User and group storage management the CMS CERN T2 centre

G. Cerminara¹, G. Franzoni¹, A. Pfeiffer¹ - for the Compact Muon Solenoid collaboration

¹ CERN, PH CMG-CO; CH-1211 Geneva 23 Switzerland
E-mail: giovanni.franzoni@cern.ch

Abstract. A wide range of detector commissioning, calibration and data analysis tasks is carried out by CMS using dedicated storage resources available at the CMS CERN Tier-2 centre. Relying on the functionalities of the EOS disk-only storage technology, the optimal exploitation of the CMS user/group resources has required the introduction of policies for data access management, data protection, cleanup campaigns based on access pattern, and long term tape archival. The resource management has been organised around the definition of working groups and the delegation to an identified responsible of each group composition. In this paper we illustrate the user/group storage management, and the development and operational experience at the CMS CERN Tier-2 centre in the 2012-2015 period.

1. Introduction

The Compact Muon Solenoid (CMS) experiment [1] is an omni-purpose detector operating at the Large Hadron Collider [2] at CERN. The central feature of the CMS apparatus is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. Within the superconducting solenoid volume are a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass/scintillator hadron calorimeter. Muons are measured in gas-ionization detectors embedded in the steel flux return yoke outside the solenoid. In addition, the CMS detector has extensive forward calorimetry. The first level of the CMS trigger system, composed of custom hardware processors, uses information from the calorimeters and muon detectors to select the most interesting events in a fixed time interval of less than 4 µs. The High Level Trigger (HLT) computer farm further decreases the event rate from around 100 kHz to around 0.5 kHz, before data storage.

A wide range of detector commissioning, calibration and data analysis tasks is carried out by members of the Compact Muon Solenoid (CMS) collaboration using dedicated storage resources available at the CMS CERN Tier-2 centre. This paper describes the usage of CMS CERN Tier-2 data storage resources to serve single users and working teams, and is structured as follows: after a brief description of the CMS CERN Tier-2, we introduce the concept of a working team and how it relates to the existing work organisation of CMS; we then present the logic and technology used to define the working teams and let them self-administer as much as possible; we conclude with an outlook on the deployment of data archival functionality and on the future application of the same administration paradigm to the CMS CERN Analysis Facility (CAF).
2. The CMS CERN Tier-2 Centre

The Tier-2 [3] based at CERN, being a computing facility of the CMS hosting laboratory, plays an unique role in CMS: it provides computing resources to users who are stationed at CERN for extended periods of time (at least one year for more than half of their time) and to activities recognised as a priority for CMS.

The storage space of the CMS CERN Tier-2 is based on EOS [4, 5], an open source distributed disk storage system in production at CERN since 2011. The two main pools of resources available at the CMS CERN T2 are:

- user space: 1 TB of storage for analysis or development tasks is made available to the CMS users who are resident at CERN. 370 collaborators have a user area, for 128 TB of space used in total.
- group space: most of the non centrally managed (i.e. via PhEDEx [3]) data at CMS is used for analysis activities shared by a team of users, e.g. detector groups (e.g. ECAL), physics object (muon) or analyses teams (exotica). The typical use case is the production of datasets, like ntuples, by one or few collaborators used by the whole group. The group resources are used by 330 users grouped in 33 working teams, for a total amount of available space of 1073 TB.

3. Working Teams

The roots of the CMS collaboration structure rest on working teams, the so called level two groups, each charged with a set of goals and deliverables. Detector (e.g. the electromagnetic calorimeter ECAL detector performance group), physics object (e.g. muon physics object group) or analyses teams (e.g. exotica analysis group) are three examples of level two groups we will deal with in this paper. Any level two group is lead by one or two convenors.

When we started administrating CMS CERN T2, see figure 1 first row, the EOS disk resources available to the working teams were organised in a directory structure reflecting the active working teams with all working teams sharing a single quota; every user with access to the CMS CERN T2 could write anywhere in the directory tree. While very simple to implement, such a structure suffered of two major pitfalls: 1) any time the overall quota would be reached, the activity of all the groups would come to a halt 2) the space available to a any given team was in no ways dimensioned to its needs, and there was no incentive to remove less relevant data to make space for new data.

To move away from those factors limiting an effective usage of the resources, we have: first formalised the correspondence between the CMS level two groups and one directory sub-tree in the EOS group space; then we deployed access restrictions by charging the level two convenors with the responsibility of defining and maintaining a list of users to be granted writing credentials. Such restrictions were implemented exploiting the EOS functionality of setting access control lists using egroups [6], the supported mechanism to manage authorisation of resources at CERN.

The next key step has been the introduction of a separate quota for each working team. In concertation with all the groups involved, the amount of resources granted to the working teams were agreed on the basis of: 1) the amount of space already occupied at the time of the quota deployment - no group could be assigned less space than it was already using 2) proportionality to the pledges for disk space at the CMS Tier-2’s [3].

While every member of a working team can remove files she/he is the owner of, it has to be possible for a working team to carry out a cleanup campaign irrespective of who has originally created them. A typical example is the need of space for newer versions of reprocessed data, which should be obtained by removing superseded versions (which were potentially written by members of the group which have left the collaboration since). In light of this use case, we
introduced the role of ’cleaners’, see the second row of figure 1: a dedicated egroup-defined access control list of members of the working team who can remove files everywhere in the directory sub-tree, irrespective of who the owner of the files is.

4. Data Popularity for Clean Up Campaigns
Determining which files are best to be removed when in need of free space can be a non trivial task in complex directory sub-trees used by working teams with a dozen(s) of members.

To facilitate the process, we have distributed, both to the single members and to the convenors and cleaners of the working teams, lists of stale directories; ”stale” directory defined as: containing files all older than four months and not accessed for more than three months. We could distribute such popularity information thanks to monitoring data provided to us by the SPC group of the CERN IT department. A dedicated set of tools mines the monitoring data and composes email messages to all the users owning stale directories, and to all the convenors and cleaners with summary views.

Bimonthly distribution of popularity reports has helped the working teams and the single users to organise cleanup campaigns in a more targeted and time efficient fashion. Due to the lack of monitoring data, the popularity reports have not been distributed after the middle of 2014; we’re in the process of restoring the process which is deemed to restart by mid 2015.

**Figure 1.** Schematic representation of the administration structure for the group storage CMS CERN T2. First row: at first all users performing group activities shared one single quota. Second row: working teams have been defined, with a team responsible (crown) and a set of team cleaners (broom). Third row: popularity reports were distributed to facilitate the clean up by removing stale files, thus regularly making room for newly produced data. Fourth row: archival functionalities are going to ease the clean up campaigns (tape).
5. Archival
One of the difficulties with freeing storage space lies with the unlikely but possible need of accessing "stale" data again in the future.

A typical example is an analysis entering the review phase: in most cases, there’s no need of going back again to access the data initially used for the measurement; yet in some cases, cross checks or additional studies may urgently be required by the paper review committee, many weeks after the first draft of the paper has been made available. To avoid a situation when the data is needed with no time available to recreate it, working teams tend to occupy a lot of their EOS with stale data.

As CMS representatives, we have discussed this use case with the CERN IT/DSS group. The devised solution has resulted in an additional EOS functionality to archive directory trees in EOS into castor [7], the CERN tape-based mass storage system. After a few rounds of technical test carried out jointly with IT/DSS, the archival functionality (subject of a report at this conference) has recently become available in the EOS instance available in production at CERN for CMS.

As a second technical solution to implement clean up campaigns, we are in the process of communicating to the working teams the instructions to use the archival EOS functionality: the cleaners will be granted the credentials to 1) move stale directory trees into a directory dedicated to the archival procedure and 2) create the archive in castor and purge the archived data from EOS, to release the space.

6. Outlook
Following the success with the CERN Tier-2, also the CMS CERN Tier-3 (deployed) and the CMS CERN Analysis Facility (CAF) (forthcoming) are to be based on the same working team driven design and the same EOS+eGroups technology choices.

The CERN Analysis Facility (CAF) consists of resources reserved for high priority and short turn-around time (low-latency) workflows, such as prompt-feedback, fast calibrations and analyses. In Run-I the CAF relied on castor storage and a pool of dedicated CPU’s; its refurbishment for Run-II, which is ongoing comprises:

- the cleanup of old/stale users, move to EOS and management of disk and CPU resources based on the paradigm of working teams
- the reorganisation of the access to the CERN batch queues, managed through dedicated LSF [8] groups. We will use the new capability of mapping e-groups into LSF groups, to have a single entry point for the self-management of the CAF and CMS CERN Tier-2 and Tier-3

7. Conclusions
We have presented the development and operational experience as administrators of the CMS CERN Tier-2 centre in the 2012-2015 period. By deploying a structure based on 1) working teams and on 2) the delegation of the management to their convenors, we have significantly increased the efficiency and reliability of the storage resources management and exploitation.

Thanks to the positive outcome, we are now restructuring the CMS CERN Analysis Facility (CAF) too, based on the same working team driven design.

References
[1] CMS Collaboration, The CMS experiment at the CERN LHC, JINST 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.
[2] Lyndon Evans and Philip Bryant, LHC Machine, 2008 JINST 3 S08001, doi:10.1088/1748-0221/3/08/S08001.
[3] CMS Collaboration, CMS computing : Technical Design Report, 2005, CERN-LHCC-2005-023.
[4] X Espinal et al, *Disk storage at CERN: Handling LHC data and beyond*, 2014 J. Phys.: Conf. Ser. 513 042017
doi:10.1088/1742-6596/513/4/042017

[5] A J Peters, *EOS as the present and future solution for data storage at CERN*, 2015, proceedings of this conference.

[6] *Using e-groups to manage authorisation to CERN resources*, https://espace.cern.ch/e-groups-help/default.aspx

[7] CASTOR homepage http://castor.web.cern.ch/

[8] Release Notes for Platform LSF Version 7.0, Nov. 2006, PLATFORM corporation