Available on CMS information server **CMS CR -2009/093**

The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

11 May 2009

Debugging Data Transfers in CMS

G. Bagliesi, S. Belforte, K. Bloom, B. Bockelman, D. Bonacorsi, I. Fisk, J. Flix, J. Hernandez, J. D'Hondt, M. Kadastik, J. Klem, O. Kodolova, C.-M. Kuo, J. Letts, J. Maes, N. Magini, S. Metson, J. Piedra, N. Pukhaeva, L. Tuura, S. Sõnajalg, Y. Wu, P. Van Mulders, I. Villella, F. Würthwein

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 called the LoadTest was designed and deployed to equip the WLCG sites that support CMS with a means for debugging, load-testing and commissioning data transfer routes among CMS computing centres. 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 sites 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 the CMS data operations team. 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.

Presented at *17th International Conference on Computing in High Energy and Nuclear Physics (CHEP*

Debugging Data Transfers in CMS

G. Bagliesi[1,2], S. Belforte[3], K. Bloom[4], B. Bockelman[4], D. Bonacorsi[5], I. Fisk[6], J. Flix[7,8], J. Hernandez[8], J. D'Hondt[9], M. Kadastik[10], J. Klem[11], O. Kodolova[12], C.-M. Kuo[13], J. Letts[14], J. Maes[9], N. Magini[2,15], S. Metson[16], J. Piedra[17], N. Pukhaeva[18], L. Tuura[19], S. Sõnajalg[10], Y. Wu[6], P. Van Mulders[9], I. Villella[9], F. Würthwein[14]

1. INFN Sezione di Pisa, Italy 2. CERN, Geneva, Switzerland 3. INFN Sezione di Trieste, Italy 4. University of Nebraska, Lincoln, NE, USA 5. Università degli Studi di Bologna and INFN Sezione di Bologna, Italy 6. Fermilab, Batavia, IL, USA 7. Port d'Informació Científica (PIC), UAB, Barcelona, Spain. 8. Centro de Investigaciones Energeticas Medioambientales y Tecnologicas (CIEMAT), Madrid, Spain 9. Vrije Universiteit Brussel, Brussels, Belgium 10. National Institute of Chemical Physics and Biophysics (NICPB), Tallinn, Estonia 11. Helsinki Institute of Physics (HIP), Helsinki, Finland 12. Institute of Nuclear Physics, M.V. Lomonosov Moscow State University, Moscow, Russia 13. National Central University (NCU), Chung-li, Taiwan 14. University of California San Diego, La Jolla, CA, USA 15. INFN – CNAF, Bologna, Italy 16. HH Wills Physics Laboratory, Bristol, UK 17. MIT, Boston, MA, USA 18. CC-IN2P3, Lyon, France 19. Northeastern University, Boston, MA USA.

Email: `jletts@ucsd.edu`

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 called the LoadTest was designed and deployed to equip the WLCG sites that support CMS with a means for debugging, load-testing and commissioning data transfer routes among CMS computing centres. 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 sites 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 the CMS data operations team. 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.

1. Introduction

The CMS experiment is a large particle physics experiment located at CERN, Geneva, Switzerland that is presently being commissioned for beginning of data taking. CMS depends on a world wide distributed data grid of about 60 computing and storage clusters to archive and analyze its data. Individual clusters vary both in size (10 TB to a few PB) as well as expertise of their operations teams. The several hundred end-to-end data transfer links between these sites needed to be commissioned. In

July 2007 CMS created a “Debugging Data Transfers” (DDT) task force to coordinate the debugging of data transfer links in the preparation period and during the CSA07 data transfer test [1]. The CSA07 service challenge was a data challenge in 2007 designed to test the transfer system at 50% of the design goal for 2008. The goal of the DDT task force was to deliver fully debugged and operational end-to-end links to the CMS data operations team. We aimed to commission the most crucial transfer routes among CMS tiers by designing and enforcing a clear procedure to debug problematic links. The 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 were achieved for that link.

This paper details the activity of this task force before, during and since CSA07. Section 2 describes the task force charge and scope. Section 3 describes some of the details of the CMS Computing Model and the CSA07 data challenge relevant to this task force. Section 4 describes briefly the system components used to transfer data across the wide area network. Section 5 details the metric used to commission links and progress in link commissioning. Section 6 concludes with a description of the current status of link commissioning in CMS.

2. Task Force Charge within CMS

The DDT task force was focused on the status of data transfer links, defined as unidirectional end-to-end data transfer between site A and site B. The responsibilities of the task force were set out to be:

- to define details on how the metrics are measured to put links in/out of production status;
- to define a procedure, including a set of steps or stages to pass that gets a link from a decommissioned state to production, including documentation of the kinds of tests and tools to use. This includes helping sites to resolve their problems by pointing them to storage element (SE) support channels for the SE they have chosen to deploy, for example. The task force is the first point of contact for the site administrators. The task force thus facilitates information exchange;
- documentation and creation of a list of known problems encountered, and instructions for solving them;
- creation of a table that keeps track of the matrix of status of all links;
- weekly reports on the status of this matrix.

3. The CMS Computing Model and Service Challenge Workflows

3.1. The CMS Computing Model

The CMS computing model [2] has three tiers of computing facilities. These sites are interconnected by high-speed networks of 1-10 Gbps. Data flows between and within each of these tiers:

- Tier-0 at CERN, used for data export from CMS and archival to tape.
- 8 Tier-1 centres, 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 sites. The Tier-1 sites are typically at national laboratories with large computing facilities and archival storage systems.
- about 50 Tier-2 facilities, where data analysis and Monte Carlo production are primarily carried out. These sites typically do not have tape backup systems, only disk storage.

The CMS computing model envisions commissioning all links between:

- CERN to Tier-1 sites, and Tier-1 sites to CERN (14 links)
- All other Tier-1 to Tier-1 cross-links (42 links)
- All Tier-1 to Tier-2 downlinks (about 400 links)
- All Tier-2 to “regional” Tier-1 uplinks (about 50 links)

Therefore, the total number of links to be commissioned in the computing model is about 500. This number will increase with the addition of new Tier-2 sites. At the beginning of the task force only about 30 transfer links were sufficiently stable to be considered commissioned.

Tier-2 to non-regional Tier-1 uplinks were not a priority in 2007 but were commissioned if the sites wished. Each Tier-2 is associated to a Tier-1 (called the "regional" or "associated" Tier-1), although in some cases this Tier-1 is not geographically near the Tier-2. The Tier-2 to Tier-2 cross-links are not part of the computing model, but in fact are used especially within the same country as in the United States, Germany and Belgium. These about 2500 links are not included in the scope of the computing model but were also considered by the DDT task force if the sites wanted to commission them. Likewise, links that begin or end at a Tier-3 site are not in the computing model. A Tier-3 site is typically a small or medium-sized computing facility associated with a Tier-1 or a Tier-2 site, usually at a university or research institute. These links were commissioned on request of the Tier-3 site. There were only six Tier-3 sites active within CMS that attempted data transfers during the CSA07 period.

3.2. CSA07 Workflows

The CSA07 data transfer challenge preparation began with the production of Monte Carlo datasets, primarily at Tier-2 sites. These datasets were transferred to CERN via the regional Tier-1 sites at which point the CSA07 workflows would begin at CERN. For this activity to take place, commissioned data transfer uplinks from Tier-2 sites to at least one Tier-1, and from those Tier-1 sites to Tier-1_CERN needed to exist.

The subsequent CSA07 workflows included the transfer of data from the Tier-0 at CERN to the 7 Tier-1 sites, data reprocessing at the Tier-1 sites, and data transfer of the resulting skimmed datasets to the Tier-2 sites, either directly or via a Tier-1 to Tier-1 cross-link. The CSA07 project is described in [1], including the various metrics measured to determine success.

Eventually it was decided that the priority of debugging effort was to be:

- CERN to Tier-1 and Tier-1 to CERN links
- All other Tier-1 to Tier-1 links
- Tier-1 to and from associated Tier-2 links, establishing at least one link per Tier-2 in each direction so that the Tier-2 site is useful to data operations for data analysis and Monte Carlo production activities
- Tier-1 to and from non-regional Tier-2 sites.

4. System Components

4.1. PhEDEx, the Data Transfer Middleware

PhEDEx [3,4] 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 CSA07 workflows and Monte Carlo production transfers. A "Debug" instance was used from early August 2007 to handle test transfers and in September became used exclusively for the test transfers.

The PhEDEx LoadTest [5,6] is the main way that data transfer links are tested within CMS. All DDT traffic is in the context of the PhEDEx LoadTest. 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 now tuneable from a web interface. 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.

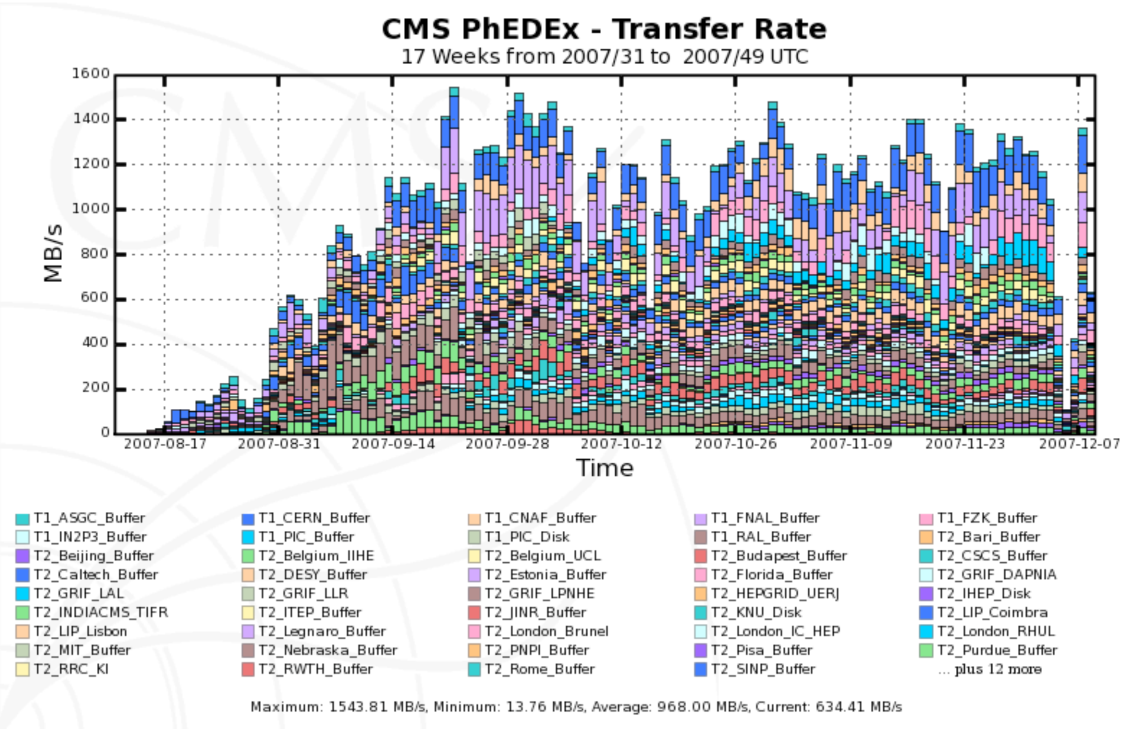

Fig. 1. Total data transfer rate for all links in the Debug instance of PhEDEx.

As seen in Figure 1, LoadTest transfers began in the Debug instance of PhEDEx in early August and ramped up before the start of CSA07 to about 1 GB/s. This rate was maintained more or less constantly over the course of CSA07 despite the addition of new links throughout September and October. This indicates that there are real limitations in the capacities of sites to transfer data in and out of their storage elements. These limitations are well below the physical network capacities of the links between sites.

4.2. Storage Elements

The CMS Tier-1 and Tier-2 sites use various storage elements (SE) systems, which are briefly described:

- The dCache storage system [7] was developed at DESY and FNAL. The system gets support from both US OSG storage support group as well as the dCache User Forum in Europe. It is used by FNAL, PIC, FZK and IN2P3 as the Tier-1 storage backend and by a majority of the active Tier-2 centres that have deployed distributed storage solutions. As the system is developed in Java, it can be deployed on different platforms and different operating systems and a number of these have been used also in the CMS associated sites.
- The CASTOR [8] storage system was developed at CERN and is used by the Tier-1 centres at CERN, CNAF, RAL and ASGC. There are no Tier-2 sites that use this storage system.
- Developed as part of the gLite middleware, DPM [9] is mostly used by smaller Tier-2 centres. About 10 Tier-2 sites use DPM as their storage solution.

The Tier-1 sites also have tape back-ends to their storage systems, which were not exercised in this project.

4.3. Transfer Protocols

Transfers between some sites are made with third party SRM-based transfers. However, where load is an issue, transfers are scheduled by FTS [9]. FTS is part of the gLite middleware and therefore used mostly by the EGEE sites. 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 the Tier-1 site. Regional Tier-2 sites have dedicated FTS channels at the servers at their associated Tier-1, while non-regional transfers use non-dedicated channels.

4.4. Networking

No major issues with networking affected the debugging of transfers. However, it is apparent that majority of links are not performing anywhere close to the speeds per stream that they should be able to achieve, and no coordinated effort has been done to identify fully the reasons. Only a handful of sites have performed testing to understand the network path between them and to try to tune their storage accordingly. The majority of storage nodes are running un-tuned default kernel configurations that do not favour high-speed long distance transfers, but are sometimes tuned for the requirements of the storage elements at the sites.

Most Tier-1 sites are interconnected through 10 Gbps networks, although they serve Tier-2 sites that are often connected through national network infrastructures with a more limited capacity (1-2 Gbps). This imposes limitations on some regions.

Some hardware issues that caused network outages in Brazil occurred during CSA07. Network bandwidth was rarely a limiting problem otherwise.

5. Commissioning Procedures and Progress

5.1. Link Commissioning Metric

The first activity of the DDT task force was to define and implement a metric by which links can become commissioned and subsequently handed over to data operations. There are several stages through which a link passes from “NOT-TESTED” to “COMMISSIONED”:

- **NOT-TESTED**: links never actually tested, i.e. links showing no successful transfer attempts within PhEDEx.
- **PENDING-COMMISSIONING**: links that have transferred successfully at least one file in PhEDEx, but have not yet passed the requirements below for link commissioning.
- **COMMISSIONED**: links that are demonstrated to work, and can be delivered to data operations. Note that this commissioning does not imply that the link or the site has met the requirements of the computing model or the current service challenge, but simply that the link has passed some minimum requirements to be considered usable for data operations. To be COMMISSIONED during 2007, a link was required to:
    - Transfer 300 GB/day for 6 out of 7 consecutive days, and transfer a total of 2.3 TB during that same 7-day period.
    - For links involving an endpoint at a Tier-2, this requirement was relaxed to 4 out of 5 days, and a total transfer volume of 1.7 TB. This is to match the service requirement of business hours only support committed to by the Tier-2 sites.

- **PROBLEM-RATE,** for links that were working but whose rate has dropped off. To remain COMMISSIONED, a link had to transfer at least 300 GB/day for a single day at least once every 7 days. Otherwise, the link had to be re-commissioned by following the procedure above.

These requirements were developed with the idea of having a higher threshold to commission than to decommission the link. These thresholds can be increased in time as networks and sites develop, since the rates implied by 300 GB/day are of the order of 3-4 MB/s per link, far below the commitments outlined in the computing model which envision Tier-2 sites being able to download a total of up to 5 TB/day from Tier-1 sites, or over 60 MB/s sustained downloads. However, the computing model also envisions that transfers will occur in bursts, not as a continuous rate over several days. The metric used during CSA07 deviated from this model to prove the stability of data transfer links.

A member of the DDT team presented a list of changes to the COMMISSIONED or PROBLEM-RATE links to the data operations team at daily and weekly meetings. Any exceptions to the commissioning or decommissioning metric were decided in this meeting and were only rarely approved. Full enforcement of these procedures began in early September 2007 in a phased manner.

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 (>20 MB/s) in a 24 hour period, or
- 422 GB (>5 MB/s) for Tier-2-Tier-1 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. During link exercising about 1% of transfer links showed problems that resulted in decommissioning. Almost all such problem links were re-commissioned in a short period of time.

5.2. Monitoring of Link Status

A tool was developed to extract transfer volume data from PhEDEx and apply the DDT commissioning criteria. This tool takes data transfers from both the Production and Debug instances of PhEDEx into consideration. Figure 2 shows an example of this DDT Matrix. Green links are those that are COMMISSIONED, red are PROBLEM-RATE, light blue are PENDING and white links are NOT-TESTED.

5.3. Link Commissioning Progress

During the CSA07 Tier-1 to Tier-1 link debugging was identified as an urgent priority, as only a few Tier-2 sites were able to commission links from multiple Tier-1s and keep them commissioned over a long time period. To allow Tier-2 sites to participate in the physics transfers phase of CSA07, the data would have to be copied from one Tier-1 to another Tier-1 before the Tier-2 could download the data from the associated Tier-1. As can be seen on Figure 3, around the beginning of October a major increase in commissioned links shows the time when DDT effort was prioritized for Tier-1 to Tier-1 matrix commissioning. In a short time period over half of the links were commissioned. No major technical obstacles were overcome during this period. This was in part legacy from previous efforts to bring up the transfer infrastructure and in part a matter of focus and attention of the Tier-1 administrators to make this activity a priority, mainly in setting up LoadTest samples, creating and approving subscriptions, etc. Although there were ongoing technical challenges before and after the links became commissioned, there was no major technical obstacle to commissioning these links in October.

The largest overall increase in commissioned links came in the period leading to the start of CSA07. At this time, we began to enforce the policy to use only commissioned links for production data transfers, which clearly encouraged sites to put in the effort to commission their links and participate in more than just the test exercises. Most of the progress in commissioning the Tier-2 data transfer links involved configuration of FTS channels or configuration of the PhEDEx transfer software at the Tier-2.

	ASGC	CERN	CNAF	FNAL	FZK	IN2P3	PIC	RAL
ASGC	⬛	🟩	🟦 🟩	🟩	⬜	🟦	🟦 ⬜	⬜ 🟦
CERN	🟩	⬛	🟩 🟦	🟩	🟩	🟩	🟩	🟩
CNAF	🟩 🟦	🟦 🟩	⬛	🟦	🟦 ⬜	🟦 ⬜	🟩 ⬜	🟦
FNAL	🟩	🟩	🟦	⬛	🟩 🟥	🟥 🟩	🟩 🟥	🟦
FZK	⬜	🟩	⬜ 🟦	🟥 🟩	⬛	🟦 🟥	🟦 🟩	🟦 🟩
IN2P3	🟦	🟩	⬜ 🟦	🟩 🟥	🟥 🟦	⬛	🟥 🟦	🟦 🟥
PIC	⬜ 🟦	🟩	⬜ 🟩	🟥 🟩	🟩 🟦	🟦 🟥	⬛	🟦 🟩
RAL	🟦 ⬜	🟩	🟦	🟦	🟩 🟦	🟥 🟦	🟩 🟦	⬛

Fig. 2. Link status matrix. The green links were COMMISSIONED, the red links those that were commissioned and developed problems, the blue in the process of commissioning, and the white untested.

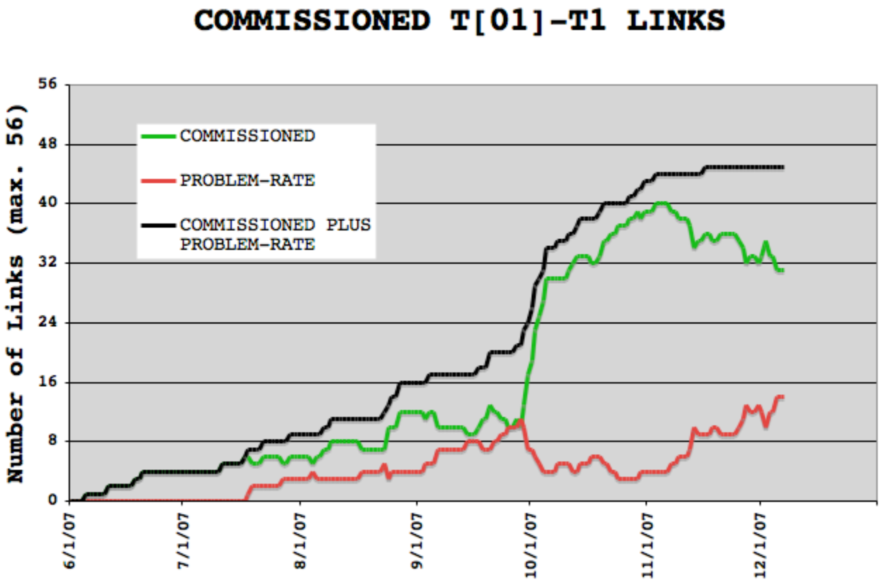

Fig. 3. History of the status of COMMISSIONED links between CERN and Tier-1 centres and Tier-1 to Tier-1 cross-links. The green line shows the number of COMMISSIONED links, the red line the number of PROBLEM-RATE links, and the black line the sum of the two, being the total number of links to ever commission during the exercise.

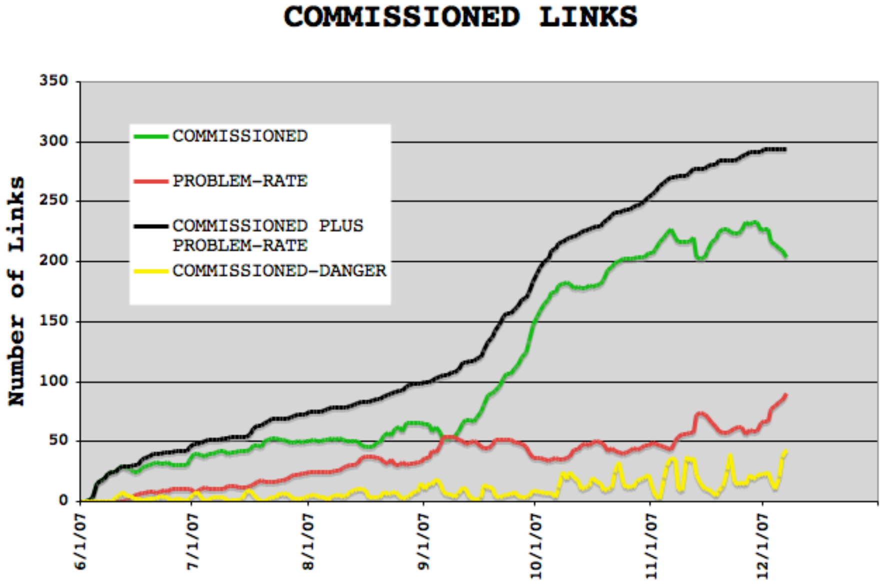

Fig. 4. History of the status of COMMISSIONED links between all sites. The green line shows the number of COMMISSIONED links, the red line the number of PROBLEM-RATE links, and the black line the sum of the two, being the total number of links to ever commission during the exercise. The yellow line shows the number of COMMISSIONED links that were in danger of decommissioning within the next two days.

6. Current Status and Conclusions

During the task force, an improvement in the number and quality of data transfer links was achieved, through the hard efforts of site administrators, PhEDEx developers, data operations and networking experts, etc. The initial mandate of the DDT task force concluded with the end of the CSA07 service challenge in November 2007. However, the effort was considered useful and continued in 2008 in a modified form. The DDT task force continues to aid sites in their data transfer link commissioning efforts. The task force undertook dedicated campaigns in 2008 that included helping sites complete the commissioning of all of their downlinks from the Tier-1 sites, for example, which are documented in a CMS web site [10]. Most sites were also encouraged to commission a second Tier-2 to Tier-1 uplink, so that Monte Carlo production data could be uploaded to more than one Tier-1 site.

Firstly, the metric was modified to more closely match the CMS computing model requirements. To commission links from 2008 onwards, a data link must transfer at a rate of at least 20 MB/s over 24 hours. Recognizing that uplinks from Tier-2 to Tier-1 sites have a lower requirement in the computing model, they are only required to transfer 5 MB/s. Secondly, it was recognized by the CMS computing management that continual exercising of transfer links placed an unnecessarily large burden on site administrators and storage systems. Within the computing model data links are foreseen to be transferring data in bursts with periods of inactivity. To more closely match the model, the requirements for links to stay commissioned was changed. A testing program was organized in 2008 in which the DDT task force attempted to exercise each link in rotation for 12 hours, trying to meet the metric goal of 20 MB/s or 5 MB/s for a Tier-2-Tier-1 uplink. Links were only decommissioned if they failed this transfer exercise for three days, or developed an obvious problem and the data operations team requested the disabling of the transfer link.

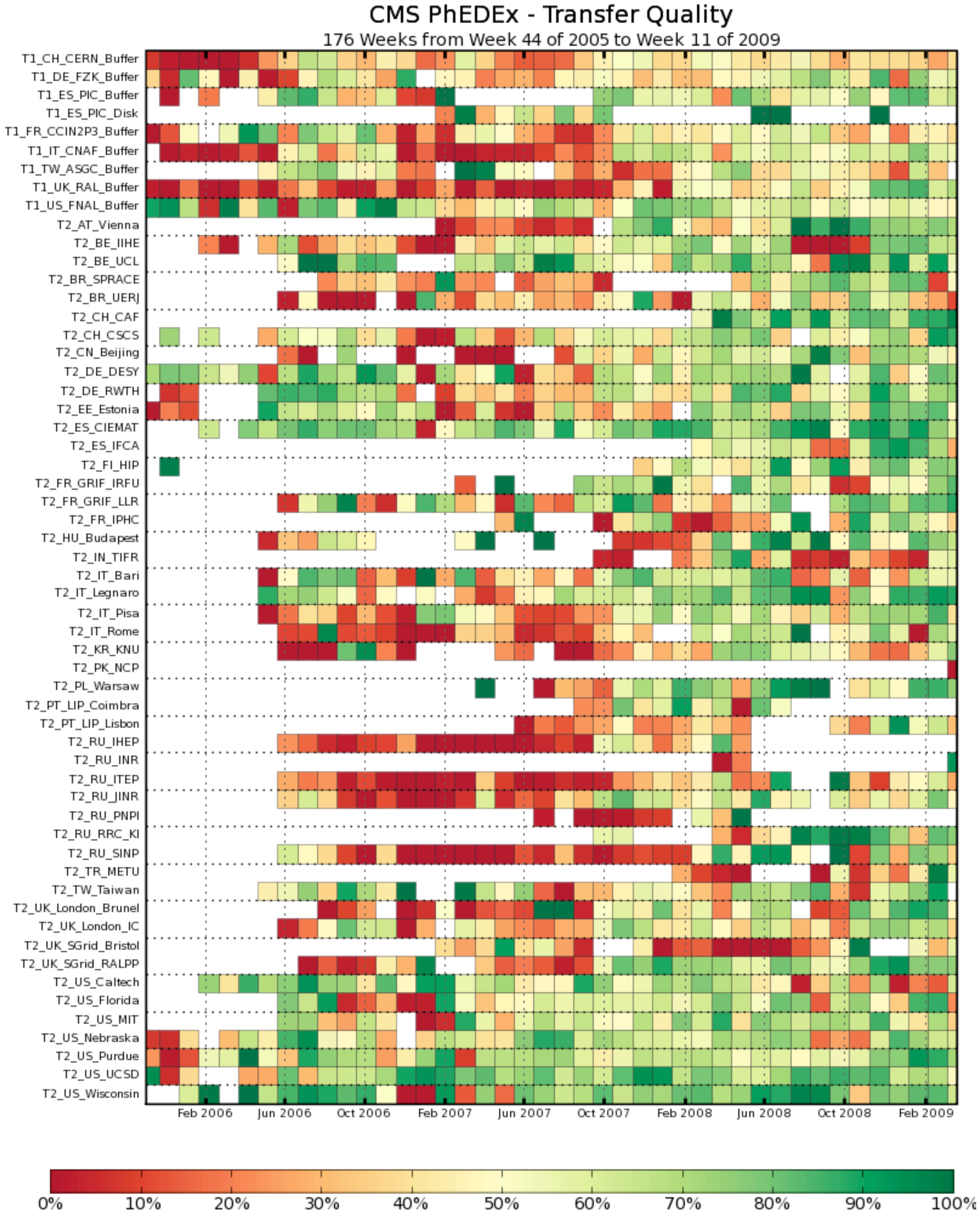

Fig 5. Improvement of the data transfer quality from 2006-2009.

The status of commissioned links as of March 2009 is as follows:

- 35 Tier-2 sites have all of their downlinks from Tier-1 sites COMMISSIONED, and 2 more have seven out of eight links COMMISSIONED
- 43 Tier-2 sites have at least two COMMISSIONED uplinks to Tier-1 sites.
- All Tier-1 to Tier-1 links are COMMISSIONED.
- Transfer quality continues to improve, as seen in Fig. 5.

The experience of the DDT task force showed that dedicated computing campaigns 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 [11,12].

In conclusion, a focused effort to debug transfer links within CMS proved to be useful in helping to maintain a working system for data transfers, documenting common problems and solutions, and alerting site administrators to problems. This effort is continuing with requirements and testing exercises that more closely match the CMS computing model and expected data transfer patterns at the start of data taking later this year.

References

[1] See for example the CSA07 website: https://twiki.cern.ch/twiki/bin/view/CMS/CSA07

[2] "The CMS Computing Model", CMS NOTE/2004-031.

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

[4] R. Egeland et al., "Data Transfer Infrastructure for CMS Data Taking", ACATier-08, Erice, Italy, November 2008.

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

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

[7] M. de Riese et al., "The dCache Book", http://www.dcache.org/manuals/Book

[8] O Bärring et al., "CASTOR2: Design and Development of a Scalable Archtecture for a Hierarchical Storage System at CERN," CHEP07, Victoria, B.C., Canada, September 2007. See also the CASTOR project website: http://castor.web.cern.ch/castor/docs.htm

[9] A. Frohner et al., "Data Management in EGEE," CHEP09, Prague, Czech Republic, March 2009.

[10] DDT project information can be found at: https://twiki.cern.ch/twiki/bin/view/CMS/DDT

[11] S. Belforte et al., "The Commissioning of CMS Computing Centres in the Worldwide LHC Computing Grid", NSS-IEEE, Dresden, 2008.

[12] S. Belforte et al., "The Commissioning of CMS Sites: Improving the Site Reliability," CHEP09, Prague, Czech Republic, March 2009.