Deployment and Operational Experiences with CernVM-FS at the GridKa Tier-1 Center

Manfred Alef, Axel Jäger, Andreas Petzold and Bernhard Verstege
Karlsruhe Institute of Technology, Steinbuch Centre for Computing, Hermann-von-Helmholtz Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
E-mail: andreas.petzold@kit.edu

Abstract. In 2012 the GridKa Tier-1 computing center hosts 130 kHS06 computing resources and 14 PB disk and 17 PB tape space. These resources are shared between the four LHC VOs and a number of national and international VOs from high energy physics and other sciences. CernVM-FS has been deployed at GridKa to supplement the existing NFS-based system to access VO software on the worker nodes. It provides a solution tailored to the requirement of the LHC VOs. We will focus on the first operational experiences and the monitoring of CernVM-FS on the worker nodes and the squid caches.

1. GridKa Tier-1 Center
The GridKa Tier-1 center is hosted by Steinbuch Centre for Computing at Karlsruhe Institute of Technology. GridKa was established in 2002 as a regional computing center for the LHC experiments which were still developing their computing models at that time and several other high energy physics experiments that were already/still taking data. Today, GridKa supports all four LHC experiment VOs and seven more VOs from High Energy Physics and Astroparticle Physics (Auger, BABAR, Belle, Belle2, CDF, Compass, DØ), and several other VOs from different fields of science.

Among the 11 WLCG Tier-1 centers GridKa provides approximately 14% of the resources available to the experiments and it is the largest center supporting all four LHC VOs. In 2012 GridKa provides 130 kHS06 of CPU resources, 14 PB of disk storage, and 17 PB of tape storage to its users.

2. CernVM File System
CernVM File System (CernVM-FS) is a read-only file system designed to deliver software and conditions data fast and reliably to worker nodes (WNs) in the Grid. All data are transported via the HTTP protocol and aggressive caching is used both locally on the WNs as well as on a hierarchy of caching proxy servers. Further details can be found in [1].
3. CernVM-FS Deployment at GridKa

At the GridKa Tier-1 center, CernVM-FS was deployed in early 2012 to supplement the traditional shared NFS software area. Currently, only ATLAS and LHCb are using CernVM-FS, however these two VOs have a nominal share of approximately 45% of the CPU resources at GridKa.

Figure 1. CernVM-FS setup used at GridKa.

Figure 1 shows the CernVM-FS setup which is used at GridKa. Each worker node in the batch farm uses a separate partition on the local hard disk to store the local cache data of CernVM-FS. In order to minimize volume of data transferred from the central Stratum-1 server at CERN, WNs connect to Squid caching proxy servers within the GridKa network to download data. These Squid servers connect to the Stratum-1 only if the requested data was not previously cached or the time-to-live of the data in the cache of the Squid server has expired before.

In order to provide load balancing and fault tolerance of the Squid service, three Squid servers are currently in production at GridKa. All WNs randomly choose one Squid server to connect to and will automatically fail-over to a different server, in case the chosen server is unavailable.

Deployment of CernVM-FS provides advantages to the Tier-1 center in the areas of reliable scaling and required hardware. Scaling the performance of the traditional NFS-based setup to the required number of clients could only be achieved by using expensive high-end servers and dedicated high-performance storage systems. The NFS setup comprised five servers with dual Intel® Xeon® E5530 CPUs, 32 GB of RAM, 10 Gb Ethernet, and 8 Gb/s Fibre Channel connections to dedicated Data Direct Networks™ storage systems were used. In contrast, the Squid servers are inexpensive machines previously used as WNs with Intel® Xeon® 5160 CPUs, 12 GB of RAM, and 1 TB of local disk space. As CernVM-FS clients are aggressively caching data on the WNs, no network or Squid performance bottlenecks were observed.

Although it is very likely that a shared NFS area will always be required at GridKa to serve the needs of VOs that do not use CernVM-FS, the setup can be simplified in the future and already now two servers have been repurposed.
4. Operational Experience at GridKa

After initial deployment of CernVM-FS at GridKa, monitoring for the Squid servers and the local cache as well as CernVM-FS version on WNs has been deployed. While no malfunctions impacting production have been observed, the monitoring has been helpful to detect problems regarding the Squid configuration and the update of CernVM-FS.

As the input/output operations of the jobs are no longer executed against high-performance storage systems but against the local hard disk drives of the WN, a high number of simultaneously starting jobs on one worker node can lead to slow access to the local CernVM-FS cache. At GridKa, the maximum number of concurrent jobs on one WN currently is 24 and no problems due to slow software setup for jobs have been observed. However, at other computing centers problems have been observed when running 64 concurrent jobs on one WN [2].

4.1. Squid Caching Proxy Setup

Three Squid caching proxy servers are used for CernVM-FS at GridKa. The CernVM-FS documentation, a recommended configuration for the Squid service is provided. This configuration is optimized for software and conditions distribution via CernVM-FS to keep the load on the Stratum-1 server as low as possible.

Accidental mis-configuration can have serious effects which are illustrated in Figure 2. The number of requests from GridKa to the Stratum-1 increased by three orders of magnitude when the size limit of cached objects was accidentally set to a wrong value on the GridKa Squid servers.

It is possible to reduce the requests from the local Squid servers to the Stratum-1 even further by peering the Squid caches. Peered servers share their cache digests and first request files from a peer if one has already downloaded the file, instead of sending the request directly to the Stratum-1. Squid peering will be enabled at GridKa in the near future.

4.2. Handling CernVM-FS Upgrades

While upgrading CernVM-FS is as easy as upgrading the installed RPM packages on the WNs, the new version will only become active for a repository (i.e. software or conditions directory of one VO), once all jobs using that repository have finished and the automount process has dropped the mount. Since WNs at GridKa have up to 24 assigned job slots and batch queues up to 120 hours are available, it can take weeks until by coincidence no jobs from one VO run
on a WN and the mount can be dropped.

The situation is illustrated in Figure 3. Within two days of upgrading, almost 90% of the WNs use the latest version of CernVM-FS for the LHCb repository. However, it takes another week until the fraction of WNs using the old version of CernVM-FS decreases below 5%.

Currently, it is not possible to force a lazy unmount of a CernVM-FS repository in order to activate the upgraded version at the next access. Thus the only available option in a batch farm where all VOs share all WNs is to drain nodes where the old CernVM-FS version is still active and put them back into production afterwards. If this coincides with upgrades of other packages which require a reboot to become active, no additional work and reduction in provided CPU resources is necessary. In other cases, additional work might be required in order to ensure a uniform working environment on all WNs.

5. Conclusion
CernVM-FS has been successfully deployed at the GridKa Tier-1 center. Although only two of the LHC VOs are currently using CernVM-FS, significant savings in hardware due to reduction in use of the traditional NFS-based software area were achieved. Since the start of deployment no major operational problems have been experienced and no reliability issues have been observed by the users. The issue with upgrading CernVM-FS is being discussed with the developers and we hope that a more convenient solution can be found.

References
[1] Blomer J, Aguado-Sánchez C, Buncic P and Harutyunyan A 2011 Journal of Physics: Conference Series 331 042003 URL http://stacks.iop.org/1742-6596/331/i=4/a=042003
[2] Kemp Y and Haupt A private communication