Software installation and condition data distribution via CernVM FileSystem in ATLAS

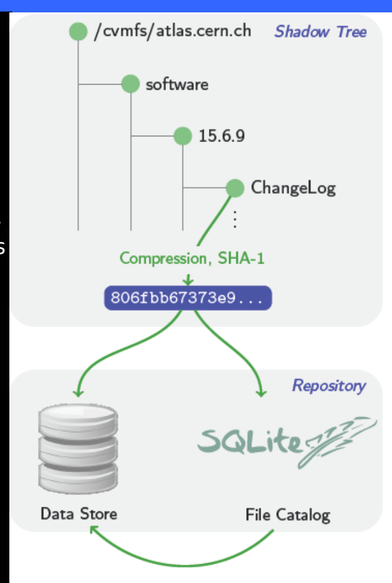
On behalf of the ATLAS collaboration

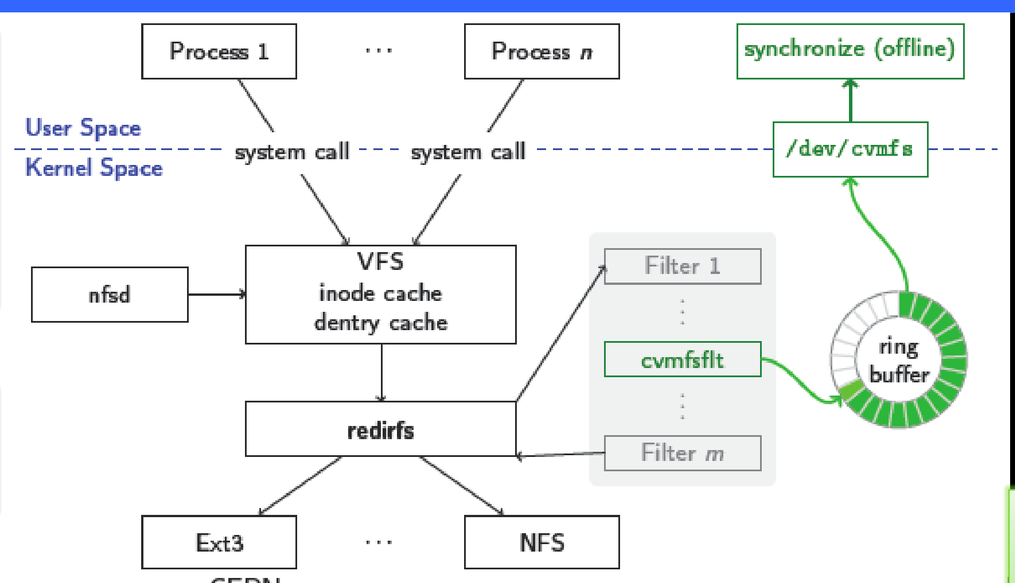
A De Salvo[1], A De Silva[2], D Benjamin[3], J Blomer[4], P Buncic[4], A Harutyunyan[4], A. Undrus[5], Y Yao[6]

The ATLAS Collaboration is managing one of the largest collections of software among the High Energy Physics Experiments. Traditionally this software has been distributed via rpm or pacman packages, and has been installed in every site and user's machine, using more space than needed since the releases could not always share common binaries. As soon as the software has grown in size and number of releases this approach showed its limits, in terms of manageability, used disk space and performance. The adopted solution is based on the CernVM FileSystem, a fuse-based http, read-only filesystem which guarantees file de-duplication, on-demand file transfer with caching, scalability and performance.

CernVM-FS

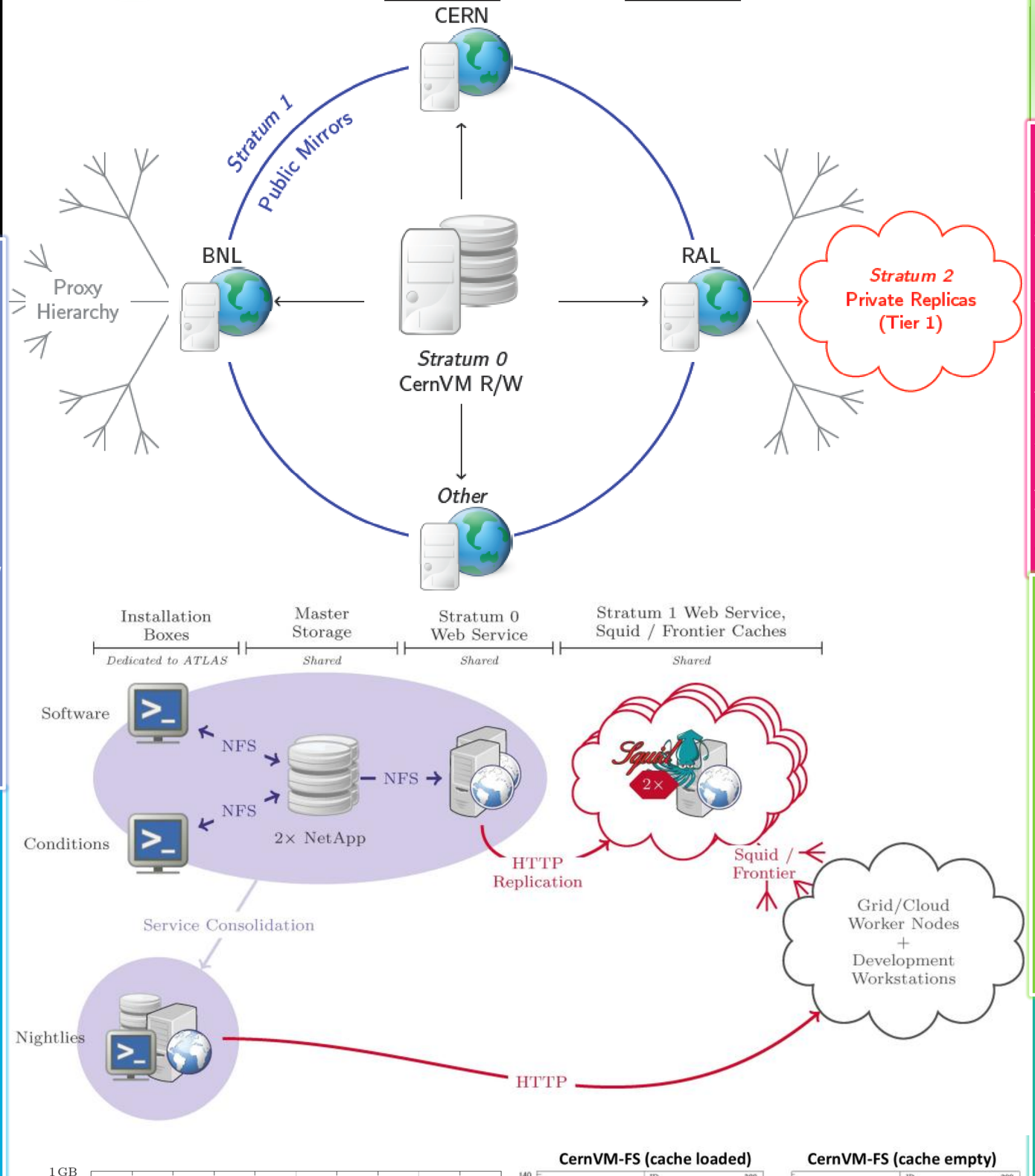
The CernVM File System (CernVM-FS) is a file system used by various HEP experiments for the access and on-demand delivery of software stacks for data analysis, reconstruction, and simulation. It consists of web servers and web caches for data distribution to the CernVM-FS clients that provide a POSIX compliant read-only file system on the worker nodes. CernVM-FS leverages the use of content-addressable storage.
Files with the same content are automatically not duplicated because they have the same content hash.
Data integrity is trivial to verify by re-calculating the cryptographic content hash.
Files that carry the name of their cryptographic content hash are immutable. Hence they are easy to cache as there is no expiry policy required.
There is natural support for versioning and file system snapshots.
CernVM-FS uses HTTP as its transfer protocol and transfers files and file catalogs only on request, i.e. on an open() or stat() call. Once transferred, files and file catalogs are locally cached.





CernVM-FS Infrastructure for ATLAS

The ATLAS experiment maintains three CernVM-FS repositories, for stable software release, for conditions data, and for nightly builds, respectively.
The nightly builds are the daily builds that reflect the current state of the source code, provided by the ATLAS Nightly Build System. Each of these repositories is subject to daily changes.
The Stratum-0 web service provides the entry point to access data via HTTP.
Hourly replication jobs synchronize the Stratum 0 web service with the Stratum 1 web services. Stratum 1 web services are operated at CERN, BNL, RAL, FermiLab, and ASGC LHC Tier 1 centers.
CernVM-FS clients can connect to any of the Stratum 1 servers. They automatically switch to another Stratum 1 service in case of failures. For computer clusters, CernVM-FS clients connect through a cluster-local web proxy. In case of ATLAS, CernVM-FS mostly piggybacks on the already deployed network of Squid caches of the Frontier service

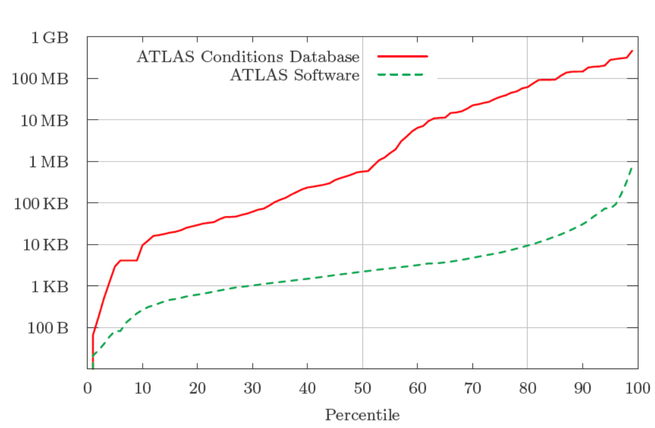


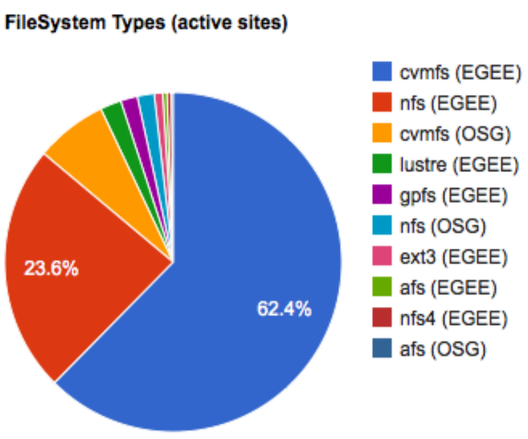
Nightly Releases

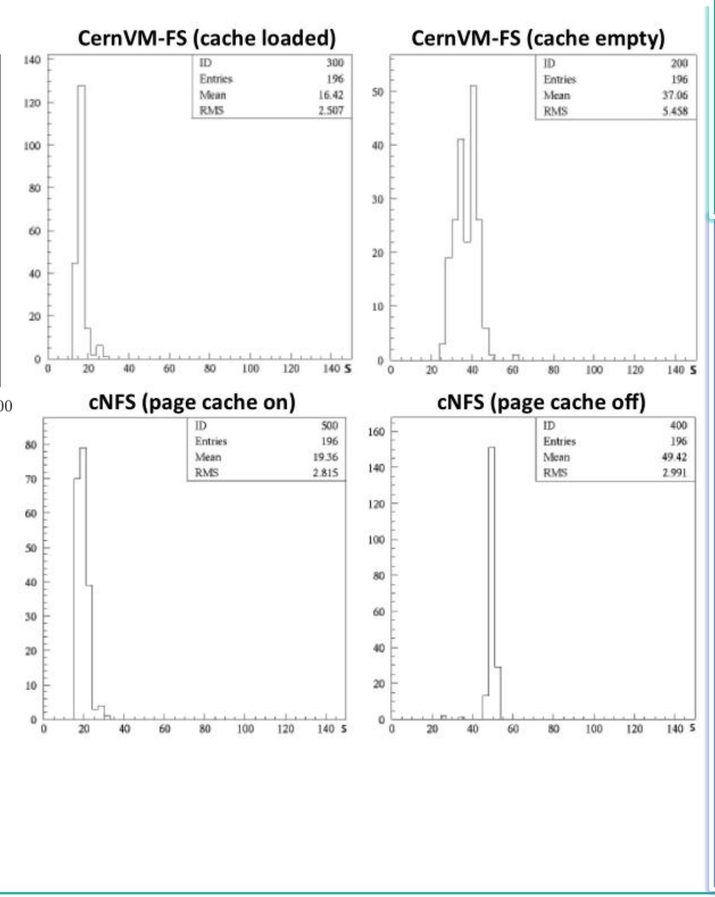
The ATLAS Nightly Build System maintains about 50 branches of the multi-platform releases based on the recent versions of software packages.
Code developers use nightly releases for testing new packages, verification of patches to existing software, and migration to new platforms, compilers, and external software versions.
ATLAS nightly CVMFS server provides about 15 widely used nightly branches.
Each branch is represented by 7 latest nightly releases. CVMFS installations are fully automatic and tightly synchronized with Nightly Build System processes. The most recent nightly releases are uploaded daily in few hours after a build completion.
The status of CVMFS installation is displayed on the Nightly System web pages .

Conditions data

ATLAS conditions data refers to the time varying non-event' data needed to operate and debug the ATLAS detector, perform reconstruction of event data and subsequent analysis. The conditions data contains all sorts of slowly evolving data including detector alignment, calibration, monitoring, and slow controls data. Most of the conditions data is stored within an Oracle database at CERN and accessed remotely using the Frontier service. Some of the data is stored in external files due to the complexity and size of the data objects.
CernVM-FS provides an ideal solution for distributing the files that would other wise have to be distributed through the ATLAS distributed data management system.
CernVM-FS allows the large computing sites simplification by reducing the need for specialized storage areas.







Changes to CernVM-FS repository are staged on an "installation box" using a read/write file system interface.
There is a dedicated installation box for each repository. CernVM-FS tools installed on these machines are used to commit changes, process new and updated files, and to publish new repository snapshots.
The machines act as a pure interface to the master storage and can be safely re-installed if required.

Installed items:

1. Stable Software Releases;
2. Nightly builds;
3. Local settings;
4. Software Tools
5. Conditions Data

Stable Software Releases

The deployment is performed via the same installation agents used in the standard Grid sites.
The installation agent is currently manually run by a release manager whenever a new release is available, but can be already automatized by running periodically.
The installation parameters are kept in the Installation DB, and dynamically queried by the deployment agent when started. The Installation DB is part of the ATLAS Installation System, and updated by the Grid installation team.
Once the releases are installed in CernVMFS, a validation job is sent to each site. For each validated releases a tag is added to the site resources, meaning they are suitable to be used for with the given release.
The Installation System can handle both CernVMFS and non-CernVMFS sites, transparently.

Software Tools

The availability of Tier3 software tools from CernVM-FS is invaluable for the Physicist end-user to access Grid Resources and for analysis activities on locally managed resources, which can range from a grid-enabled Data Center to a user desktop or a CernVM Virtual Machine.
The software, comprising multiple versions of Grid Middleware, front-end Grid job submission tools, analysis frameworks, compilers, data management software and a User Interface with diagnostics tools, are first tested in an integrated environment and then installed on the CernVM-FS server by the same mechanism that allows local disk installation on non-privileged Tier3 accounts.
The User Interface integrates the Tier3 software with the Stable and Nightly ATLAS releases and the Conditions Data, the latter two residing on different CernVM-FS repositories.

Local Settings

The local parameters needed before running jobs, ATLAS currently uses an hybrid system, where the generic values are kept in CernVM-FS, with a possible override at the site level from a location pointed by an environment variable, defined by the site administrators to point to a local shared disk of the cluster. Periodic installation jobs write the per-site configuration in the shared local area of the sites.
The fully enabled CernVM-FS option, where the local parameters are dynamically retrieved by the ATLAS Grid Information System (AGIS) is being tested.

Performance

Some measurements on CernVM-FS access times against the most used file systems in ATLAS have shown that the performance of CernVM-FS is either comparable or better performing when the system caches are filled with pre-loaded data.
The results of a metadata-intensive test performed at CNAF, Italy, show the comparison of performance between CernVM-FS and cNFS over gpfs . The tests are equivalent to the operations that are performed by each ATLAS job in real life. The average access times with CernVM-FS are ~16s when the cache is loaded and ~37s when the cache is empty, to be compared with the cNFS values of ~19s and ~49s respectively with page cache on and off.
Therefore CernVM-FS is generally even faster than standard distributed file systems.
According to other measurements, CVMFS also performs consistently better than AFS in wide area networks with the additional bonus that the system can work in disconnected Mode.

[1] Istituto Nazionale di Fisica Nucleare, Sezione di Roma1; [2] TRIUMF (CA); [3] Duke University (US); [4] CERN (CH); [5] Brookhaven National Laboratory (US); [6] LBNL (US)