Large scale commissioning and operational experience with tier-2 to tier-2 data transfer links in CMS

J Letts¹, N Magini²
¹ Department of Physics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093 USA.
² CERN, 1211 Geneva 23, Switzerland.

E-mail: jletts@ucsd.edu

Abstract. Tier-2 to Tier-2 data transfers have been identified as a necessary extension of the CMS computing model. The Debugging Data Transfers (DDT) Task Force in CMS was charged with commissioning Tier-2 to Tier-2 PhEDEx transfer links beginning in late 2009, originally to serve the needs of physics analysis groups for the transfer of their results between the storage elements of the Tier-2 sites associated with the groups. PhEDEx is the data transfer middleware of the CMS experiment. For analysis jobs using CRAB, the CMS Remote Analysis Builder, the challenges of remote stage out of job output at the end of the analysis jobs led to the introduction of a local fallback stage out, and will eventually require the asynchronous transfer of user data over essentially all of the Tier-2 to Tier-2 network using the same PhEDEx infrastructure. In addition, direct file sharing of physics and Monte Carlo simulated data between Tier-2 sites can relieve the operational load of the Tier-1 sites in the original CMS Computing Model, and already represents an important component of CMS PhEDEx data transfer volume. The experience, challenges and methods used to debug and commission the thousands of data transfers links between CMS Tier-2 sites world-wide are explained and summarized. The resulting operational experience with Tier-2 to Tier-2 transfers is also presented.

1. Evolution of the CMS computing model

The CMS computing model [1] has three tiers of computing facilities connected by high-speed networks of 1 to 10 Gbps. Data flows between and within each of these tiers. These include the Tier-0 at CERN used for data export from CMS and archival to tape and 7 Tier-1 centers used for the tape backup and large-scale reprocessing of CMS data and the distribution of data products to the Tier-2 centers. There are about 50 Tier-2 facilities, where data analysis and Monte Carlo production are primarily carried out. These centers are typically at universities and do not have tape backup systems, only disk storage.

The original CMS computing model envisioned the commissioning of data transfer links between Tier-2 sites and their “associated” Tier-1 site, usually within the local region. In 2008, this was expanded to include commissioning links to and from all Tier-1 sites to and from all Tier-2 sites. This represented the commissioning of only a few hundred data transfer links, which was accomplished during the original “Debugging Data Transfers” (DDT) project in 2007 and 2008 [2].

Commissioning of links between Tier-2 sites themselves was not part of the model at that time. However, in 2009 it became clear that commissioning the entire mesh of Tier-2 to Tier-2 data transfer paths would be necessary [3]. The motivations for this commissioning project were several:
The introduction of the /Store/Results service [4], in which user-generated samples needed to be transferred between Tier-2 sites. Tape archiving was not a requirement. Therefore, transfers via a Tier-1 site where tape archiving is the default was not desired.

Implementation of a fall-back local stage out of analysis results in CRAB [5,6], the current analysis framework of CMS. When user analysis jobs finish successfully, the job results are usually staged out to a remote Tier-2 site where the user has a dedicated storage allocation. This destination Tier-2 may be different from the site where the job ran. When this stage out fails, CMS implemented in 2009 a fall-back to store the job output at the Tier-2 site where the job executed. In the future it is foreseen to implement a method to transfer the results to the proper destination using the existing CMS data transfer infrastructure, thus invoking a need for all Tier-2 to Tier-2 data transfer links to be operational.

Direct Tier-2 to Tier-2 file sharing, which can reduce the operational load on intermediate Tier-1 sites and speed up transfers of detector data and Monte Carlo files in general.

2. The data transfer infrastructure of CMS

PhEDEx [7-11] is the data transfer middleware of the CMS experiment. Within PhEDEx there are several “instances”, which generally means separate databases, accounting, etc. The “Production” instance is for commissioned links only, and carries out CMS production workflows. A “Debug” instance is used exclusively to handle test transfers for debugging and commissioning purposes.

The PhEDEx LoadTest [12,13] is the main way that data transfer links are tested within CMS. The procedure is to inject files at a certain rate into the database and queue them for transfer over the various links. Injection rates are tunable on the Web. The injected logical files map to 256 real files at each site, so that files are transferred multiple times without the need to constantly create new files.

Transfers between some sites are made with third party SRM transfers [14] using GridFTP. However, where load is an issue, transfers are scheduled by FTS [15], part of the gLite middleware. Its main features include submission of data transfer jobs, which are scheduled by an FTS server based on the settings of the channel utilized for that specific source/destination combination. It allows sites to set limitations on the number of files in transfer, number of streams used etc. Transfers to or from a Tier-1 site use the FTS server at that Tier-1 site. Regional Tier-2 sites have dedicated FTS channels at the servers at their associated Tier-1, while “non-regional” transfers use “T1-STAR” channels. Each Tier-2 also has a “STAR-T2” channel on the FTS server at the associated Tier-1 center that is used to import data from other Tier-2 sites.

3. Commissioning procedures and strategies

In order to consider a data transfer link to be “commissioned” and therefore ready for production data transfers, the link must pass a certain metric. In 2008-2009 these commissioning metrics were revised to more closely match the Computing Model requirements for higher rates and transfers in bursts. Links with sources at the Tier-0 or a Tier-1 site are required to transfer at least 1.65 TB (>20MB/s) in a 24 hour period, but only 422 GB (>5MB/s) for a link with a source at a Tier-2 site, i.e. Tier-2 to Tier-2 links, as well as uplinks from Tier-2 to Tier-1 sites.

For the current project only Tier-2 sites that were otherwise fully commissioned [16] for CMS analysis and with at least 1 Gbps network connections were considered, leaving 40 Tier-2 sites, for 1,560 links in total. Some of these were commissioned prior to the beginning of this project in March 2010. For example, during the “October Exercise,” an LHC-wide data challenge, about 200 Tier-2 to Tier-2 links were commissioned to meet the immediate data transfer needs of the exercise.
The difficulties in commissioning over a thousand transfer links are quite different from dealing with a couple of hundred links as we did in previous years. Strategies that worked in 2007-2008, debugging individual links with the help of individual site administrators, do not scale to thousands of transfer paths. For this reason, we chose to debug only patterns of problems that affected the transfer success to or from a particular site for many links, thus avoiding the trap of trying to debug many seemingly unrelated individual problems. In this way, the effort needed scales as the number of sites rather than the number of links (which increases as the number of sites squared). Support for the project in the form of Tier-2 site administrators likewise scales with the number of sites.

**Figure 1.** Number of Tier-2 to Tier-2 data transfer links commissioned per day during the project.

**Figure 2.** Cumulative number of Tier-2 to Tier-2 data transfer links commissioned during the project.
Problems were tracked through the Savannah [17] ticketing system at CERN. During this project, 86 tickets were issued by DDT or by others for problems that directly affected Tier-2 to Tier-2 data transfers. Typical problems found and fixed were networking routing issues, configuration problems with the PhEDEx agents, storage authentication or other general storage problems, configuration problems with FTS channels, etc. By the end of the project in September 2010, all but 6 of the tickets had been closed. The remaining transfer links that were not commissioned can largely be traced to issues detailed in one or more of the remaining open Savannah tickets. In general Tier-2 site administrators were extremely responsive to requests for help in solving problems. Up to 30 transfer links per day were commissioned, with an average of 7 per day, as seen in Figure 1. The cumulative progress of link commissioning was steady during the summer, as shown in Figure 2.

![StageOut Errors from Tier-2 Sites (Number of Analysis Jobs)](image)

**Figure 3.** Number of CMS analysis jobs per week that failed to stage out job output.

### 4. Beneficial effects

Several beneficial side-effects of commissioning Tier-2 to Tier-2 data transfer routes were seen during the summer of 2010. In particular, the number of CMS analysis jobs that ended in an error transferring results from one Tier-2 to the destination site, usually another Tier-2 site, declined steadily throughout the period, as seen in Figure 3. Changes were also made to the CRAB analysis software in June 2010 that significantly improved stage out error rates, by appending a random string to the output file name to avoid overwriting a pre-existing output file, a common failure mode. However, the commissioning of PhEDEx data transfer links also spurred sites to solve many storage and networking problems on the paths between Tier-2 sites, and a steady decline in error rates can be seen throughout the period.

The commissioning of data transfer links between Tier-2 sites also made possible direct file sharing between sites, bypassing the need for intermediate transfers with a Tier-1. These transfer paths were heavily used for production data transfers in 2010, as can be seen in Figure 4, at about half the total data transfer volume of the Tier-1 to Tier-2 links. In general, the transfer quality as measured by the success rate of transfer attempts was very good, as seen in Figure 5. Load-balancing within the PhEDEx data transfer software ensures that if some transfer paths ceased to be operational, other paths are chosen to move the data.
5. Conclusions
In general CMS was very successful at commissioning and subsequently using Tier-2 to Tier-2 data transfers links in a production environment. Over 95% of the links destined for commissioning passed metrics and were put into production. Much effort from Tier-2 site administrators to fix problems quickly was appreciated. Debugging activities focused on finding patterns of failure that affected many links to or from particular sites, rather than on individual links, and so the effort required for the project scaled as the available personnel resources at Tier-2 sites themselves.

Tier-2 to Tier-2 file sharing represents a useful extension of the original CMS Computing Model, allowing direct peer-to-peer file sharing of physics results as well as the transfer of detector data and Monte Carlo simulation files.

Figure 4. Transfer volume per week to Tier-2 destinations from Tier-1 sites (red) and from other Tier-2 sites (yellow).

6. References
[1] “The CMS Computing Model”, CMS NOTE/2004-031.
[2] Bagliesi G et al., “Debugging Data Transfers in CMS”, J. Phys.: Conf. Ser. 219 (2010) 062055, presented at CHEP09, Prague CZ, March 2009.
[3] Bonacorsi D et al., “Experience with the CMS Computing Model from Commissioning to Collisions”, PS01-3-176 presented at this conference.
[4] Giffels M et al., “Design and Early Experience with Promoting User-created Data in CMS,” PO-MON-008 presented at this conference.
[5] C. Codispoti et al., “CRAB: a CMS Application for Distributed Analysis”, IEEE Trans. Nucl. Sci. 56 (2009) 2850-2858.
[6] Vaandering E et al., “CMS Distributed Analysis Infrastructure and Operations: Experience with the First LHC Data,” PS25-2-212 presented at this conference.
[7] Magini N, Bonacorsi D and Rossman P, “Distributed Data Transfers in CMS,” PO-MON-40 presented at this conference.
**Figure 5.** Production PhEDEx data transfer quality for Tier-2 source sites to Tier-2 destination sites, varying color scale from red denoting 0% success to dark green for 100% success.

[8] Magini N et al., “Improving CMS Data Transfers Among its Distributed Computing Facilities,” PS23-3-204 presented at this conference.

[9] Bonacorsi D et al., “PhEDEx High Throughput Data Transfer Management System,” CHEP06, Bombay, India, February 2006.

[10] R. Egeland et al., “Data Transfer Infrastructure for CMS Data Taking”, ACAT08, Erice, Italy, November 2008.

[11] L. Tuura et al., “Scaling CMS Data Transfer System for LHC Start-Up,” J. Phys. Conf.: Ser. 119 (2008) 072030, presented at CHEP07, Victoria, BC, Canada.

[12] Bagliesi G et al., “The CMS LoadTest 2007: An Infrastructure to Exercise CMS Transfer Routes among WLCG Tiers”, CHEP07, Victoria, B.C., Canada, September 2007.

[13] Magini N et al., “The CMS Data Transfer Test Environment in Preparation for LHC Data Taking,” NSS-IEEE, Dresden 2008.

[14] For more information, please see https://srm.fnal.gov/twiki/bin/view/SrmProject

[15] Frohner A et al., “Data Management in EGEE,” J. Phys. Conf.: Ser. 219 (2010) 062012, presented at CHEP09, Prague, CZ, March 2009.

[16] Hernandez J et al., “Monitoring the Readiness and Utilization of the Distributed CMS Computing Facilities During the First Year of CMS Running,” PS42-2-202 presented at this conference.

[17] Perrin Y, Orellana F, Roy M, Feichtinger D, “The LCG Savannah software development portal,” in Proc. CHEP04, CERN-2005-002, p. 609-612.