20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013)

Contribution ID: **29** Type: **Oral presentation to parallel session**

ECFS: A decentralized, distributed and fault-tolerant FUSE filesystem for the LHCb online farm

Monday 14 October 2013 13:30 (20 minutes)

The LHCb experiment records millions of proton collisions every second, but only a fraction of them are useful for LHCb physics.
In order to filter out the “bad events” a large farm of x86-servers (˜2000 nodes) has been put in place. These servers boot from and run from NFS, however they use their local disk to temporarily store data, which cannot be processed in real-time (”data-deferring”). These events are subsequently processed, when there are no live-data coming in. The effective CPU power is thus greatly increased. This gain in CPU power depends critically on the availability of the local disks. For cost and power-reasons, mirroring (RAID-1) is not used, leading to a lot of operational headache with failing disks and disk-errors or server failures induced by faulty disks.

To mitigate these problems and increase the reliability of the LHCb farm, while at same time keeping cost and power-consumption low, an extensive research and study of existing highly available and distributed file systems has been done. While many distributed file systems are providing reliability by “file replication”, none of the evaluated one supports erasure algorithms.

A decentralised, distributed and fault-tolerant “write once read many” file system has been designed and implemented as a proof of concept providing fault tolerance without using expensive - in terms of disk space - file replication techniques and providing a unique namespace as a main goals.

This paper describes the design and the implementation of the Erasure Codes File System (ECFS) and presents the specialised FUSE interface for Linux.
Depending on the encoding algorithm ECFS will use a certain number of target directories as a backend to store the segments that compose the encoded data. When target directories are mounted via nfs/autofs - ECFS will act as a file-system over network/block-level raid over multiple servers.

Author: RYBCZYNSKI, Tomasz (AGH University of Science and Technology (PL))

Co-authors: BONACCORSI, Enrico (CERN); NEUFELD, Niko (CERN)

Presenter: RYBCZYNSKI, Tomasz (AGH University of Science and Technology (PL))

Session Classification: Data Stores, Data Bases, and Storage Systems

Track Classification: Data Stores, Data Bases, and Storage Systems