Contribution ID: **10** Type: **Oral Presentation**

Explore New Computing Environment for LHAASO offline data analysis

Friday, 23 March 2018 10:00 (30 minutes)

The exploitation of a new computing environment has become an urgent practice to overcome a series of challenges with the development of the new generation of High Energy Physics(HEP). LHAASO(Large High Altitude Air Shower Observatory) is expected the most sensitive project to studies the problems in Galactic cosmic ray physics, and requires massive storage and computing power. Efficient parallel algorithms/frameworks and High IO throughput are key to meet the scalability and performance requirements of LHAASO offline data analysis. Though Hadoop has gained a lot of attention from scientific community for its scalability and parallel computing framework for large data sets, it is still difficult to make LHAASO data processing tasks run directly on Hadoop. In this paper we explore ways to build a new computing environment using Hadoop to make LHAASO jobs run on it transparently. Particularly, we discuss a new mechanism to support LHAASO software to random access data in HDFS. Because HDFS is streaming data stored only supporting sequential write and append. It cannot satisfy LHAASO jobs to random access data. This new feature allows the Map/Reduce tasks to random read/write on the local file system on data nodes instead of using Hadoop data streaming interface. This makes HEP jobs run on Hadoop possible. We also develop diverse MapReduce model for LHAASO jobs such as Corsika simulation, ARGO detector simulation (Geant 4) and MK2A data processing. And we wrap the models to make them transparent to users. In addition, we provide the real-time cluster monitoring in terms of cluster healthy, number of running jobs, number of finished jobs and number of killed jobs. Also the accounting system is included. This work has been in production for LHAASO offline data analysis to gain about 40,000 CPU hours per month since September, 2016. The results show the efficiency of IO intensive job can be improved about 46%. Finally, we describe our current work of data migration tools to serve the data move between HDFS and other storage system or Tape.

Primary author: Dr HUANG, Qiulan (Institute of High Energy of Physics, Chinese Academy Sciences)

Co-author: Prof. SUN, Gongxing (Institute of High Energy of Physics,Chinese Academy of Sciences)

Presenter: Dr HUANG, Qiulan (Institute of High Energy of Physics, Chinese Academy Sciences)

Session Classification: Networking, Security, Infrastructure & Operation Session

Track Classification: Networking, Security, Infrastructure & Operations