Improving CMS data transfers among its distributed computing facilities

J Flix\(^1,2\), N Magini\(^3\) and A Sartirana\(^4\)

\(^1\) Port d’Informació Científica (PIC), Bellaterra, Spain
\(^2\) Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain
\(^3\) CERN, Geneva, Switzerland
\(^4\) École Polytechnique, Palaiseau, France

E-mail: jflix@pic.es, Nicolo.Magini@cern.ch, sartiran@llr.in2p3.fr

Abstract. CMS computing needs reliable, stable and fast connections among multi-tiered computing infrastructures. For data distribution, the CMS experiment relies on a data placement and transfer system, PhEDEx, managing replication operations at each site in the distribution network. PhEDEx uses the File Transfer Service (FTS), a low level data movement service responsible for moving sets of files from one site to another, while allowing participating sites to control the network resource usage. FTS servers are provided by Tier-0 and Tier-1 centres and are used by all computing sites in CMS, according to the established policy. FTS needs to be set up according to the Grid site’s policies, and properly configured to satisfy the requirements of all Virtual Organizations making use of the Grid resources at the site. Managing the service efficiently requires good knowledge of the CMS needs for all kinds of transfer workflows. This contribution deals with a revision of FTS servers used by CMS, collecting statistics on their usage, customizing the topologies and improving their setup in order to keep CMS transferring data at the desired levels in a reliable and robust way.

1. Introduction

The CMS Experiment [1] is one of 4 large particle physics experiments that have been recording proton-proton collisions at the centre-of-mass energy of 7 TeV since March 2010 (and lead ion-lead ion collisions at the centre-of-mass energy of 2.76 TeV per nucleon pair since November 2010) at the LHC accelerator at CERN, Geneva, Switzerland. To archive and analyse its data, CMS and the other LHC experiments depend on the Worldwide LHC Computing Grid (WLCG) [2], a worldwide distributed data grid of over 150 compute and storage clusters varying in size, from tens of TBs to several PBs. The CMS computing model [3] has three tiers of computing facilities:

- Tier-0 at CERN (T0), used for prompt processing, tape backup and export to Tier-1 centres of CMS data.
- 7 Tier-1 (T1) centres, used for the tape backup and large-scale reprocessing of CMS data, and distribution of data products to the Tier-2 centres. The Tier-1 centres are typically at national laboratories with large computing facilities and archival storage systems.
- 50 Tier-2 (T2) facilities, where data analysis and Monte Carlo production are primarily carried out. These centres are typically at universities and do not have tape backup systems.
only disk storage. Each Tier-2 is “associated” for support to a Tier-1, usually on the basis of geographical proximity, but is not restricted to sharing data with this Tier-1 centre.

These sites are interconnected by high-speed networks of 1-10 Gbps and the CMS computing model envisions commissioning all links between all of these sites for production data transfers [4].

2. PhEDEx

In order to meet the CMS data distribution requirements, the Physics Experiment Data Export (PhEDEx) [5] project was designed to facilitate and manage global data transfers over the grid.

PhEDEx is based on a high-availability Oracle database cluster hosted at CERN (Transfer Management Data Base, or TMDB) acting as a “blackboard” for the system state (data location and current tasks). PhEDEx software daemon processes or “agents” connect to the central database to retrieve their work queue, and write back to TMDB the result of their actions. Central agents running at CERN perform data routing and transfer task creation. The download agents running at the sites receive these tasks and initiate the data transfers using technology-specific backends interacting with the WLCG middleware.

Finally, PhEDEx provides two interfaces for data operations management, and for transfer monitoring: an interactive web site, and a web Data Service to retrieve information from TMDB in machine-readable formats.

Since the beginning of the 2010 LHC physics run in March, CMS has been steadily transferring data with PhEDEx with peaks in global speed between all sites exceeding 2.8 GB/s, distributing 21 PB of replicas over all sites [6].

3. FTS

The core data-movement engines of PhEDEx are the “backend” modules of the file download agent. The backend used for transfers between most sites on the WLCG is the FTS backend, which is responsible for splitting the file transfer queue into discrete chunks that are submitted as transfer jobs to the File Transfer Service (FTS) [7], periodically checking the status of those transfer jobs in FTS, and uploading the results to the PhEDEx database.

The File Transfer Service, developed as part of the gLite middleware [8], is a data movement service for transferring files between storage elements. It was designed to balance resource usage at sites used by multiple experiment Virtual Organizations (VOs), prevent network or storage overload, and enforce prioritization in transfer jobs.

FTS exposes a web service interface to submit asynchronous bulk requests for parallel file transfers (transfer jobs). Transfer requests are stored in an Oracle database and assigned to unidirectional queues, called “channels”. A channel is not tied to a physical network path but is defined between endpoints, where each endpoint might be:

- a site
- a group of sites (“cloud”)
- a catch-all endpoint for all sites (“star”).

The configuration of each channel defines the resource restrictions such as the maximum number and relative share of transfers for each of the VOs using the channel.

FTS interacts with storage elements through the Storage Resource Manager (SRM) [9] interface, a middleware service providing uniform transparent access to storage management capabilities irrespective of the underlying technology. The File Transfer Agents for each FTS channel can delegate transfers to one of the SRM servers involved as an srmCopy request, or execute a third-party transfer using the GridFTP [10] transfer protocol, an extension of the FTP protocol supporting secure Grid authentication and multiple parallel streams. Protocol
Figure 1. Tier-1 to Tier-1 average transfer rates per stream, with RMS as error bars. Statistics over 30 days up to 05/10/2010. The number of transfers in the statistics is reported above the error bar.

and protocol parameters, such as timeouts or number of simultaneous GridFTP streams, are set in the channel configuration. The resulting throughput on a site-to-site channel is the transfer rate for a GridFTP stream, times the number of GridFTP streams per file, times the number of files in active transfer on the channel.

The FTS servers are deployed at the Tier-0 at CERN and at each of the Tier-1 sites. The CERN FTS server manages transfers on dedicated channels between the Tier-0 and the Tier-1s. The FTS server at each Tier-1 centre is typically configured to manage:

- Imports on dedicated channels from other Tier-1s
- Exports and imports on dedicated channels to and from the associated Tier-2s
- Exports and imports on shared channels (“cloud” and “star”) to and from the non-associated Tier-2s
- Third-party transfers to the associated Tier-2s from other Tier-2s on shared channels.

4. The FTS Monitor
The FTS Monitor [11] is a web-based monitoring system developed at the CC-IN2P3 Tier-1 providing a graphical view of the FTS activity. The service retrieves data directly from the FTS backend database to generate summary statistics and to provide detailed reports about transfer activities. FTS Monitor web pages display channel configuration, statistics about transfers in the last 14 days on each channel, and detailed information on all jobs submitted in the last 24 hours, including the status and throughput of each individual file transfer. The file-level details are also published in machine-readable XML format. The FTS Monitor has been deployed at 6 of the Tier-1s supporting CMS, with CC-IN2P3 running the latest version.

5. FTS Monitor Parser
The database of each FTS server contains detailed information about all transfers performed in all provided channels: transfer rates per file and per stream, SRM response times, etc., a wealth of information that can be extremely useful to spot issues and debug problems. As
mentioned, these data are exposed in full detail on the FTS Monitor pages which are of great help to site administrators and transfer experts for the investigation of ongoing issues. However, the FTS Monitor view is specific of each FTS server and it does not offer historical data, while having a global and historical view may be of extreme importance in spotting out structural or optimization problems which are not point-wise in time and do not concern a single FTS server. The FTS Monitor Parser answers this need by extracting full daily statistics from the FTS Monitors around the world, collecting them in a repository and digesting them into meaningful plots and summaries.

Figure 1 and figure 2 show examples of how FTS Monitor Parser can provide global (i.e. multi-FTS) and historical views of transfer performances. Figure 1 plots the average rate per stream for all channels between Tier-1 sites. Figure 2 plots the daily average rate per stream in the transfers from Tier-1 CC-IN2P3 to Tier-1 PIC. All plots are based on data collected over 30 days up to 05/10/2010 and considering only CMS transfers. Data is collected from the FTS Monitors at 5 of the 7 Tier-1 centres supporting CMS, namely PIC, INFN-CNAF, KIT, RAL and ASGC.

6. Use of FTS statistics in transfer operations
In the following we give some examples of usage in operations of the information provided by the FTS Monitor Parser.

6.1. FTS channel optimization
One important issue that can be addressed by exploiting the data provided by the FTS Monitor Parser is the identification of slow links in the FTS common channels, for example those which provide a common queue for all transfers from a given Tier-1 to all non-regional Tier-2 sites. The corresponding links may have very different performances (as measured by rate per stream), and transfers in slow links keep the channel slots busy for a longer time, increasing the queuing time for all others. Therefore, in order to optimize transfer submission, it is useful to identify the slow and fast links and assign them to two or more separate cloud channels. Such optimization has already been performed at the PIC Tier-1. As shown in figure 3, the rate per stream in PIC exports to non-regional Tier-2s may vary from few kB/s up to 5 MB/s. On the basis of these data the PIC administrators took the links below a given threshold and assigned them to
**Figure 3.** Average transfer rates per stream from the PIC Tier-1 to Tier-2s, with RMS as error bars. Statistics over 30 days up to 05/10/2010. The number of transfers in the statistics is reported above the error bar.

**Figure 4.** Daily rate plot for PIC Tier-1 exports to Tier-2s. 7 days up to 07/10/2010. The change in FTS configuration was applied on 06/10/2010.

A different channel, freeing more transfer slots for the fast moving transfers. Figure 4 shows the increase in overall transfer rate as effect of the change.

**6.2. Identifying infrastructural issues**

Another important use case is the possibility to spot out site or infrastructural issues exploiting the statistics over a long time period. An example in this sense can be seen looking at the plot of figure 5 which reports the ratio between the export and import rates between the PIC Tier-1 and other sites. This ratio is lower than 1 for most of the sites, i.e. PIC export rates are almost always systematically lower than the import rates. This fact is general and independent from the distance which seems to suggest a structural issue. One possible explanation is a known limitation of the Solaris kernel used on the disk servers at PIC which penalizes the outgoing transfers when the other endpoint runs a Scientific Linux 5 kernel. This issue has been brought to the attention of the PIC site administrators and still requires investigation and dedicated testing.
7. Outlook

At present, the FTS Monitor Parser project has developed the basic tools and tested the first use cases. Some of the results have been reported in this paper. However, more work is required in order to exploit the full potential of this tool in terms of analysis of the distributed data transfer infrastructure. First of all, as mentioned, at the moment FTS Monitor Parser does not gather data from all FTS servers. We are planning to include all servers with the next version of the tool so that the statistics may be complete. We are also identifying the relevant statistics to be computed and published in a complete and meaningful set of plots. These will be presented in a daily updated web page and will be at disposal of the expert teams as well as of the site administrators. With the help of such statistics the central team of distributed data transfers of CMS is planning to start a campaign to spot out issues and address them to the sites. Regular usage in operations will, of course, give more feedback on which are the most relevant statistics to gather. Some new statistics are already planned to be included, e.g. distribution by file size, transfer preparation time, FTS queuing times, etc. Moreover, at the moment FTS Monitor Parser is gathering data only for the CMS VO, but in the future it can be opened to other VOs.

Acknowledgments

We would like to thank Lionel Schwarz from CC-IN2P3 for prompt support in fixing issues and including new features in the FTS Monitor. We would also like to thank the local contacts and FTS server administrators at all Tier-1 centres for the deployment of the FTS Monitor at their sites and for the feedback on operating the service.

References

[1] The CMS Collaboration 2008 The CMS experiment at the CERN LHC JINST 3 S08004
[2] Knoblock J et al. 2005 LHC Computing Grid Technical Design Report CERN-LHCC-2005-024
[3] Bonacorsi D 2007 The CMS computing model Nucl. Phys. B (Proc. Suppl.) 172 53-56
[4] Letts J 2010 Large scale commissioning and operational experience with Tier-2 to Tier-2 data transfer links in CMS these proceedings CMS CR-2010/224
[5] Egeland R et al. 2008 Data transfer infrastructure for CMS data taking PoS(ACAT08)033
[6] Rossman P Distributed data transfers in CMS these proceedings
[7] Frohner A et al. 2010 Data management in EGEE J. Phys.: Conf. Ser. 219 062012
[8] gLite project website: http://glite.web.cern.ch/glite/
[9] Abadie L et al. 2007 Storage Resource Managers: Recent international experience on requirements and multiple co-operating implementations 24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007) 47–59
[10] Allcock B et al. 2002 Data management and transfer in high-performance computational grid environments Parallel Computing 28 749-771
[11] FTS Monitor project website: https://forge.in2p3.fr/projects/show/ftsmonitor