CernVM-FS at RAL Tier-1 Status and Developments

Catalin Condurache

STFC UK Research and Innovation

RO-LCG 2018, Cluj Napoca, Romania, 17-19 October 2018

Outline

- UK GridPP collaboration and RAL Tier-1 centre
- CernVM-FS - introduction
- Brief history
- CernVM-FS infrastructure @RAL
- The users
- Recent developments
- Plans

Outline

- UK GridPP collaboration and RAL Tier-1 centre
- CernVM-FS - introduction
- Brief history
- CernVM-FS infrastructure @RAL
- The users
- Recent developments
- Plans

GridPP UK

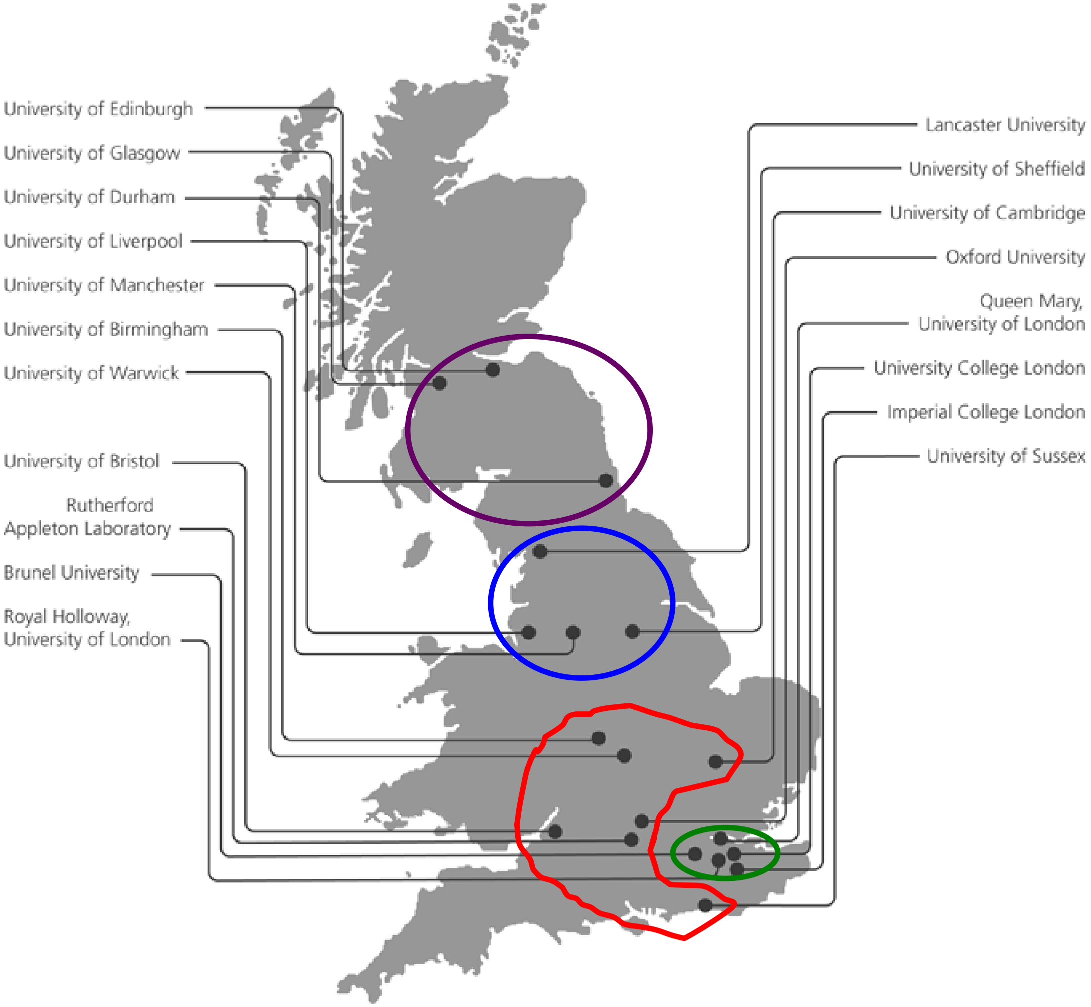
- The GridPP Collaboration is a community of particle physicists and computer scientists based in the United Kingdom and at CERN
- It consistently delivers world-class computing in support of all LHC experiments and many more user communities in a wide variety of fields

GridPP UK

- ~10% of WLCG
- Collaborating

Institutes

- ScotGrid
- NorthGrid
- SouthGrid
- LondonGrid



New UK Research Organisation

- UK Research and Innovation, launched 1st April 2018, is the new funding organisation for research and innovation in the UK
- It brings together the seven UK research councils, Innovate UK and a new organisation, Research England, working closely with its partner organisations in the devolved administrations
  - Includes STFC, which runs RAL
- UK Research and Innovation intends to be an outstanding organisation that ensures the UK maintains its world-leading position in research and innovation

Rutherford Appleton Laboratory - RAL

- 15 miles south of Oxford on Harwell Campus
- Run by STFC
- Multi-discipline centre supporting university and industrial research in big facilities:

  Neutron Science, Lasers, Space Science, Computing
- Hosts UK LHC Tier-1 Facility (RAL Tier-1, RAL-LCG2)
  - Also RALPP Tier-2

RAL Tier-1 Centre

- CPU: ~236k HS06 (~22k cores)
  - Latest procurement ~91k HS06
- Castor: ~16.5 PB useable
  - Dropping as older HW retired
- Ceph: ~20 PB raw / ~13 PB configured
  - Latest procurement (19.5 PB raw / 14.2 PB configured) in acceptance testing
- Tape: 10k slot SL8500
  - 80PB capacity (T10KD)
  - ~30PB physics data

Outline

- UK GridPP collaboration and RAL Tier-1 centre
- CernVM-FS - introduction
- Brief history
- CernVM-FS infrastructure @RAL
- The users
- Recent developments
- Plans

Introduction – CernVM File System?

- Read-only, globally distributed file system optimized for scientific software distribution onto virtual machines and physical worker nodes in a fast, scalable and reliable way
- Some features - aggressive caching, digitally signed repositories, automatic file de-duplication
- Built using standard technologies (fuse, sqlite, http, squid and caches)
- Files and directories are hosted on standard web servers and get distributed through a hierarchy of caches to individual nodes – POSIX like access

Introduction – CernVM File System?

- Software needs one single installation, then it is available at any site with CernVM-FS client installed and configured
- Mounted in the universal ***/cvmfs*** namespace at client level
- The method to distribute HEP experiment software within WLCG, also adopted by other computing communities outside HEP
- Can be used everywhere (because of http and squid) i.e. cloud environment, local clusters (not only grid)
  - Add CernVM-FS client to a VM image => ***/cvmfs*** space automatically available

Outline

- UK GridPP collaboration and RAL Tier-1 centre
- CernVM-FS - introduction
- Brief history
- CernVM-FS infrastructure @RAL
- The users
- Recent developments
- Plans

Brief History

- Summer 2010 – RAL was the first Tier-1 centre to test CernVM-FS at scale and worked towards getting it accepted and deployed within WLCG

Brief History

- Summer 2010 – RAL was the first Tier-1 centre to test CernVM-FS at scale and worked towards getting it accepted and deployed within WLCG
- February 2011 – first global CernVM-FS Stratum-1 replica for LHC VOs in operation outside CERN

Brief History

- Summer 2010 – RAL was the first Tier-1 centre to test CernVM-FS at scale and worked towards getting it accepted and deployed within WLCG
- February 2011 – first global CernVM-FS Stratum-1 replica for LHC VOs in operation outside CERN
- September 2012 – non-LHC Stratum-0 service at RAL supported by the GridPP UK project
  - Local installation jobs used to automatically publish the Stratum-0
  - Shared Stratum-1 initially

Brief History

- Aug - Dec 2013 – Stratum-0 service expanded to EGI level
  - Activity coordinated by the EGI CVMFS Task Force
  - ‘gridpp.ac.uk’ space name for repositories
  - Web interface used to upload, unpack tarballs and publish
  - Separated Stratum-1 at RAL
  - Worldwide network of Stratum-1s in place (RAL, CERN, NIKHEF, OSG) – it followed the WLCG model

Brief History

- Aug - Dec 2013 – Stratum-0 service expanded to EGI level
  - Activity coordinated by the EGI CVMFS Task Force
  - 'gridpp.ac.uk' space name for repositories
  - Web interface used to upload, unpack tarballs and publish
  - Separated Stratum-1 at RAL
  - Worldwide network of Stratum-1s in place (RAL, CERN, NIKHEF, OSG) – it followed the WLCG model
- March 2014 – 'egi.eu' domain
  - Public key and domain configuration became part of standard installation (as for 'cern.ch')

Brief History

- December 2014 – HA 2-node cluster for non-LHC Stratum-1
  - It replicates also ‘opensciencegrid.org’, ‘desy.de’, ‘nikhef.nl’ repos

Brief History

- December 2014 – HA 2-node cluster for non-LHC Stratum-1
  - It replicates also ‘opensciencegrid.org’, ‘desy.de’, ‘nikhef.nl’ repos
- January 2015 – CVMFS Uploader consolidated
  - Grid Security Interface (GSI) added to transfer and process tarballs and publish - based on DN access, also VOMS Roles
  - Faster and easier, programmatic way to transfer and process tarballs

Brief History

- December 2014 – HA 2-node cluster for non-LHC Stratum-1
  - It replicates also ‘opensciencegrid.org’, ‘desy.de’, ‘nikhef.nl’ repos
- January 2015 – CVMFS Uploader consolidated
  - Grid Security Interface (GSI) added to transfer and process tarballs and publish - based on DN access, also VOMS Roles
  - Faster and easier, programmatic way to transfer and process tarballs
- March 2015 – 21 repos, 500 GB at RAL
  - Also refreshed Stratum-1 network for ‘egi.eu’ – RAL, NIKHEF, TRIUMF, ASGC

Brief History

- Sep 2015 – single consolidated HA 2-node cluster Stratum-1
  - 56 repos replicated from RAL, NIKHEF, DESY, OSG, CERN
- …<*fast forward*>…

Outline

- UK GridPP collaboration and RAL Tier-1 centre
- CernVM-FS - introduction
- Brief history
- CernVM-FS infrastructure @RAL
- The users
- Recent developments
- Plans

CernVM-FS Infrastructure @RAL

- Stratum-0 service @ RAL (EGI, STFC)
  - Maintains and publishes the current state of the repositories
  - 32GB RAM, 12TB disk, 2x E5-2407 @2.20GHz
  - cvmfs-server v2.5.1 (includes the CernVM-FS toolkit)
  - 35 repositories
  - egi.eu
    - *auger, biomed, cernatschool, chipster, comet, config-egi*
    - *dirac, eosc, extras-fp7, galdyn, ghost, glast, gridpp, hyperk, km3net*
    - *ligo, lucid, mice, neugrid, pheno, phys-ibergrid, pravda, researchinschools*
    - *skatelescope, solidexperiment, snoplus, supernemo, t2k, wenmr, west-life*
  - gridpp.ac.uk
    - *londongrid, scotgrid, northgrid, southgrid, facilities*

CernVM-FS Infrastructure @RAL

- Operations Level Agreement for Stratum-0 service
  - between STFC and EGI.eu
  - provisioning, daily running and availability of service
  - service to be advertised through the EGI Service Catalog

CernVM-FS Infrastructure @RAL

- CVMFS Uploader service @ RAL (EGI, STFC)
  - In-house implementation that provides upload area for *egi.eu* (and *gridpp.ac.uk*) repositories
  - Currently 1.9 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

- Stratum-1 service (WLCG, EGI, STFC)
  - Standard web server (+ CernVM-FS server toolkit) that creates and maintains a mirror of a CernVM-FS repository served by a Stratum-0 server
  - Part of the worldwide network of servers (RAL, NIKHEF, TRIUMF, ASGC, IHEP) replicating the *egi.eu* repositories
  - RAL - 2-node HA cluster (cvmfs-server v2.5.1)
    - each node – 64 GB RAM, 55 TB storage, 2xE5-2620 @2.4GHz

CernVM-FS Infrastructure @RAL

- Stratum-1 service (WLCG, EGI, STFC)
  - RAL - 2-node HA cluster (cvmfs-server v2.5.1)
    - it replicates 80 repositories – total of 28 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
    - very recent request
      - GGUS#137752 – Replicate OSG CVMFS repos to EGI Stratum-1s
      - 12 OSG repos to be replicated – 615GB
      - part of Fermilab VO

CernVM-FS Infrastructure @RAL

- 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
- The EGI Staged Rollout
  - RAL is an early Adopter for cvmfs client, cvmfs server and frontier-squid

Outline

- UK GridPP collaboration and RAL Tier-1 centre
- CernVM-FS - introduction
- Brief history
- CernVM-FS infrastructure @RAL
- The users
- Recent developments
- 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, comet, pheno*
- Space and Earth Sciences
  - *auger, extras-fp7*
- Biological Sciences
  - *chipster, enmr*

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
- *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

- UK GridPP collaboration and RAL Tier-1 centre
- CernVM-FS - introduction
- Brief history
- CernVM-FS infrastructure @RAL
- The users
- Recent developments
- Plans

Developments – ‘protected’ CernVM-FS Repositories

- Repositories natively designed to be public with non-authenticated access
  - One needs to know only minimal info - access to the 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/posibility for some years

Developments – ‘protected’ CernVM-FS Repositories

- 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
  - “Accessing Data Federations with CVMFS” (CHEP 2016 - https://indico.cern.ch/event/505613/contributions/2230923/)
- We took the opportunity and looked to make use of this new feature by offering 'secure' CernVM-FS to interested user communities

Developments – ‘protected’ CernVM-FS Repositories

- Working prototype at RAL
  - Stratum-0 with mod_gridsite, https enabled
    - ‘cvmfs_server publish’ operation incorporates an authorization info file (DNs, VOMS roles)
    - access based on .gacl (Grid Access Control List) file in *<repo>/data/* directory that has to match the required DNs or VOMS roles
  - CVMFS client + cvmfs_helper package (enforces authz to the repository)
    - obviously '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 is not possible for HTTPS

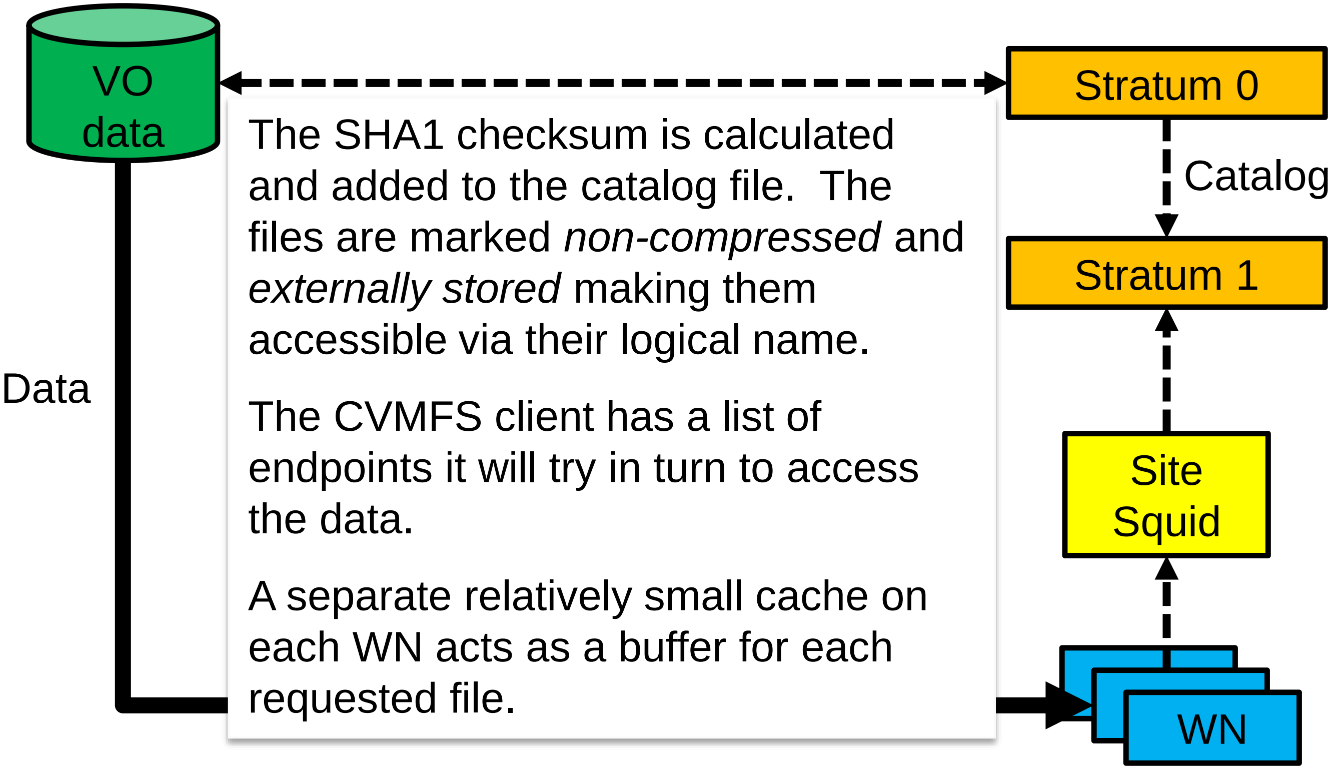
Plans – ‘protected’ CernVM-FS Repositories

- Cloud environment - good starting point for a use case
  - Multiple VMs instantiated at various places and accessing the ‘secure’ repositories provided by a Stratum-0
  - A VM is not shared usually, it has a single user (which has root privileges as well)
  - The user downloads a certificate, creates a proxy and starts accessing the ‘secure’ repo
  - Process can be automated by using ‘robot’ certificates
    - and better by downloading valid proxies
- Another possible use case
  - Access from shared UIs, worker nodes
- No effort allocated in last 6-9 moths though…

Developments – 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
  - CVMFS clients are configured to be pointed at a non-CVMFS data
  - i.e. external XROOT storage can be referred by a CVMFS repository and accessed in a POSIX-like manner (‘ls’, ‘cp’ etc)
- Work in early stage at RAL (for LIGO – incl X.509 read-access authorization)

Developments – Large-Scale CVMFS



The SHA1 checksum is calculated and added to the catalog file. The files are marked *non-compressed* and *externally stored* making them accessible via their logical name.

The CVMFS client has a list of endpoints it will try in turn to access the data.

A separate relatively small cache on each WN acts as a buffer for each requested file.

Alastair Dewurst et al – LS-CVMFS and Dynafed - CHEP 2018

Thank you!

Questions, comments?