Predicting Bounds on Queuing Delay in the EGEE grid

Julien Perez, Cecile Germain-Renaud, Balázs Kégl

1. Predicting Bounds on Queuing Delay in the EGEE grid

Contribution ID: 51 Authors: J. Perez (LRI LRI, CNRS-INRIA-University Paris-Sud 11); B. Kégl (LAL LRI, CNRS-University Paris-Sud 11); C. Germain-Renaud (LRI).

Description of the activity

Predicting the performance of schedulers is a notoriously difficult task. As a consequence, grid users might be tempted to work around the standard grid middleware by designing specific strategies, which would be counterproductive if generally adopted. On the other hand, Machine Learning has been successfully applied to performance prediction in distributed and shared environments. This paper reports on experiments on predicting the basic parameters of scheduling in the EGEE framework.

Grid added value

The expected running time (RT) of jobs and expected queuing delay (QD) are important inputs for grid global schedulers. Within gLite, QD is dynamically published by the Computing Elements into the grid information system, which is in turn queried by the scheduling agents called the brokers. At this time, little is known about the accuracy of the prediction of QD. In ordinary production, gLite uses the published QD for minimizing the expected job turnaround time, and errors in this prediction impact grid utilization. gLite also considers all jobs being equivalent, so it is difficult (without reconfiguring the site schedulers) to raise the priority of certain classes of jobs in situations such as social emergency, important events for a scientific community, or software prototyping. To overcome these problems, reinforcement learning has been proposed as a solution for time-constrained scheduling by coupling efficient prediction of QD and scheduling decisions.

Experience or proposed activity

We carried out preliminary statistical analysis (including summary statistics, density estimation, and time series analysis) on scheduler logs of a site of the EGEE grid (the LAL node). We show that the experimental arrival process and service times are extremely far from simple standard models (the classical M/M/N Kendall queue model with Poissonian arrival times and exponential service time), and might in fact exhibit long-range correlation and periodic behaviour. The failure of linear auto regression suggests that non-linear methods are more appropriate in the time series analysis of the expected queuing delay. We are currently investigating such methods (neural networks, Gaussian processes and hidden Markov models), which can be able to take into account both inter-arrival time and load.

Future evolution

The major pitfall in similar analyses is the possible lack of significance of the data. Further research could greatly profit from an easier access to the existing isolated experiments. Furthermore, easier access would also enhance motivation for developing analysis software.