Performance studies and improvements of CMS Distributed Data Transfers

Jose Flix Molina for the CMS Collaboration

Abstract

CMS computing needs reliable, stable and fast connections among multi-tiered computing infrastructures. CMS experiment relies on File Transfer Services (FTS) for data distribution, 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 centers and used by all the computing sites in CMS, subject to established CMS and sites setup policies, including all the virtual organizations making use of the Grid resources at the site, and properly dimensioned to satisfy all the requirements for them. Managing the service efficiently needs good knowledge of the CMS needs for all kind of transfer routes, and the sharing and interference with other Virtual Organizations using the same FTS transfer managers. This contribution deals with a complete revision of all FTS servers used by CMS, customizing the topologies and improving their setup in order to keep CMS transferring data to the desired levels in a reliable and robust way, as well as complete performance studies for all kind of transfer routes, including overheads measurements introduced by SRM servers and storage systems, FTS server misconfigurations and identification of congested channels, historical transfer throughputs per stream for site-to-site data transfer comparisons, file-latency studies, among others... This information is retrieved directly from the FTS servers through the FTS Monitor webpages and conveniently archived for further analysis. The project provides a monitoring interface for all these values. Measurements, problems and improvements in CMS sites connected to LHCOPN are shown, where differences up to x100 are visible, constant performance measurements of data flowing from Tier-0 to Tier-1s, comparison to other existing monitoring tools (PerfSonar, LHCOPN dashboard), as well as the usage of the graphical interface to understand, among others, the effects for sites when connecting to LHCONE network. Given the multi-VO added value of this tool, this work is serving as a reference for building up the WLCG FTS monitoring tool, which will be based on the FTS messaging system.

Presented at CHEP 2012: International Conference on Computing in High Energy and Nuclear Physics
Performance studies and improvements of CMS distributed data transfers

D Bonacorsi¹, J Flix²,³, R Kaselis⁴, J Letts⁵, N Magini⁴, A Sartirana⁶

¹ University of Bologna, INFN, Italy
² Port d’Informació Científica (PIC), Spain
³ Centro de Investigaciones Medioambientales y Energéticas (CIEMAT), Spain
⁴ CERN, Switzerland
⁵ University of California San Diego, La Jolla, CA, USA
⁶ LLR/E.Polytechnique, Paris, France

E-mail: jflix@pic.es

Abstract. CMS computing needs reliable, stable and fast connections among multi-tiered distributed infrastructures. CMS experiment relies on File Transfer Services (FTS) for data distribution, 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 centers and used by all the computing sites in CMS, subject to established CMS and sites setup policies, including all the virtual organizations making use of the Grid resources at the site, and properly dimensioned to satisfy all the requirements for them. Managing the service efficiently needs good knowledge of the CMS needs for all kind of transfer routes, and the sharing and interference with other VOs using the same FTS transfer managers. This contribution deals with a complete revision of all FTS servers used by CMS, customizing the topologies and improving their setup in order to keep CMS transferring data to the desired levels, as well as performance studies for all kind of transfer routes, including overheads measurements introduced by SRM servers and storage systems, FTS server misconfigurations and identification of congested channels, historical transfer throughputs per stream, file-latency studies,... This information is retrieved directly from the FTS servers through the FTS Monitor webpages and conveniently archived for further analysis. The project provides an interface for all these values, to ease the analysis of the data.

1. Introduction

The Large Hadron Collider (LHC), located at CERN, became operational in spring 2010 with proton-proton collisions at centre-of-mass energy of 7 TeV, increasing up to 8 TeV in 2012. In 2011, the collected data was about two orders of magnitude compared to 2010, and in April 2011 the instantaneous luminosity went beyond the maximum ever obtained at that date by any hadron collider. Additionally, Pb-Pb collisions were as well recorded during these two years of operations. The Compact Muon Solenoid (CMS) [1] is one of the four detectors that observes the collisions and it generates PetaBytes of data per year. To scientifically exploit the data, the data processing requires the use of computing and storage resources from several centres outside CERN. These are in fact coordinated by the Worldwide LHC Computing Grid (WLCG) [2], which at most sites exploits the
computing infrastructure provided by other Grid projects, like EGI, Open Science Grid and NorduGrid, a worldwide distributed data grid of over 150 compute and storage clusters varying in size, from tens of TBs to several PBs.

In the case of the CMS collaboration, around 60 sites from about 20 countries are involved, ~100 sites if considering Tier-3 facilities. They are organized with a tiered structure, where different tier levels correspond to different functions:

- A Tier-0 at CERN, which takes care of the prompt event reconstruction and detector calibration, the distribution of raw and processed data to external sites and the backup storage of the raw data.
- 7 Tier-1 sites run the subsequent reprocessing, including data skimming, keep an active copy of the raw data and store the Monte Carlo (MC) generated at Tier-2 sites. The Tier-1 centres are typically at national laboratories with large computing facilities and archival storage systems.
- ~50 Tier-2 sites get samples of the skimmed data for analysis and are used to run the MC simulation. 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.

A complete description of the CMS computing model and its services can be found elsewhere [3]. 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. Given the complexity of the data transfer infrastructure, it is important to measure its performance in a continuous way to identify and correct problems. CMS has developed a tool intended to provide with all necessary monitoring to analyze and improve the data transfer connections among the distributed computing sites. The following sections describe the CMS data management tool, the use of the File Transfer System (FTS), its monitoring to understand the performance of data transfers, and how this understanding has helped improving the overall data transfer picture in CMS.

2. CMS Data Management tool: PhEDEx

In order to meet the CMS data distribution requirements, the Physics Experiment Data Export (PhEDEx) [4] 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 4.5 GB/s, in an hourly bin, distributing 65 PB of replicas over all sites [5]. Figure 1 shows the typical 2011 values in the CMS data transfer topology, for different transfer routes among the tiered structured, excluding Tier-3s.

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) [6], periodically checking the status of those transfer jobs in FTS, and uploading the results to the PhEDEx database.

Figure 1. Data Transfers in the CMS topology (Tier-3s excluded). Typical 2011 values are shown.

3. File Transfer Service (FTS)

The File Transfer Service, developed as part of the gLite middleware [7], 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 tasks.

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) [8] interface, a middleware service providing uniform transparent access to storage management capabilities irregardless 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 [9] transfer protocol, an extension of the FTP protocol supporting secure Grid authentication and multiple parallel streams. Protocol 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/imports on dedicated channels to and from the associated Tier-2s;
- exports/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

Although FTS comes with an integrated monitor for FTS admins, a more convenient web-based monitoring system was developed at the CC-IN2P3 Tier-1, the FTS Monitor [10], providing a graphical view of the FTS activity, which was soon accepted and deployed by many sites running FTS servers. 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. For CMS, the FTS Monitor has been deployed at 6 of the Tier-1s supporting CMS, and the Tier-0 at CERN.

Figure 2. A view of the FTS Monitor showing details of a finished transfer in a FTS job.
5. The FTS Monitor Parser: a CMS tool for data transfer performance studies

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 [11] answers this need by extracting full daily statistics from the FTS Monitors around the world, collecting them in a dedicated DDBB and digesting them into meaningful plots and summaries. The tool is collecting data since mid 2010, and the graphical user interface was further improved by mid 2011.

The next figures show examples of how FTS Monitor Parser can provide global (i.e. multi-FTS) and historical views of transfer performances. Figure 3 shows the average rate per stream for all channels between Tier-1 sites from 1\textsuperscript{st} January 2012 to 10\textsuperscript{th} May 2012. The numbers point to the total transfers analyzed, and the mean and rms are shown. All these sites connect to the same network infrastructure, the LHCOPN, and differences of more than one order of magnitude are seen, on average, in this time period. Figure 4 plots the daily average rate per stream in the transfers from PIC Tier-1 to other CMS Tier-1s (mean and rms plotted). Figure 5 shows the SRM overheads in preparing/releasing transferred files, both on source and destination, for transfers from PIC Tier-1 to CNAF Tier-1, from 1\textsuperscript{st} January to 31\textsuperscript{st} March 2012. One can clearly see that, in average, CNAF SRM server dominates the overheads on PIC to CNAF transfers. All plots are based on data collected considering only CMS transfers from the FTS Monitors at 6 of the 7 Tier-1 centres supporting CMS, namely PIC, IN2P3, INFN-CNAF, KIT, RAL and ASGC.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{ftsmonitor_ranking.png}
\caption{Tier-1 to Tier-1 average transfer rates per stream, from 1/1/2012 to 11/5/2012 (mean and rms are displayed, as well as the number of files transferred in the period).}
\end{figure}
Figure 4. Daily average rate per stream from PIC Tier-1 to other CMS Tier-1s, from 1/1/2012 to 11/5/2012 (*mean* and *rms* are displayed).

Figure 5. SRM overheads in preparing/releasing transferred files, both on source and destination, from PIC Tier-1 to CNAF Tier-1, from 1st January to 31st March 2012.
6. Use of FTS statistics in transfer operation

In the following sections 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 FTS 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 several CMS Tier-1s, for which exports to non-regional Tier-2s may vary from few kB/s/stream up to 5 MB/s/stream. After the separation of slow-speed and high-speed links into two separate channels, for all of the cases there was an increase, a factor 2-4, in overall transfer rate as effect of the FTS configuration change.

Also, the FTS Monitor Parser knows at any time when the files have been submitted to the FTS servers, when the transfers indeed start and finish. With this information, file latency studies can be performed, and in particular one can detect congested FTS channels, i.e. channels that are not able to process the files in the queue, due to a misconfiguration of the channel (not too many files in parallel allowed), or due to slow-throughput performance of the transfers on the channel, or due to a saturation on many sites using cloud or star channels. This monitoring is extremely useful for site admins and data transfer experts who want to optimize file latencies.

Figure 6 shows an example: the channel occupancy on CERN-FNAL channel on CERN FTS server, which manages transfers from CERN Tier-0 to FNAL Tier-1. All queued transfers are rapidly digested and processed (the maximum number of parallel transfers is set to 271 in the channel). Note how the transfers queue during an FTS upgrade, how PhEDEx does not submit more transfers, until the upgrade finishes, the queue is processed and new transfers are submitted to CERN FTS server.

![CERN-FNAL Ch. (FTS PILOT)](image)

**Figure 6.** CERN-FNAL FTS channel occupancy on CERN FTS server.
Not intrinsic to FTS, but important for the use of FTS servers, is related to PhEDEx configuration. The measured rates in a link might be low due a misconfiguration on PhEDEx agents at the destination sites, which are the responsible to submit the tasks to the corresponding FTS servers. This normally is seen in the FTS Monitor Parser as underused channels (not filled, or filled in bunches) or low transfer attempts to FTS servers. The FTS Monitor Parser tool helps identifying all these cases, and letting sites and experts to know, so they can better tune increasing the number of parallel files from some sources on PhEDEx agents configuration, or number of files in the queue for fast digesting FTS channels, or reducing the configuration time to get new transfers from central agents, for very fast transfers links,... An example will be shown in section 6.3.

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. Some of these problems might be related to problematic disk pools or problematic site switches, either on the source or destination, problems on the intermediate *on-the-fly* routers managing the actual transfer, or any misconfiguration of these components, which might degrade or even block data transfers between a source and a destination site.

Some examples have been detected with CMS since we started looking into the data:

- systematical *asymmetries* between export and import rates for a given computing site. It was detected a limitation of the Solaris kernel used on some disk servers which penalized the outgoing transfers when the other endpoint runs a Scientific Linux 5 kernel [11].
- identification of slow links to distant regions and optimizing the *TCP/IP parameters* to increase the performance on the data transfer connections;
- habilitation of *fumbo frames* (Ethernet frames with more than 1500 bytes of payload) on sites, which can carry up to 9000 bytes of payload. This however implies the covariant scaling of numerous intermediating logic circuits along the network path to accommodate the maximum transmission unit (MTU) required. Sometimes this is not supported by Internet service providers, or changed by misconfigurations, and it reflects in poor performing transfers (routers reconverting to MTU 1500) or blocked transfers if intermediate routers cannot handle frames of MTU 9000.

![FTS Monitor History Plot](image)

**Figure 7.** Transfers from IN2P3 Tier-1 to US Tier-2s affected by a RENATER-GEANT misconfiguration.
Figure 7 shows an example of a RENATER-GEANT misconfiguration, related to the Less than Best Effort (LBE) service which is deployed on GEANT: “a very small percentage of network capacity is allocated to LBE so that, under congestion, the BE traffic and any higher priority traffic classes are protected from LBE traffic”. A misconfiguration applied on the routers made all transfers from IN2P3 Tier-1 to US Tier-2 sites to be LBE tagged, and hence having very low priority and reduced bandwidth, even having high amount of packets lost, compared to other transfers. This problem was affecting the US transfers for months, however, we only noticed the problem when the site was under heavy load exporting data to many CMS Tier-2s. Once the problem was corrected in GEANT, transfers jumped up into more reasonable levels.

6.3. Understanding transfers from Tier-0 to Tier-1s

A detailed study [12] was performed beginning 2012 to verify that transfers from Tier-0 to Tier-1 sites could cope with the maximum traffic expected during this year 2012 of data taking (1.2 GB/s sustained, distributed among custodial tape pledges). The plan was to achieve the following target rates: KIT (155 MB/s), PIC (66 MB/s), IN2P3 (98 MB/s), CNAF (142 MB/s), RAL (114 MB/s), FNAL (572 MB/s), and ASGC (55 MB/s).

By means of a dedicated LoadTest, this test was made incremental and using an special FTS instance at CERN. The FTS monitoring tool was used to debug the transfers and the evolution of the test: some modifications on remote download PhEDEx agents for fast links were needed to increase the throughputs to the sites. For example, for FNAL Tier-1 the deadtimes on its remote PhEDEx agents were reduced, letting the agent to better get more requests from central PhEDEx agents to send to FTS server, and then optimizing the submission of transfers to FTS, keeping the channel full with some backpressure. Also, changes on rates were observed due to a CASTOR upgrade at CERN, positively for all sites, which pointed to the importance to measure these values continuously when there are major upgrades at the storage level at sites.

![Figure 8. Tier-0→Tier-1 tests: measured transfer rates.](image-url)
Figure 9. Tier-0→Tier-1 tests: measured transfer qualities.

Figure 8 shows the transfer rates from Tier-0 to Tier-1s, and Figure 9 shows the transfer qualities, as a result of the test, which spanned for 1 month and a half, starting on February 2012, and collecting many statistics per link to draw conclusions and help on the customization (of about 100,000 transfers of 2.5 GB files were done on each Tier-0→Tier-1 link). Table 1 shows all the averaged values from the FTS Monitor for the test period. Note that, apart from the differences observed on the transfer rates (MB/s/streams), the SRM overheads introduced on the actual transfers span from a few % up to 30% of the total transfer time for a given file, indicating that there is a window to optimize the SRM servers on some of the sites, reducing this overhead.

Table 1. Tier-0→Tier-1 tests: FTS Monitoring averaged values for the test period.
6.4. Understanding the migration to LHCONE

The LHCONE project aims to provide effective entry points into a network infrastructure that is intended to be private to the LHC Tiers. This infrastructure is built to complement the LHCOPN, but addressing the connection needs of the LHC Tier-2 and Tier-3 sites.

In the context of preparing for LHCONE to go into production, CMS has launched an activity to measure in detail the performance, quality and latency of large-scale data transfers among CMS Tier-2 sites [5], before and after connecting to LHCONE, and the FTS Monitor values are used to provide feedback in the design and commissioning of new transfer infrastructure.

Some load tests were made, still done at the moment, for a selection of CMS Tier-2s, based on their network performance, to check for gains and or losses when connecting through LHCONE. So far, around 175 transfer links between Tier-2s have been tested, measuring the maximum rate, latency for transferring all the files of the test set, and the transfer success rate (or quality) in hourly bins. Additionally, all the values which can be measured via FTS Monitor Parser are as well provided, for each of these individual tests.

Figure 10 shows an example of the values measured in the transfer test from London Imperial College Tier-2 to Indian TIFR Tier-2, before connecting to LHCONE.

![Figure 10. Performance of a Tier-2↔Tier-2 connection, before connecting to LHCONE.](image-url)
7. Outlook

At present, the FTS Monitor Parser project has developed some basic monitoring and historical archive of data which exposed daily by FTS Monitors deployed at Tier-1 sites. The data is presented in a dedicated web page and it is at disposal of the expert teams as well as of the site administrators.

Some of the results have been reported in this paper, and they are encouraging to understand and improve the distributed data transfers between CMS sites, and in particular among any computing sites. Although we do not gather data from all FTS servers, FNAL Tier-1 is missing, the proof of the concept has been demonstrated, as well as its positive use, as the collected statistics were crucial to spot problems on the current framework, results were instrumental to understand the performance of data transfers, they helped improving the overall system, and the tool is used regularly used in Operations to understand data transfer issues.

The FTS Monitor Parser is gathering data only for the CMS VO, but the tool can be used by any other VO. In particular, these type of studies inspired the WLCG Global Transfer Monitoring [13], in which the CMS team cooperates closely with the Dashboard Team, joining efforts to develop and deploy a more powerful common WLCG tool. In particular, the new tool does not use the FTS Monitor anymore, as it catches the information which is sent by each FTS server via dedicated message bus.

These results and the understanding of FTS servers, configurations and optimizations, provided as well feedback to the development of the new FTS (3.0) [14].

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. The authors wish to thank especially the Dashboard Team at CERN, for the fruitful collaboration towards a common tool development for the benefit of the whole WLCG collaborators.

The Port d’Informació Científica (PIC) is maintained through a collaboration between the Generalitat de Catalunya, CIEMAT, IFAE and the Universitat Autònoma de Barcelona. This work was supported in part by grants FPA2007-66152-C02-01/02 and FPA2010-21816-C02-01/02 from the Ministerio de Educación y Ciencia, Spain.

References

[1] CMS Collaboration, “The Compact Muon Solenoid Technical Proposal”, CERN/LHCC 94-38, 1994
[2] Worldwide LHC Computing Grid. Technical Design Report. E-ref: http://lcg.web.cern.ch/LCG/TDR/LCG TDR v1 04.pdf
[3] CMS Collaboration, “CMS Computing Project: Technical design report”, CERN-LHCC 2005-023, 2005
[4] R. Egelund et al. “Data transfer infrastructure for CMS data taking”, PoS(ACAT08) 033, 2008
[5] R. Kaselis et al., “CMS Data Transfer operations after the first years of LHC collisions”, in these proceedings
[6] A. Frohner et al., “Data management in EGEE”, J. Phys. Conf. Ser. 219 062012, 2010
[7] gLite project website: http://glite.cern.ch/
[8] L. Abadie et al., “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, 2007
[9] B. Allcock et al., ”Data management and transfer in high-performance computational grid environments”, Parallel Computing 28 749-771, 2002
[10] FTS Monitor project website: https://forge.in2p3.fr/projects/show/ftsmonitor
[11] J Flix et al, “Improving CMS data transfers among its distributed computing facilities”, J. Phys. Conf. Ser. 331 052013, 2011
[12] See https://twiki.cern.ch/twiki/bin/viewauth/CMS/CMST0T1TestsFTSMonitorResults
[13] J. Andreeva et al., “Providing WLCG Global Transfer monitoring”, in these proceedings
[14] Z. Molnár et al., “Next generation WLCG File Transfer Service (FTS)”, in these proceedings