Ensuring data consistency over CMS distributed computing system

To cite this article: Paul Rossman and (on behalf of the CMS Computing and Offline Projects) 2010 J. Phys.: Conf. Ser. 219 072050

View the article online for updates and enhancements.
Ensuring Data Consistency Over CMS Distributed Computing System

Paul Rossman\(^1\) on behalf of the CMS Computing and Offline Projects

\(^1\) Fermi National Accelerator Laboratory, CMS/CD MS 234, P.O.Box 500, Batavia, IL, 60510, USA

E-mail: rossman@fnal.gov

Abstract. CMS utilizes a distributed infrastructure of computing centers to custodially store data, to provide organized processing resources, and to provide analysis computing resources for users. Integrated over the whole system, even in the first year of data taking, the available disk storage approaches 10 petabytes of space. Maintaining consistency between the data bookkeeping, the data transfer system, and physical storage is an interesting technical and operations challenge. In this paper we will discuss the CMS effort to ensure that data is consistently available at all computing centers. We will discuss the technical tools that monitor the consistency of the catalogs and the physical storage as well as the operations model used to find and solve inconsistencies.

1. Introduction

The Large Hadron Collider (LHC) at CERN, Geneva, Switzerland [1] is expected to start operations in 2009. The Compact Muon Solenoid experiment (CMS) [2] in one of the two general purpose detectors at the LHC. CMS utilizes a tiered and distributed infrastructure of computing centers to perform analysis on collected data [3]. Data is stored only on disk at Tier-2 centers and on tape with disk caches at Tier-1 centers. To handle the movement of data between these computing centers, a data transfer management system named PhEDEx (Physics Experiment Data Export) was developed. PhEDEx provides site managers and users a realtime view of the global CMS data transfer state along with a centralized system for making data movement decisions [4]. PhEDEx also automates for CMS many of low level tasks related to data handling typically found in HEP experiments such as large-scale data replication, tape migration and data consistency [5]. Since 2006, CMS has transferred over 60 petabytes across its network (see Fig. 1). During the first year of LHC data taking, it is expected that CMS will transfer 200 terabytes or 100,000 files per day between its primary computing centers [3].

2. The 3 Stages of Maintaining CMS Data Consistency

During the first year of data taking, the available disk storage integrated across all tiers is expected to approach 10 petabytes. Maintaining consistency between the data transfer management system and physical storage is an important operational task which is required to guarantee uninterrupted data availability.
Figure 1. CMS has been transferring data for several years to exercise and validate the hundreds of components necessary for successful data taking. Since 2006, CMS has transferred over 60 petabytes across its network.

2.1. During data transfer

All CMS sites are required to use two PhEDEx components to verify data consistency. Upon data arrival the first component, FileDownloadVerify, is used to verify the file by comparing the on disk file size with the cataloged file size. If FileDownloadVerify discovers a variation in file size the file is removed and retransferred.

The second component, BlockDownloadVerify, accepts requests from a central PhEDEx component at CERN that looks for transfers which have been stuck in transit for 3 days or more. This component can also be invoked on demand for verification of any block (a logical grouping of files) [6] on any site at any time. BlockDownloadVerify manages access load to the storage elements of the sites to prevent overloading the systems. Results from BlockDownloadVerify (see Fig. 2) are available on the PhEDEx web-based management interface [7].

Figure 2. Results from BlockDownloadVerify as shown on the PhEDEx web-based management interface
For CMS Tier-1 sites a third component is required. The FileMSSMigrate component verifies that data has migrated properly to the Tier-1s mass storage tape system. If problems arise during migration, additional steps will need to be taken by the local site administrators for resolution.

2.2. During analysis
CMS uses ProdAgent [8], a central production and processing infrastructure and CRAB [9], a user tool providing distributed access to datasets for analysis. CMS is currently expanding the functionality of ProdAgent and CRAB to report data inconsistencies to the CMS Dashboard, the central monitoring system of CMS [10].

2.3. Periodically on storage
All CMS sites routinely perform consistency checks on data. Several tools are available to site administrators for this purpose. Three utilities are distributed with PhEDEx.

- The BlockConsistencyCheck tool uses the local storage namespace of the site and DBS [11], the CMS Data Bookkeeping Service, to find files missing at the site. Missing files and file size mismatches at the site are reported. Inconsistencies identified by this tool are provided to central CMS administrators via a ticketing system for resolution.
- The StorageConsistencyCheck tool also uses the local storage namespace of the site and finds orphan files which are not registered in the central catalogs. The output from this tool is a list of files known to the local storage namespace but not the catalogs. Resolution of inconsistencies are again reported via a ticketing system to central CMS administrators.
- The BlockDownloadVerify-injector tool allows operators and shifters to perform consistency checks on the block level of datasets without direct access to the local mass storage. In contrast to the above mentioned site administrator possibilities, the injector uses the BlockDownloadVerify component to verify the consistency of a specific block of a datasets. This is often used by non-local operators and shifters to invoke quick evaluations at remote sites if problems have been reported or noticed in the central monitoring. Results are available via the PhEDEx web-based management interface as shown in Fig. 2.

3. Checksum vs. File Size
CMS is currently investigating greater adoption of checksum based file verification. Since 2006, CMS has averaged over 176,000 successful data transfers per month from the Tier-0 to Tier-1 centers based on file size verification alone (See Fig. 3). However there have been isolated instances where data files have been transferred passing the file size check yet having invalid checksums compared to the catalog. This could have a negative avalanche effect where a file with an invalid checksum is transferred successfully to other sites therefore multiplying the problem. If noticed, a situation like this would create unnecessary work for site administrators who need to remove the corrupt data files and request retransfer of the data.

At present, most CMS sites verify the data file after transfer with PhEDEx using only the files size and its cataloged file size. For Tier-2 and Tier-3 sites, if the sizes match, PhEDEx marks the transfer as successful and the file becomes available for analysis at the destination site. For Tier-1 sites, PhEDEx uses an additional step (the FileMSSMigrate component) to verify that the transferred file was written to tape.

Fermilab, using dCache as its disk cache in front of the Enstore tape system [12], developed a solution to use cksum checksums as an additional consistency check. This solution is lightweight and scalable across hundreds of storage nodes.

It is based on the centrally calculated cksum checksum determined soon after every data file is generated. This checksum is stored in DBS along with the file size.
Figure 3. CMS has averaged over 176,000 successful data transfers per month since 2006 from the Tier-0 to Tier-1 centers. Several campaigns have taken place and are being planned to develop solutions intended to minimize the transfer errors we encounter.

Upon file arrival, the FileDownloadVerify component queries the namespace of the Fermilab mass storage to determine the dCache pool node the file resides on. A second query is then made directly to the dCache pool node and a checksum is calculated on the local file using cksum. PhEDEx then compares the value to the cataloged file checksum and declares the transfer a success or failure. In the case of a failure, the file is deleted. PhEDEx automatically schedules a retransfer of the file.

There are several advantages to this method of verification.

- This solution is fast, efficient and scalable because the checksum is calculated on the pool node where the file resides and therefore scales with the number of pool nodes.
- As the initial checksum is calculated when the file is generated and stored in DBS, only one checksum has to be calculated when the file arrives thus improving scalability further.
- The checksum calculation happens some minutes after the transfer completes to avoid calculating the checksum while the file is still in the disk controller’s memory buffer on the pool node. We are therefore more likely to be checking the actual file on disk.

4. Summary
The existing consistency tools are continuously improving with additional functionality. The goal is to provide an open namespace framework that generalizes the consistency tools and simplifies their usage. As LHC startup approaches, CMS is making every effort to ensure that all data is consistently available at its computing centers worldwide.

5. Acknowledgments
We would like to thank the funding agencies under which this work was conducted and all CMS collaborators for their help, support and contributions.
References

[1] O. Bruning, P. Collier, P. Lebrun, S. Myers, R. Ostojcic, J. Poole and P. Proudlock, *LHC design report. Vol. I: The LHC main ring*, CERN-2004-003, 2004

[2] R. Adolphi et al. [CMS Collaboration], *The CMS experiment at the CERN LHC*, JINST 0803, S08004 (2008) [JINST 3, S08004 (2008)]

[3] CMS Collaboration, *The CMS Computing Model*, CERN/LHCC 2004-035, 2004

[4] L. Tuura et al, *Scaling CMS data transfer system for LHC start-up*, Journal of Physics, Conference Series (119) 072030, 2008

[5] J. Rehn et al, *PhEDEx high-throughput data transfer management system*, Proceedings of Computing in High-Energy and Nuclear Physics (CHEP06), Mumbai, India, 13 Feb - 17 Feb 2006

[6] CMS Collaboration, *The CMS Computing Project Technical Design Report*, CERN/LHCC 2005-023, 2005

[7] The PhEDEx web-based management interface is located at http://cmsweb.cern.ch/phedex/

[8] D. Evans, A. Fanfani, et. al. *The CMS Monte Carlo Production System: Development and Design*, CMS CR-2007/057, 2007

[9] D. Spiga et al, *The CMS remote analysis builder (CRAB)*, LNCS 4873:580-586 (2007)

[10] The CMS Dashboard is located at http://dashboard.cern.ch/cms/

[11] A. Afaq, A. Dolgert, Y. Guo, C. Jones, S. Kosyakov, V. Kuznetsov, L. Lueking, D. Riley, V. Sekhri, *The CMS Dataset Bookkeeping System*, Journal of Physics, Conference Series (119) 072001, 2008

[12] G. Oleynik et al, *Fermilab's multi-petabyte scalable mass storage system*, Proceedings of the 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies (April 11 - 14, 2005), (MSST 2005)