CMS Experience with a World-Wide Data Federation

For the CMS collaboration: Ken Bloom[1], Tommaso Boccali[2], Brian Bockelman[1], Dan Bradley[3], Sridhara Dasu[3], Federica Fanzago[4], Igor Sfiligoi[5], Matevž Tadel[5], Carl Vuosalo[3], Frank Würthwein[5], Marian Zvada[1]

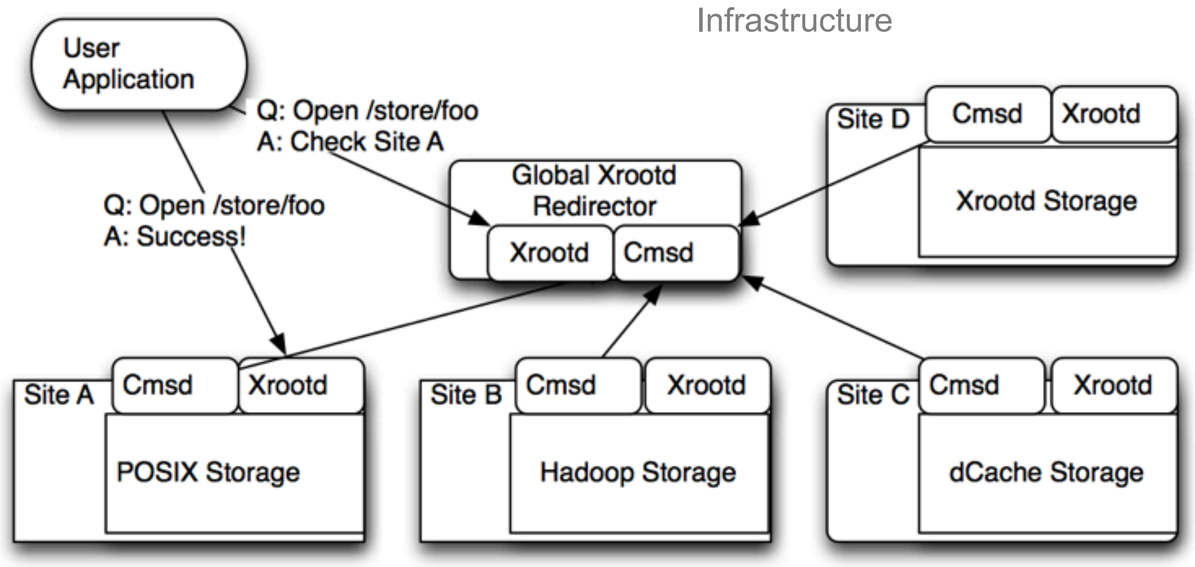
Major goals

Project focuses on development and deployment of storage resources that are accessible for **any data**, **anywhere** at **anytime**. We focus on CMS use case as a test bed towards full production integration into the CMS experiment, but keep general deployment for any application as our goal (especially for other high energy physics organizations). For CMS specifically, the goal is to **enable easier**, **faster** and more **efficient** processing and analysis of data recorded at the Large Hadron Collider (LHC) through more flexible use of computing resources anywhere, whether owned by CMS or opportunistically available through cloud-like facilities.

As **LHC Run 2** begins, the **AAA** data federation is **fully integrated** and ready for the influx of new data. CMS physicists use it automatically and transparently for location-independent processing of the full range of tasks from the event display of a particular collision to the large scale skimming and selection of billions of events. The success of the "Computing, Software and Analysis" (CSA14) challenge shows that CMS users are embracing the system and will rely on it for the rapid production of LHC physics results.

Major activities

The AAA Federated Storage Infrastructure



We have been deploying an XRootD infrastructure that spans all of the Tier-1, Tier-2 (and few Tier-3) sites in CMS (~56 sites). Each site's XRootD server is interfaced with the local storage system, allowing it to export the CMS namespace. Current **storage systems** include sites with **dCache** (22), **DPM** (18), **Hadoop** (7), **StoRm** (6), **Lustre** (2) and **Castor** (1). The site servers subscribe to a local redirector; the local redirectors from each site then subscribe to a redundant US regional redirector hosted by FNAL and UNL. A similar topology of servers and redirectors exists in Europe (Bari, LLR, Pisa). This forms a large tree-structured federated storage infrastructure that overlays on top of existing storage systems, allowing users access to any on-disk data without knowing its location.

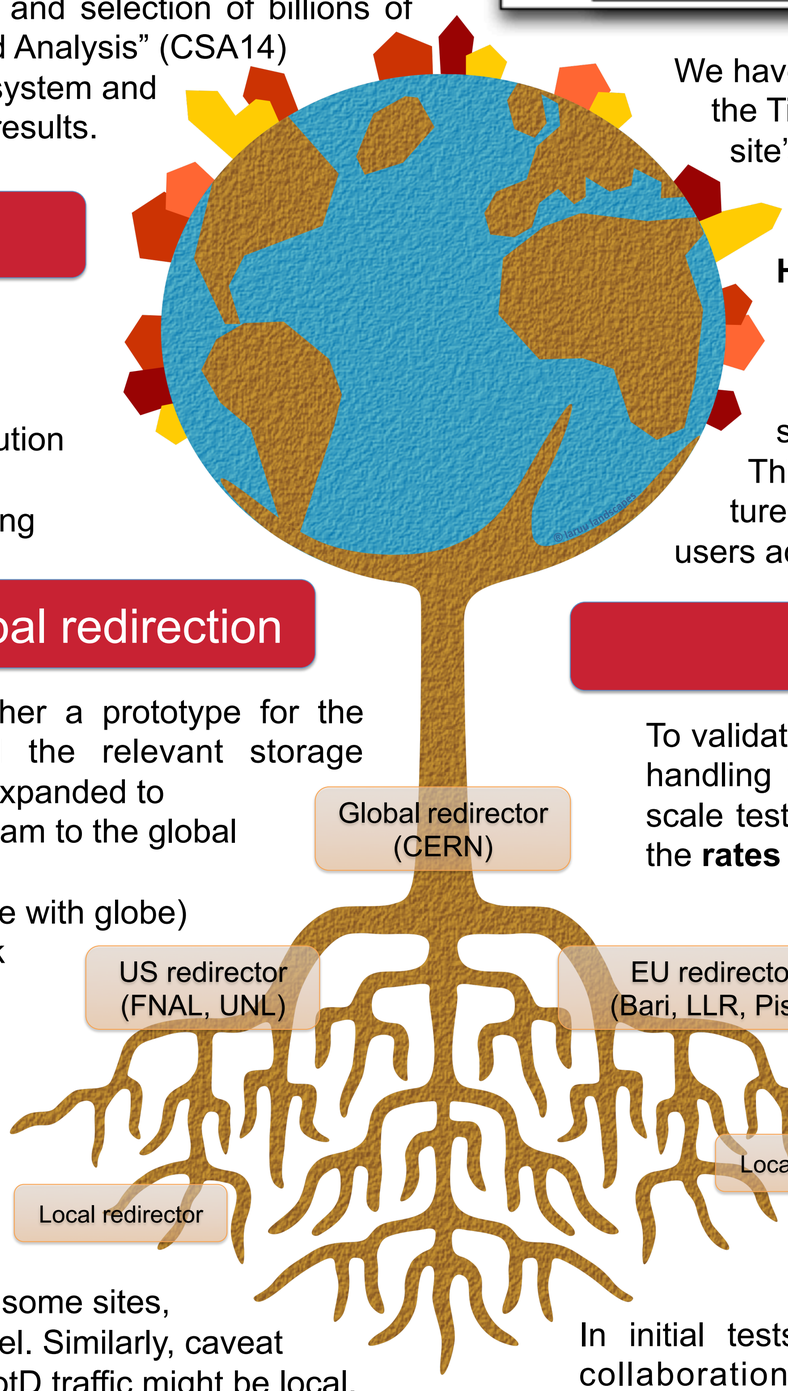
Core technologies

- **XRootdD** for federating storage
- **HTCondor** for job management
- CernVM File System (**CVMFS**) with **Parrot** Virtual File System for software distribution
- **Gratia** for resource accounting
- **MonALISA** and **CMS dashboard** for monitoring

Architecture: local, regional and global redirection

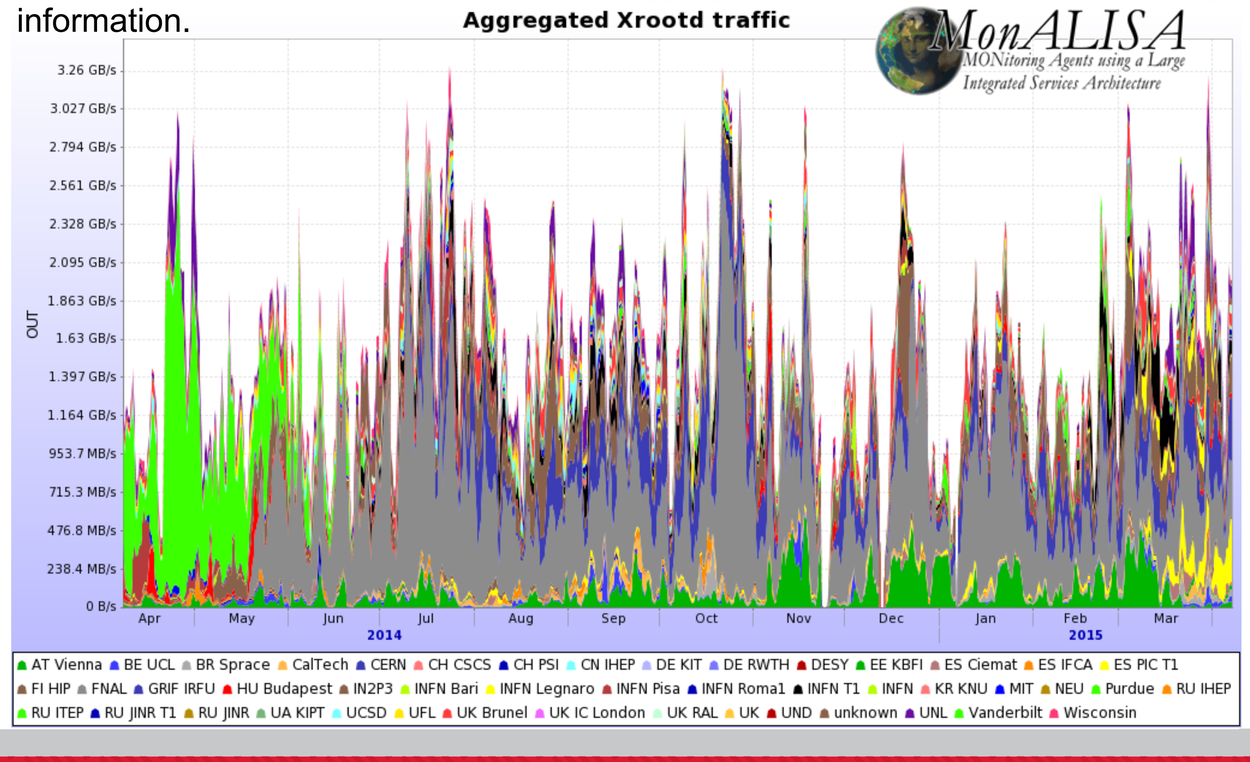
To explore the XRootD architecture, we put together a prototype for the WLCG, involving CMS sites worldwide and all the relevant storage technologies. This prototype has been success and expanded to a regional redirector-based system subscribed upstream to the global redirector as peak of the redirectors hierarchy.

This injects another layer into the hierarchy (see figure with globe) which will make sure requests keep in a local network region if possible or send upstream otherwise.



Infrastructure and monitoring

To give a sense of scale, bottom right dashboard figure shows a month-long graph of daily average transfer rates, sorted by source site. Note, that this graph includes both internal and external transfers at some sites, and daily volumes depend heavily on user activity level. Similarly, caveat is present in the summary monitoring that some XRootD traffic might be local. The summary monitoring and MonALISA are used as the only aggregators for all CMS summary monitoring. Parts of summary monitoring functionality are being replicated by the CMS dashboard by aggregation of the detailed monitoring information.

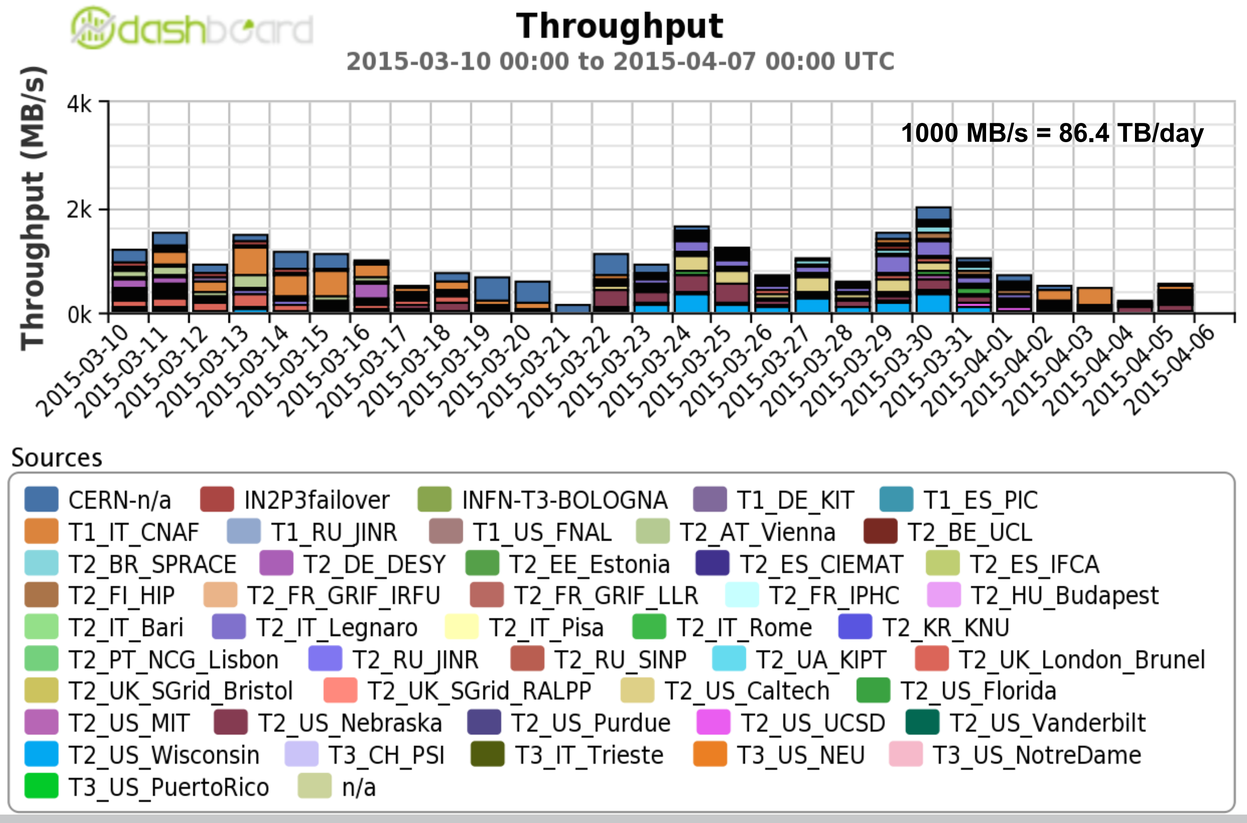


Scale testing

To validate that the AAA component of the infrastructure is capable of handling the expected traffic required to process the new data, a scale testing campaign has been undertaken. These **tests measure** the **rates** that **files** at a CMS site can be **opened** and **read**.

- Minimum benchmarks were determined based upon a historical study of CMS jobs.
  - An average job opens a new file once per 1000 seconds and reads from a file at an average rate 0.25MB/s.
  - Assuming a worst case of 100,000 jobs opening files at a site at once gives a benchmark of 100Hz for file-opening rates.
  - For file reading, an assumption of 600 simultaneous jobs actively reading files from a site gives a total rate of 150MB/s.

In initial tests, many sites failed to achieve these benchmarks, but collaboration with site administrators and optimization of site configurations improved performance to match the benchmarks.



[1] University of Nebraska-Lincoln, USA
[2] INFN Sezione di Pisa, ITALY
[3] University of Wisconsin-Madison, USA
[4] INFN Sezione di Padova, ITALY
[5] University of California-San Diego, USA