PROCEEDINGS OF SCIENCE

INFN Tier-1 experiences with Castor-2 in CMS computing challenges

Daniele Bonacorsi*
(on behalf of INFN-CNAF Tier-1 and the CMS Collaboration)
INFN-CNAF, Viale B. Pichat 6/2, 40127, Bologna, Italy
E-mail: Daniele.Bonacorsi@cnaf.infn.it

The CMS combined Computing, Software and Analysis challenge of 2006 (CSA06) is a 50 million event exercise to test the workflow and dataflow associated with the data handling model of CMS. It was designed to be a fully Grid-enabled, 25% capacity exercise of what is needed for CMS operations in 2008. All CMS Tier1's participated, and the INFN Tier-1 - located at CNAF, Bologna, Italy - joined with a production Castor-2 installation as a Hierarchical Storage Manager solution to address data storage, dat access and custodial responsibility. After the prompt reconstruction phase at the Tier-0, the data was distributed to all participating Tier-1's, and calibration/alignment, re-reconstruction and skimming jobs ran at the Tier-1's. Output of skimming jobs were propagated to the Tier-2's, to allow physics analysis job submissions. The experience collected by the INFN Tier-1 storage group during the pre-challenge Monte Carlo production, the preparation and the running of the CSA06 exercise - as well as the Tier-1 preparation activities for next CMS Computing challenges in 2007 - are reviewed and discussed.

XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research
April 23-27 2007
Amsterdam, the Netherlands

*Speaker.

PoS(ACAT)018

1. Introduction

The CMS experiment constructed a baseline Computing Model [1] and a Technical Design [2] for the computing system which is expected to be needed in the first years of the LHC running. The architectural design is a distributed system of computing services and resources that interact with each other in a Grid-enabled manner. The baseline architectures for the data management have been identified, along with the workflows involving the computing centres. It together comprises the computing, storage and connectivity systems that CMS will need to perform data transfer and archiving, data processing, Monte Carlo event generation, and any other computing-related activities. Significant attention was focused on the development of a data model with heavy streaming at the level of the RAW data based on trigger physics selections. We expect that this will allow maximum flexibility in the use of distributed computing resources.

2. Tiered architecture of computing resources

The CMS Computing Model makes use of the hierarchy of computing Tiers as proposed in the MONARC [3] working group and in the first Review of LHC Computing [4, 5]. The CMS computing resources are:

- a Tier-0 centre plus a CMS CERN Analysis Facility (CMS-CAF) located at CERN;
- 7 Tier-1 centres located at large regional computing sites;
- about 30 Tier-2 centres.

The computing centres available to CMS around the world are distributed and configured in a tiered architecture that functions as a single coherent system. Each of the three tier levels provides different resources and services, as outlined in Table 2.1).

2.1 Tier-0 and CMS CERN Analysis Facility (CMS-CAF)

Unique and located at CERN, the Tier-0 (T0) has several tasks: *i)* it accepts RAW data from the CMS Online Data Acquisition and Trigger System; *ii)* it archives the RAW data to tape; *iii)* it groups them into data streams and feeds the prompt first-pass reconstruction, producing the full event (FEVT) and in some cases also extracts a first-pass Analysis Object Data (AOD); *iv)* it classifies reconstructed (RECO) data into about 50 "primary" datasets, according to their physics content (e.g. trigger path); *v)* it makes such datasets available for transfers to the next Tier stage resources (i.e. the Tier-1s); *vi)* it maintains the CMS-CAF, that performs latency-critical activities (like detector diagnostics, trigger performance services, derivation of calibration and alignment constants). At CERN, the Tier-0 and the CMS CERN Analysis Facility (CMS-CAF) provide complementary functions for CMS computing, as different "logical" entities within a unique large "physical entity". Together, they provide the required CMS computing at CERN.

The Tier-0 is used for highly organized workflow to deal with quasi-realtime data-flows, including prompt calibration, prompt reconstruction, re-processing of express data streams, etc. It accepts RAW data from the CMS Online Data Acquisition and Trigger System; it archives the RAW data to tape; it groups them into data streams and feeds the prompt first-pass reconstruction,

Centre	CPU [MSI2k]	Disk [PB]	Tape [PB]
Tier-0	4.6	0.4	4.9
CMS CAF	4.8	1.5	1.9
Tier-1	15.2	7.0	16.7
Tier-2	19.3	4.9	-

Table 1: Resources needed at CMS Tiers to process and analyze data in 2008.

producing the full event (FEVT) and in some cases also extracts a first-pass Analysis Object Data (AOD); it classifies reconstructed (RECO) data into about 50 "primary" datasets, according to their physics content (e.g. trigger path); it makes such datasets available for transfers to the next Tier stage resources (i.e. the Tier-1's).

The CMS-CAF is for latency-critical activities and high-priority asynchronous access to data coming from Tier-0. It hence performs verification of detector and trigger performance and calibration, data quality assurance;it also allows rapid analysis of high-priority or Ôexpress lineÕ physics (5-10% of all data taken), and access for data analysis by users at CERN.

2.2 Tier-1

The Tier-1 centres provide the dedicated computing facilities to store a given share of CMS real and simulated data, which is "actively" used for reprocessing, skimming, event serving and other large-scale tasks requiring fast access to the bulk data. A major role of the Tier-1 centres is to provide permanent storage archiving capabilities for data and simulated data and allow high-throughput access for providing selected subsets of data to Tier-2 centres, where individual users will run analysis jobs.

In CMS collaborating countries, 7 large computing centres act as Tier-1 (T1) sites (ASGC in Taiwan, GridKA in Germany, INFN-CNAF in Italy, IN2P3 in France, PIC in Spain, RAL in U.K., FNAL in U.S.A.) and each of them is associated with a group of smaller Tier-2 (T2) centres. The tasks of a T1 centre in summary are: *i)* to receive some subsets of the primary datasets from the T0; *ii)* to provide tape archiving capability for FEVT; *iii)* to provide substantial CPU power for scheduled data-intensive tasks (i.e. skimming, re-reconstruction, calibration, AOD extraction); *iv)* to distribute RECOs, skims, AODs to a part of the next Tier stage resources (i.e. its associated group of T2s).

2.3 Tier-2

The CMS Computing Model requires a numerous set of computing centres (about 30), with consistent CPU resources but limited disk space (and no tape archiving), to provide the computing capacity for user analysis, calibration studies, and Monte Carlo production: these are the realm of the Tier-2 centres. T2's rely upon T1's for access to datasets and for secure storage of the newly produced data.

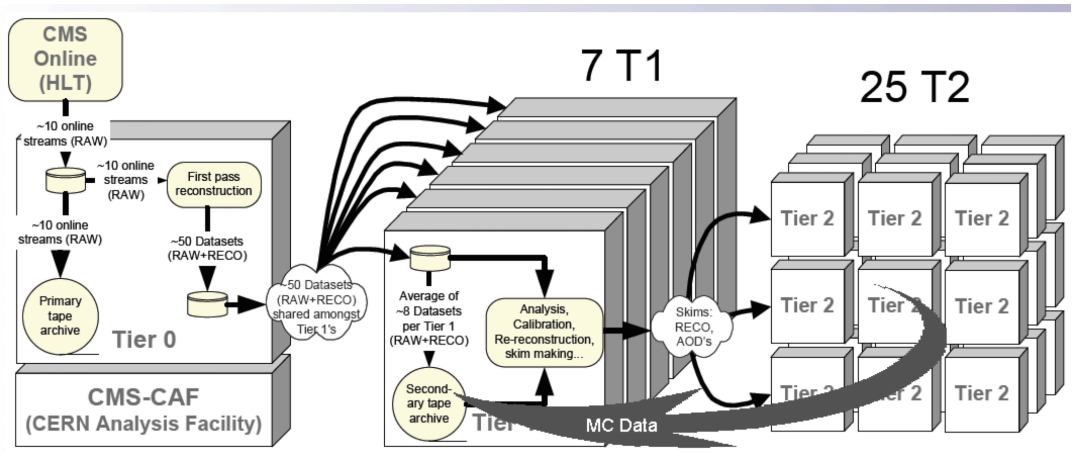

Figure 1: Pictorial view of Tiers in the CMS Computing Model.

The basic functions supported by a Tier-2 include: *i)* fast and detailed Monte Carlo event generation; *ii)* data processing for physics analyses, including late stage analysis requiring very fast data access; *iii)* data processing for calibration and alignment tasks, and detector studies.

3. CMS computing services and operations

The set of computing services and their behaviour together provide the CMS distributed computing system as part of the Worldwide LHC Computing Grid (WLCG). The general approach of CMS to developing the computing system is an iterative process of developing the system components, and integrating them together at successive steps of scale, in major "challenges" (both WLCG Service Challenges [6] and Data Challenges - DC04/CSA06/CSA07, Magnet Test and Cosmic Challenge - MTCC, readiness for data taking, etc) [7]. CMS adopted a loosely coupled system of services that can be improved upon and replaced with higher-performance, more-functionalities versions, while specifying well-defined interfaces and delegating functionality across the software stack. This approach allows CMS to commission increasingly functional and scalable systems even in the absence of a fully-defined engineering blueprint of all the components.

3.1 Guiding principles and system overview

In this section, the baseline CMS computing model [2] is described, regarding the guiding principles of the architecture and how the system components and computing services interoperate to address the basic experiment use cases and workflows needed to achieve the CMS goals at LHC turn-on.

Since CMS data (as in general event data in HEP) is written once, never modified and subsequently read many times, the optimization in the CMS computing system development must be for read access. Optimisation for the large bulk case, but without limiting a user from accomplishing basic tasks, must be implemented. Minimisation of the dependency of the jobs on the Grid Worker Node (WN) is implemented, so that the overall throughput of the jobs in the distributed computing system (running 24 hours a day, 7 days a week) can be made more stable and fault-tolerant. As

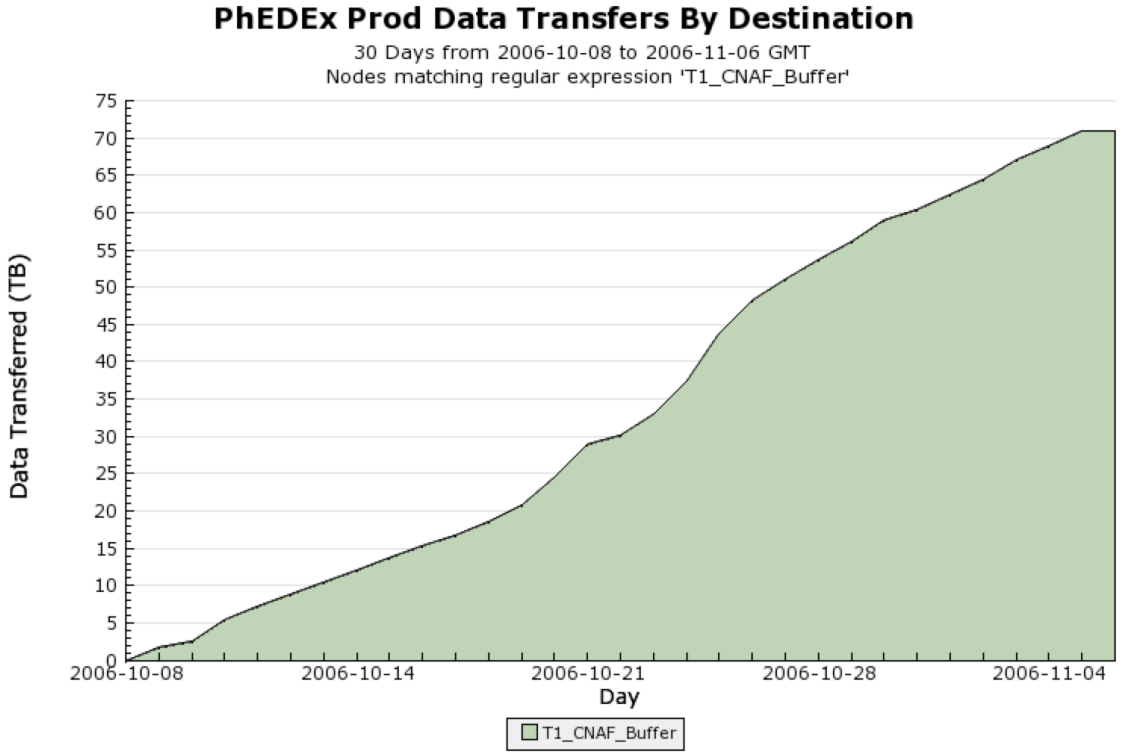

Figure 2: Overall amount of data transferred in the T0->INFN-T1 route during CSA06 (incremental view).

a requirement on the software framework and the computing infrastructure, it must be possible to track the provenance of datasets produced. Site-local configuration information should remain site-local, to add flexibility for the local site system administrator to configure and evolve the local system as needed without any synchronisation to the rest of CMS. A "keep the solution simple" principle is applied, to avoid paying the cost of complexity unless actually needed.

The overall architecture of the CMS computing system along with the most important systems and services can be divided into a *Grid Workload Management System*, a *CMS Data Management system* and other CMS-specific services, needed to support specific needs of experiment applications and software. A *CMS Workflow Management system* holds all of these pieces together into a coherent system supporting all CMS necessary workflows (data (re-)reconstruction, calibration activities, Monte Carlo production, AOD production, skimming and general user analysis) and shields users/operators of these systems from the full complexity of the underlying architecture.

3.2 Data transfers, data access and storage systems in CMS

CMS relies on the PhEDEx [8] project as a reliable, scalable dataset replication system. It manages transfer operations by automating many low level tasks and without imposing constraints on choice of Grid or other distributed technologies. The PhEDEx project addresses the CMS functional requirements for a system able to guarantee managed and structured data flow, multiple transfer modes, multiple priorities and scopes. The PhEDEx infrastructure comprises a set of "nodes", transfer points in the overall topology, operating independently as logical entities given the task of hosting "software agents", i.e. persistent stateless processes, responsible for undertaking

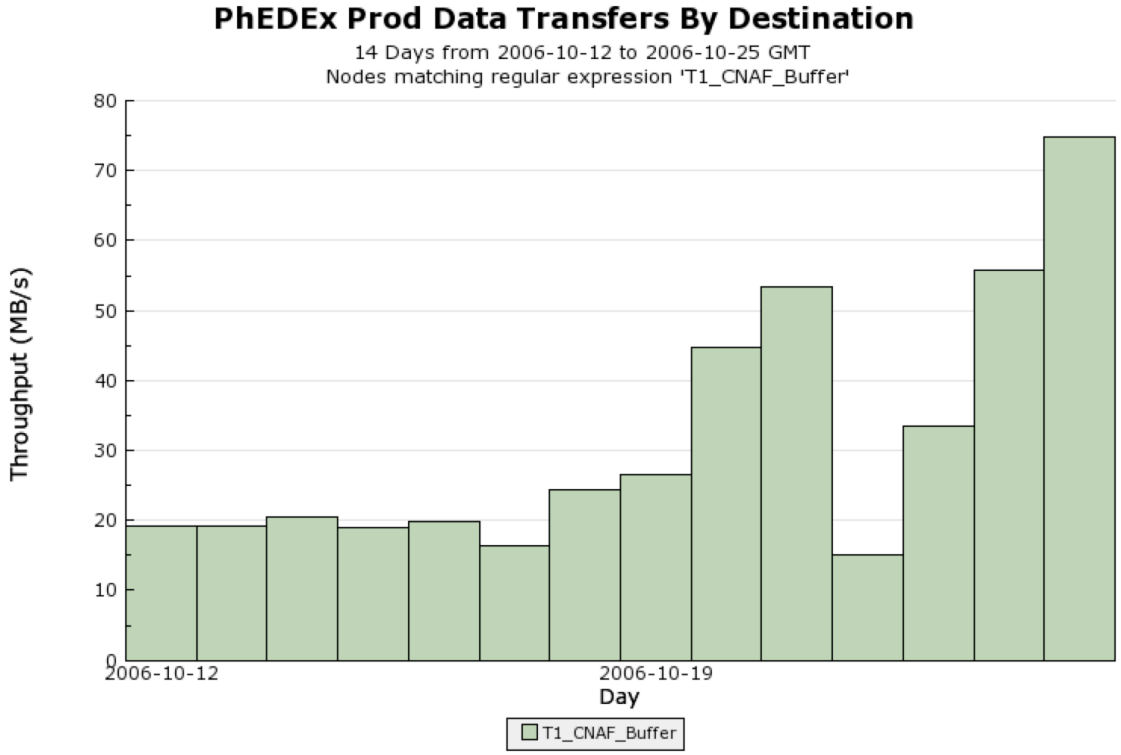

Figure 3: Daily data transfers over a 2-weeks period (12-25 October) during CSA06: two ramp-up periods can be seen (see text).

specific tasks (they exchange information about system state through a central 'blackboard', which contains dataset-replica mappings and locations, dataset subscriptions and allocations, replica set metadata, transfer states).

CMS relies on storage systems located at Tiers to provide access to files, including internal management of replication in a disk cache and/or tape systems. The baseline storage systems that sites have must provide a SRM [9] storage management interface. CMS systems interface to site storage from the Grid side through the aforementioned PhEDEx system, possibly through a layer of file transfer services or directly through the SRM interface. CMS applications running in jobs will interface to storage through a POSIX-like interface, where file-open commands may require the specific syntax of Storage URL's (SURLs). Storage systems have an internal catalogue (or even just a file system) that implements a local namespace. The use of SURL addressing allows an abstraction of physical storage. Sites, and in particular T1 sites, will provide storage systems technically capable of providing long-term, custodial storage of CMS data, and for this responsibility we expect that sites will make Service Level Agreements specifying the availability, throughput, error rate, etc [10]. For T2 sites CMS will instead allow for more lightweight storage systems. These will be used in particular for placement of datasets used for analysis that can relatively easily be replaced through re-generation or reimport from storage systems at T1 sites.

4. INFN Tier-1 storage systems and CMS computing challenges

At a Tier-1 centre, the CMS needs for data storage and data access capabilities are addressed by hierarchical mass storage systems that use tape library back-ends to ensure the custodial data storage function, with high-throughput cache disks and in many cases full datasets "pinned" in front-end disk storage systems.

In this paper we focus on the INFN Tier-1 centre located at CNAF (Bologna, Italy). The set-up of the storage resources at INFN-CNAF is briefly described in the following section (a full description can be found elsewhere [11]). The following sections will instead focus on the actual use of such storage resources by CMS, and to some operational implications seen in CMS Computing, Software and Analysis challenges.

4.1 Storage set-up at INFN-CNAF

The INFN-CNAF resources for data storage consist of a Castor HSM system, namely one StorageTek L5500 silos library partitioned in 2 form-factor slots, i.e. about 2000 slots LTO-2 form and about 3500 slots 9940B form respectively. In total, 6 LTO-2 drives and 4 9940B drives, both with 2Gb/s FC interface, are installed, plus 3 more which were added for WLCG Service Challenges requirements. A number of 10 tapeservers, namely 1U Supermicro 3 GHz 2GB with 1 Qlogic 2300 FC HBA, STK CSC Development Toolkit provided by CERN (with licence agreement with STK), are installed and directly connected with the FC drive output. At the moment, the system offers a tape custodial capacity of about 560 TB (not all in use by CMS though).

The Castor central machine is a IBM x345 2U machine 2x3GHz Intel Xeon, raid1 with double power supply, with Red Hat A.S. 3.0. It runs all central Castor services (Nsdaemon, vmgrdaemon, Cupvdaemon, vdqmdaemon, msgdaemon) and the Oracle client for the central database. The Oracle machine is a 1 IBM x345, Red Hat A.S. 3.0. It runs Oracle DB 9.i rel 2 (more resources are allocated for system back-up and are not described here). A Dell 1650 R.H 7.2 runs Castor monitoring service (Cmon daemon) and Nagios central service for monitoring and notification, plus interface commands to the tapeservers. A 1U Supermicro 3 GHz 2GB with 1 Qlogic 2300 FC HBA accessing the CNAF SAN and runnig the Castor-2 stager, namely Cdbdaemon, stgdaemon and rfiod. A set of diskservers, namely 1U Supermicro 3 GHz 2GB with 1 Qlogic 2300 FC HBA accessing the CNAF SAN and runnig rfiod are installed.

4.2 Preparation for CSA06

The Castor system has been fully upgraded from version 1 to version 2 at CNAF Tier-1 in July 2006, and was tested by CMS since then. CMS proficiently also ran MonteCarlo production against Castor-2 in August 2006. Today, Castor-2 is used in production by all VOs hosted at the Tier-1, but not all VOs use the same set-up. Before and during CSA06, CMS opted for a Castor-2 set-up with two separate service classes, namely a 'tape' class for imported data to be sent to tapes, and a 'disk-only' class for data whose custody was not required. The number of diskservers inserted in the CNAF Castor pools and statically allocated to serve the CMS VO vary accordingly.

During CSA06 preparation and testing, several operational issues arose, the most critical being addressed just before CSA06 started and allowing Tier-1 to join the challenge. Pre-CSA06 operations by CMS, together with other experiments transfer and access activities, raised serious

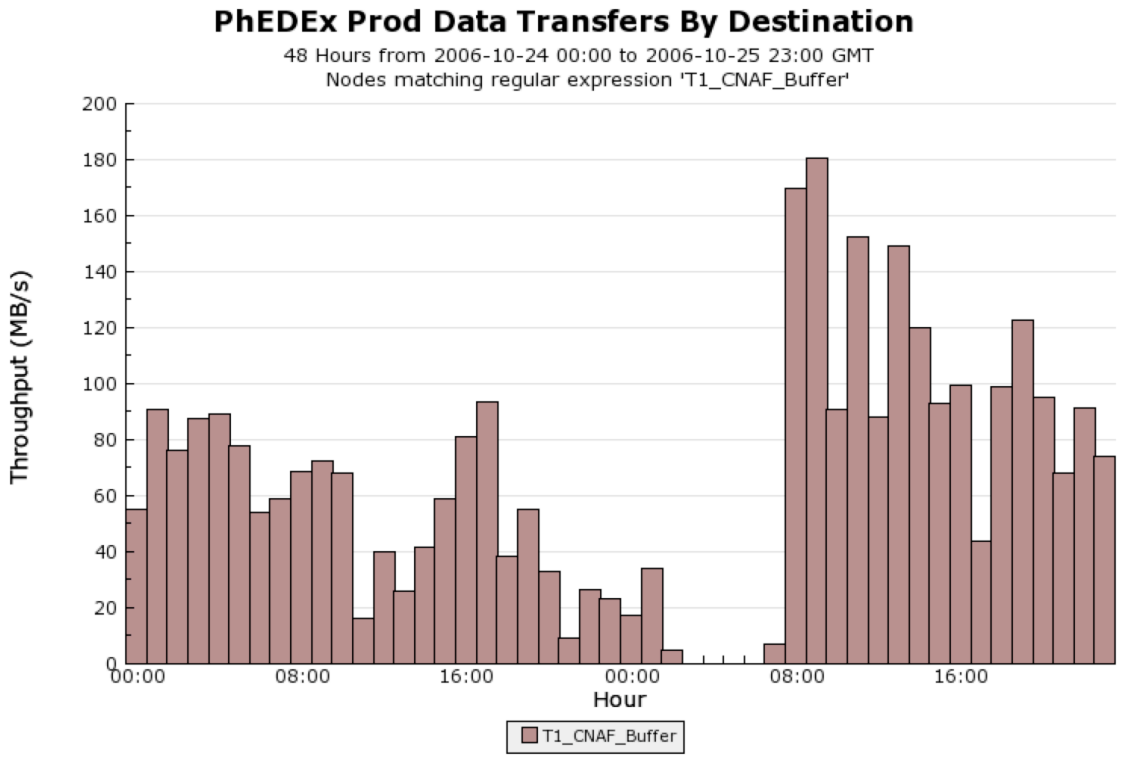

Figure 4: Hourly data transfers over a 2-days period (24-25 October) during CSA06: peaking performances are clearly visible (see text).

concerns on the Castor-2 reliability under the actual load of a production-quality or challenge-like activity by any experiment hosted at the Tier-1. Slow responsiveness of the stager commands, invoked inherently in any activity against the system, were the main symptom and caused the system to be unusable for some days just after a Castor release upgrade. Actions were taken to address the crisis, and a task force was created at CNAF, working in close collaboration with the Castor team and CERN-IT people.

A joint task force with the Castor team and CERN DBA were started up, with the charge to investigate the problems with the stager database. Oracle usage statistics from CMS at CERN were imported into the system configuration at CNAF. Soon after, the stager query command ('stager_qry') was identified as the one showing the worst performance loss, with response times ranging from 20-30 seconds up to more than one minute, and an accumulation on the stager database of many pending requests for queries (up to 3000) was observed. In such conditions of slow responsiveness, the system has to be considered down, since the SRM layer could not perform any put/get request (all of them rely on a stager query afterwards). A SQL tuning of the stager query command was then performed: its Oracle execution plan was found not to be the optimal one; the overall Oracle instance was restarted, and corruption on many indexes was found also. The database tuning continued with the support of CERN DBA: the automatic gathering of the statistics was stopped , and it invalidated the previously imported statistics as well as another reimport performed later, and new statistics were imported from ATLAS (at CERN all DBs share the same set of statistics). The SQL tuning continued, and AWR statistics ('awrrpt') were collected

and tuning by cardinality feedback was started. The feeling of the involved parties at this debugging step was quite different: the Castor team addressed the SQL tuning mainly, whereas CNAF storage/database experts and CMS contancts at the Tier-1 underlined the importance to closely debug the stager's code and its interaction with the DB schema under such load conditions. To shorten the tuning loops, a sysdba temporary access on the CNAF Oracle instance was arranged, so that the CERN staff could start to make AWR reports themselves, after long debugging by the Oracle DBA at CNAF, and investigations focussed on the cardinality of the subqueries. The situation did not improve much, though. Later, a change on the schema at CNAF was done, with just a few unused tables removed, and latest statistics from ATLAS were imported also. SQL tuning hence restarted, by changing the hints in the query itself. The execution plan changed with the new statistics, and after a short period of good performance the database slowed down again, with an average response time of some seconds per each stager query; because of that, the whole tuning process takes quite some time before having some statistics to work on. Then, after moving the indexes to a different table space (which took quite some time but it demonstrated to be useless), the Castor team restarted investigating the latest execution plan, discovering that indeed in some circumstances the Castor-2 system was spending a lot of time in a specific join between 'CastorFile' and 'DiskCopy' tables. This turned out to be the main source of problems, because several (3000) 'DiskCopies' with correspondent 'SubRequests' were present for few 'CastorFiles' and, despite a proper 'UNIQUE' in the statement, Oracle was performing a full join (3000 by 3000) for a given 'CastorFile', leading to millions of rows that had to be collapsed to a single one. This very well explained the high average Oracle response time. The source of the 'DiskCopies' accumulation itself was in the nameserver alias introduced, while the table in the db and the physical files on the disk servers have the original nameserver host as part of the key to identify them. The cure was to clean-up all these 'DiskCopies' (which were in the 'Failed' status): the procedure was thoroughly documented on the Castor web forum for further reference. After the service went up again, a final problem was found regarding many orphaned 'tapeCopies', resulting from previous misuses in SC4 by some LHC experiments, leading to many failing migrations due to the absence of the correspondent 'diskCopy': they were all cleaned up and also the 'CastorFiles' entries had to be deleted, then the system went back into the operational status.

During the intervention period, namely 2 full weeks on the second half of September, 4 people at CERN devoted about 9 FTE-days to helping solve these problems, of which 1.5 FTE-days from the CERN DBA expert. At CNAF, 3 storage experts, 1 DBA expert and 2 experiment people (CMS and LHCb) spent part/all of their time on this work for its full duration (1 storage expert, the DBA expert and the CMS person at the Tier-1 de-prioritized any other work and focused only on this for the full duration of the debugging work. The overall system was extensively tested by CMS at the Tier-1, for 2 days, just before the start of the CSA06 exercise, to verify a satisfactory level of reliability and to allow the INFN Tier-1 to be able to enter into the challenge, and the final testing ended positively.

4.3 Operational issues and performances in CSA06

During CSA06 operations, interesting results were achieved and some critical aspects were identified and are being addressed and followed up since then.

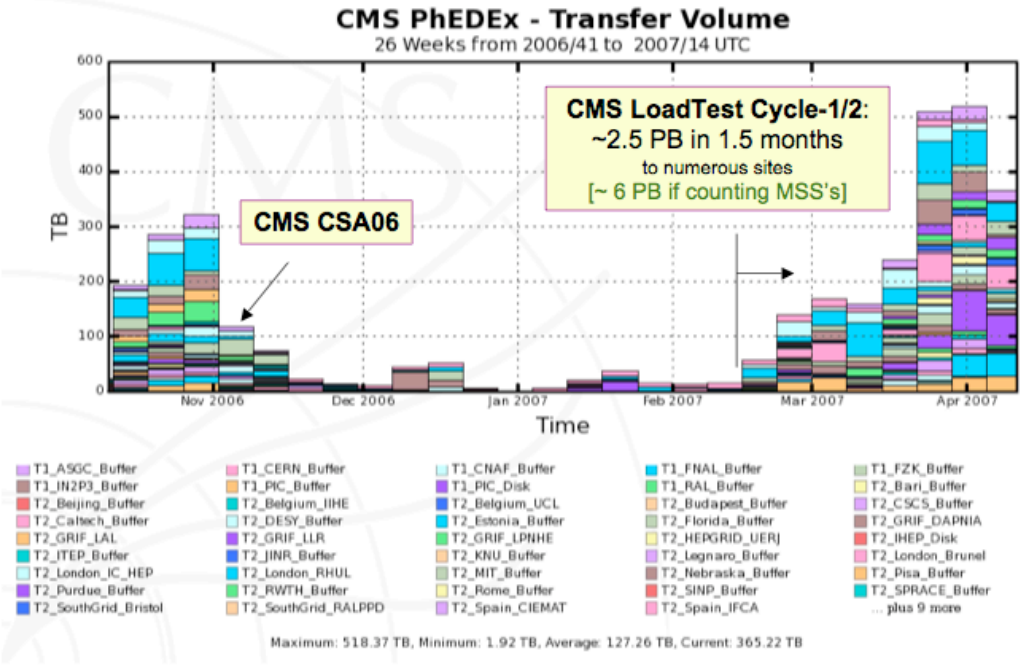

Figure 5: Data transfer volume from autumn 2006 to spring 2007. The 2006 peak corresponds to the CSA06 challenge exercise. The ramp-up in 2007 is due to the CMS LoadTest activity (see text).

Past Service Challenges experiences and ad-hoc pre-CSA06 testing by CMS people at the Tier-1 showed and quantified how the performances of the overall system may be adjusted by acting on the amount of allocated resources and their configurations, in particular of the number of diskservers and tapeservers used by the CMS VO, and on the configuration of the Castor pool. Despite the minimal amount of disk space required to enter CSA06 was 70 TB, the INFN Tier-1 planned to allocate to CMS only about 40 TB. This quota did not significantly improve during the CSA06 operational period, and had strong implications on the operational load and on the achieved performances. Nevertheless, even with a much lower disk space availability, CMS tuned the system accordingly and succeeded in imported 70 TB of CSA06 data samples from the Tier-0 (see Fig. 2). An example of the operational implications of the lack of allocated storage resources is the Garbage Collector behaviour: during operations, this Castor component was invoked far more frequently than expected to delete data already migrated to tapes and to free disk space for new coming data, and the consequent high load evidenced some features and fragilities of its implementation. This 'operational debugging' was very useful for the Castor team, who was active and fast in putting bug fixes into the subsequent Castor-2 sub-releases.

The amount of data imported from the T0 to INFN T1 in CSA06 can be seen in the incremental plot shown in Fig 2: a total of about 70 TB of data were transferred to Castor-2 at CNAF during the CSA06 exercise. The Castor-2 day-by-day import performances can be disentangled in Fig 3, where the daily data transfers over a 2-weeks period is shown: two clear ramp-up periods can be seen, and are the effect of babysitting over problematic diskservers and of import-oriented optimization in the Castor-2 CMS pools configuration. A zoom over a 2-days period is also shown in Fig 4. The number of diskservers inserted in the Castor pools and statically allocated to serve

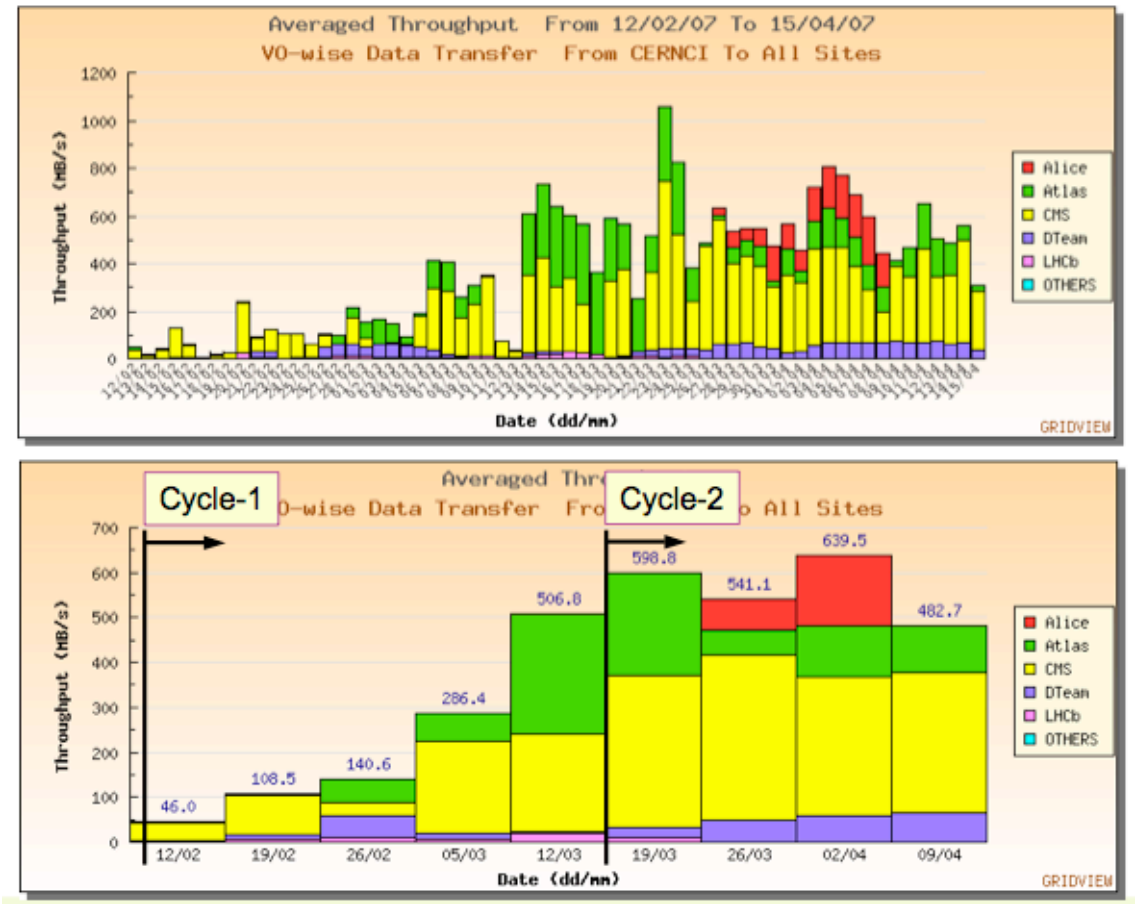

Figure 6: Average throughput from CERN to Tier-1's (aggregate) during WLCG multi-VO tests (first plot is daily, second plot is weekly). The biggest contribution among LHC experiments comes from CMS. The "Cycle" labels refer to the LoadTest activity, and improvements on stability are clearly seen (see text).

the CMS VO turned out to be critical for the performances. It changed during the pre-CSA06 and CSA06 period, and only an average of 4 (2) diskservers were always up and running in the CSA06 time slot and were proficiently running the challenge on the 'tape' ('disk-only') service class respectively. It can be concluded that, with the current set-up and the current number of allocated resources:

- >120-160 MB/s could be sustained for few hours;
- >90 MB/s could be sustained for about 10 hours, with peaks at 170 MB/s;
- >40-60 MB/s could be sustained for many hours, with peaks >100 MB/s

All the subsequent activities of data access and skimming had to face with stringent limitations due to both amouint of resources and superimposing load from other LHC experiments on the shared part of the Castor set-up (e.g. the Castor LSF queues in running data access requests, as well as the concurrent multi-VO traffic on the available tapeservers to get data back from tapes to disk buffer for re-reconstruction).

4.4 Preparation for CSA07

An overall review of the Castor-2 set-up and amount of resources, on the basis of the CSA06 experience, is in progress to address the computing needs of CMS in the forthcoming years. The first testing ground for this program of work is the Computing challenge CSA07, starting in September 2007.

The CMS operational experience on Castor-2 at INFN-CNAF showed no evident advantage in modifying current Castor-2 infrastructure. A unique stager db - as suggested by CERN - may be a limitation, though: even if it's not a concern for scalability, for sure it is in case of scheduled VO-specific downtimes, (when to solve a problem on the stager for one experiment, you stop all experiments). The CMS experience showed that atleast 3 Castor-2 major components appear as points of failures, namely the name server, the stager and the castor-lsf components. Even if the time to fully reconfigure ns/stager in case of hardware failure is about 1 day only, it is quite critical to be needed to stop all experiments in case it happens.

Many operational issues arose from actual management of the system. Frequent expert actions on stager db via SQL are needed. Some daemons are not stable and actually limit the scalability. Managing hardware failure is in general not well supported: the diskserver dismission and file draining is still hard, disk-to-disk copy among pools rely on poor logic (just rely on GC), and all this has an overall impact on stability and performances (these problems are known, and being addressed by Castor developers).

Some improvements after CSA06 may easily be achieved by allocating more disk resources to CMS, by consolidating the past experience in solid support procedures and troubleshooting paths and by constantly exercise the transfers to/from Castor to gain more operational experience. The CMS "LoadTest" program was launched for this; it is organized as repeated cycles of test transfers, and it is evidently increasing the data volume of successful transfers among CMS Tiers, as well as the overall stability of the aggregated rates (see Fig 5). The outcome of this are very promising, as can also be seen by the WLCG multi-VO transfer tests (from CERN to Tier-1's) done in the first quarter of 2007 (see Fig 6), where the CMS contribution is evident, and improvements can be easily seen going from LoadTest Cycle-1 to LoadTest Cycle-2.

A recommendation from the Castor team is that a focussed monitoring system should be implemented as soon as possible at INFN-CNAF, possibly with Lemon to leverage CERN expertise on using it, to ensure that such problems are identified earlier in the future and can be addressed before the service reaches a critical state. This program is currently ongoing at INFN-CNAF and will be fully finalized within July 2007, when also the lemon-cli will be available on User Interfaces and VO-boxes.

As a final remark, since Castor-2 reliability at off-CERN sites is intermittent still, the CMS experience strongly raises the need to centrally improve the release cycle mechanism to "synchronize" off-CERN Castor-based Tier-1's, and to improve the efficiency of troubleshooting by joining efforts among involved communities.

5. Summary

As a grand summary, many actions have been taken by CMS since pre-CSA06 phase to help

the INFN-CNAF storage group to improve the maturity, stability and usability of the Castor set-up at the INFN Tier-1.

As a "Castor-2 survival kit", the CMS operational experience suggests to *i)* allocate adequate resources (add diskservers to reduce the impact of misbehaviours/failures, plus acquire more tape drives, up to the needed scale); *ii)* keep the system as simple as possible (i.e. simplifying the implementation to using one service class only); *iii)* keep the system as empty as possible (i.e. constantly clean-up the stager db and trigger CMS-driven cleanup actions on un-needed files); *iv)* join your efforts with other Tier-1's adopting Castor (since due to lack of manpower at CERN, the operational support to Castor off-CERN sites is an issue: on the CMS side, the CMS Facilities/Infrastructure Operations project being started in 2007 is also addressing this issue).

The Tier-1 is continuing the work towards a higher stability, with the precious help coming from the experimental communities. In addition, the migration to a more stable Castor release is foreseen in July 2007, and is strongly asked by CMS to be done as early as possible, to be prepared for the CSA07 exercise starting in September 2007.

References

[1] CMS Collaboration, "The CMS Computing Model", CERN LHCC 2004-035

[2] CMS Collaboration, "The CMS Computing Project Ð Technical Design Report", CERN-LHCC-2005-023

[3] M. Aderholz et al., "Models of Networked Analysis at Regional Centres for LHC Experiments (MONARC), Phase 2 Report", CERN/LCB 2000-001

[4] LHC Computing Grid (LCG) project: http://www.cern.ch/lcg/

[5] S. Bethke et al., "Report of the Steering Group of the LHC Computing Review", CERN/LHCC 2001-004 (2001).

[6] D. Bonacorsi, "WLCG Service Challenges and Tiered architecture in the LHC era", IFAE, Pavia, April 2006

[7] D. Bonacorsi *et al*, Towards the operation of the INFN Tier-1 for CMS: lessons learned from CMS DC04, in: Proc. ACAT05, DESY Zeuthen, 2005

[8] T. Barrass *et al*, Software agents in data and workflow management, in: Proc. CHEP04, Interlaken, 2004. See also http://www.fipa.org

[9] Storage Resource Management (SRM) project website, http://sdm.lbl.gov/indexproj.php?ProjectID=SRM

[10] WLCG Memorandum Of Understanding, CERN-C-RRB-2005-01/Rev, March 2006

[11] P.Ricci et al, "Experience with Fabric Storage Area Network and HSM Software at the INFN-CNAF Tier-1", this Conference