Contribution ID: **51** Type: **Oral**

Towards a statistical model of EGEE load

Wednesday, 13 February 2008 11:35 (25 minutes)

Preliminary results indicate that EGEE job traffic shares some properties with the Internet traffic: the distributions of the inter-arrival times seem to be heavy-tailed and the time series of the loads indicate long-range power-law correlations. Precise characterizations are currently investigated on two aspects.
a) Marginal distributions. Modeling the distributions at different spatial and temporal scales will provide an insight into the way the flow of jobs is actually dispatched on the resources. Relevant statistical approaches are parametric modeling of the nominal behavior as well as the tail behavior of the distributions, and specifically, extreme value theory.
b) Time-dependent structures in the time series. We explore two kinds of well-known stochastic models: Poisson processes, with a possibly non homogeneous or stochastic intensity; and self-similar stochastic models such as fractional Gaussian noises (FGN) and fractional ARIMA processes (ARFIMA).

1. Short overview

The comprehensive monitoring data provided by EGEE makes it possible to analyze from a statistical point of view two characteristics of the activity on the grid, namely the frequency of the arrivals of the jobs on resources and the load on the computer elements. The results of this analysis are relevant in various areas, such as resource dimensioning, providing differentiated Quality of Service (QoS), middleware-level and user-level scheduling.

3. Impact

All of the attempts to provide a differentiated QoS to the EGEE user community share two common problems: 1) accurate, complete publishing of the state of the grid resources and 2) propagation of the scheduling policies implemented on the constituent CEs. Both the state and policy are required by the various scheduling systems at work on the EGEE infrastructure to determine the optimal resource for a particular task. This work addresses the first of these issues. The WMS, and workflow enactors or overlay systems as well, may exploit our results in order to get a more accurate estimation of the expected waiting time at a CE. On the other hand, confirming our initial observations about heavy-tailed distributions and long-range power-law correlations should impact Quality Insurance and Control by proposing concise and meaningful indicators that capture the dynamics of both the collective behavior of users (input flow), and the reaction of the middleware services to these requests.

4. Conclusions / Future plans

The data have been gathered by the GridPP Real Time Monitor. The MATLAB analysis tools will be released through the future Grid Observatory activity, together with updated data from the same source. The statistical characteristics of usage and load will likely undergo significant changes in the near future (LHC activity, communities joining or expanding). The public availability of data and tools will help tracking these evolutions.

Provide a set of generic keywords that define your contribution (e.g. Data Management, Workflows, High Energy Physics)

Job management, Grid Observatory, Statistical Models

Primary authors: GERMAIN-RENAUD, Cecile (Unknown); COLLING, David (Imperial College London); VAZQUEZ, Emmanuel (Supelec)

Presenter: GERMAIN-RENAUD, Cecile (Unknown)

Session Classification: Workflow and Parallelism

Track Classification: Existing or Prospective Grid Services