Contribution ID: **142** Type: **Demonstration**

Bioinformatics portal on Grid: the GPSA - Grid Protein Sequence Analaysis - case.

Tuesday, February 12, 2008 4:00 PM (0 minutes)

Bioinformatics analysis of data produced by high-throughput biology, for instance genome projects, is one of the major challenges for the coming years. Two of the requirements for this analysis are access to up-to-date databanks (of sequences, patterns, 3D structures, etc.) and access to relevant algorithms (for sequence similarity, multiple alignment, pattern scanning, etc.). Since 1998, we have been developing the Web server NPS@ (Network Protein Sequence Analysis), that provides the biologist with many of the most common resources for protein sequence analysis, integrated into a common workflow. We have adapted this portal to the EGEE Grid. The bioinformatics grid portal GPS@ ("Grid Protein Sequence Analysis") simplifies and automates the EGEE grid job submission and data management mechanisms using XML descriptions of available bioinformatics resources: algorithms and databanks.

3. Impact

One major problem with a grid-computing infrastructure is the distribution of files and binaries, as for the BLAST or ClustalW algorithms, through the job submission process. Sending a binary of the algorithm to a node on the grid is quite simple because of its size (few kilobytes) and can be done at each execution. Putting on the grid a databank, ranging from tens of megabytes (as Swiss-Prot) to gigabytes (as EMBL), consumes a large part of network bandwidth, and greatly increases the execution time if done inappropriately. The GPSA 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, the GPS@ portal simplifies the bioinformatic grid submission, and provides biologists with the benefits of the EGEE grid infrastructure to analyze large biological datasets.

If demonstration is requested please explain what visual or interactive aspects of the contribution necessitate a demonstration rather than a presentation or poster?

A demo can be interactive with the visitors (what could not possible with an oral or poster presentation), and will show in real time the efficiency and the online-availibility of the GPSA portal through different bioinformatics scenarii.

URL for further information:

http://gpsa-pbil.ibcp.fr

4. Conclusions / Future plans

The GPS@ grid Web portal (Grid Protein Sequence Analysis) is a bioinformatic integrated portal that provides a biologist with a user-friendly interface to the grid resources (computing and storage) made available by the EU-EGEE project. The GPS@ portal will be used as case study in the context of the EGEE PORTAL group to implement the recommendations raised by this group.

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

Bioinformatics, Portal.

1. Short overview

Although grid computing offers great potential for executing large-scale bioinformatics applications, practical utilization is constrained by the middleware's ease-of-use. Biologists are generally unwilling to use command-line interfaces or complex toolkits consisting of numerous components, such as most current grid middlewares. Integrating the required applications in a Web portal is then an efficient way to bring these scientists to the grid.

Primary author: Dr BLANCHET, Christophe (CNRS IBCP)

Co-authors: Mr MICHON, Alexis (CNRS IBCP); Mrs ELOTO, Christelle (CNRS IBCP); Dr COMBET, Christophe (CNRS IBCP)

Presenter: Dr BLANCHET, Christophe (CNRS IBCP)

Session Classification: Demonstrations

Track Classification: Application Porting and Deployment