Contribution ID: **9** Type: **Oral**

Distributed system for genetic linkage analysis using EGEE and BOINC

Tuesday, 12 February 2008 14:00 (20 minutes)

Tasks are submitted via web and parallelized into thousands or even millions of CPU-bound jobs ranging from a few seconds to a few minutes long. Efficient and reliable execution is complicated due to unbounded queuing times, high execution and scheduling overheads, high job failure rates and insufficient scalability of the EGEE middleware.
Our solution is to first submit lightweight clients which, when started on remote resources, fetch the actual jobs from the central job server and execute them. For this purpose we adopt open-source BOINC platform, used in the last few years for large-scale cycle-stealing such as SETI@HOME and many others. Built for volatile desktop environments, BOINC is capable of efficiently managing billions of jobs and millions of unreliable clients, yielding high performance through sophisticated scheduling mechanisms to overcome network, hardware and software faults. Furthermore, BOINC is firewall friendly and has a built-in accounting functionality.

3. Impact

Our system decouples the application logic from the job submission and management mechanisms, essentially building on demand a virtual dedicated cluster from EGEE resources .
The system has two main components. One application-independent part maintains the required amount of active BOINC clients in EGEE (i.e. the number of resources in the virtual cluster) by monitoring and actively rescheduling stuck, failed or evicted BOINC clients back into the grid. A thin wrapper over publicly-available BOINC clients is used to enable their execution in EGEE.
Another part, based on BOINC server, maintains the queue of the actual application jobs and accommodates the partial results. The jobs and results are communicated in a secure way, the integrity and validity are checked and user-specified routines are invoked to produce the final result. The system can efficiently execute even seconds-long jobs, as BOINC clients run them back-to-back, caching the executable and constant data remotely.

URL for further information:

http://bioinfo.cs.technion.ac.il/superlink-online

4. Conclusions / Future plans

Execution of over million jobs, each ranging from a few seconds to minutes, completed within 30 days on BIOMED VO CPUs, consuming about 2 TFLOPs on 300 (average) concurrently executing clients (from 100 to 700). The run was fully-automated and completed despite the failures of the BOINC server hardware, UI and broker nodes.
The system is generic and will facilitate porting other applications. The use of BOINC allows us to effortlessly integrate the clusters and desktop grids outside of EGEE.

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

Bioinformatics, Job management, short jobs, BOINC, pilots

1. Short overview

Genetic linkage analysis is a statistical tool used to seek for disease-provoking genes. However many analyses are infeasible due to the high computational demands. Superlink-online web portal enables such demanding

analysis tasks through their automated parallelization, submission, and execution on thousands of BIOMED VO CPUs. We designed a system which efficiently and reliably executes millions of jobs, overcoming high scheduling overheads, unbounded queuing times and job failures.

Primary author: Mr SILBERSTEIN, Mark (Technion - Israel Institute of Technology)

Co-authors: Mr SHAROV, Artyom (Technion - Israel Institute of Technology); Prof. SCHUSTER, Assaf (Technion - Israel Institute of Technology); Prof. GEIGER, Dan (Technion - Israel Institute of Technology)

Presenter: Mr SHAROV, Artyom (Technion - Israel Institute of Technology)

Session Classification: Life Sciences

Track Classification: Application Porting and Deployment