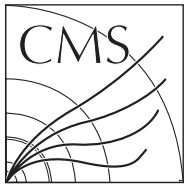
Available on CMS information server

CMS CR -2011/037



The Compact Muon Solenoid Experiment

Conference Report

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

15 February 2011 (v2, 16 February 2011)

Distributed Data Transfers in CMS

Paul Rossman for the CMS Collaboration

Abstract

The multi-tiered computing infrastructure of the CMS experiment at the LHC depends on the reliable and fast transfer of data between the different CMS computing sites. Data have to be transferred from the Tier-0 to the Tier-1 sites for archival in a timely manner to avoid overflowing disk buffers at CERN. Data have to be transferred in bursts to all Tier- 2 level sites for analysis as well as synchronized between the different Tier-1 sites. The data transfer system is the key ingredient which enables the optimal usage of all distributed resources. The operation of the transfer system consists of monitoring and debugging of transfer issues to guarantee a timely delivery of data to all corners of the CMS computing infrastructure. Further task of transfer operation is to guarantee the consistency of the data at all sites, both on disk and on tape. Procedures to verify the consistency and to debug and repair problems will be discussed.

Presented at *CHEP2010: International Conference on Computing in High Energy and Nuclear Physics 2010*

Distributed data transfers in CMS

Nicolo Magini[1], Natalia Ratnikova[2], Paul Rossman[3], Alberto Sánchez-Hernández[4] and Tony Wildish[5]

[1] CERN, Switzerland
[2] Institut für Experimentelle Kernphysik, KIT, Germany
[3] Fermi National Accelerator Laboratory, United States
[4] CINVESTAV, Mexico City, Mexico
[5] Princeton University, United States

E-mail: ratnik@ekp.uni-karlsruhe.de,rossman@fnal.gov

Abstract. The multi-tiered computing infrastructure of the CMS experiment at the LHC depends on the reliable and fast transfer of data between the different CMS computing sites. Data have to be transferred from the Tier-0 to the Tier-1 sites for archival in a timely manner to avoid overflowing disk buffers at CERN. Data have to be transferred in bursts to all Tier-2 level sites for analysis as well as synchronized between the different Tier-1 sites. The data transfer system is the key ingredient which enables the optimal usage of all distributed resources. The operation of the transfer system consists of monitoring and debugging of transfer issues to guarantee a timely delivery of data to all corners of the CMS computing infrastructure. Further task of transfer operation is to guarantee the consistency of the data at all sites, both on disk and on tape. Procedures to verify the consistency and to debug and repair problems will be discussed.

1. Introduction

The Large Hadron Collider (LHC) at CERN, Geneva, Switzerland [1] started operations in 2010. The Compact Muon Solenoid experiment (CMS) [2] is 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 and simulated data [3]. Data are stored on tape with disk caches at Tier-1 centers and on disk only at Tier-2 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 real-time view of the global CMS data transfer state along with a centralized system for making data movement decisions [4]-[5]. 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, verification of migration to tape, and data consistency checks [6].

2. Data transfers

CMS uses a hierarchical architecture of tiered centers with a single Tier-0 center at CERN, seven Tier-1 centers at national computing facilities, 51 Tier-2 and 52 Tier-3 centers at institutes and universities worldwide.

Raw CMS detector data are transferred from the Tier-0 at CERN to Tier-1 centers for archival storage, reconstruction, and skimming. This is done in a timely manner to avoid overflowing

disk buffers at CERN. Skimmed reconstructed data are then transferred for analysis to all Tier-2 centers as well as synchronized between the Tier-1 centers. Simultaneously, simulated Monte Carlo data are produced at Tier-2 centers and distributed to Tier-1 centers for archival storage.

Diagram on Figure 1 shows distribution of CMS data files produced and cataloged since March 2004, including Monte Carlo simulated data and data collected by the CMS detector.

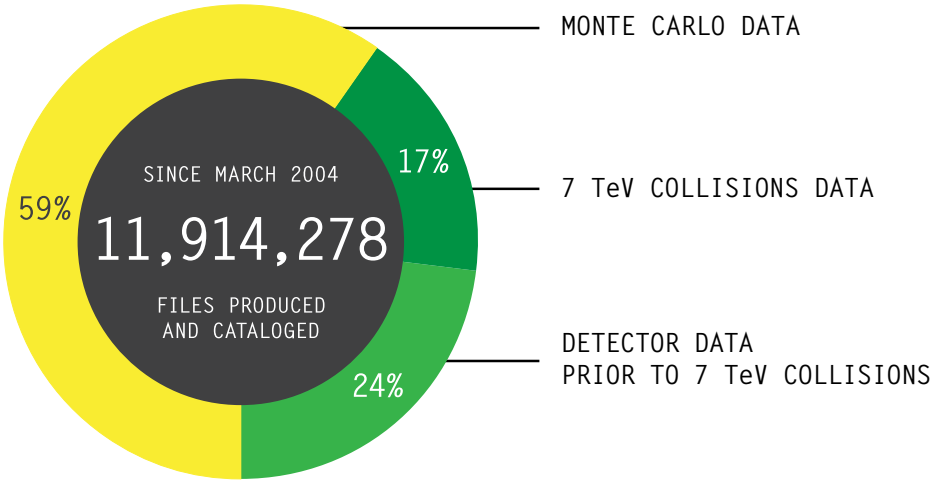


Figure 1. Almost 12 million files produced and cataloged since March 2004 include detector data recorded and reconstructed during the preparation to the collision runs, collisions data recorded and reconstructed after start of LHC physics program at $\sqrt{s} = 7\ TeV$ on 30 March 2010, and simulated data of different kinds.

Figure 2 shows average daily data transfer rate and cumulative transferred volume of CMS data for the last six years.

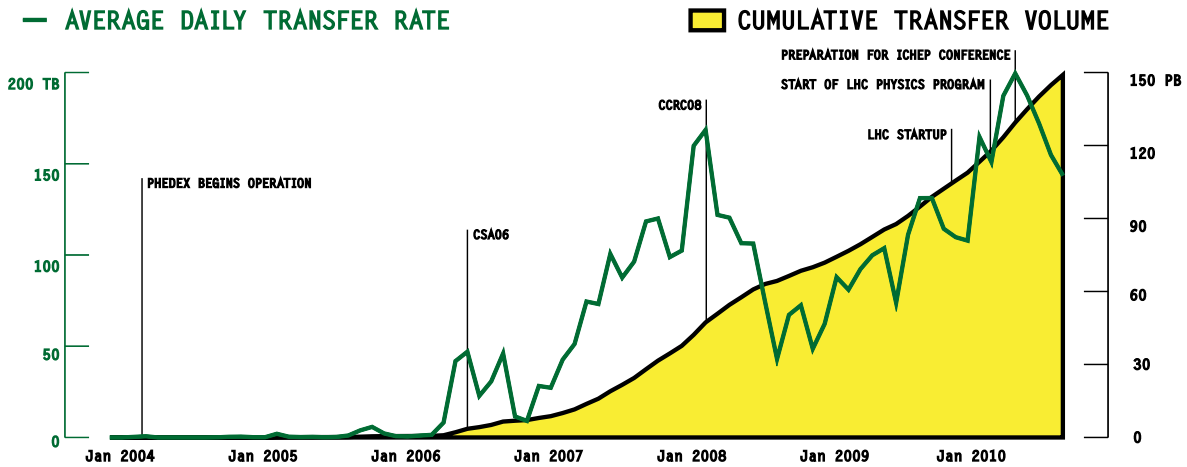


Figure 2. Average daily data transfer rate and cumulative transferred volume of CMS data for the six years of PhEDEx operations. Also shown special events causing spikes in data transfers.

The cumulative volume of CMS transfers currently exceeds 150 petabytes of data since 2004. Almost 12 million files have been produced and cataloged.

3. Data consistency project

Ever increasing data volumes mean that rare problems are becoming more common. We are entering a period where physicists require access to greater amounts of physics data. Proactively checking consistency of storage is therefore becoming increasingly important.

Scalability of storage-checking is a challenging problem. Complete consistency checks at a Tier-1 site which take 4 hours today may take an entire day by the end of next year. Checking consistency of large data volumes can be a costly operation that puts additional load on the storage system and databases. These loads compete for resources needed for regular data operation tasks.

The goal of our project is to provide CMS-wide tools for the management of consistency checks intended to minimize data operations risks. Table 1 shows typical data inconsistencies and associated risks.

Table 1. Typical inconsistencies and associated risks

Inconsistency Type	Example	Associated Risks
Missing file	File is assigned to the site in the central catalog, but not found in storage	- application failure
Orphaned file	File is present in storage, but not known to the central data catalogs	- inefficient usage of storage resource
Corrupted file	File attribute (size, checksum) in storage and in the central catalogs differ	- application failure - wrong results

While orphaned files may lead to just inefficient use of storage resources, missing or corrupted files may cause an application to fail or even affect the results.

3.1. Problem description

The CMS Data Management system relies on two Central Data Catalogs:

- Dataset Bookkeeping Service DBS [7] keeps detailed information about all produced datasets
- Transfer Management Data Base TMDB [8] keeps information about data transfers and data location

In addition every site maintains Local Storage Element (SE) Catalogs, accessible by various protocols depending on the storage technology.

While the CMS data transfer system is designed to propagate all state changes of the file transparently, in a complex and highly distributed environment inconsistencies may still occur due to various factors. Failure of an individual component, human error during manual intervention, latencies on a heavily loaded system may affect the normal workflow and result in failures to synchronize the catalogs. Another potential source of inconsistencies are transitions due to reorganization of the storage systems. In these cases a full check of all data at a site is required to confirm a successful completion of the transition.

3.2. CMS approach

Sites have the local expertise to optimize access to their SEs for consistency checking, but do not often have manpower to develop a complete tool chain.

Our solution is to provide CMS-wide tools for management of consistency checks, with site-local plugins that minimize the overhead to the sites' developers. This tool-chain is extensible and flexible enough that it can and will be used in other domains (space-management, performance monitoring, alerts).

3.3. Implementation

PhEDEx uses an agent-based approach [12], where every site deploys own PhEDEx agents, each performing a certain task related to data replication.

Data consistency checks are performed by the BlockDownloadVerify agent. Interactions with a SE are implemented as plugins [13] for various storage protocols: CASTOR [9], dCache [10], SRM [11], posix to name a few. It is simple to add a new SE-access protocol, and to use local-access methods for efficient interaction with the SE.

The agent contacts TMDB at regular intervals to check if test requests were injected. To optimize the throughput, the amount of tests fetched per duty cycle can be dynamically adjusted. Results are uploaded in bulk, to minimize expensive database connections. Smart caching and internal sorting help to optimize the performance.

The only mandatory parameter in the agent configuration in the type of the storage technology. Other configuration parameters can be adjusted locally within some hard limits introduced for safety.

4. Using storage dumps for performance boost

SE providers can dump a snapshot of their SE contents at a given moment into a file in some standardized format, such as Syncat [14], which is a storage technology independent format developed for the purpose of synchronizing file catalogs. Experiments can use this information to check the contents of the SE on very short time scales without direct calls to the storage system.

SE experts can produce the storage dump at a frequency that will insure minimal impact to the SE itself, using whatever highly optimized site-specific techniques they have available to them. Experiment's consistency checking tools at the site can be configured to use information from the dump file if it is available, and to go back to native tools otherwise, or if more up-to-date information is needed.

Application of this approach to consistency checks at the German CMS Tier-1 at KIT has shown a dramatic improvement in speed. Storage dumps were produced using the chimera-dump tool provided by dCache, which produces XML files in Syncat format. A consistency check of file sizes for all data on site, which would normally take more than half a day, now completes in about three hours.

File size information from the storage dumps is also used for efficient accounting of the occupied storage space.

5. Operations

CMS consistency tools are deployed on all CMS sites as a part of the PhEDEx release. Site-local agents are constantly monitored by the CMS central Computing Shift Person. Test requests can be injected by various means:

- locally using an injector script
- remotely using a central injector agent running at CERN

- internally scheduled tests initiated upon completion of a particular operation, e.g. file download.
- via a "dropbox" method

The "priority" and "expiration time" parameters are provided to allow additional control on scheduling of tests.

All test results are stored centrally and are available on the PhEDEx web page, by command-line tool, or via data service APIs.

The detected inconsistencies are filed in the problem reporting and tracking system and are further handled by data operations experts. Monthly checks for the orphaned files are scheduled at all Tier-1 sites.

6. Summary

150 petabytes of data are accumulated by CMS since 2004. Almost 12 million of files are produced and cataloged. Average daily rate of data transferred by PhEDEx tool since 2007 exceeds 110 terabytes. Amount of CMS data as well as the number of data transfers between CMS sites will continue to grow during the forthcoming continuous period of LHC operations and data taking. Important task of the data operations is to ensure availability and reliability of the stored data, including data consistency checks and synchronization of the data catalogs across multiple sites and varying storage technologies. In our project we provide CMS-wide tools for managing data consistency, which combine the use of the standard CMS infrastructure of PhEDEx agents with the site-local plugins that implement interactions with the local Storage Element. Particular attention is paid to the performance and load management of the consistency checking tools, which compete for the resources with the regular data operations tasks. Massive tests at several CMS Tier-1 centers have shown good performance. Utilization of storage dumps for dCache technology at KIT resulted in additional factor of four speed-up while reducing the number of interactions with the storage system. This method is highly recommended to other storage technologies.

Acknowledgments

We would like to thank the transfer infrastructure development team, the site administrators and facility teams, operation teams and all other CMS collaborators who contributed to a successful CMS data transfer operation. We are also grateful to the data storage systems developers, especially the dCache team for their interest and help in our work on the data consistency project. We thank the international funding agencies amongst those the Department of Energy, the National Science Foundation, the German Research Foundation DGF and the Bundesministerium für Bildung und Forschung BMBF for their support of this work.

References

[1] Bruning O *et al.* 2004 LHC design report. Vol. I: The LHC main ring. CERN-2004-003
[2] CMS Collaboration 1994 CMS, the Compact Muon Solenoid: Technical proposal, CERN-LHCC-94-38
[3] CMS Collaboration 2005 CMS computing technical design report, CERN-LHCC-2005-023
[4] Rehn J *et al.* 2006 PhEDEx high-throughput data transfer management system *Proc. of CHEP06, Mumbai, India*
[5] Tuura L *et al.* 2008 Scaling CMS data transfer system for LHC start-up. *J. Phys.: Conf. Ser.* **119** 072030
[6] Egeland R, Metson S, Wildish T 2008 Data transfer infrastructure for CMS data taking *Proc. of ACAT'08, Erice, Italy*
[7] Afaq A et al 2008 The CMS Dataset Bookkeeping Service *J. Phys.: Conf. Ser.* **119** 072001
[8] Barrass T, et al 2004 Software agents in data and workflow management *Computing in High Energy Physics (CHEP04), Interlaken*
[9] Bärring O et al 2007 CASTOR2: Design and Development of a Scalable Architecture for a Hierarchical Storage System at CERN *Computing in High Energy Physics (CHEP07), Victoria, B.C., Canada*

[10] *The dCache Book* http://www.dcache.org/manuals/Book
[11] 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)* 4759
[12] *IEEE Foundation for Intelligent Physical Agents* http://www.fipa.org/
[13] *Consistency checking tools* https://twiki.cern.ch/twiki/bin/view/CMS/PhedexProjConsistency
[14] Millar P et al 2010 Dealing with orphans: Catalogue synchronisation with SynCat *J. Phys.: Conf. Ser.* **219** 062060
[15] Serfon C 2010 Data management tools and operational procedures in ATLAS : Example of the German cloud *J. Phys.: Conf. Ser.* **219** 042053