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

Contribution ID: **385** Type: **Poster presentation**

Evaluation of Apache Hadoop for Parallel Data Analysis with ROOT

Monday, 14 October 2013 15:00 (45 minutes)

The Apache Hadoop software is a Java based framework for distributed processing of large data sets across clusters of computers using the Hadoop file system (HDFS) for data storage and backup and MapReduce as a processing platform. Hadoop is primarily designed for processing large textual data sets which can be processed in arbitrary chunks, and must be adapted to the use case of processing binary data files which can not be split automatically. However, Hadoop offers attractive features in terms of fault tolerance, task supervision and controlling, multi-user functionality and job management.

For this reason, we have evaluated Apache Hadoop as an alternative approach to PROOF for root data analysis. Two alternatives in distributing analysis data are discussed: Either the data is stored in HDFS and processed with MapReduce, or the data is accessed via a standard Grid storage system (dCache Tier-2) and MapReduce was used only as execution back-end.

The focus in the measurements are on the one hand to safely store analysis data on HDFS with reasonable data rates and on the other hand to process data fast and reliably with MapReduce.

For evaluation of the HDFS, data rates for writing to and reading from local Hadoop cluster have been measured and are compared to normal data rates on the local NFS. For evaluation of MapReduce, realistic ROOT analyses have been used and event rates were compared to PROOF.

Primary author: LEHRACK, Sebastian (LMU Munich)

Co-authors: DUCKECK, Guenter (Experimentalphysik-Fakultaet fuer Physik-Ludwig-Maximilians-Uni); Dr EBKE, Johannes (Ludwig-Maximilians-Univ. Muenchen (DE))

Presenters: DUCKECK, Guenter (Experimentalphysik-Fakultaet fuer Physik-Ludwig-Maximilians-Uni); Dr EBKE, Johannes (Ludwig-Maximilians-Univ. Muenchen (DE)); LEHRACK, Sebastian (LMU Munich)

Session Classification: Poster presentations

Track Classification: Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization