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

Contribution ID: **333** Type: **Oral presentation to parallel session**

Testing SLURM open source batch system for a Tier1/Tier2 HEP computing facility

Thursday 17 October 2013 14:38 (22 minutes)

In this work the testing activities that were carried on to verify if the SLURM batch system could be used as the production batch system of a typical Tier1/Tier2 HEP computing center are shown. SLURM (Simple Linux Utility for Resource Management) is an Open Source batch system developed mainly by the Lawrence Livermore National Laboratory, SchedMD, Linux NetworX, Hewlett-Packard, and Groupe Bull. Testing was focused both on verifying the functionalities of the batch system and the performance that SLURM is able to offer.

We first describe our initial set of requirements. Functionally, we started configuring SLURM so that it replicates all the scheduling policies already used in production in the computing centers involved in the test, i.e. INFN-Bari and the INFN-Tier1 at CNAF, Bologna. Currently, the INFN-Tier1 is using IBM LSF (Load Sharing Facility), while INFN-Bari, an LHC Tier2 for both CMS and Alice, is using Torque as resource manager and MAUI as scheduler.

We show how we configured SLURM in order to enable several scheduling functionalities such as Hierarchical FairShare, Quality of Service, user-based and group-based priority, limits on the number of jobs per user/group/queue, job age scheduling, job size scheduling, and scheduling of "consumable resources". We then show how different job typologies, like serial, MPI, multi-thread, whole-node and interactive jobs can be managed. Tests on the use of ACLs on queues or in general other resources are then described. A peculiar SLURM feature we also verified is triggers on event, useful to configure specific actions on each possible event in the batch system.

We also tested highly available configurations for the master node. This feature is of paramount importance since a mandatory requirement in our scenarios is to have a working farm cluster even in case of hardware failure of the server(s) hosting the batch system.

Among our requirements there is also the possibility to deal with pre-execution and post-execution scripts, and controlled handling of the failure of such scripts. This feature is heavily used, for example, at the INFN-Tier1 in order to check the health status of a worker node before execution of each job. Pre- and post-execution scripts are also important to let WNoDeS, the IaaS Cloud solution developed at INFN, use SLURM as its resource manager. WNoDeS has already been supporting the LSF and Torque batch systems for some time; in this work we show the work done so that WNoDeS supports SLURM as well.

Finally, we show several performance tests that we carried on to verify SLURM scalability and reliability, detailing scalability tests both in terms of managed nodes and of queued jobs.

Authors: Mr ITALIANO, Alessandro Italiano (INFN-CNAF); SALOMONI, Davide (Universita e INFN (IT)); DONVITO, Giacinto (Universita e INFN (IT))

Presenter: DONVITO, Giacinto (Universita e INFN (IT))

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

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