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

Contribution ID: **284** Type: **Poster presentation**

Job Scheduling in Grid Farms

Monday, October 14, 2013 3:00 PM (45 minutes)

The vast majority of jobs in the Grid are embarrassingly parallel. In particular HEP tasks are divided into atomic jobs without need for communication between them. Jobs are still neither multi-threaded nor multi-core capable. On the other hand, resource requirements reach from CPU-dominated Monte Carlo jobs to network intense analysis jobs.

The main objective of any Grid site is to stably operate its Grid farm while achieving a high job slot occupancy, an optimal usage of the computing resources (network, CPU, memory, disk space) and guaranteed shares for the VOs and groups. In order to optimize the utilization of resources, jobs must be distributed intelligently over the slots, CPUs, and hosts. Although the jobs resource requirements cannot be deduced directly, jobs are mapped to POSIX user/group ID based on their VOMS-proxy. The user/group ID allows to distinguish jobs, assuming VOs make use of the VOMS group and role mechanism.

The multi-VO Tier-2 site at DESY (DESY-HH) supports ~20 VOs on federated computing resources, using an opportunistic resource usage model. As at many EGI/WLCG sites, the Grid farm is based on the queuing system PBS/TORQUE, which was deployed from the EMI middleware repositories. Initially, the scheduler MAUI was used. It showed severe scalability problems with 4000 job slots as soon as the number of running plus queued jobs approached 10000. Job scheduling became slow or even blocked. In addition, MAUI's many configuration options appeared to be hard to control.

To be able to further increase the number of worker nodes as requested by the VOs (to currently 8000 job slots), DESY-HH needed a scalable and performing scheduler, which runs in conjunction with PBS/TORQUE. In the course of studying alternative scheduling models, a home-made scheduler was developed (working title: MySched), which is tailored to then needs of the DESY-HH Grid farm and uses the C-API of PBS/TORQUE. It is based on a simple scheduling model without support for multi-core jobs and job parallelism and is optimized for high job slot occupancy and intelligent distribution of jobs to the worker nodes. Furthermore, it allows for a fine-grained adjustment of limits and parameters on VO and group level.

In the contribution to CHEP 2013 we will discuss the impact of a classification of jobs according to their potential resources requirements on scheduling strategies. Subsequently, we will describe our home-made implementation and present operational results.

Primary author: Dr GELLRICH, Andreas (DESY)

Presenter: Dr GELLRICH, Andreas (DESY)

Session Classification: Poster presentations

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