Journal of Physics: Conference Series **664** (2015) 042056 doi:10.1088/1742-6596/664/4/042056

Pooling the resources of the CMS Tier-1 sites

A Apyan[1], J Badillo[2], J Diaz Cruz[3], S Gadrat[4], O Gutsche[3], B Holzman[3], A Lahiff[5], N Magini[3], D Mason[3], A Perez[6], F Stober[7], S Taneja[8], M Taze[9] and C Wissing[10] for the CMS Collaboration

[1]Massachusetts Institute of Technology, 77 Mass Ave, Cambridge, MA 02139, USA
[2]Universidad de Los Andes, Carrera 1E No 18A 10, Bogota, Colombia
[3]Fermilab, PO Box 500, Batavia IL 60510-5011, USA
[4]Lyon CC, Centre de Calcul de l'Institut National de Physique Nucleaire et de Physique des Particules (IN2P3), 29, Boulevard du 11 novembre, F-69622 Villeurbanne Cedex, France
[5]Rutherford Appleton Laboratory, Didcot OX11 0QX, United Kingdom
[6]CIEMAT, Division de Fisica de Particulas, Avenida Complutense 40,E-28040 Madrid, Spain
[7]Karlsruhe Institut fuer Technologie, Postfach 6980 D-76128 Karlsruhe, Germany
[8]Universita di Bologna, Sezione dell' INFN, Viale C. Berti Pichat, I-40127 Bologna, Italy
[9]Cukurova University, Fen-Ed Fakultesi,TR-01330 Adana, Turkey
[10]DESY, Notkestraße 85, 22607 Hamburg, Germany

E-mail: `christoph.wissing@desy.de`

Abstract. The CMS experiment at the LHC relies on 7 Tier-1 centres of the WLCG to perform the majority of its bulk processing activity, and to archive its data. During the first run of the LHC, these two functions were tightly coupled as each Tier-1 was constrained to process only the data archived on its hierarchical storage. This lack of flexibility in the assignment of processing workflows occasionally resulted in uneven resource utilisation and in an increased latency in the delivery of the results to the physics community.

The long shutdown of the LHC in 2013-2014 was an opportunity to revisit this mode of operations, disentangling the processing and archive functionalities of the Tier-1 centres. The storage services at the Tier-1s were redeployed breaking the traditional hierarchical model: each site now provides a large disk storage to host input and output data for processing, and an independent tape storage used exclusively for archiving. Movement of data between the tape and disk endpoints is not automated, but triggered externally through the WLCG transfer management systems.

With this new setup, CMS operations actively controls at any time which data is available on disk for processing and which data should be sent to archive. Thanks to the high-bandwidth connectivity guaranteed by the LHCOPN, input data can be freely transferred between disk endpoints as needed to take advantage of free CPU, turning the Tier-1s into a large pool of shared resources. The output data can be validated before archiving them permanently, and temporary data formats can be produced without wasting valuable tape resources. Finally, the data hosted on disk at Tier-1s can now be made available also for user analysis since there is no risk any longer of triggering chaotic staging from tape.

In this contribution, we describe the technical solutions adopted for the new disk and tape endpoints at the sites, and we report on the commissioning and scale testing of the service. We detail the procedures implemented by CMS computing operations to actively manage data on disk at Tier-1 sites, and we give examples of the benefits brought to CMS workflows by the additional flexibility of the new system.

1. Introduction

The CMS experiment is one of the two multi-purpose experiments located at the LHC storage ring at CERN. CMS observes proton-proton collision provided by the LHC. The center of mass energy of proton-proton collisions has been 7 TeV and 8 TeV in the first running period "Run1" during the years 2010 until 2012.

1.1. Distributed Computing Infrastructure

The distributed computing infrastructure, which gets used for storing, processing and analysing the data of the experiment, builds on top of the Worldwide LHC Computing Grid (WLCG) [1]. WLCG is setup in a tiered structure. There is one Tier-0 site located at CERN. The main tasks of the Tier-0 are the prompt reconstruction of the CMS data and the long term storage of the data on tape media. An additional long term copy of the data is placed at one of the Tier-1 sites. CMS is supported by a number of Tier-1 sites, namely KIT in Germany, PIC in Spain, CCIN2P3 in France, CNAF in Italy, JINR in Russia[1], ASGC in Taiwan[2], RAL in the United Kingdom and FNAL in the US. Besides data storage the Tier-1 sites are primarily used for centrally organized processing. About 50 Tier-2 sites around the world support the CMS experiment. Tier-2 sites provide no tape archive but only disk space for storing of data. The Tier-2 sites are mainly used for centrally organized production of Monte Carlo events and for analysis jobs sent by individual physicists.

1.2. Restrictions in Using Tape Resources

Tape archives are usually set up as a Hierarchical Storage Manager (HSM), composed of the tape archive and a disk buffer. Clients usually interact only with the buffer and the HSM manages the transfers of files from and to the tapes. Files remain on the buffer disks until space is needed either to store files that are supposed to be written to tape or for files that should be read from tape.

For the scheduled processing of a dataset it is desirable to pre-stage the required data to disk buffer prior to the submission of the processing jobs. The initialization of the pre-stage can be done through the SRM interface provided by the sites. The SRM protocol [2] provides a number of functionalities, that are independent of the actual storage technology installed at the sites. A common practice during Run1 was also to send a support ticket to the sites with a request to pre-stage a data set. The same holds for the so-called 'pinning' of files, which prevents the automatic removal from the buffer disk. This can be achieved via SRM commands, but in practice often happened through support tickets.

Since the actual archiving to tape happened through the HSM, the processing had to happen at the same site where the archiving of the data products was planned. That introduced a restriction in flexibility where to perform the processing, because free CPU cycles where not always immediately available at the archiving location.

The coupling of disk buffer and tape archive led also to a restriction of analysis user access to the Tier-1 sites. Since it was not obvious which files were on disk and which only on tape, large amount of unintended tape staging was possible. Therefore CMS required a dedicated VOMS role *t1access* for the execution of analysis jobs. The role was granted only to "expert users", that were made familiar with the special situation of possible tape operation during file access.

For Run2 CMS has commissioned a data federation [3] based on the xrootd [4] technology. All files existing in the federation can be accessed transparently without knowing the site that is actually hosting the files. That includes the access of files via the wide area network. In order

[1] Support from JINR as with a Tier-1 has been established in 2013
[2] ASGC supported CMS as a Tier-1 until 2012

to include the Tier-1 storage into the federation, CMS set the prerequisite that all files that are potentially accessible in the federation must be available on disk.

2. Separating disk from tape at Tier-1s

Starting in 2010, CERN developed EOS [5] as a scalable solution to provide a high availability disk-based storage for user analysis, in place of the traditional CASTOR [6] HSM. In 2011, CMS started the migration to EOS of the input and output data for its activities at CERN, phasing out CASTOR access completely by the end of Run1 except for Tier-0 prompt reconstruction and custodial archive.

The successful example of EOS, coupled with the experience already matured by other LHC experiments, resulted in 2012 in the recommendation of the WLCG Technical Evolution Groups on Data and Storage Management to separate disk caches from tape archives [7]. To overcome the discussed limitations, CMS explored the possibility to adopt this recommendation at the RAL Tier-1, where the separate disk endpoint was enabled for production in April 2013, and consequently formed a plan to commission new disk endpoints at all Tier-1s by 2014.

2.1. Requirements

CMS set the requirement for the Tier-1s to be able to operate on disk and on tape as two independent logical namespaces, meaning that a file written to or deleted from disk should not be automatically transferred to or deleted from tape, and vice versa.

The disk and tape archive namespaces would be published as different nodes in the CMS Data Management system, PhEDEx [8]. As for any other PhEDEx node, data subscribed should be resident until explicitly deleted; thus, actions in PhEDEx could replace the traditional functions of the HSM:

- Subscribing data to disk is the equivalent of recalling from tape and pinning,
- Subscribing data to archive triggers the migration to tape,
- Deleting from disk releases the cached replica on disk.

Batch jobs and interactive accesses can only target the disk node to read and write files; only PhEDEx transfers are allowed to access the tape node, via a small disk buffer for automated migrations and recalls. The nodes are connected to each other and to any other node (other Tier-1s, Tier-2s, etc.) in the PhEDEx topology through standard transfer links.

2.2. Technical implementations

CMS did not prescribe any particular technical implementation, leaving sites with the freedom to adopt the solution which was most suitable for their storage system. At least the following options were considered feasible:

- Deploying two different storage systems, one for disk and one for tape,
- Using two different trees in the namespace of a single storage system: one with automated migration to tape, and one without.

In either case, transfers between the two nodes can be performed using standard tools such as FTS [9] or xrootd copy, but using a single storage also opens up the possiblity to perform internal operations on the file metadata instead of a physical transfer to copy the file between the two namespaces.

3. Deployment of the Tier-1 disk nodes

3.1. RAL

As mentioned, RAL was the first Tier-1 to deploy a separate disk node in April 2013. The site chose to use CASTOR as technology for both disk and tape, using two separate namespaces on a single storage. RAL is currently evaluating different alternatives for a new disk based solution.

3.2. CNAF

The CNAF Tier-1 deployed a separate namespace for disk-only files on their StoRM [10] storage in August 2013. With StoRM, CNAF can profit of additional flexibility, because space can be reallocated dynamically between the two namespaces if they are deployed on the same filesystem.

3.3. KIT, CCIN2P3, PIC

The three European Tier-1s which had adopted the dCache [11] storage for tape worked together with the dCache developers to identify a common solution for the disk-only storage; eventually all three sites chose a deployment with two namespaces on a single dCache instance. For dCache, the development team proposed to perform transitions between disk and tape buffer using hard links in the namespace rather than internal transfers; this functionality has not yet been needed in production however. KIT commissioned the disk endpoint in November 2013, followed by CCIN2P3 in January 2014 with PIC joining in February 2014.

3.4. FNAL

The FNAL Tier-1 is running a very large dCache instance for tape, which in 2013 had over 9 PB of data already pinned on disk. In order to minimize the disruption, the site chose to deploy a second dCache instance dedicated to disk, using fake PhEDEx transfers to register disk-resident files in the namespace of the new instance. Even so, the migration to the new namespace took a few months, and eventually the disk servers were physically reattached to the new instance to make the files available on the disk node in March 2014.

3.5. JINR

The Russian Tier-1 at JINR has started its commissioning during the Long Shutdown and has been configured in disk-tape-separated mode from the beginning. As a first step a disk-only endpoint based on dCache technology was included into the CMS system. The site plans to add a second dCache instance for the tape archival part towards the start phase of Run2. In the end the setup is expected to look conceptually similar to the FNAL configuration.

3.6. Commissioning the new system

Following the deployment of each new endpoint, the local configuration at the corresponding site needed to be changed to ensure that jobs would only interact with the disk node for reading and writing; this required a simple update of the mapping rules in the Trivial File Catalogs [12] at the site. The CMS Computing Operations team submitted test workflows to verify the correct functionality.

The new nodes also needed to be connected to the rest of the CMS storage endpoints. The links were commissioned in PhEDEx according to the standard testing procedure [13], and once they were enabled over 10 PB of data were transferred between April 2013 and March 2014 to populate the disk nodes as shown in Figure 1. In early 2015, the links between the tape and disk endpoints have been successfully tested at the target rates required to stage data for processing during LHC Run 2. The results are displayed in Figure 2.

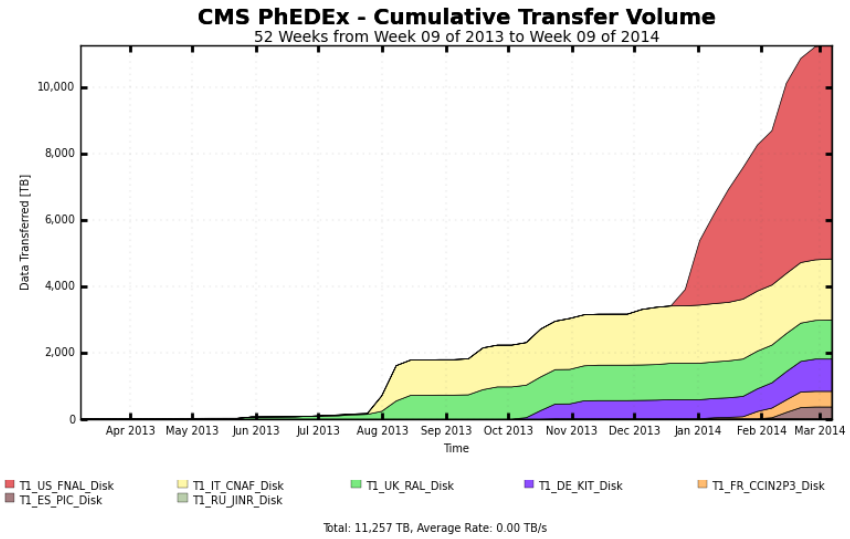

Figure 1. Population of separated disk endpoints as a function of time.

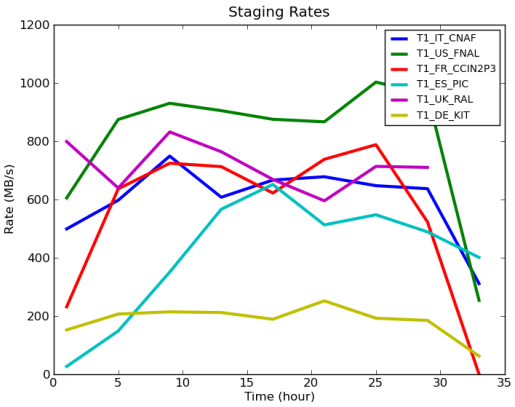

Figure 2. Tape staging rates at the CMS Tier-1 sites.

4. Improvements in operations

After performing the actual separation of disk and tape endpoints, the re-configuration of sites and its recommissioning in the new setup, the shortcomings mentioned in Chapter 1.2 were overcome.

With jobs running at a site accessing only the disk endpoint, accidental tape stage staging can be avoided. That has allowed CMS to open the Tier-1 sites for any CMS user. A fairshare of $\approx 5\%$ is allocated at each site. All data residing on the disk endpoint can also be published into the data federation because they are immediately accessible. These two improvements have significantly shortened the time for users to access the output of a processing campaign. Users can either run jobs directly on the CPUs of the Tier-1 where the data got produced or they can access the data from basically any remote location employing the data federation.

Staging and pinning of data has become a subscription in the PhEDEx data transfer system, while in the past the site admins often needed to be included to ensure a proper result of such requests. Through the PhEDEx monitoring the process remains traceable for the site administrators.

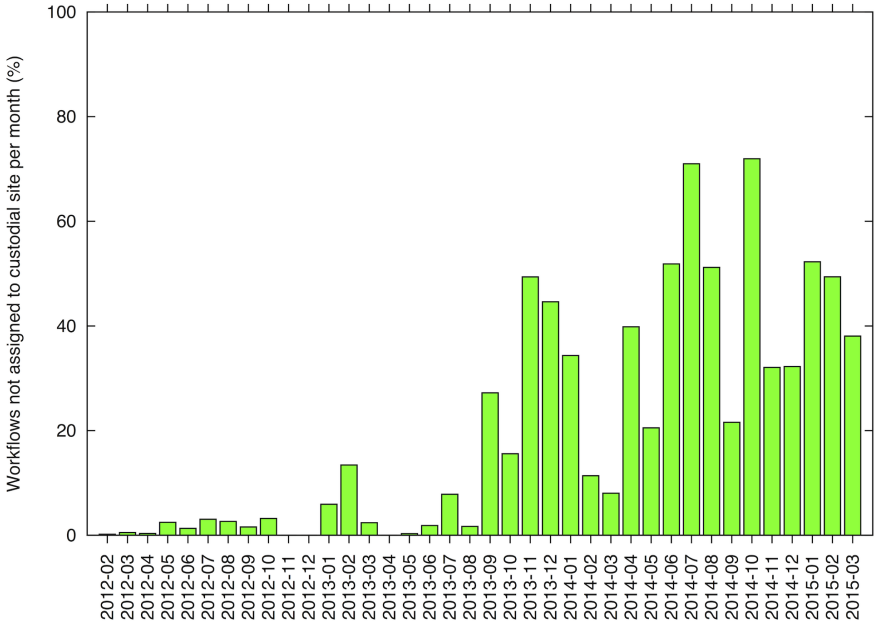

Figure 3. Percentage of workflows assigned to the custodial site as a function of time.

The flexibility where to execute workflows has been increased. Processing and custodial location are now decoupled and jobs can be sent to any site including other CPU resources than Tier-1 sites. The output date are written to disk endpoints first and a subscription to a tape endpoint leads to the final archiving. Figure 3 shows the fraction of workflows that did not get assigned to the custodial site. With more and more sites having disk and tape separated an increased fraction of workflows got run at other than the custodial sites.

5. Conclusions

During LHC Run1 in the years from 2010-2012 the tape archives of the CMS Tier-1 sites were strictly coupled to buffer disks in front of them. Hierarchical Storage Manager (HSM) systems managed the migrations from and to tape. That led to a number of inefficiencies. It introduced a coupling of processing and archiving location and basically prevented CMS user analysis jobs from execution at Tier-1 sites. In the years of the Long Shutdown 1 (LS1) it was decided to separate disk and tape endpoints. Technically this has been achieved by either splitting the name, where one part got only disks attached and the other got connected to the tape archives or by introducing two different storage instances, one serving a large disk pool the other providing the archive. After re-configuring and recommissioning the sites in the new setup, the observed restrictions have been overcome. Tier-1 sites can now be accessed by user jobs and the organized processing can essential be done on any CPU resource. Archiving and staging operations have become a subscription in the data management system of CMS.

6. Acknowledgments

The authors want to thank the development teams of the various storage technologies. They have been always very supportive during the whole process. We also want to thank the storage system administrators at the various Tier-1 sites. We are grateful for the support that has been received from the various funding agencies.

References

[1] Knobloch J *et al.* 2005 LHC Computing Grid Technical Design Report CERN-LHCC-2005-024
[2] Abadie L *et al.* 2007 Storage Resource Managers: Recent international experience on requirements and multiple co-operating implementations *24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007)* pp 47-59
[3] Bloom K (for the CMS Collaboration) 2014 *J. Phys.: Conf. Series* **513** 042005 doi:10.1088/1742-6596/513/4/042005
[4] `http://xrootd.slac.stanford.edu`
[5] Peters A and Janyst L 2011 *J. Phys.: Conf. Series* **331** 052015 doi:10.1088/1742-6596/331/5/052015
[6] Lo Presti G *et al.* 2007 CASTOR: A Distributed Storage Resource Facility for High Performance Data Processing at CERN *IEEE / NASA Goddard Conference on Mass Storage Systems and Technologies 2007* pp 275-280 doi:10.1109/MSST.2007.7
[7] Bell T *et al.* 2012 Report of the WLCG Technology Evolution Groups in Data and Storage Management URL `http://wlcg.web.cern.ch/news/teg-reports`
[8] Egeland R *et al.* 2008 Data transfer infrastructure for CMS data taking PoS ACAT **08** (2008) 033
[9] Frohner A *et al.* 2010 Data management in EGEE *J. Phys.: Conf. Series* **219** 062012
[10] Cavalli A *et al.* 2010 StoRM-GPFS-TSM: a new approach to Hierarchical Storage Management for the LHC experiments *J. Phys.: Conf. Series* **219** 072030
[11] The dCache book `http://www.dcache.org/manuals/book.shtml`
[12] Giffels M *et al.* 2014 The CMS Data Management System *J. Phys.: Conf. Series* **513** 042052 doi:10.1088/1742-6596/513/4/042052
[13] Bagliesi G et al. 2010 Debugging data transfers in CMS *J. Phys.: Conf. Series* **219** 062055 doi:10.1088/1742-6596/219/6/062055