GSI-HPC Cluster

W. Schön[1], *C. Huhn*[1], *V. Penso*[1], *T. Roth*[1], *J. Trautmann*[1], *the HPC department*[1], *and the IT-Security division*[1]

[1]GSI, Darmstadt, Germany

Abstract:The GSI HPC cluster is mainly used by international collaborations to analyse experimental data (GSI/LHC). The cluster is being extended to be prepare for the needs of FAIR. New concepts to improve cluster scalbility are developed and deployed in new clusters.

Introduction

The FAIR project needs more powerfull computing in terms of FLOPs, I/O rates and online storage capacities. For efficient use of a cluster it is crucial to identify and eliminate bottlenecks. With increasing complexitivity of such a cluster, monitoring, anomaly detection and management are also crucial. In addition, IT security becomes an more and more demanding problem - especially taking into account the international usage of the cluster and the existence of broad bandwidth connections to remote sites.

Compute Power

In an environment typical for GSI, LHC or FAIR experiments, the pure compute power is the most easy problem to solve. Due to the independence of the physics events, parallelization can be done easily by analyzing events simultaneously on a number of cores. No communication between the cores is necessary. The simplest solution is to increase the number of cores in the cluster to enhance the number crunching power. However, more indirect limitations will occur: I/O per node, management of a huge number of cores and nodes, power density, cooling can become critical.

I/O Capacity and Meta Data Perfomance

I/O is the crucial factor in analysing data of large experiments (e.g. CBM or PANDA). If the I/O capacity is not sufficient, the jobs will be I/O bound (jobs will wait for I/O instead of analysing data). Typical simulation jobs or theory calculations are much less I/O demanding. The I/O problem is solved employing parallel cluster file systems at GSI (Lustre [1]). These clusters aggregate the capacity of single RAID groups to a filesystem with a global view to the whole data space. The clusters are designed, built and operated by the HPC department. GSI participates in the development and testing of the Lustre (and is a founding member of the EOFS [2]). In the past three years, the production cluster with about 125 file servers and a 7 Petabyte capacity has provided a fast and reliable storage for all of GSI's working groups. The next generation cluster running the current stable Lustre version an providing 7.7 PB capacity has been set up recently. It will employ two important new features: the data object are now stored on disk using the ZFS [4] file system, which provides protection against data corruption, continuous integrity checking and automatic repair. And the metadata can now be distributed across several servers, improving on one of the general bottleneck of Lustre, metadata performance. In particular, heavy load on the meta data server can severely limit the overall cluster performance. Especially massively parallel concurrent access of jobs to one and the same file can cause problems, slowing down the entire cluster and thus the HPC farm. User code is often making very inefficient use of Lustre, e.g. by writing and reading huge numbers of small files instead of writing one large file. Excessive meta data usage can also be point to bad user code (e.g. nested infinite loops querying meta data). It is therefore necessary to monitor and detect anomal meta data operations to protect the whole system.

Managing the GSI Cluster

The GSI clusters are now being managed by the Chef configuration management tool [5] with great success. Meanwhile, old NFS services and servers have been phased out. Remaining physics data that was not moved to the Lustre clusters has been pooled up in few modern file servers, while distribution of scientific software has been entirely moved to CernVM-FS [6]. The goal to improve stability [3] by avoiding common shared resources for user code and system code apart from Lustre was attained, as was the goal to improve scalability by distributing the data in many instances.

To scale out to the planned size of FAIR T0-Computing, a new open source job scheduler for the GSI batchfarm is being tested [7].

References

[1] Schoen et. al. I/O Optimised Cluster for Data analysis, GSI scientific report 2010

[2] www.eofs.org

[3] Huhn et. al. Comparison ...., GSI scientific report 2011

[4] Open ZFS, http://open-zfs.org/

[5] www.chef.io

[6] cernvm.cern.ch

[7] http://slurm.schedmd.com

DOI:10.15120/GR-2015-1-INFRASTRUCTURE-IT-13