Contribution ID: **94** Type: **Poster**

Early failure detection: a method and some applications

Tuesday, February 12, 2008 4:00 PM (0 minutes)

The complexity of the hardware/software components, and the intricacy of their interactions, defeat attempts to build fault models only from a-priori knowledge. A black-box approach, where we observe the events to spot outliers, is appealing by its simplicity, and large body of experience in quality control. The general challenge is to detect anomalies as soon as possible. Much better solutions than simple thresholding are routinely used in e.g. clinical trials and the supervision of production lines. In the case of abrupt changes, the Page-Hinkley statistics provides a provably efficient method, which minimizes the time to detection for a prescribed false alarm rate. We have applied this method to quantities (e.g. number of arrived and served jobs per unit of time) that are easily computed from the output of existing services. The main result is that we are able to efficiently detect failures of very different origins (e.g. some software bugs, blackholes) without human tuning.

3. Impact

Fast and reliable detection of failures can both raise alarms bringing operator intervention, as well as trigger automatic reaction, e.g. avoid job submission to blackhole sites. The proposed method is quite general, and can be applied at various points in the middleware, including the site level, or by end-user software. Nonetheless, gLite Logging and Bookkeeping service, which concentrates information on the job processing, would be the most effective target. The approach of affecting job scheduling by LB-computed statistics had been used before. Experimental validation and comparison is thus desirable: a significant dataset of "challenge examples" should be available. Examples tagged by system administrators are rare. The Job Provenance (archive of LB data and more) provides the required information from two aspects: easy access to filtered L&B data, and valuable information for calibrating and evaluating failure detection methods wrt. known and well-understood past events.

4. Conclusions / Future plans

The implementation of the statistics per-se is fairly straightforward. The codes for exploiting the test on archived data, including both the extraction of the quantities of interest and the test itself, will be released through the Grid Observatory, in order to demonstrate the performance and scalability levels required for the production environment. Full integration into gLite raises the usual technical issues, and appropriate tools (triggering alarms etc.) remain to be developed.

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

Fault detection, Statistics

1. Short overview

Both Grid middleware services and applications face failures, and the more widely deployed they are, the higher is the price for not detecting the failures early (lost jobs, wasted resources ...). Automated detection, diagnosis, and ultimately management, of software/hardware problems define autonomic dependability. This work report on a generic mechanism for autonomic detection of EGEE failures involving abrupt changes in the behaviour of quantities of interest, and on some applications.

Primary authors: KRENEK, Ales (Masaryk Univeristy); GERMAIN-RENAUD, Cecile (Unknown)

Presenters: KRENEK, Ales (Masaryk Univeristy); GERMAIN-RENAUD, Cecile (Unknown)

Session Classification: Posters

Track Classification: Existing or Prospective Grid Services