



Contribution ID: 229

Type: oral presentation

Dynamic Data Management for the Distributed CMS Computing System

Monday 13 April 2015 16:30 (15 minutes)

The Dynamic Data Management (DDM) framework is designed to manage the majority of the CMS data in an automated fashion. At the moment 51 CMS Tier-2 data centers have the ability to host about 20 PB of data. Tier-1 centers will also be included adding substantially more space. The goal of DDM is to facilitate the management of the data distribution and optimize the accessibility of data for the physics analyses. The basic idea is to replicate highly demanded data and remove data replicas that are not often used. There are four distinct parts that work together on achieving this goal: newly created datasets are inserted into the managed data pool, highly popular datasets are replicated across multiple data centers, data centers are cleaned of the least popular data, and temporarily unavailable sites are accounted by creating replicas of missing datasets. We are also developing various metrics that show whether the system behaves as expected. The CMS Popularity Service is the authoritative source of the popularity information – CPU time, number of accesses, number of users/day, etc. of datasets, blocks and files. The information is collected from various resources namely directly from the CMS Software, from the CMS Remote Analysis Builder (CRAB) and from files accessed via xrootd. It is subsequently summarized and exposed using a REST API to other CMS services for further use. This paper describes the architecture of these components and details the experience of commissioning and running the system over the last year.

Primary author: PAUS, Christoph (Massachusetts Inst. of Technology (US))

Presenter: PAUS, Christoph (Massachusetts Inst. of Technology (US))

Session Classification: Track 5 Session

Track Classification: Track5: Computing activities and Computing models