Contribution ID: **116** Type: **Oral**

Improved Task Scheduling for Distributed Data Analysis Systems

Thursday 15 April 2010 10:00 (20 minutes)

In this work, we present an evaluation of a late-binding scheduler for HEP distributed analysis, where data is distributed globally with multiple replicas in a non-uniform way. We employ the late-binding technique in order to optimize job placements thereby minimizing the per-job time-to-completion. We evaluate different scheduling and prioritization strategies, and evaluate the approach using a prototype system implemented with the Ganga end-user tool and the DIANE scheduling framework applied to ATLAS distributed analysis.

Detailed analysis

In the context of ATLAS distributed analysis, the presently used workload management systems do not optimally place jobs. At worst jobs are pre-assigned to a single execution site at submit time, and at best jobs are assigned to run on a limited number of closely-located sites (so-called “Atlas Clouds”). This preassignment of jobs to sites leads to two suboptimal behaviours:

1. While a job is waiting in a queue at a site, the data and resource availability can change and therefore the job’s placement at that site becomes less and less ideal.
2. The placement of an entire task (all jobs in the task) at a single or just a few sites can lead to other sites sitting idle or being utilized by lower priority jobs.

With the recent startup of the LHC, the urgency of achieving scientific results in a timely manner is on the critical path. Enabling quasi-interactive analysis on the grid is therefore essential.

Conclusions and Future Work

Making effective use of resources for distributed user analysis is an important challenge with the potential to improve the user experience
substantially. In the HEP community alone, physicists numbering in the thousands use the grid facilities to run jobs numbering in the hundreds of thousands daily.

In the current economic model of the Grid based on public funding and SLAs, resource utilization of a site is a primary metric of its success. Therefore it is a responsibility of the user communities to apply job distribution strategies which reward sites according to their quality of service

Impact

We analyze the impact of job scheduling strategies on handling of job output and subsequent implications for overall data-management strategy for ATLAS collaboration.

To achieve a quasi-interactive distributed analysis system, we consider a few metrics:

1. Time-to-completion. This is a measure of both overall performance for all users at all sites, and for individual users. The variance of this measure is used to evaluate the stability of the schedules. Timely delivery of partial results is also desirable to permit users to take corrective actions as soon as possible.
2. The correlation of job priority with time-to-completion. This is a measure of how well the priority values are respected by the workload management system.
3. Fairness to sites is an important property of the scheduling system as sites are rewarded for running jobs successfully. Therefore, sites should receive a share of the global jobs proportional to their quality, where quality relates to their efficiency and the popularity of their resident data.

Keywords

data management, scheduling

URL for further information

http://cern.ch/diane

Primary authors: Dr VAN DER STER, Daniel (CERN); Mr MOSCICKI, Jakub (CERN); Dr LAMANNA, Massimo (CERN)

Presenter: Mr MOSCICKI, Jakub (CERN)

Session Classification: Data Management

Track Classification: Scientific results obtained using distributed computing technologies