CernVM-FS – beyond LHC computing

C Condurache, I Collier
STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot, OX11 0QX, UK
E-mail: catalin.condurache@stfc.ac.uk

Abstract. In the last three years the CernVM File System has transformed the distribution of experiment software to WLCG sites. CernVM-FS removes the need for local installations jobs and performant network file servers at sites and it often improves performance at the same time. Now established and proven to work at scale, CernVM-FS is beginning to perform a similar role for non-LHC computing. The deployment of CernVM-FS service at RAL Tier-1 is presented, as well as the proposed development of a network of Stratum-0 and Stratum-1 replicas somewhat modelled upon the infrastructure developed to support the WLCG computing. A case study of one non-LHC Virtual Organization is also included, describing their use of the CernVM-FS Stratum-0 service, along with a web interface intended to be used as a tool to upload software at Stratum-0 sites.

1. What is CernVM-FS?
The CernVM File System (CernVM-FS) [1] is a read-only, globally distributed file system optimized for software distribution. Originally developed for the CERN Virtual Machine project [2], it was designed to solve the problem of distributing frequently changing experiment software to virtual machines (VM) that might not have access to traditional software servers without having to include it within the VM images. It is built using standard technologies, being implemented as a POSIX read-only file system in the user space (a FUSE module) [3]. Internally, a CernVM-FS repository is defined by its file catalog, which is a SQLite database [4]. CernVM-FS uses the HTTP protocol, which allows data replication to multiple web servers and data caching by standard web proxies such as Squid [5].

The replication source server is the CernVM-FS Stratum-0 and contains a read-writeable copy of the repository. In a typical setup, repositories are replicated to a network of web servers, which form the CVMFS Stratum-1 service. While a CernVM-FS Stratum-0 repository server is able to serve clients directly, a large number of clients is better served by a set of Stratum-1 replica servers. Multiple Stratum-1 servers improve the reliability, reduce the load and protect the Stratum-0 master copy of the repository from direct accesses.

In the traditional way, the experiment software is distributed using installation jobs with shared software areas on common distributed file systems, such as Network File System (NFS) [6] or Andrew File System (AFS) [7]. This entire process is error-prone and manpower intensive, as a typical software release is several Gigabytes in size, consists of tens of thousands of files and has to be stored at each site supporting a specific Virtual Organization (VO).

Now CernVM-FS removes the need for local installation jobs and conventional Network File System (NFS) servers at every site and its use simplifies the environment across the Computing Grid. For any Virtual Organization, the software needs one single installation and then is available at any site with CernVM-FS client installed.
Once the signed file catalog has been downloaded on a worker node at a site and mounted using FUSE, the metadata operations require no further network access. Together with the file based de-duplication (since the CernVM-FS repository is a form of content-addressable storage), this makes CernVM-FS very efficient in terms of disk usage and network traffic.

The data validity and integrity of the file system are ensured by the combination between the digitally signed file catalog and the cryptographic content hash used as a method to locate a file within the catalog.

2. CernVM-FS History at RAL Tier-1

The Tier-1 centre at the Rutherford Appleton Laboratory (RAL) [8] provides large-scale computing facilities for UK Particle Physics experiments and it is used through the GridPP [9] and European Grid Infrastructure (EGI) [10] projects for UK and wider European access.

During the summer 2010 the RAL Tier-1 was the first Worldwide LHC Computing Grid (WLCG) [11] site involved in testing CernVM-FS at scale and worked towards getting it accepted and deployed within WLCG. Then in February 2011 the first global CernVM-FS Stratum-1 replica for LHC Virtual Organizations became operational outside CERN at the RAL Tier-1.

In December 2012, a new initiative to offer CernVM-FS Stratum-0 services to non-LHC VOs was started at RAL, and first repositories have been setup for the *mice* [12] and *na62* [13] VOs supported by the GridPP UK project. Between January and August 2013 the RAL CernVM-FS Stratum-0 service has been extended to other international small VOs.

As at October 2013 the initiative is entering the ‘EGI service’ phase with more VOs supported or requesting access to this software distribution service and a separate CernVM-FS Stratum-1 service is being offered at RAL for non-LHC VOs only. In collaboration with the Open Science Grid (OSG) [14], which is running a similar CernVM-FS infrastructure, it is proposed that experiments with worldwide presence will be given access to a consolidated software distribution mechanism across both organizations.

In parallel, the RAL Tier-1 group is involved in the verification of CernVM-FS client software at scale before it is released into production.

3. *enmr.eu* CernVM-FS Stratum-0 at RAL Tier-1

The CernVM-FS Stratum-0 deployed at RAL Tier-1 is offering its services to non-LHC experiments belonging not only to the High Energy Physics (HEP) community, but also to other communities like Life or Earth Sciences.

WeNMR [15] is a Virtual Research Community (VRC) supported by EGI which aims to build a worldwide e-Infrastructure for Nuclear Magnetic Resonance (NMR) spectroscopy and Structural Biology. Its VO – *enmr.eu* – is currently the largest in the area of Life Sciences with over 570 registered users and a steady growth rate. *enmr.eu* has an active presence across the Grid at more than 25 sites, which makes the distribution of new software releases time consuming.

The deployment and configuration of the CernVM-FS Stratum-0 repository for *enmr.eu* at RAL Tier-1 has reduced the effort required to maintain multiple NFS software areas at all other supporting sites. Installation jobs run by the *enmr.eu/Role=Manager* at the RAL Tier-1 batch farm upload and configure new software releases within a standard NFS area, then a synchronization mechanism updates the `/cvmfs/wenmr.gridpp.ac.uk` directory within the Stratum-0 repository.

Work is ongoing to consolidate the *enmr.eu* Stratum-0 at RAL and to move toward a CernVM-FS only environment.

4. CernVM-FS Task Force

The CernVM-FS Task Force [16] is a working group setup by EGI to establish a CernVM-FS infrastructure that allows EGI VOs to use it as a standard method of distribution of their software across the EGI computing resources. It has an active role in promoting the use of CernVM-FS technology by VOs and in creating a network of sites providing the core services (Stratum-0, Stratum-
1, Squid). It also encourages the institutions running local Stratum-0 and Stratum-1 servers to become part of the EGI CernVM-FS infrastructure by offering their services at a regional or larger level.

The Task Force also promotes cooperation with other organizations, such as OSG and the WLCG, on monitoring tools for CernVM-FS and by cross-replicating repositories for VOs supported by multiple collaborations.

5. A ‘Relaxed’ Topology for the EGI CernVM-FS Infrastructure Deployment

The CernVM-FS Task Force’s main task is the development by the beginning of 2014 of a network of Stratum-0 repositories and Stratum-1 replicas modelled upon the infrastructure established to support the WLCG computing. The CernVM-FS software distribution model for the LHC VOs is shown in Figure 1.

In this model the software is installed by the LHC VOs at the Stratum-0 instance hosted at CERN and replicated to Stratum-1 replica servers hosted by WLCG Tier-1 sites. The CernVM-FS clients connect to one of the Stratum-1 services via a distributed hierarchy of proxy servers.

The proposed CernVM-FS infrastructure model for the EGI community is presented in Figure 2. This model consists of more than one Stratum-0 instances which are disjoint and represent the source repositories where software is installed by the EGI VOs. Stratum-0 and Stratum-1 servers can be geographically co-located or not, depending on the resources made available by a specific site. Another main characteristic of this model is the fact that a Stratum-1 can replicate a whole Stratum-0 or can partially replicate, depending again on the site resources and on what VOs are being supported.

This is the ‘relaxed’ model designed for the EGI experiments, where the software repositories are spread across several Stratum-0 locations and where Stratum-1 instances can participate with less than full commitment in the replication process. Of course, similar to the WLCG model, a distributed and consolidated network of Stratum-1s makes CernVM-FS software distribution more resilient, so each EGI Stratum-1 should aim to replicate as much as possible.

It is worth mentioning one advantage of the CernVM-FS EGI model which is that it can make use of the existent hierarchy of proxy servers used for the LHC software distribution.
6. CernVM-FS Stratum-0 Web Frontend

The CernVM-FS Stratum-0 repositories at RAL were initially maintained by the means of installation jobs in the case of VOs that had a CPU allocation at RAL Tier-1. As more VOs without any CPU
allocation started to use the CernVM-FS service, manual handling of software tarballs was a temporary solution and alternatives were investigated. One option was to give VO Software Grid Managers (SGMs) privileges to login into Stratum-0 servers and upload the software (method used elsewhere at other Stratum-0 sites), but this was not agreed at RAL as a good practice policy because of potential security issues with non-local user accounts.

To remove the need for privileged roles and jobs, a web interface which allows VO Software Managers to upload and publish their software to the Stratum-0 has been developed during summer 2013 (see Figure 3). It authenticates users (mainly VO SGMs) based on their X509 certificates that are managed by a web server and it is due to be made available to all VOs by the end of 2013.

The application allows uploading and unpacking tarballs within the `/cvmfs/<repo_name>` ‘space’, also users can remove obsolete software or visualize the entire directory structure. At the backend a synchronization process with the real CernVM-FS Stratum-0 repository is regularly triggered whenever new changes are detected.

7. Summary

As to date (October 2013) CernVM-FS has proved itself very scalable in distributing the LHC experiment software at nearly 150 WLCG sites with more than 400,000 logical CPUs. Following this successful model, the EGI CernVM-FS Task Force is coordinating efforts to develop a similar infrastructure that will serve the EGI supported VOs and not only. The RAL Tier-1 centre is leading this activity by hosting Stratum-0 and Stratum-1 services, but it is envisaged that a consolidated EGI CernVM-FS infrastructure will include additional Stratum-0 and Stratum-1 sites and also will make use of the similar existing services within OSG.

All HEP and non-HEP communities will overall benefit from this project as it standardizes the computing environment across the Grid and frees valuable resources at supporting sites to be reused for other purposes.

Acknowledgments

Many thanks to Jakob Blomer for his advices during CernVM-FS Stratum-0 initial implementation at the RAL Tier-1. Also many thanks go to Michal Knapik for his invaluable work in developing the CernVM-FS Stratum-0 web frontend.

References

[1] Blomer J, Aguado-Sanchez C, Buncic P and Harutyunyan A 2011 Distributing LHC application software and conditions databases using the CernVM file system Journal of Physics: Conference Series 331 042003
[2] CERN Virtual Machine Project - http://cernvm.cern.ch
[3] FUSE - http://fuse.sourceforge.net
[4] SQLite - http://www.sqlite.org
[5] Squid - http://www.squid-cache.org
[6] NFS - http://nfs.sourceforge.net
[7] AFS - http://www.openafs.org
[8] RAL Tier-1 - http://www.stfc.ac.uk
[9] GridPP UK - http://www.gridpp.ac.uk
[10] EGI - http://www.egi.eu
[11] WLCG - http://wlcg.web.cern.ch
[12] MICE - http://mice.iit.edu
[13] NA62 - http://na62.web.cern.ch
[14] OSG - http://www.opensciencegrid.org
[15] WeNMR - http://www.wenmr.eu
[16] EGI CernVM-FS Task Force - https://wiki.egi.eu/wiki/CVMFS-TF:CVMFS_Task_Force