



Contribution ID: 48

Type: **not specified**

## Optimizing the usage of multi-Petabyte storage resources for LHC experiments

*Tuesday, 27 March 2012 12:10 (20 minutes)*

### Description of the Work

During the first two years of data taking, the CMS experiment has collected over 20 PetaBytes of data and processed and analyzed it on the distributed, multi-tiered computing infrastructure on the WorldWide LHC Computing Grid. Given the increasing data volume that has to be stored and efficiently analyzed, it is a challenge for LHC experiments to fully profit of the available network and storage resources and to facilitate daily computing operations.

We have developed the CMS Popularity Service that tracks file accesses and user activity on the grid and will serve as the foundation for the evolution of their data placement. We have deployed a fully automated, popularity-based site-cleaning agent in order to scan Tier2 sites that are reaching their space quota and suggest obsolete, unused data that can be safely deleted without disrupting analysis activity.

Current work is to demonstrate dynamic data placement functionality based on this popularity service and integrate it in the data and workload management systems: as a consequence the pre-placement of data will be minimized and additional replication of hot datasets will be requested automatically.

### Conclusions

Experiment and user activities keep increasing and oblige the experiments to ensure the future scalability of the system by automating manual operations and optimizing the usage of available resources. The strategies we are presenting go exactly in this direction. The popularity and cleaning systems are the first step towards the implementation of an optimized data placement model, where the number of dataset copies kept on grid sites is directly related to their popularity.

### Impact

Given the scale of the CMS grid infrastructure, it is a complex problem to control and optimize the usage of the storage. The CMS physics community consists of over 20 physics-groups that have pledged on over 50 Tier2s, resulting in over 124 physics-group Tier2 associations. At the moment it takes considerable human effort to control the evolution of the space and to verify that the groups are not exceeding their pledges. We have provided tools to monitor which data is actually being used and suggest data that can be safely removed. The monitoring tools we have developed allow controlling the evolution of the storage space on sites, reducing considerably the manual effort and improving day-to-day operations. The ideas in this contribution can be extended to other scientific domains that makes usage of the grid for their data analysis and that wants to learn how their community is making usage of the available data and eventually implement automatic strategies to optimize their distribution.

## Overview (For the conference guide)

In the last two years of LHC operation, the experiments have made a considerable usage of grid resources for the data storage and offline analysis. To achieve the successful exploitation of these resources a significant operational human effort has been put in place and it is the moment to improve the usage of the available infrastructure.

In this respect, the CMS Popularity project aims to track the experiment's data access patterns (frequency of data access, access protocols, users, sites and CPU), providing the base for the automation of data cleaning and data placement activity on grid sites. As well, the popularity-based Site Cleaning Agent has been developed to monitor the evolution in time of the used and pledged space and remove unused data replicas at full Tier2s.

This presentation will give an insight into the development, validation and production process of these systems. We will analyze how the framework has influenced resource optimization and daily operations in CMS.

**Primary authors:** Dr SPIGA, Daniele (CERN); Dr GIORDANO, Domenico (CERN); Dr KARAVAKIS, Edward (CERN); Mr HARALD BARREIRO MEGINO, Fernando (CERN); Dr GIRONE, Maria (CERN); Mr CINQUILLI, Mattia (CERN); Dr MAGINI, Nicolo (CERN); Mrs MANCINELLI, Valentina (CERN)

**Presenters:** Dr SPIGA, Daniele (CERN); Dr GIORDANO, Domenico (CERN); Mr HARALD BARREIRO MEGINO, Fernando (CERN); Dr GIRONE, Maria (CERN)

**Session Classification:** Services for Data Management and Messaging

**Track Classification:** Operational services and infrastructure