The CMS Data Transfer Test Environment in Preparation for LHC Data Taking

G. Bagliesi[1,2], L. Bauerdick[3], S. Belforte[4], K. Bloom[5], B. Bockelman[5], D. Bonacorsi[6], C. Brew[7], J. D'Hondt[8], R. Egeland[9], S. Elgammal[10], F. Fassi[11], I. Fisk[3], J. Flix[12], J. M. Hernandez[13], M. Kadastik[14], J. Klem[15], O. Kodolova[16], C.-M. Kuo[17], J. Letts[18], J. Maes[8], N. Magini[6,2], S. Metson[19], J. Piedra[20], N. Pukhaeva[11], G. Qin[21], P. Rossman[3], A. Sartirana[6], J. Shih[21], S. Sõnajalg[14], D. Teodoro[22], A. Trunov[23], L. Tuura[24], P. Van Mulders[8], T. Wildish[25], Y. Wu[3], F. Würthwein[18]

***Abstract*—The CMS experiment is preparing for LHC data taking in several computing preparation activities. In distributed data transfer tests, in early 2007 a traffic load generator infrastructure 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 Centres. The LoadTest is based upon PhEDEx as a reliable, scalable dataset replication system. In addition, a Debugging Data Transfers (DDT) Task Force was created to coordinate the debugging of data transfer links in the preparation period and during the Computing Software and Analysis challenge in 2007 (CSA07). 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 experiences with the overall test transfers infrastructure within computing challenges - as in the WLCG Common-VO Computing Readiness Challenge (CCRC'08) - as well as in daily testing and debugging activities are reviewed and discussed, and plans for the future are presented.**

I. Introduction

THE CMS Experiment [1] is one of 4 large particle physics experiments at the LHC accelerator at CERN, Geneva, Switzerland that is presently being commissioned to resume data taking in 2009. To archive and analyze its data, CMS and the other LHC experiments depend on the Worldwide LHC Computing Grid (WLCG) [2], a worldwide distributed data grid of over 150 compute and storage clusters. Individual clusters vary both in size (10 TB to few PB) as well as expertise of their operations teams. The several hundred end-to-end links between these sites needed to be commissioned. 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 [3]. 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 operational end-to-end links to the CMS Data Operations team 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. Activity of the DDT Task Force resumed in 2008 and continued in the period leading up to and beyond the WLCG Common Computing Readiness Challenge (CCRC'08) [4]. The CCRC'08 challenge was designed to test the readiness of the WLCG to sustain the workflows of all 4 LHC experiments concurrently at the full design scale for the start of LHC data taking in 2008.

This note details the activity of the DDT task force. Section II describes the task force charge and scope. Section III describes some of the details of the CMS Computing Model relevant to this task force. Section IV describes briefly the system components used to transfer data across the wide area network. Section V details the metric used to commission links. Section VI discusses the documentation effort within DDT and related projects like PhEDEx [5], which is the main data transfer tool in CMS. Section VII categorizes problems found in commissioning links or keeping them commissioned over time. Section VIII discusses the activities of the DDT Task Force during 2008. In Section IX, the performance of transfer routes between CMS sites during CCRC'08 transfer tests is presented. Conclusions and future plans are discussed in Section X.

Manuscript received November 14, 2008.

1. INFN-Pisa, Italy 2. CERN, Switzerland 3. Fermilab, USA 4. INFN-Trieste, Italy 5. University of Nebraska-Lincoln, USA 6. INFN-CNAF, Italy 7. Rutherford Appleton Laboratory (RAL), Didcot, UK 8. Vrije Universiteit Brussel (VUB), Belgium 9. University of Minnesota, USA 10. Université Libre de Bruxelles (ULB), Belgium 11. CC-IN2P3, Lyon, France 12. Port d'Informació Científica (PIC), UAB, Barcelona, Spain 13. Centro de Investigaciones Energeticas Medioambientales y Tecnologicas (CIEMAT), Madrid, Spain 14. National Institute of Chemical Physics and Biophysics (NICPB), Tallinn, Estonia 15. Helsinki Institute of Physics, Finland 16. Institute of Nuclear Physics, M.V. Lomonosov Moscow State University, Moscow, Russia 17. National Central University (NCU), Chung-li, Taiwan 18. UCSD, La Jolla, CA USA 19. HH Wills Physics Laboratory, Bristol, UK 20. MIT, Boston, USA 21. Academia Sinica Grid Computing (ASGC), Taiwan 22. Universidade do Estado do Rio de Janeiro (UERJ), Brasil 23. Forschungszentrum Karlsruhe (FZK), Germany 24. Northeastern University, Boston, USA 25. Princeton University, USA

II. 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,
- Definition of the procedure to commission a link, 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,
- Creating a table that keeps track of the matrix of status of all links,
- Reporting weekly on the status of this matrix.

III. The CMS Computing Model

The CMS computing model [6] 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 (T0), 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. The T1 centers are typically at national laboratories with large computing facilities and archival storage systems.
- 45+ Tier 2 (T2) 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.

There are 44 "active" T2 centers, meaning that a T2 site successfully tested at least one data transfer link according to the procedures described in the following. There are also additional CMS T2 centers that have not yet succeeded in testing at least one link.

The CMS computing model envisions commissioning all links between:

- CERN to T1 sites, and T1 sites to CERN (14 links)
- All other T1-T1 cross-links (42 links)
- All T1 to T2 downlinks (352 links)
- All T2 to "regional" T1 uplinks (44 links)

Therefore, the total number of links to be commissioned in the computing model is 452. This number will increase with the addition of new sites, 9 links per new T2 site.

T2 to non-regional T1 uplinks are not a priority but were commissioned if the sites wished, or when needed by the CMS Data Operations team. Each T2 is associated to a T1 (called the "regional" or "associated" T1), although in some cases this T1 is not geographically near the T2.

T2-T2 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 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.

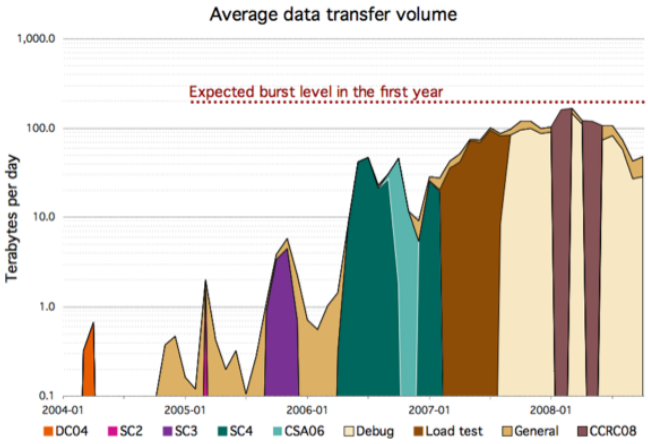

Fig. 1. Total data transfer rate for all links in all instances of PhEDEx. The LoadTest/Debug traffic and the CCRC'08 challenges are highlighted.

Likewise, links that begin or end at a Tier 3 (T3) site are not in the computing model. A T3 site is typically a small or medium-sized computing facility associated with a T1 or a T2, usually at a university or research institute. These links were commissioned on request of the T3 site.

The overall order of the program of work for the DDT Task Force was discussed at the management level, and after several iterations it was decided that the priority of debugging effort was to be:

- CERN-T1 and T1-CERN links
- All other T1-T1 links
- T1 to and from associated T2 links, establishing at least one link per T2 in each direction so that the T2 site is useful to Data Operations for data analysis and Monte Carlo production activities
- T1 to non-regional T2 sites
- T2 to non-regional T1 sites.

IV. System Components

A. PhEDEx, the Data Transfer Middleware

PhEDEx [5] 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 CMS 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 [7] 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 tunable from a web interface. The files that are injected are in fact linked to a data set typically of 256 files at each site, so that files are transferred multiple times without the need to constantly create new files.

Figure 1 shows the total transfer rate in PhEDEx in all instances. The activation of the LoadTest and the Debug

instance in 2007 brought a transition from a challenge-driven traffic to a constant load.

B. Storage Elements

The CMS T1 and T2 sites use various storage element (SE) systems, which are briefly described:

- The dCache storage system [8] was developed at DESY and FNAL. The system gets support from both US OSG storage support group as well as dCache User Forum in Europe. It is used by FNAL, PIC, FZK and IN2P3 as the T1 storage backend and by a majority of the active T2 centers 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 [9] storage system was developed at CERN and is used by the T1 centers at CERN, CNAF, RAL and ASGC. There are no T2 sites that use this storage system.
- Developed as part of the gLite middleware, DPM [10] is mostly used by smaller T2 centers. About 10 T2 sites use DPM as their storage solution.
- StoRM [11] was developed by INFN and is a grid Storage Resource Manager for disk-based storage systems. StoRM is designed to work over native parallel filesystems, but supports also standard Posix file systems. Currently, StoRM is used by 2 T2 sites and by the T1 center at CNAF for disk-only storage.
- BeStMan [12] was developed by Lawrence Berkeley National Laboratory, for disk-based storage sytems and mass storage systems. It works on top of existing disk-based unix file systems. It is currently used at 1 T2 site and at some T3 sites in the United States.

The T1 sites also have tape backends to their storage systems, which were not exercised in this project.

All of these different storage element systems are accessible through the Storage Resource Manager (SRM) [13] interface, a middleware service providing uniform transparent access to storage management capabilities irregardless of the underlying technology. The SRM implementation used by CMS in 2007 was version 1.1. Deployment within CMS of SRM version 2.2 started near the end of CSA07, and was completed during CCRC'08 phase 1 in February 2008.

C. Transfer Protocols

File replicas located on SRMs are identified by a Storage URL (SURL) detailing the SRM hostname and the location of the file on the storage element (possibly in a virtual filesystem). When a file transfer between two SRMs is requested, each of the SRMs provides a Transfer URL (TURL) for the requested protocol which is only valid only for the duration of the transfer. The protocol normally used for transfers between SRMs is the GridFTP protocol, offering the functionalities of FTP, but with additional support for grid security and for parallel stream data transfer.

Transfers between some sites are made with 3-rd party transfers submitted directly to the SRMs. However, where load is an issue, transfers are scheduled by the File Transfer Service (FTS) [14]. FTS was developed as part of gLite middleware and is 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 that is utilized for this specific source/destination combination. FTS allows sites to set limitations on the number of files in transfer, number of streams used etc. FTS channel managers may specify shares on a channel to balance transfer rates between different VOs. Each FTS channel can operate in one of two modes: 3rd-party SRM batch transfer (SRMCOPY) and 3rd-party single file GridFTP transfer (URLCOPY). In SRMCOPY the transfer is handed to one of the SRM servers as a srmCopy request. The srmCopy request is handled by the source SRM in SRMCOPY PUSH mode, while in SRMCOPY PULL mode it is handled by the destination SRM. In URLCOPY FTS mediates the transfer by negotiating transfer URLs from both SRM servers, and then requests a 3rd-party GridFTP transfer from one of the SRM servers for the file. In CMS, FTS SRMCOPY was used at CNAF and FNAL.

FTS servers are deployed at CERN and T1s.

D. Networking

No major issues with networking handicapped 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 the reasons fully. 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 untuned default kernel configurations that do not favor high-speed long distance transfers, but are sometimes tuned for the requirements of the storage elements at the sites.

Most T1 sites are interconnected through 10 Gbps networks, although they serve T2 sites that are often connected through national network infrastructures with a more limited capacity (1-2 Gbps). This imposes limitations on some regions. In addition, some T2 sites in Russia and India have Gbps connectivity only to CERN, preventing commissioning of links with other T1s.

V. Commissioning Progress

A. 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

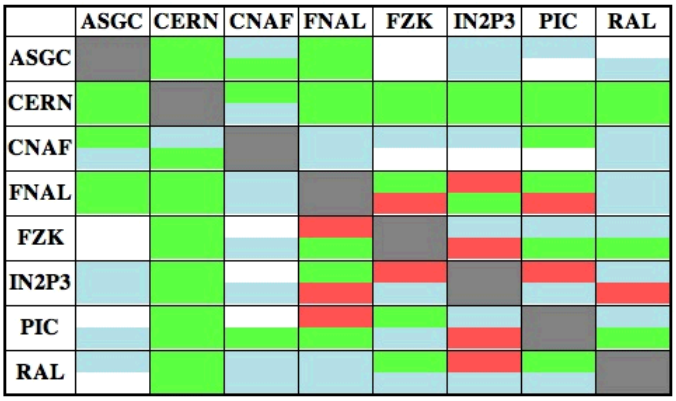

Fig. 2. Link Status Matrix. The green links were COMMISSIONED, the red links those that were commissioned and developed problems, the blue in the process of commissioning, and the white untested.

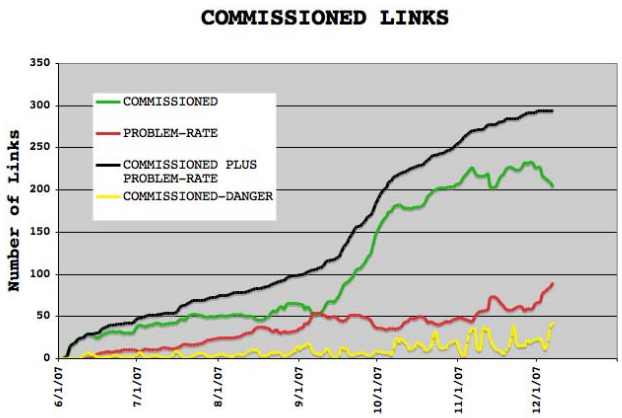

Fig. 3. History of the status of COMMISSIONED links between all sites. The green line shows the number of COMMISSIONED links, the red line the number of PROBLEM-RATE links, and the black line the sum of the two, being the total number of links to ever commission during the exercise. The yellow line shows the number of COMMISSIONED links that were in danger of decommissioning within the next two days.

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, a link must:

- 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 T2, this requirement is 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 T2 sites.

- **PROBLEM-RATE**, for links that were working but whose rate has dropped off. To remain COMMISSIONED, a link must transfer at least 300 GB/day for a single day at least once every 7 days. Otherwise, the link must 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.

B. Monitoring of Link Status

A tool was developed by B. Bockelman and S. Sõnajalg, a CERN summer student, 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.

C. History of link commissioning

The number of commissioned links fell at each of the following major events, as can be seen in Figure 3:

- August Vacations, showing that systems needed constant monitoring in order to function stably.
- Transfer of all LoadTest activities to the Debug instance of PhEDEx in September.
- Major change to the injection procedure for the LoadTest in early October.

The largest 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.

The quality of PhEDEx transfers in the instance running the LoadTest is shown in Figure 4 for a time period before DDT and CSA07 and for the last month of CSA07. Clear improvement in the number of active links and the overall transfer quality is seen.

VI. Documentation

One of the major charges of the DDT task force was the documentation of common problems and solutions. This section details the documentation produced by DDT. A detailed study of the different types of problems encountered is given in the following section.

A CERN Hypernews forum was used to keep a record of DDT related activities and discussions [15], in particular the link debugging work and the organization of the task force activity. The forum formed into a troubleshooting knowledge base as it was the main channel for reporting DDT-related data transfer problems and the solutions discovered.

The DDT Twiki [16] was the nexus for general documentation and in particular of the information directed at the participating sites. Procedures developed by the task force had their own pages and the reports given by the task force were collected in a single location. The most challenging task for a site administrator, bringing up the first link, was devoted a set of pages. The subjects included setting up the transfer links, configuring FTS transfers and creating the LoadTest file samples in PhEDEx.

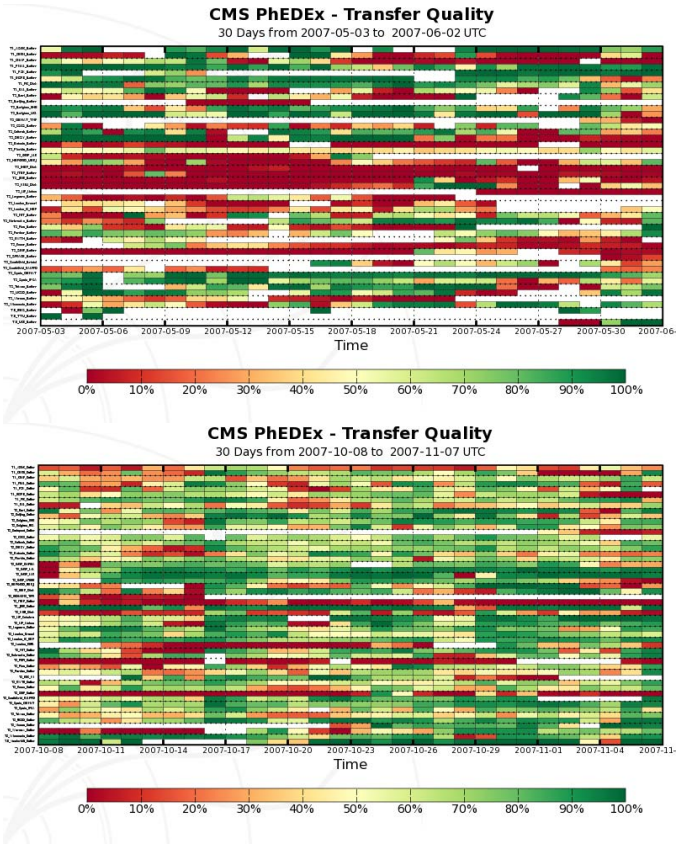

Fig. 4. Data transfer quality in the month before the DDT Task Force project began (upper plot) and during the last month of CSA07 (lower plot). A clear improvement in the number and quality of data transfer links is seen. The plots show the percentage of error-free file transfers.

Outside of DDT, an effort was started by the PhEDEx developers to clarify the possible outcomes of the PhEDEx transfers. In general, better error messaging in PhEDEx, FTS and SRM would be a big help, and at least with regard to PhEDEx some of the work has been done.

VII. Common Challenges, Problems and Solutions

The DDT Hypernews was used for alerts and reminders to sites, and to discuss real technical problems that arose and affected data transfers. There are some common problems that were solved, and some that remain and still affect data transfers regularly.

A. Storage Element I/O Limitations

Considering first challenges that are ongoing, the most important is the limitations on storage systems reported by the sites. This affects other aspects of the data transfers including the management of the LoadTest and created a lot of manual work for the sites, especially the T1 sites. T1 sites have different limitations on I/O into their storage elements. For example, if a T1 site has a real limit of 80 MB/s for CMS transfers, and wants to commission 14 links to T1s at 5 MB/s each in the LoadTest, this exhausts the capacity of the site for transfers. Generally such a site would have at least a 1 Gbps if not a 10 Gbps network connection, so it's clear that the limitation comes from the SE and/or SRM, not the network. Such limitations caused sites to juggle LoadTest traffic between their required links, causing a lot of extra manual work for the site administrator. Since an aggregate transfer rate of 80 MB/s is well below CMS requirements, its clear that this is an important area for improvement. This type of problem affects sites that run dCache as well as CASTOR.

B. Other Storage Element Problems

The following problems were common with dCache:

- Under heavy load directories ended up owned by "root", disabling access to subdirectories. This has been fixed in more recent dCache versions.
- Certificate revocation lists (CRLs) update problems resulted in "Unknown CA" errors. Requires disk pool restarts.
- Longer than expected downtime for dCache version upgrades.
- Disruptions from hardware failures and pool losses.
- Corrupted and lost files, including partial or total loss of entire data sets and even entire file systems.

The following problems were common with CASTOR:

- Long disruptions at CASTOR T1's due to problematic upgrades.
- FTS channels became congested by transfers lingering in "Cancelled" state. One significant cause was GridFTP doors having problems canceling transfers.
- Stager interface and overload errors.
- Direct srmCopy transfers, especially those by srmcp, tended to overwhelm CASTOR. Such problems were rare during CSA07 as all CASTOR sites required the use of FTS and URLCOPY mode for transfers.

The main DPM-specific issue encountered in the DDT forum was that DPM does not currently support srmCopy requests, however it can carry out transfers in SRMCOPY mode if they are started by the SRM server at the other end of the channel. Therefore FTS channels ending at a DPM destination site should be configured in URLCOPY or SRMCOPY PUSH mode.

C. Centralized Injection of LoadTest Files

The introduction of centralized LoadTest injection in October 2007 was one of the major improvements to LoadTest and solved most of the related problems. Before October there were many threads in the DDT Hypernews concerning LoadTest injection problems and blocked or full transfer queues caused by too many injected files. The introduction of centralized injections with an easy-to-use web interface almost completely eliminated this type of problem.

D. PhEDEx Agent Misconfiguration

Another type of problem that arose very often concerned PhEDEx configuration. Specifically, there were problems involving:

- Agent misconfiguration/inconsistencies

- Multiple agents running the same tasks
- Expired credentials
- Agents down

Although the PhEDEx web pages provide extensive monitoring information, there is no active monitoring. In the absence of any standard active monitoring, many sites have developed their own private PhEDEx active monitoring, informing the site administrator of agents that go down, proxies that are about to expire, etc.

In addition, sanity checks of agent configuration could be useful for site administrators. However, such inconsistencies are usually problems that occur once and are not repeated at the same site. We have been made aware of work in progress already on this point.

E. SRM and FTS Timeouts

SRM and FTS transfer timeouts are a very common problem for sites, in particular transfers that start but never complete. These are not exclusively problems with the FTS or SRM server in question, but also result from problems with the other SRM party, another service at either site, or the network. The issues are difficult to understand and diagnose mainly due to a lack of good error messages, and in particular a lack of detail on the action that was being executed, which computer and service was attempting to talk to which other computer and service and why. Better error reporting and some use of SRM and FTS error codes would be useful. CMS began a classification of PhEDEx error codes during CSA07.

Common problems with SRM included:

- Long delays and timeouts in case of high load.
- Most of the problems with acquiring source or destination TURL were storage subsystem specific issues, not related to SRM implementation itself.

Common problems with FTS:

- Complexity of the system has caused a number of issues (it does need Oracle and has a number of components which can cause problems, tracking them down has always been very time consuming and not straightforward). FTS is also quite complicated to configure the first time.
- Canceling problem: if for what ever reason the storage job can not be cancelled fully in the first attempt it tended to remain in the "Canceling" state and hence be counted as an active job, which lead to channels filling up with canceling jobs and causing the transfers to grind to a halt. A number of causes have been identified including FTS bugs, but in general it caused a lot of problems during CSA07 preparation and the challenge itself
- Poor logging access: for the end user it is almost impossible to get adequate reasons from FTS on why a job failed, the only adequate logs are those which are available on the server side of FTS, which makes debugging very tedious.
- "Error in bind()", in which FTS cannot retrieve the proxy from the myproxy server because of too many connections.
- Appearance of SRM authentication errors caused by the corruption of the delegated credentials on the FTS server.
- Site discovery problems (configuration issue): as FTS is native to gLite and not to OSG, we had a number of issues with site discovery as the different US sites didn't show up in the FTS server configuration due to requirements on data in the site BDII which was not by default configured in OSG. A large number of tests and a long cycle of debugging which took weeks finally resolved the issue down to the specific details allowing the US sites to participate in FTS transfers

F. Authentication Errors

Another source of error is authentication, usually read access at a remote site SE. Transfers fail because the DN of the transferring site administrator is not valid on a source site. For example, T2_US_UCSD allows read access to the SE for all DNs with a valid CMS VOMS extension. However, this is not a standard. The creation of a standard requirement for this was discussed during and after CSA07 and would be very helpful.

G. Other Problems

Most other problems concerned various services or hardware failures, and were typically a single occurrence at sites. Data corruption, deletion, operator error, hardware failures, network outages etc. will always occur. Monitoring of such problems, either within PhEDEx or general site monitoring is the only way to communicate the occurrence of such failures to administrators. DDT tried to monitor things that affected data transfers, but it is clear that the best monitoring is monitoring closer to the source of the problem, i.e. at the sites themselves. The more successful sites were those that managed to find and fix problems promptly. This is as much an issue of monitoring as effort and manpower.

VIII. Debugging Data Transfers in 2008

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 work of the DDT Task Force concluded with the end of the CSA07 service challenge in November 2007. However, the effort was considered useful and resumed in 2008 in a modified form.

Firstly, the metric was modified to more closely match the CMS computing model requirements. To commission links in 2008, a data link must transfer at a rate of at least 20 MB/s averaged over 24 hours. Recognizing that uplinks from T2 to T1 sites have a lower requirement in the computing model, they are only required to transfer 5 MB/s. A steady influx of new links commissioned according to the 2008 metric, leading to the following status in September 2008:

- 56 CERN-T1 and T1-T1 crosslinks are currently COMMISSIONED, representing 100% of the total
- 233 T1-T2 downlinks are in COMMISSIONED status, representing 66% of the total mesh. 36 T2s have at least 2 commissioned downlinks as required for CMS site commissioning

- 41 T2-T1 regional uplinks are COMMISSIONED, representing 85% of the total. In addition, 74 non-regional T2-T1 links are also COMMISSIONED
- 9 T2-T2 cross-links are also COMMISSIONED

In total, 413 links had achieved commissioning according to the 2008 metrics in September 2008. The major events leading to commissioning of new links were:

- Commissioning of the missing T1-T1 links was quickly completed by March
- Many new non-regional T1-T2 downlinks were commissioned in April in preparation for the CCRC'08 phase 2 Data Transfer Tests, in collaboration with T1 site administrators
- A concerted effort by the DDT team was carried out in June to complete commissioning of missing non-regional links to T2s which had already most of their links commissioned, and to bring new sites to commission their first links.

Secondly, it was recognized by the CMS computing management that continual exercising of transfer links placed a large burden on site administrators and storage systems that is not required in the computing model. In fact, 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 have been changed. In 2008, a testing program was organized 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 T2-T1 uplink. Links were only decommissioned if they failed this transfer exercise for three days, or developed an obvious problem and a request came from Data Operations to disable the transfer link. Link exercises were suspended during periods of Production activity, holidays and announced site downtimes. Following a first test round of link exercises in February 2008, three full rounds of link exercises were carried out by the DDT task force: a first round in February-March 2008, a second round in April 2008, and a third round in June-July 2008. The third round was carried out injecting LoadTest files at 50 MB/s to gather statistics on how many links were able to meet this target rate, although the metrics to mantain commissioning stayed at 20 MB/s. During the first round of transfer exercises, about 1/3 of the links under exercise had real issues, but most of the sites were able to solve them during the exercise time window of 3 working days, or to recommission the link in the following weeks, while 7% of the links were unable to meet the new target metrics and were decommissioned. In the second round in April, fewer links encountered problems during the exercises, leading to only 2% of the links being decommissioned. In the final round of link exercises in June-July, only 2 T1-T2 downlinks failed to meet the base metrics of 20 MB/s to mantain commissioning, while 20% of the T1-T2 downlinks under testing exceed the extra-target of 50 MB/s. The periodic link exercising ended in July 2008, and the LoadTest injection rate on commissioned links is now set to 0.25 MB/s (corresponding to 1-2 files transferred every couple of hours) for monitoring.

IX. Data Transfer tests during the Common Computing Readiness Challenge

The Common Computing Readiness Challenge (CCRC'08) was established to test the readiness of the Worldwide LHC Computing Grid to sustain the production needs of all 4 LHC experiments concurrently. The challenge ran in two phases in February and May 2008, respectively. During the February test, the workflows were exercised at a reduced scale compared to the data taking requirements. In addition, some of the middleware components were still in the process of deployment. In particular, many sites were still running a storage element that only supported version 1 of the SRM protocol at the beginning of February; nonetheless, by the end of February most sites had upgraded their storage elements to support version 2.2 of the SRM protocol and were able to participate successfully in the challenge.

During the second phase in May, the experiment workflows were exercised at full scale using a nearly final version of the deployed middleware. All aspects of the workflows of the 4 LHC experiments were tested during the challenge; in this section, we will focus on the Distributed Data Transfer tests performed by CMS during the period of the challenge.

A. T0-T1 Transfer Tests

In phase 1 in February [17], T0-T1 transfers were tested at a rate of 50% of the nominal rates required for data taking in 2008. Three target performances were set for transfers from T0 to the disk buffers at T1s: a minimum target at 25%, a base target of 40%, and an optimal target of 50% of the nominal rate. Migration from disk buffers to tape at the T1s should exceed 25% of the nominal rates, sustained for a minimum of 2 days (3 days optimally). In addition, T1s should demonstrate to be able to sustain the full chain of transfers from T0 to disk buffers to tape for a constant flow of an amount of data equivalent to 3 days of data taking at 40% of nominal rates. All 7 T1s demonstrated the ability to sustain data transfers at the required rates and met the target metrics. In the second phase of CCRC'08 in May [18], the target was raised to the nominal rate of 600 MB/s for 2008 data taking, with the aim to reach an extra-target of 850 MB/s in some periods of the challenge. The goal was to sustain the target rate for 3 days in a row during 2 distinct weeks in May. Individual target rates for each of the 7 T1s were specified and measured, but in May the main target was to mantain the overall outbound rate from CERN. Figure 5 shows the overall performance of transfers from CERN in May. Both the target and extra-target overall rates were achieved. During CCRC'08, the ability to sustain concurrent transfers between the different experiments was also verified, with a total outbound rate from CERN of 2 GB/s achieved for 2 days.

B. T1-T1 Transfer Tests

Transfer tests between T1s in phase 1 aimed at achieving a wide participation of T1 sites, and aggregate inbound/outbound targets were defined. T1s were required to demonstrate 50% of their overall 2008 outbound and inbound

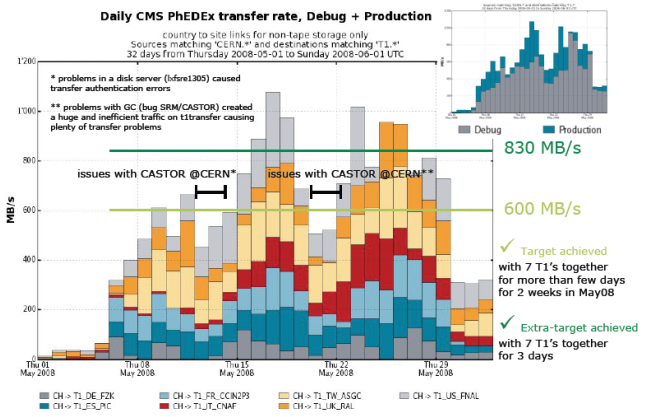

Fig. 5. Data transfer rate from CERN to T1s by destination site (main) and PhEDEx instance (inset) during CCRC'08 phase 2 in May 2008.

rates to at least 3 T1s, among which at least 1 not in their continent. The target rates were required to be sustained for a minimum of 2 days with an optimal target of 3 days. All of the T1s were able to meet the target rate in phase 1 except for FNAL, where export rates similar to those from CERN would have been required to meet the metrics. Therefore, in preparation for phase 2 the T1-T1 replication schema was revised and rebalanced to more closely match the experiment requirements. In particular, it was realized that it is not necessary for an individual source T1 to demonstrate the nominal metrics in order to distribute the reprocessed datasets to the other T1s in a timely manner: as soon as a fraction of the replicated dataset is available on some of the destination T1s, PhEDEx will start to reroute transfers and all T1s will contribute to the replication. In phase 2, the target for T1-T1 transfers was then changed to a latency target: each T1 was required to replicate to all other T1s within 4 days a dataset of appropriate size. All of the T1 sites achieved the extra-target of replicating the full dataset, and monitoring of the fraction of the transfers that happened through rerouting revealed that indeed other T1s contributed to the transfers.

C. T1-T2 Transfer Tests

In T1-T2 tests in phase 1, we aimed to demonstrate an aggregate T1-outbound target to regional T2s only. The LoadTest infrastructure was used to meet the target figures, which were based on the amount of storage resources at the T1s and associated T2s. All of the regions involved in the test nearly met or exceeded the targets, except for the Asian region, where the infrastructure at most of the T2s was not yet ready to transfer from the regional T1 at ASGC. In phase T2, the tests for T1-T2 transfers were extended to the full mesh of regional and non-regional links. In addition, Monte Carlo data from the Production instance of PhEDEx was used instead of the fake load generated by the LoadTest. This required the exclusive use of links commissioned by DDT, pushing many T2 sites to complete the commissioning. Target figures were aggregated by destination T2 region, for example from the IN2P3 T1 to all T2s in the Spain-Portugal region. Two rounds of transfer tests were peformed during May, involving more T2s and T2 regions in the second round. In the end, 90% of the links commissioned by DDT were involved in the transfer tests, and 53 out of 64 possible T1-"T2 region" combinations were tested successfully, while the remaining combinations had no (or not enough) commissioned links available for the test. In addition, since T1-T2 transfers are expected to happen in burst patterns, peak rates were also measured, and peaks of up to 38x above average rates were observed.

D. T2-T1 Transfer Tests

In both phase 1 and phase 2, uplinks from T2s to regional T1s were tested with the LoadTest, achieving in all cases rates well above the targets required for the upload of the results of Monte Carlo production to T1s for archival storage.

X. Conclusions

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 current problems. This effort continued in 2008 with requirements and testing exercises that more closely match the CMS computing model and expected data transfer patterns when data taking will resume next year. In addition, the CCRC'08 has demonstrated that the data transfer infrastructure has reached the scale of performance required for LHC data taking.

References

[1] CMS Collaboration, *CMS, the Compact Muon Solenoid: technical proposal*, CERN-LHCC-94-38.
[2] J. Knobloch et al., *LHC Computing Grid Technical Design Report*, CERN-LHCC-2005-024.
[3] CSA07 website: https://twiki.cern.ch/twiki/bin/view/CMS/CSA07
[4] CCRC'08 website: https://twiki.cern.ch/twiki/bin/view/LCG/WLCGCommonComputingReadinessChallenges
[5] *PhEDEx – CMS Data Transfers*, PhEDEx documentation website: http://cmsweb.cern.ch/phedex/documents.html
[6] C. Grandi et al., *The CMS Computing Model*, CMS NOTE/2004-031.
[7] G. Bagliesi et al., *The CMS LoadTest 2007: An Infrastructure to Exercise CMS Transfer Routes among WLCG Tiers*, submitted to CHEP'07 Conference, Victoria B.C., Canada, 2-9 September 2007
[8] M. de Riese et al., *The dCache Book*, http://www.dcache.org/manuals/Book
[9] CASTOR project website: http://castor.web.cern.ch/castor/docs.htm
[10] DPM project website: https://twiki.cern.ch/twiki/bin/view/LCG/DataManagementDocumentation
[11] StoRM project website: http://storm.forge.cnaf.infn.it/home
[12] BeStMan project website: http://datagrid.lbl.gov/bestman/
[13] SRM project website: http://sdm.lbl.gov/srm-wg/
[14] FTS project websites: http://egee-jra1-dm.web.cern.ch/egee-jra1-dm/FTS/, https://twiki.cern.ch/twiki/bin/view/EGEE/FTS
[15] The Hypernews list address is hn-cms-ddt-tf@cern.ch and the web address is https://hypernews.cern.ch/HyperNews/CMS/get/ddt-tf.html.
[16] https://twiki.cern.ch/twiki/bin/view/CMS/DebuggingDataTransfers
[17] CCRC'08 phase 1 CMS Distributed Data Transfers website: https://twiki.cern.ch/twiki/bin/view/CMS/CCRC08-PhaseI-TestTransfersOperations
[18] CCRC'08 phase 2 CMS Distributed Data Transfers website: https://twiki.cern.ch/twiki/bin/view/CMS/CCRC08-DataTransfers