

Recent Developments and Plans

‘Confidential’ CernVM-FS repositories

- Repositories natively designed to be public with non-authenticated access – minimal info needed: public signing key and repository URL
- Widespread usage of technology (beyond LHC and HEP) led to use cases where software needed to be distributed was not public-free
  - Software with specific license for academic use
  - Communities with specific rules on data access
- Questions raised at STFC and within EGI about availability of this feature/possibility in recent years
- Work done within US Open Science Grid (OSG) added the possibility to introduce and manage authorization and authentication using security credentials such as X.509 proxy certificate

‘Confidential’ CernVM-FS repositories

Working prototype at RAL

- Stratum-0 with *mod_gridsite*, *https* enabled
  - ‘*cvmfs_server publish*’ operation incorporates an authz info file
  - Access based on .gacl (Grid Access Control List) file in *<repo>/data/* directory that has to match the approved DNs or VOMS roles
- CVMFS client + *cvmfs_x509_helper* to enforce authz to the repository
  - Also some globus-* packages need installed!
  - *root* can always see the namespace and the files in the client cache
- Client connects directly to the Stratum-0
- No Stratum-1 or squid in between – caching not possible for HTTPS

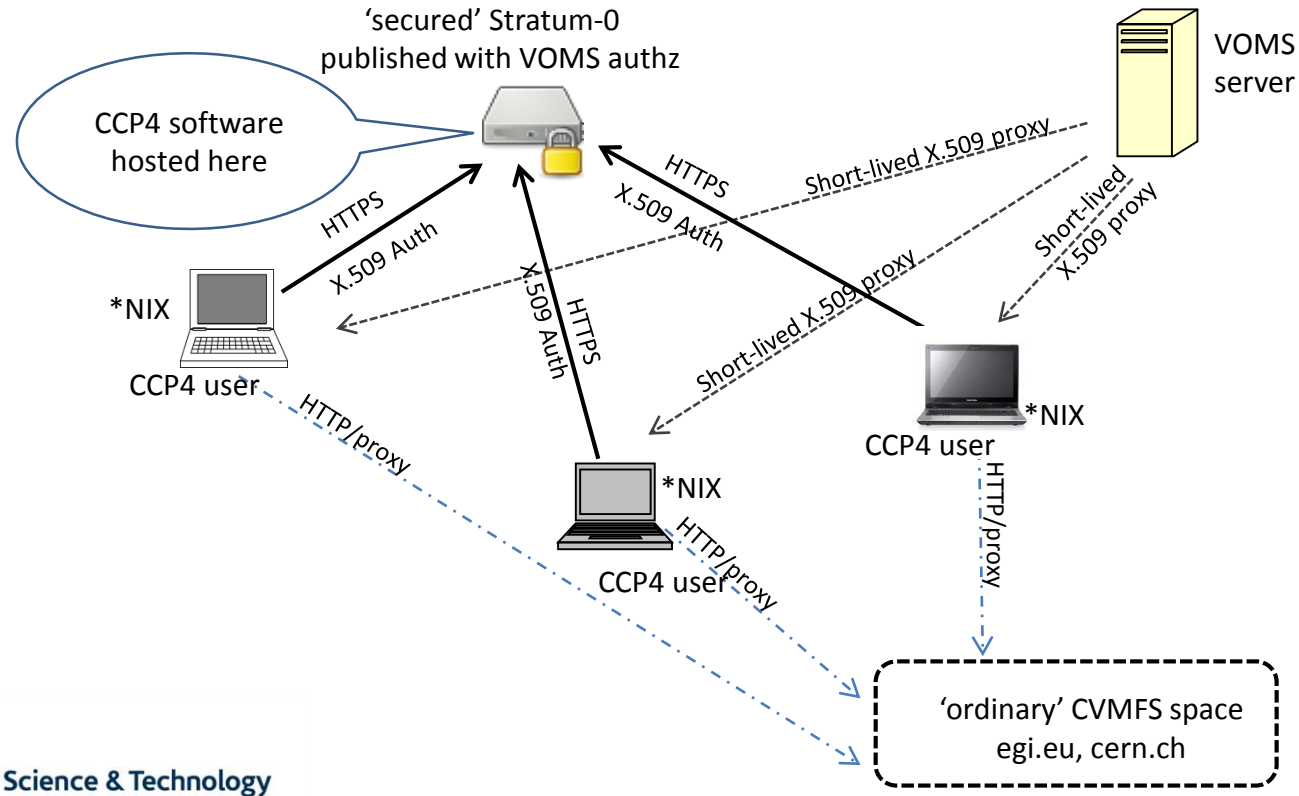
‘Confidential’ CernVM-FS repositories

Working prototype at RAL

- Repository for Collaborative Computational Project No.4
  - www.ccp4.ac.uk
  - To host CCP4 Software for Macromolecular X-Ray Crystallography
  - (probably) /cvmfs/ccp4-sw.stfc.uk
- Also exploratory discussions with ELI-NP
  - Extreme Light Infrastructure – Nuclear Physics Project www.eli-np.ro

‘Confidential’ CernVM-FS repositories

CCP4 use case



(Immediate) Plans

Re-enable support for LIGO jobs on batch farm(s)

- ‘Secure’ CVMFS client on WNs
  - It used to work back in 2017…
  - Not sure if (non-secure CVMFS) “Hello world!” LIGO jobs work at RAL
  - It might be a problem of ‘/osg/ligo’ VOMS permissions – under investigations

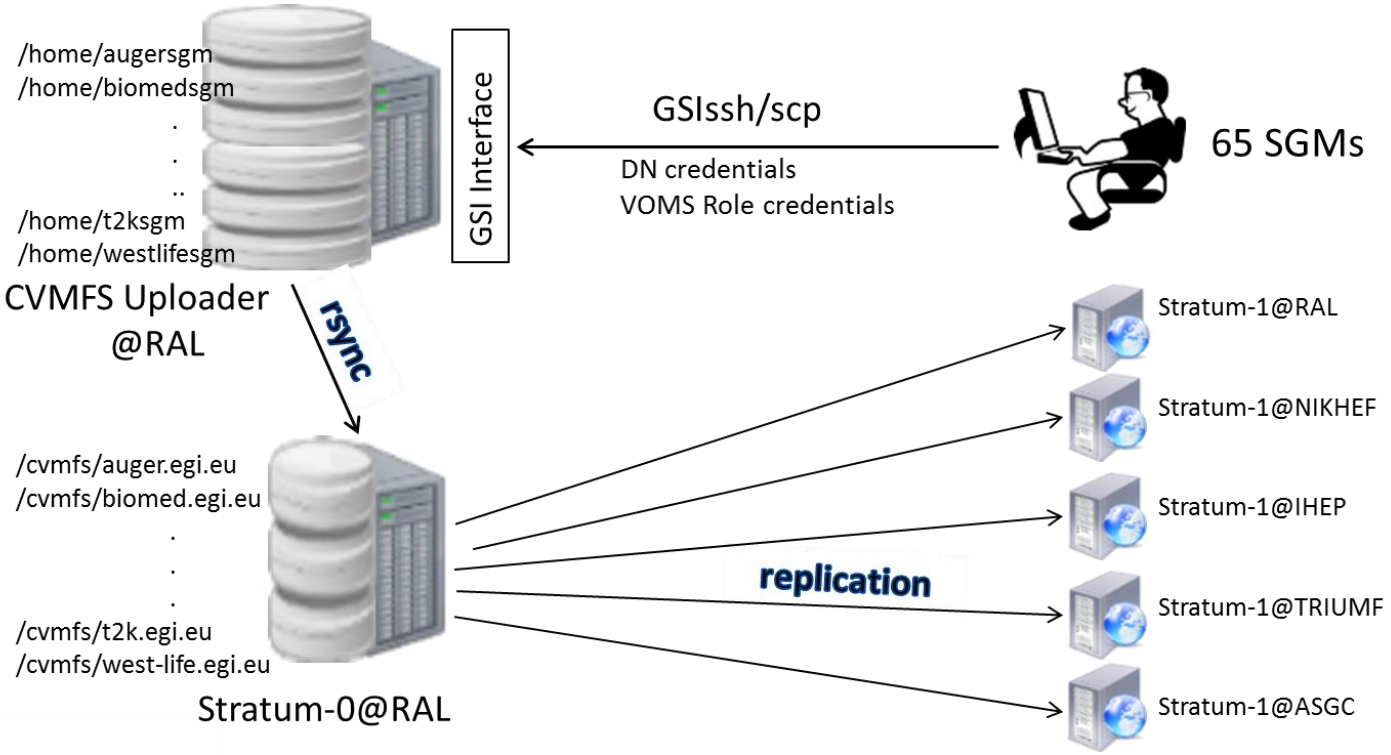
Plans

Repository Gateway and Release Managers

- Replacement for the existing Uploader service
- Makes use of recently introduced CernVM-FS features
- Gateway (GW)
  - Machine running an instance of the CVMFS repository gateway – it has access to the authoritative storage of the managed repositories
  - *cvmfs*, *cvmfs-server*, *cvmfs-gateway* packages
- Release manager (RM)
  - Machine running CVMFS server tools – can request leases from a GW and publish changes to different repositories
  - *cvmfs*, *cvmfs-server* packages

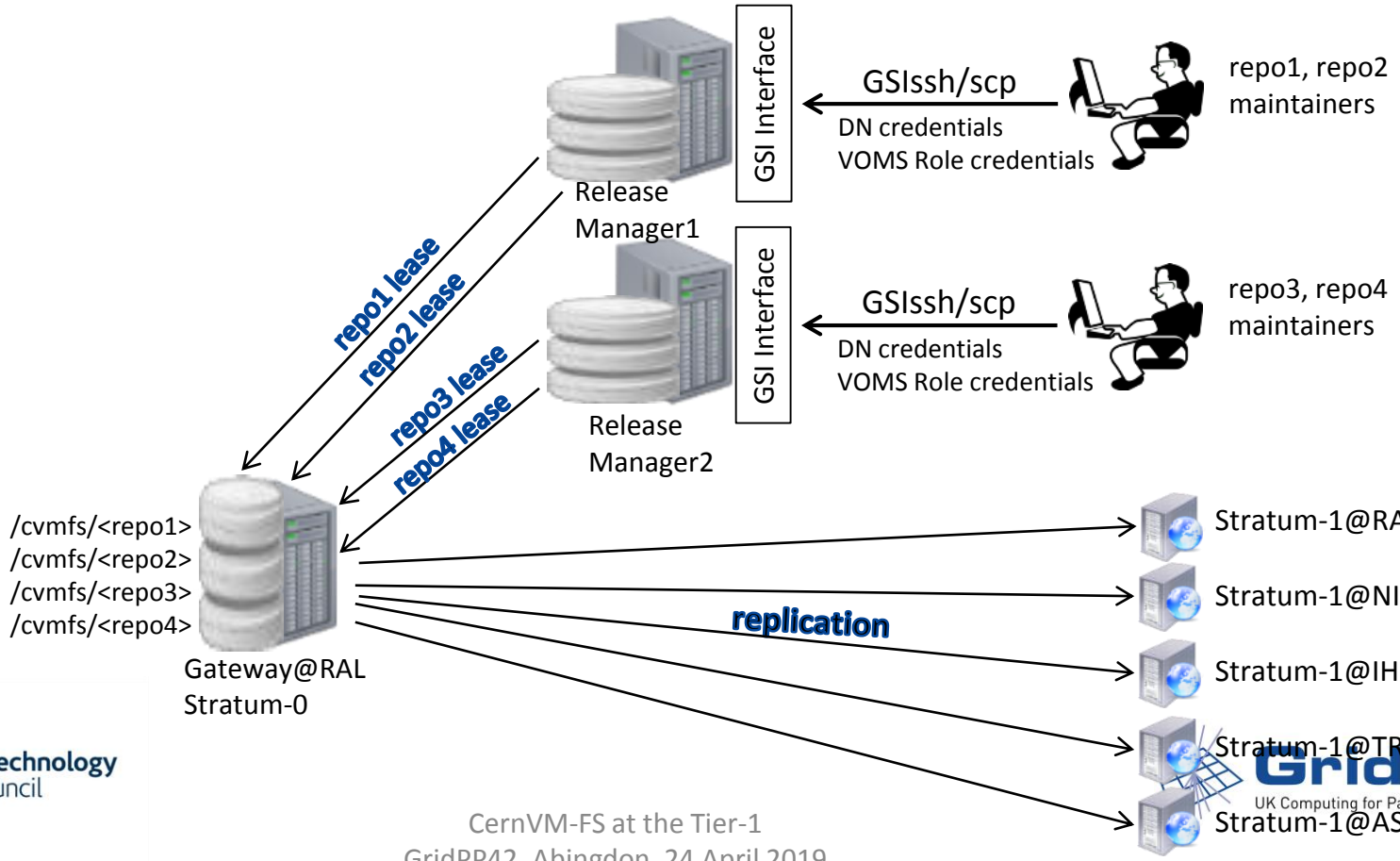
CernVM-FS Infrastructure @RAL

Repository uploading mechanism



Plans

Repository Gateway and Release Manager

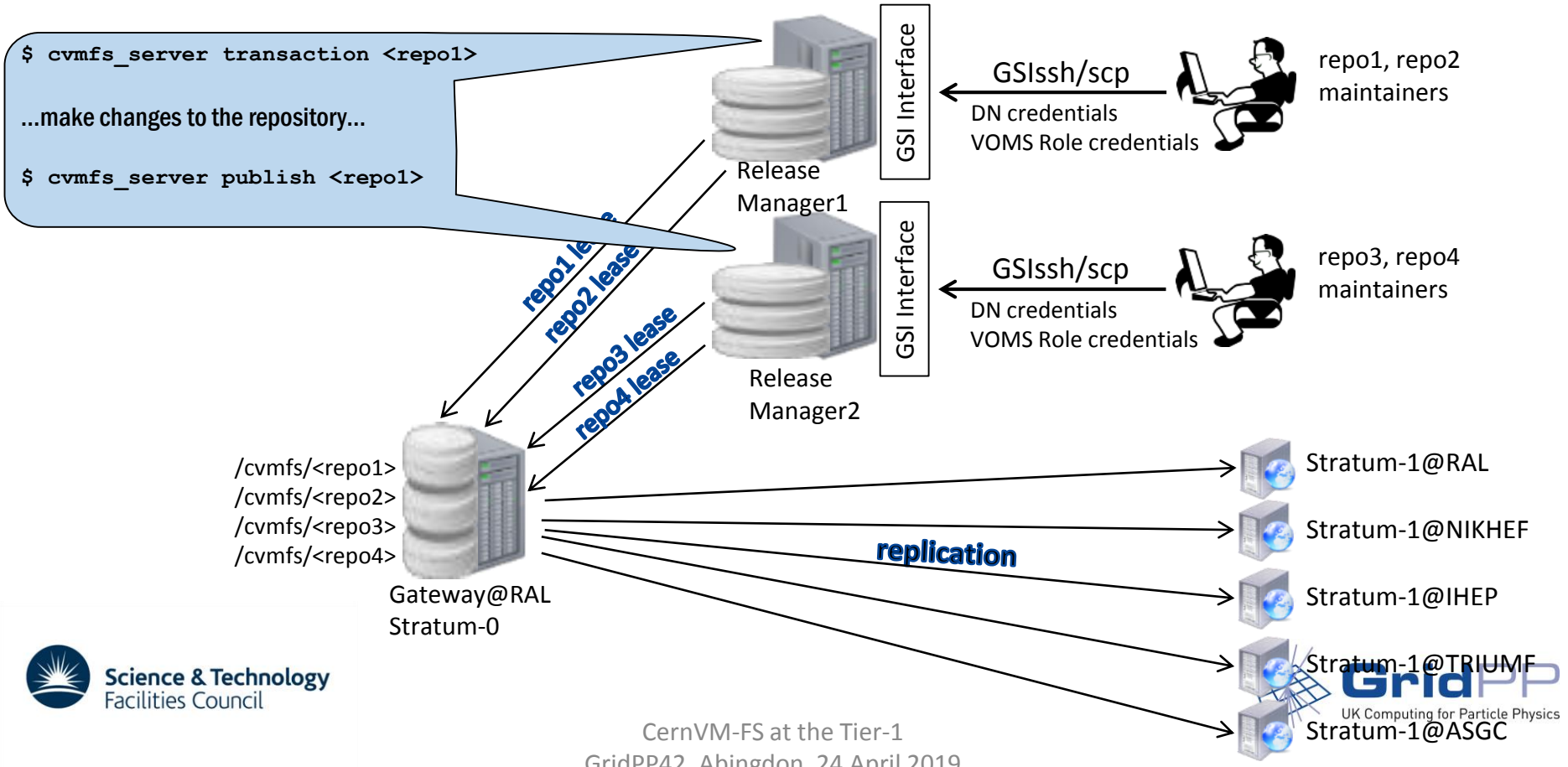


Science & Technology Facilities Council

GridPP UK Computing for Particle Physics

Plans

Repository Gateway and Release Manager

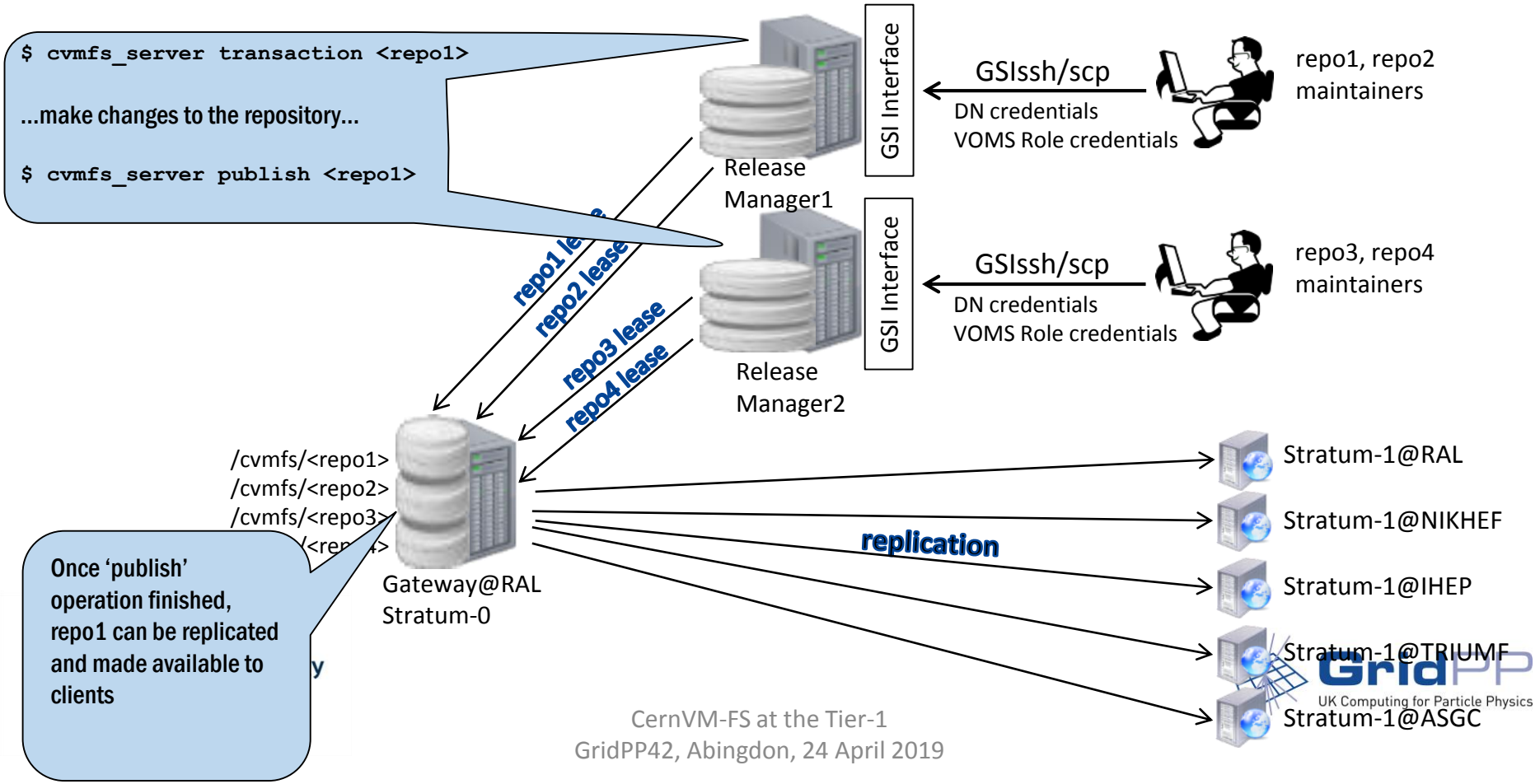


Science & Technology Facilities Council

GridPP UK Computing for Particle Physics

Plans

Repository Gateway and Release Manager



Plans

(GW + RM) vs Uploader service

- With the Uploader…
  - 60min+ delay between last repo maintenance operation and publishing at Stratum-0 level (*rsync* run by hourly cronjob)
  - Repo maintainers do not have to use cvmfs related commands
- With the GW + RM…
  - Repo maintainers need to use the *cvmfs_server {transaction,publish,abort}* commands
  - Repo is published and ready for replication once *cvmfs_server publish <repo>* is executed
- Transition period
  - Existing repos need to be gradually migrated from old Uploader (+ Stratum-0) mechanism to new GW + RM configuration
  - Reverse proxy configuration for each migrated repo to be used on Stratum-0 server - migration will be transparent at replication level
  - Initially *config-egi.egi.eu* and **.gridpp.ac.uk* repos

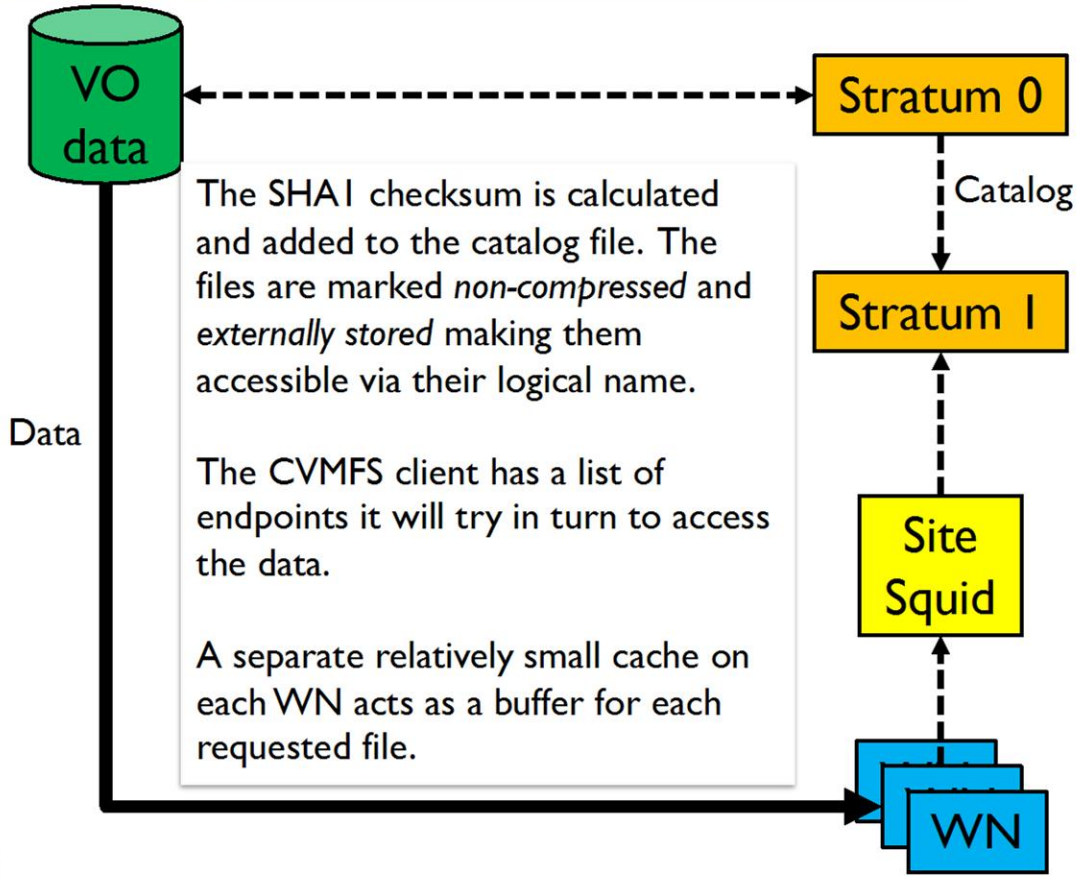
(Also) Plans

Large Scale CVMFS

- CVMFS primarily developed for distributing large software stacks (GB)
- Colleagues from OSG developed extensions to CVMFS software that permit distribution of large, non-public datasets (TB to PB)
- Data is not stored within the repository - only checksums and the catalogs
  - Data is externally stored
  - CVMFS clients are configured to be pointed at a non-CVMFS data storage
  - i.e. external XROOT storage can be referred by a CVMFS repository and accessed in a POSIX-like manner (‘ls’, ‘cp’ etc)

Large Scale CVMFS

6



Alastair Dewhurst, 30th November 2017

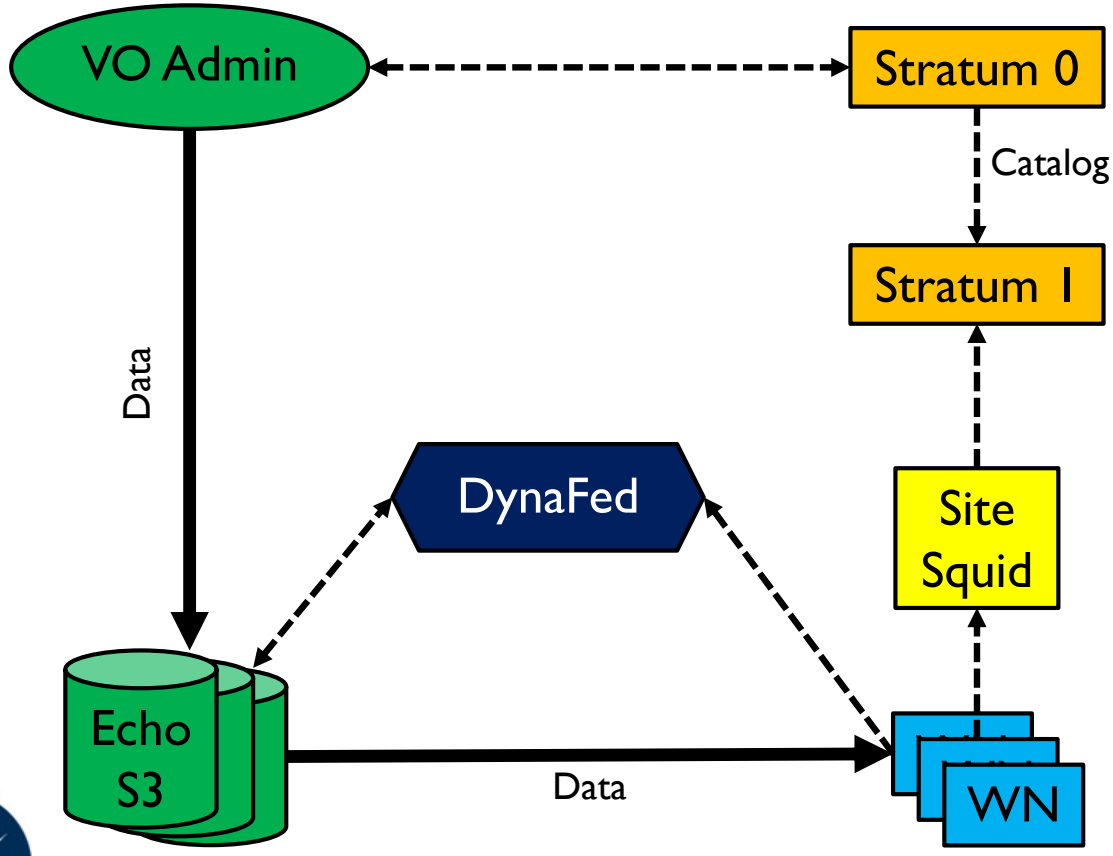
Large Scale CVMFS

Work in early stage at RAL (for LIGO – incl X.509 read-access authorization)

- Some LIGO data at RAL (thanks to Paul Hopkins) - https://dynafed.stfc.ac.uk/gridpp/ligo/frames
- Plan to have a significant share of LIGO datasets at RAL
  - LIGO is to use RUCIO
  - RUCIO can put data directly into Ceph S3 (authz via Dynafed)
- Depending on funding, Large Scale CVMFS can be a very useful thing for small VOs

Large Scale CVMFS

27



Alastair Dewhurst, 30th November 2017

Thank you!

Questions?