Contribution ID: **181** Type: **Poster**

Large scale Bioinformatics portal on EGEE Grid : the GPSA example

Describe the scientific/technical community and the scientific/technical activity using (planning to use) the EGEE infrastructure. A high-level description is needed (neither a detailed specialist report nor a list of references).

Bioinformatics analysis of data produced by high-throughput biology, for instance genome projects, is one of the major challenges for the next years. Some of the requirements of this analysis are to access up-to-date databanks and relevant algorithms. Since 1998, we are developing the Network Protein Sequence Analysis (NPS@) Web server , that provides the biologist with some most common resources for protein sequence analysis, integrated into several predefined and connected workflows.

Report on the experience (or the proposed activity). It would be very important to mention key services which are essential for the success of your activity on the EGEE infrastructure.

The bioinformatics portal GPSA was improved with the adding of MPI capabilities for some of its applications: BLAST and CLUSTALW. This work has been done with the help of the MPI group , and according to the recommendations produced during the activity of this group. CNRS IBCP has followed these recommendations, and some time adapted them, to be able to run MPI jobs on the EGEE productin platform. The bioinformatics applications that can now be ran on EGEE grid are mpiBLAST and mpiCLUSTALW. There are both famous applications in the Bioinformatics field and are ran daily by lots of scientists. These MPI tools have been validated on the EGEE production platform mainly on the current most usable MPI-site at IN2P3 LAL in Orsay. Some other tests has been done on few others MPI-sites but not with the same success rate.

Describe the added value of the Grid for the scientific/technical activity you (plan to) do on the Grid. This should include the scale of the activity and of the potential user community and the relevance for other scientific or business applications

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 usual submission Web pages. Then simply pressing the "submit" button, they get the results of executing these jobs on the EGEE grid platform. All the EGEE workload management is encapsulated into the GPS@ low-level layers: submitting, monitoring and getting the results of the bioinformatics jobs.
Running MPI-parallelized bioinformatics applications on EGEE production platform was still a challenge, and especially MPI-BLAST that has raised in the past several tricky points from MPI and EGEE sites configuration. One other applications is the CLUSTALW multiple alignment engine. Both programs are C/C++ writen, and could be compiled with gcc. (GCC release available from SL3.0.9)

Abstracts for online demonstrations must provide a summary of the demo content. Places for demos are limited and this summary will be used as part of the selection procedure. Please include the visual impact of the demo and highlight any specific requirements (e.g. network connection). In general, a successful demo is expected to have some supporting material (poster) and be capable of running on a single screen or projector.

GPS@ grid web portal (Grid Protein Sequence Analysis, http://gpsa-pbil.ibcp.fr) is the port of our Bioinformatics integrated portal, the NPS@ protein analysis portal, and would provide the biologist with a user-friendly interface for the GRID resources (computing and storage) made available by the project EGEE (2004-2008). This genomic grid user interface hides the mechanisms involved for the execution of Bioinformatics analyses on the grid infrastructure. The bioinformatics algorithms and databanks have been distributed and registered on the EGEE grid and GPS@ runs its own EGEE interface to the grid. In this way, GPS@ portal simplify the Bioinformatics grid submission, and provide biologist with the benefit of the EGEE grid infrastructure to analyze large biological datasets with parallelized bioinformatics software: for example doing a large multiple alignment with CLUSTALW, or analysing large protein sequence sets, with BLAST, on the Grid.

Primary author: Dr BLANCHET, Christophe (CNRS IBCP)

Co-authors: Mr MICHON, Alexis (CNRS IBCP); Dr COMBET, Christophe (CNRS IBCP); Prof. DELÉAGE, Gilbert (CNRS IBCP)

Presenter: Dr BLANCHET, Christophe (CNRS IBCP)

Track Classification: Demo and Poster session