Contribution ID: **91** Type: **Oral contribution**

GPS@: Bioinformatics grid portal for protein sequence analysis on EGEE grid

Wednesday 1 March 2006 14:00 (15 minutes)

One of current major challenges in the bioinformatics field is to derive valuable information from the complete genome sequencing projects, which provide the bioinformatics community with a large number of unknown sequences. The first prerequisite step in this process is to access up-to-date sequence and 3D-structure databanks (EMBL, GenBank, SWISS-PROT, Protein Data Bank…) maintained by several bio-computing centres (NCBI, EBI, EMBL, SIB, INFOBIOGEN, PBIL, ⋯). For efficiency reasons, sequences should be analyzed using the maximal number of methods on a minimal number of different Web sites. To achieve this, we developed a Web server called NPS@ [1] (Network Protein Sequence Analysis) that provides biologists with many of the most common tools for protein sequence analysis through a classic Web browser like Netscape, or through a networked protein client software like MPSA [2]. Today, the genomic and post-genomic web portals available have to deal with their local cpu and storage resources. That's why, most of the time, the portal administrators put some restrictions on the methods and databanks available. Grid computing [3], as in the European EGEE project [4], will be a viable solution to foresee these limitations and to bring computing resources suitable to the genomic research field.
Nevertheless, the current job submission process on the EGEE platform is relatively complex and unsuitable for automation. The user has to install an EGEE user interface machine on a Linux computer (or to ask for a account on a public one), to remotely log on it, to init manually a certificate proxy for authentication reasons, to specify the job arguments to the grid middleware using the Job Description Language (JDL) and then to submit the job through a command line interface. Next, the grid-user has to check periodically the resource broker for the status of his job: "Submitted", "Ready", "Scheduled", "Running", etc. until the "Done"status. As a final command, he has to get his results with a raw file transfer from the remote storage area to his local file system.
This mechanism is most of times off-putting scientist that are not aware of advanced computing techniques. Thus, we decide to provide biologists with a user-friendly interface for the EGEE computing and storage resources, by adapting our NPS@ web site. We have called this new portal GPS@ for "Grid Protein Sequence Analysis", and it can be reached online at http://gpsa.ibcp.fr, yet for experimental tests only. In GPS@, we simplify the grid analysis query: GPS@ Web portal runs its own EGEE low-level interface and provides biologists with the same interface that they are using daily in NPS@. They only have to paste their protein sequences or patterns into the corresponding field of the submission web page. Then simply pressing the "submit"button launches the execution of these jobs on the EGEE platform. All the EGEE job submission is encapsulated into the GPS@ back office: scheduling and status of the submitted jobs. And finally the result of the bioinformatics jobs are displayed into a new Web page, ready for other analyses or for results download in the appropriate data format.

[1] NPS@: Network Protein Sequence Analysis. Combet C., Blanchet C., Geourjon C. et Deléage G. Tibs, 2000, 25, 147-150.
[2] MPSA: Integrated System for Multiple Protein Sequence Analysis with client/server capabilities. Blanchet C., Combet C., Geourjon C. et Deléage G. Bioinformatics, 2000, 16, 286-287.
[3] Foster, I. And Kesselman, C. (eds.) : The Grid 2 : Blueprint for a New Computing Infrastructure, (2004).
[4] Enabling Grid for E-sciencE (EGEE), online at www.eu-egee.org

Primary author: Dr BLANCHET, Christophe (CNRS IBCP)

Co-authors: Dr COMBET, Christophe (CNRS IBCP); Prof. DELEAGE, Gilbert (CNRS IBCP); Mr LEFORT, Vincent (CNRS IBCP)

Presenters: Dr BLANCHET, Christophe (CNRS IBCP); Mr LEFORT, Vincent (CNRS IBCP)

Session Classification: 1a: Life Sciences

Track Classification: Life Science