Debugging Data Transfers in CMS

G. Bagliesi (INFN Pisa Italy, and CERN), **S. Belforte** (INFN Trieste, Italy), **K. Bloom** (Univ. of Nebraska-Lincoln, USA), **D. Bonacorsi** (INFN-CNAF, Italy), **I. Fisk** (Fermilab, US), **J. Flix** (CIEMAT, PIC, Barcelona, Spain), **J. Hernandez** (CIEMAT, Spain), **M. Kadastik** (NICPB, Tallinn, Estonia), **J. Klem** (HIP, Helsinki, Finland), **O. Kodolova** (Moscow State Univ., Russia), **C.-M. Kuo** (NCU, Chung-li Taiwan), **J. Letts** (Univ. of California San Diego, US), **N. Magini** (CERN), **S. Metson** (HH Wills Physics Laboratory, Bristol, UK), J. Piedra (MIT, Boston, US), **N. Pukhaeva** (CC-IN2P3, Lyon, France), **L. Tuura** (Northeastern Univ., Boston, US), **S. Sõnajalg** (NICPB, Tallinn, Estonia), **Y. Wu** (Fermilab, US), **F. Würthwein** (Univ. of California San Diego, US)
On behalf of CMS Offline and Computing

ABSTRACT

The CMS experiment at CERN is preparing for LHC data taking in several computing preparation activities. In early 2007 a traffic load generator infrastructure for distributed data transfer tests was designed and deployed to equip the WLCG Tiers which support the CMS Virtual Organization with a means for debugging, load-testing and commissioning data transfer routes among CMS Computing Centers. The LoadTest is based upon PhEDEx as a reliable, scalable data set replication system. The Debugging Data Transfers (DDT) Task Force was created to coordinate the debugging of the data transfer links. The Task Force aimed to commission most crucial transfer routes among CMS tiers by designing and enforcing a clear procedure to debug problematic links. Such procedure aimed to move a link from a debugging phase in a separate and independent environment to a production environment when a set of agreed conditions are achieved for that link. The goal was to deliver one by one working transfer routes to Data Operations. The preparation, activities and experience of the DDT Task Force within the CMS experiment are discussed. Common technical problems and challenges encountered during the lifetime of the taskforce in debugging data transfer links in CMS are explained and summarized.

CMS COMPUTING MODEL

The CMS computing model [1] has three tiers of computing facilities. Data flows between and within each of these tiers:

- Tier 0 at CERN, used for data export from CMS,
- 8 Tier 1 (T1) centers, including one at CERN, used for the tape backup and large-scale reprocessing of CMS data, and distribution of data products to the Tier 2 centers, and
- ~50 Tier 2 (T2) facilities, where data analysis and Monte Carlo production are primarily carried out.

Each of these sites is connected by high-speed networks of 1-10Gbs, but in a few cases less than this. Transfer links of Tier 3 sites were not explicitly considered by the Task Force.

The CMS Computing Model envisions commissioning all data transfer links between:

- Between CERN T0 and all T1 sites (14 links)
- All other T1-T1 cross-links (42 links)
- All T1 to T2 downlinks (~400 links)
- T2 to regional T1 uplinks (~50 links)

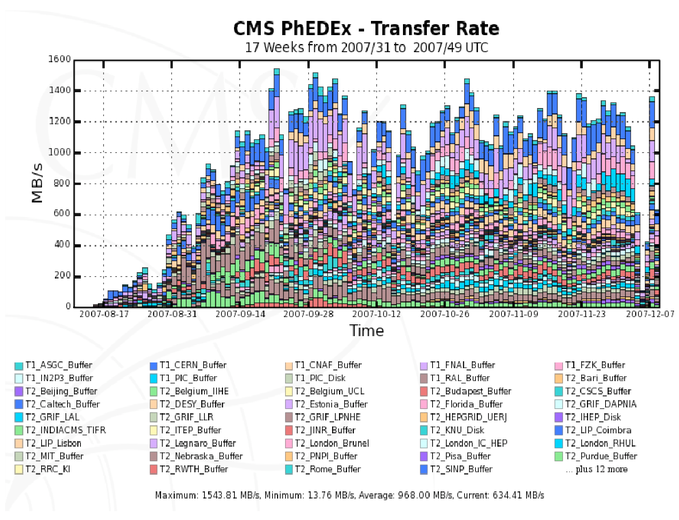
T1-T1 cross-links and T1-T2 downlinks are used primarily for the distribution of data products to the T2 sites for eventual analysis, while the uplinks from T2 to T1 sites are used to upload Monte Carlo data produced at the T2 sites for backup at the T1s.

At the beginning of the Task Force, only ~30 data transfer links were sufficiently stable to be considered COMMISSIONED.

INFRASTRUCTURE

PhEDEx [2] 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 the data challenge workflows and Monte Carlo production transfers. The DEBUG instance was created in early August 2007 to handle test transfers and in September became used exclusively for the test transfers.

The PhEDEx LoadTest [3] is used to exercise the data transfer links with test transfers. The procedure is to inject files at a certain rate into the database and queue them for transfer over the various links. Rates are tunable from a web interface.



LoadTest Data Transfer Rates to All Tiers in 2007

COMMISSIONING STEPS

The DDT Task Force defined the steps that links must pass through to achieve COMMISSIONED status:

- NOT-TESTED: no transfers ever attempted
- PENDING-COMMISSIONING: links which have transfer attempts but have not passed the METRIC defined below for commissioning
- COMMISSIONED: links which have passed the commissioning METRIC
- PROBLEM-RATE: links which after commissioning develop problems to transfer according to the METRIC
- DECOMMISSIONED: links which after commissioning subsequently fail the METRIC during testing

COMMISSIONING METRIC

To be COMMISSIONED in 2007, a link had to transfer:

- 2.3 TB in a 7 day period
- 1.7 TB in a 5 day period for links with a T2 endpoint

To remain COMMISSIONED, the link had to transfer at least 300 GB of data once every 7 days. Regular link exercises were performed on every commissioned link. Transfers in the PRODUCTION and DEBUG instances of PhEDEx were summed and counted towards metric goals.

In 2008-2009 these commissioning metrics were revised to more closely match the Computing Model requirements for higher rates and transfers in bursts. Links were required to transfer at least:

- 1.65 TB (>20MB/s) in a 24 hour period, or
- 422 GB (>5MB/s) for T2-T1 uplinks only

To remain commissioned, all links were periodically exercised in a random order and were required to meet half of the above metric goals within 3 calendar days.

T2-T2 cross-links, although not in the Computing Model, were commissioned on request of the sites.

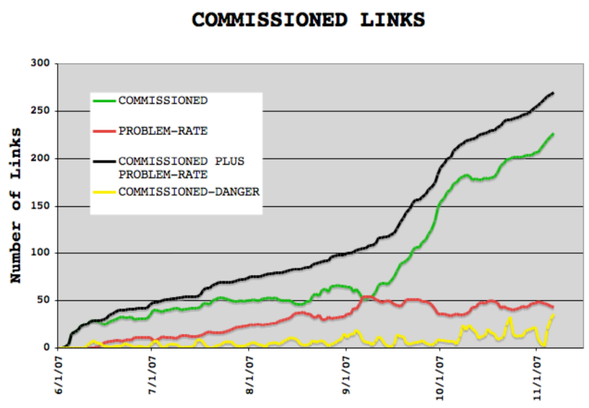
METRIC TOOLS

A tool was developed to visualize the commissioned link status of all data transfer links in CMS based on the above criteria. The following plot shows the status of the T1-T1 cross-links before a dedicated campaign to commissioned them was undertaken in September 2007. Green links are classified as COMMISSIONED, blue are PENDING-COMMISSIONING and red are PROBLEM-RATE. All of the T1-T1 links were successfully commissioned over several weeks in September and October 2007.

	ASGC	CERN	CNAF	FNAL	FZK	IN2P3	PIC	RAL
ASGC								
CERN								
CNAF								
FNAL								
FZK								
IN2P3								
PIC								
RAL								

DDT Metric Visualization Tool, September 2007

COMMISSIONING PROGRESS



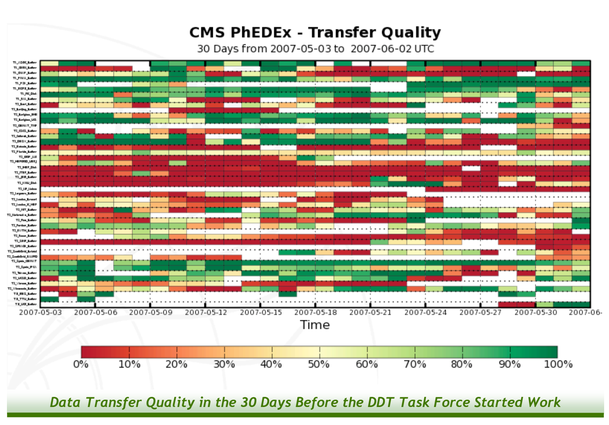
Number of Commissioned and Problem-Rate Links.

During the initial phase of the Task Force, rapid progress was made in link commissioning. Most problems were encountered in the initial link commissioning so that gains made during this time were generally retained. The number of PROBLEM-RATE links remained a somewhat constant percentage of the total number of COMMISSIONED links.

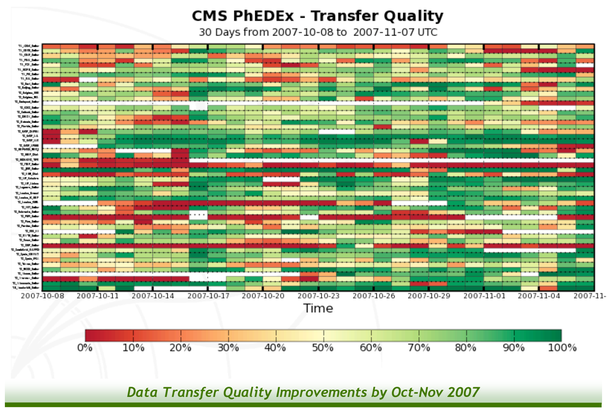
QUALITY IMPROVEMENTS

The DDT Task Force aided sites in their debugging efforts. Most communication was through a mailing list.

Before the Task Force concentrated attention on the quality and stability of data transfers, the "transfer quality" or success rate of transfers was poorer than after the completion of the first months of the work, as the following two plots show.



Data Transfer Quality in the 30 Days Before the DDT Task Force Started Work

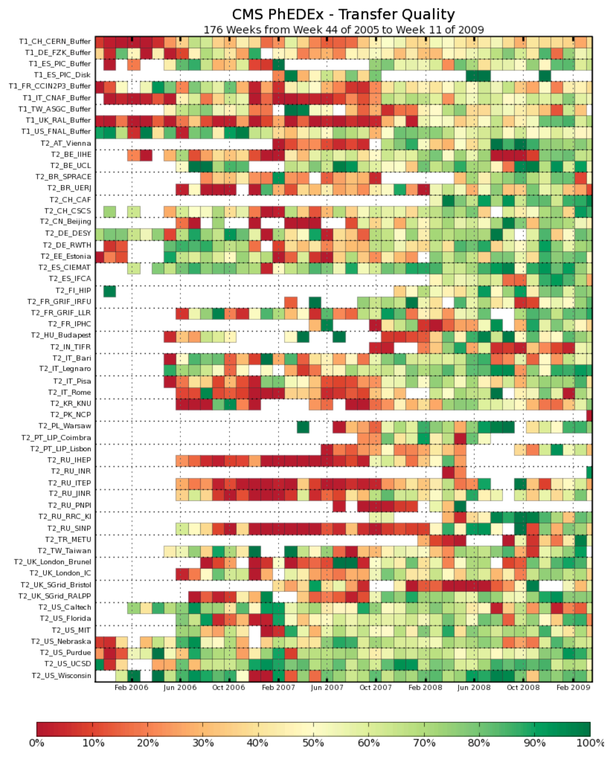


Data Transfer Quality Improvements by Oct-Nov 2007

CURRENT STATUS

The DDT Task Force continues to aid sites in their data transfer link commissioning efforts. Dedicated campaigns in 2008 included helping sites complete the commissioning of all of their downlinks from the T1 sites.

- 35 T2 sites have all of their downlinks from T1 sites COMMISSIONED, and 2 more have 7/8 links COMMISSIONED
- 43 T2 sites have at least two COMMISSIONED uplinks to T1 sites.
- All T1-T1 links are COMMISSIONED.
- Transfer quality continues to improve:



Improvement in PhEDEx Data Transfer Quality from November 2005 to March 2009

CONCLUSIONS

The experience of the DDT Task Force showed that dedicated computing campaigns can result in rapid progress in commissioning data transfer links and improving permanently the quality of data transfers, for example.

The procedures and metrics developed by the DDT Task Force are now part of the commissioned links overview by the CMS Site Commissioning project [4,5].

References:

[1] CMS Collaboration, "The CMS Computing Model," CERN LHCC 2004-035.
[2] D. Bonacorsi, et al., "PhEDEx High-throughput Data Transfer Management System," CHEP06, Bombay, India, February 2006.
[3] N. Magini et al., "The CMS Data Transfer Test Environment in Preparation for LHC Data Taking," NSS-IEEE, Dresden, 2008.
[4] J. Flix et al, "The Commissioning of CMS Computing Centres in the Worldwide LHC Computing Grid," - NSS-IEEE, Dresden, 2008.
[5] See the poster presented at this conference by J. Flix et al.,"The commissioning of CMS sites: improving the site reliability".

Poster presenter: J. Letts
DDT Task Force Co-coordinator
Department of Physics
University of California, San Diego, US
Email: jletts@ucsd.edu