Contribution ID: **21** Type: **not specified**

CESGA Experience with the Grid Engine batch system

Wednesday 21 April 2010 09:30 (30 minutes)

Grid Engine (GE) is an open source batch system with a complete documentation and support for advanced features like the possibility to configure a shadow master host for failover purposes, support for up to 10.000 nodes per master server, application level checkpointing, array jobs, DRMAA, fully integrated MPI support, a complete administration GUI, a web-based accounting tool (ARCo), etc.

CESGA has been using GE in its systems during more than 5 years. Currently it is the only batch system used at CESGA both for the supercomputers and for grid clusters.

One of the last challenges was the integration of GE with the FINIS TERRAE supercomputer installed at CESGA, it is formed by 142 Itanium nodes of 16 CPUs and 128GB of memory each one. Several stress tests were done to check SGE behaviour in a large cluster, results will be shown. It also will be explained some special SGE configurations like:

- HP-MPI and Intel MPI integration: most of the jobs that run in Finis Terrae are MPI jobs.
- Checkpointing. It has been configured in the queue configuration.
- Exclusivity: possibility to request a node exclusively.
- SSH-less node configuration: remote connections between nodes are done using qrsh instead of ssh. This is very important in the case of MPI jobs to avoid jobs expanding outside the nodes they have been assigned or trying to use more resources in a node than the ones assigned to the job.
- Interactive jobs: jobs run interactively using a shell interface.
- Application integration with GE: for example Gaussian, Gaussian is used only under the batch system, the batch job requirements are taken accordingly to Gaussian input.

In order to manage special user requirements, CESGA has developed a qsub wrapper to implement some additional functionalities like special resources, i.e. the possibility to request additional resources for the jobs than the established limits (memory, CPU time, number of processors, space on scratch disk, ⋯)

Some kind of jobs (challenges, priority agreements, ⋯) need to be prioritized, it will be explained how these jobs are prioritized using GE functionalities.

Thanks to the efforts of IC, LIP and CESGA currently GE is fully supported by gLite middleware.

In a standard configuration it requires a Computer Element (CE) for each grid infrastructure. CESGA has done several modifications to the standard configuration to support different grid projects (EGEE, EELA, int.eu.grid, Ibergrid, and other regional grid projects) using just one centralized batch system. This way resources can be shared among different grids with minimal system administration overhead.

The batch system is shared using one single GE qmaster server and a shadow qmaster for fault tolerance purposes. Jobs are submitted from different sources but all jobs are collected in a single batch server that distributes them between all the available WN.

Authors: Mr SIMON GARCIA, Alvaro (CESGA); Dr FERNANDEZ SANCHEZ, Carlos (CESGA); Mr FREIRE GARCIA, Esteban (CESGA); Dr LOPEZ CACHEIRO, Javier (CESGA); Mr REY MAYO, Pablo (CESGA); Mr DIEZ LAZARO, Ruben (CESGA); Mr DIAZ MONTES, Sergio (CESGA)

Presenter: Mr FREIRE GARCIA, Esteban (CESGA)

Session Classification: Grid and WLCG

Track Classification: Grid and WLCG