Contribution ID: **29** Type: **oral presentation**

BiG: A Grid Service to Distribute Large BLAST Runs

Friday 11 May 2007 10:00 (20 minutes)

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 community use habitually computers for their large computationally-intensive genomic and proteomic analysis. Bioinformatics is a consolidated community in the frame of EGEE since phase I, who is using the resources through applications such as the GPS@ portal.
The main challenge of this community (in the frame of the present work) is the identification and annotation of the function of protein and nucleotide sequences through their comparison with other well-known databases.

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 BiG service is already operational in the EELA infrastructure. It is being used through the Ibero-American Bioinformatics Portal, (http://portal-bio.ula.ve/) and runs on the resources of the Technical University of Valencia and Federal University of Rio de Janeiro, being currently other resource providers under negotiation. Registered users can access the Grid service through the portal. Other source of users is the community of biologists using B2GO tool (http://www.blast2go.de/), mainly interested on agricultural biology. The average usage has been of 1.5 CPU/days per day since its release in June, but jobs take sequentially more than 3 CPU/days, and results are obtained in a few hours. This peak demand of resources has fitted very well the usage fashion of Grids. This service has been used for analysis of segments of pathogens causing diseases in humans and agriculture.

With a forward look to future evolution, discuss the issues you have encountered (or that you expect) in using the EGEE infrastructure. Wherever possible, point out the experience limitations (both in terms of existing services or missing functionality)

The main problems we have experimented have been related to the reliability of the services and the proxy renewal. Automatic proxy renewal and VOMS credential through myproxy repositories seems to be still not completely solved. Other problems are not expected, since the system has extensively run on a similar production infrastructure.

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

BLAST-enabled Grid services do exist. However, most of them use sequential BLAST as their processing engine. However, MPI-enabled versions of BLAST do

exist, and performance of large runs could be extensively improved. Then, the objective has been to develop an off-line, robust service to BLAST, based on MPI-Blast and delivering higher performance. Currently, this services are used daily by hundreds of users for short runs.
The service can be integrated in portals or as a part of an application, since it exports an interface based on a statefull web service that enables launching the execution, disconnecting and reconnecting later on. The service splits the input sequences and the target database in multiple jobs and MPI processes, being possible to launch several simultaneous analyses on different protein and nucleotides databases.

Authors: Mr APARICIO, Gabriel (UPV); Dr BLANQUER, Ignacio (UPV); Prof. HERNÁNDEZ, Vicente (UPV)

Presenter: Dr BLANQUER, Ignacio (UPV)

Session Classification: Workflow

Track Classification: Workflow