Performance studies and improvements of CMS Distributed Data Transfers

#261

Daniele Bonacorsi, Universita e INFN, Italy; **Josep Flix**, PIC/CIEMAT, Spain; **Rapolas Kaselis**, CERN, Switzerland; **James Letts**, University of California, San Diego, USA; **Nicolò Magini**, CERN, Switzerland; **Andrea Sartirana**, LLR/E.Polytechnique, France;
On behalf of **CMS Computing**

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.

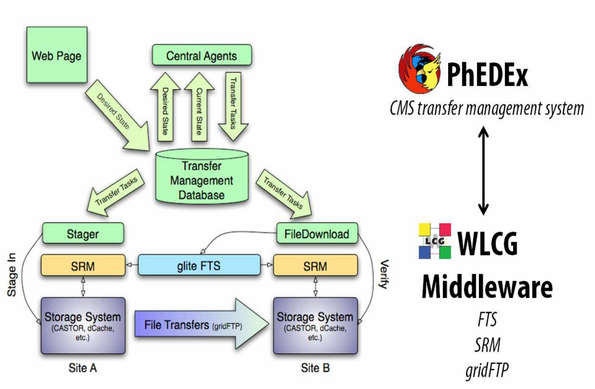
This contribution deals with a complete revision of all FTS servers used by CMS, customizing the topologies and improving their setup in order to keep CMS transferring data to the desired levels, as well as performance studies for all kind of transfer routes, including overheads measurements introduced 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.

CMS Data Transfers

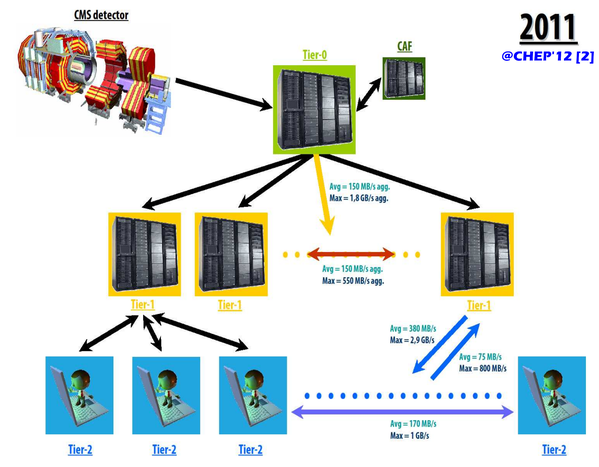
CMS computing infrastructure consist of a Tier-0, 7 Tier-1 sites and around 50 Tier-2 sites, which are all interconnected to transfer relevant data products for analysis. Including Tier-3s, CMS has a site-topology of around 100 sites. **PhEDEx** (Physics Experiment Data Export) is a reliable, scalable dataset replication system used by CMS. Each site runs a set of **software agents**, coupled & highly-specialized components which fulfill a specific "simple" task: **@CHEP'12 [1]**

- **Central agents:** routing, task assignment, ... [Run at CERN];
- **site-specific agents:** download, export, mass storage staging and migration. [Run at every CMS site on a gLite VOBOX or UI].

A transfer request is placed through PhEDEx web interface. Central agents create and distribute transfer tasks to site agents, taking into account the aggregated throughput and errors on the links in the routing algorithm. PhEDEx Download agent submits transfer batch job to FTS server, FTS chosen according to CMS policy. FTS executes the transfer as third-party with gridFTP.



PhEDEx - CMS transfer management system coupled to gLite middleware

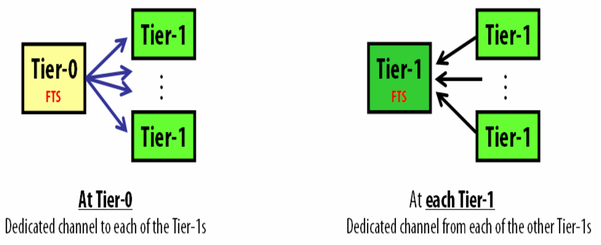


Data Transfers in the CMS topology (Tier-3s excluded). Typical 2011 values are shown

CMS FTS configuration

Glite FTS provides scheduling of multiple asynchronous file transfers on **CHANNELS**: single direction transfer queue between two endpoints, not tied to a physical network path. Each endpoint (source and destination) can be:

- A single site, e.g. **CERN-RAL**;
- a group of sites ("cloud"), e.g. **RALLCG2-CLOUDCMSITALY**;
- all sites ("star"), e.g. **CNAF-STAR**.



Typical FTS server configuration in CMS

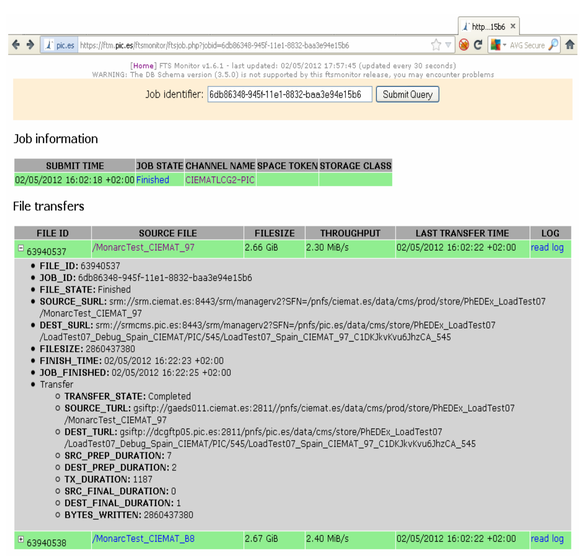
Each FTS channel defines the resource restrictions such as the maximum number and relative share of transfers for each of the VOs using the channel. FTS interacts with Storage Elements through SRM, providing uniform transparent access to storage management capabilities irregardless of the underlying technology. FTS channel configuration defines:

- Transfer limits: maximum number of concurrent active transfers;
- transfer priorities;
- transfer parameters: # of paralell TCP streams, buffer size, timeouts, ...

The FTS Monitor

https://forge.in2p3.fr/projects/ftsmonitor

The FTS Monitor is a web-based monitoring system developed at the CCIN2P3 Tier-1 providing a graphical view of the FTS activity. The service retrieves data directly from the FTS backend DDBB to generate summary statistics a provide detailed reports about transfer activities, for the last 24h. The file-level details are XML published. The FTS Monitor is currently deployed on 6 CMS Tier-1s and the Tier-0.



A view of the FTS Monitor showing details of a finished transfer in a FTS job

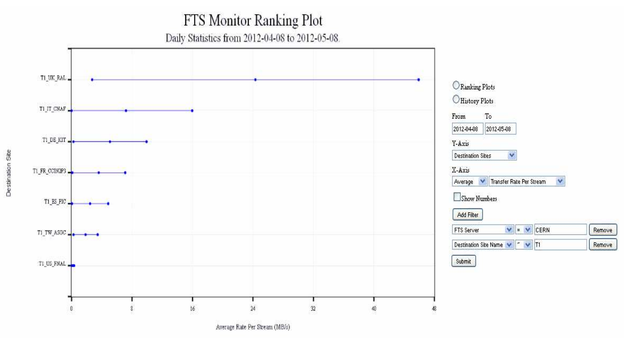
Parsing the information

http://polywww.in2p3.fr/~sartiran/FTMParser/ftmplots.html

The FTS Monitors at the Tier-1s provide detailed information about all the transfers performed in all provided channels, a wealth of information that can be extremely useful to spot issues and debug problems. However, these monitors only offer access to last 24h of transfer history, insufficient in some cases to spot out infrastructural or optimization problems, hence keeping the historical for longer periods is needed. CMS parses the information exposed by the FTS Monitors, daily, store all data in a DDBB, and provide a GUI to properly show and analyze the data. The GUI runs in GRIF Tier-2, and a replica of the whole data is as well obtained by PIC Tier-1.

The most important values are:

- Transfer rates per file and stream, point-to-point;
- FTS Channels performances;
- SRM overheads;
- transfer durations;
- FTS channel occupancies;



An example plot that can be obtained from the CMS FTS Monitor parser

LHCONE Studies

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, 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.

Performance of a T2↔T2 connection, before connecting to LHCONE

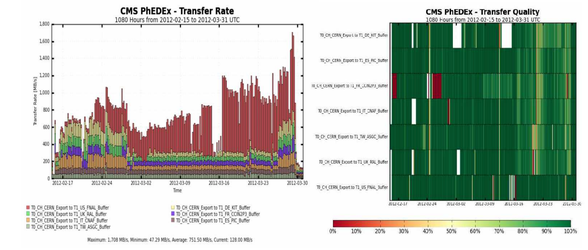
T0→T1 Studies

A detailed study 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 of data taking (1.2 GB/s sustained, distributed among pledges). By means of a dedicated LoadTest, we used the FTS monitoring tool to debug the transfers: some modifications on remote download PhEDEx agents were needed and we also observed changes on rates due to a CASTOR upgrade at CERN.

Measured in FTS Monitor

	FTS PEAK Config (Streams)	(Files)	(Timeout s)	CERN SRM Prep (s) μ	σ	CERN SRM End (s) μ	σ	Dest SRM Prep (s) μ	σ	Dest SRM End (s) μ	σ	Tx Duration (s) μ	σ	SRM Overheads (s) μ	SRM Ov. % of Active μ	Rate/Stream (MB/s/stream) μ	σ	Files Transf. #
CERN→KIT*	5	100	3600	10.8	24.0	0.2	1.9	4.4	3.2	0.8	1.4	122	294	16	16%	7.9	3.6	105983
CERN→PIC*	5	80	3600	11.3	24.7	0.2	2.4	2.4	5.8	1.5	3.9	142	123	15	11%	5.1	2.7	76691
CERN→CCIN2P3*	5	100	3600	11.1	24.3	0.2	1.5	0.6	0.8	0.6	2.0	245	180	13	6%	3.2	2.2	80868
CERN→CNAF*	3	100	3600	10.4	22.7	0.2	2.0	14.4	7.7	4.3	4.1	136	121	29	22%	9.8	6.1	111824
CERN→ASGC*	10	30	14400	11.3	25.0	0.2	2.0	19.8	14.3	5.0	1.0	98	131	36	30%	3.7	1.3	76216
CERN→RAL*	1	65	3600	10.8	24.0	0.2	3.0	14.7	13.2	2.5	1.9	94	109	28	28%	49.3	29.2	99745
CERN→FNAL**	10	284	7200	5.1	40.7	0.0	0.0	17.2	9.5	7.5	3.2	647	675	30	7%	0.5	0.2	146034
			Avg:	9.8	28.2	0.1	2.0											

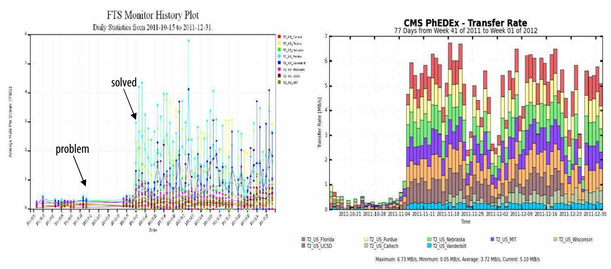
*: URLCOPY Ch. **: SRMCOPY Ch.



T0→T1 tests: Summary table with FTS Monitoring values, rates and transfer qualities

Other Studies

For shared FTS channels, links may have very difference performances. It is well known that slow links keep the FTS channel slots busy for a longer time, increasing the queuing time for all others, hence reducing throughput capabilities. By means of this tool, we can identify those links, and sites can assign them to two or more **separate cloud channels**. The tool has also spotted out **IN/OUT assimetries** in sites, helping tuning their configurations. Additionally, the differences observed in links connected to **LHCOPN are investigated**, the tool helped in detecting **infrastructural problems**, and **congested FTS channels**, among others...



Transfers from IN2P3 Tier-1 to US Tier-2s affected by a RENATER-GEANT misconfiguration

Conclusions

The CMS FTS Monitor parser project developed a basic tool which is being instrumental to understand the performance of data transfers, help improving the overall system, and is used in regular Operations:

- This tool can be used by any VO. In particular, these type of studies inspired the WLCG Global Transfer Monitoring, in which the CMS team cooperates closely, joining efforts to develop/deploy a more powerful tool; **@CHEP'12 [3]**
- These results provided feedback to the development of the new FTS (3.0). **@CHEP'12 [4]**

@CHEP'12: [references]

[1] See poster # 188 → T. Wildish et al. "From toolkit to framework - the past and future evolution of PhEDEx"
[2] See talk # 341 → R. Kaselis et al. "CMS Data Transfer operations after the first years of LHC collisions"
[3] See poster # 289 → J. Andreeva et al. "Providing WLCG Global Transfer monitoring"
[4] See talk # 436 → Z. Molnar et al. "Next generation WLCG File Transfer Service (FTS)"

The authors wish to thank especially Lionel Schwarz from CC-IN2P3 and the Dashboard Team at CERN.

Poster presenter
J. Flix (jflix@pic.es)