Contribution ID: **295** Type: **Oral**

GRID Storage Optimization in Transparent and User-Friendly Way for LHCb datasets

Wednesday 12 October 2016 11:15 (15 minutes)

The LHCb collaboration is one of the four major experiments at the Large Hadron Collider at CERN. Petabytes of data are generated by the detectors and Monte-Carlo simulations. The LHCb Grid interware LHCbDIRAC is used to make data available to all collaboration members around the world. The data is replicated to the Grid sites in different locations. However, disk storage on the Grid is limited and does not allow to keep replicas of each file at all sites. Thus, it is essential to determine the optimal number of replicas in order to achieve a good Grid performance.

In this study, we present an approach of data replication and distribution strategy based on data popularity prediction different from that previous presented at CHEP2015[1]. Each file can be described by the following features: age, reuse time interval, access frequency, type, size and some others parameters. Based on these features and access history, the probability that the file will be accessed in the long-term future can be predicted using machine learning algorithms. In addition, time series analysis and access history can be used to forecast the number of accesses to the file in the short-term period in the future. We describe a metrics that combines these predictions. This metrics helps to determine for which files the number of replicas can be increased or decreased depending on how much disk space is available or how much space needs to be freed. Moreover, the metrics indicates when all replicas of the file can be removed from disk storage. The proposed approach is being tested in LHCb production. In the study, we show the results of the simulation studies and results of the tests in the production. We demonstrate that the method outperforms our previous study, while it requires a minimal number of parameters and gives more easily interpretable predictions.

Reference:

[1] Mikhail Hushchyn, Philippe Charpentier, Andrey Ustyuzhanin "Disk storage management for LHCb based on Data Popularity estimator" 2015 J. Phys.: Conf. Ser. 664 042026, http://iopscience.iop.org/1742-6596/664/4/042026

Tertiary Keyword (Optional)

Secondary Keyword (Optional)

Distributed data handling

Primary Keyword (Mandatory)

Storage systems

Author: HUSHCHYN, Mikhail (Yandex School of Data Analysis (RU))

Co-authors: USTYUZHANIN, Andrey (Yandex School of Data Analysis (RU)); HAEN, Christophe (CERN); CHARPENTIER, Philippe (CERN)

Presenter: HUSHCHYN, Mikhail (Yandex School of Data Analysis (RU))

Session Classification: Track 4: Data Handling

Track Classification: Track 4: Data Handling