Contribution ID: **180** Type: **poster**

Grid efficiency for high throughput and data-intensive analysis of bacterial genomic sequences

Wednesday 9 May 2007 17:30 (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).

This work handles the grid performances of high throughput comparison of genomes, producing in parallel tens of Gigabytes of data. The community is composed of Bioinformatics and grid partners. Input/output data were stored on Storage Elements as flat files, according to a logical/physical directory structure. The grid efficiency was depending on the balance between wide parallel resource access and data access bottle necks. Methods and algorithms were designed for supporting high throughput.

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.

For supporting both high parallelism for independent runs and manageable data format for a large and complex information, I/O data were stored on grid Storage Elements according to a specific name-space structure of Logical File Names, which are mapped on their physical paths by a grid File Catalog.
Whenever possible, the high computing parallelism was reduced by properly grouping a set of elementary steps in single executable nodes, running on single grid Worker Nodes. Once defined parallel executable nodes, the grid middleware allows the users to group and manage many of them through a single structure. The grid Workload Management System (WMS) supports jobs of Collection and DAG type, where many nodes are grouped in a single job and are distributed by the WMS on different grid computing resources. In this task we successfully used both few Collection JDLs from hundreds to

thousands of nodes, and hundreds of DAG JDLs including thousands of nodes.

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 grid File Catalog was able to store millions of paths: large file divisions of output support high parallelism for independent runs. However, as output files become input files next task steps, multiple file access on shared storages could cause reading bottle necks. The grid efficiency depends on the balance between wide resource parallel access and emergence of reading bottle necks. This balance can be optimized by specific algorithms or by future improvings of storage access.

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

Technologies for genome sequencing disclosed large amount of knowledge about genes and proteins. The accumulation of genomic data in public databases allows a comprehensive comparative large scale investigation with the aim of solving many important tasks in molecular biology, such as functional and structural genomic annotation. Starting from these comparative analysis we can also infer phylogenetic relationships between species. In this context the grid added value consists in distributed parallel computing resources and data storage for making possible data-intensive and high throughput comparisons which cannot be supported by local farm. Once extracted the genomic data from the public database of the National Center for Biotechnology Information, our task included two massively parallel steps: one dedicated to protein sequence comparisons and the second checking the consistency of the output. They required about 2 millions and 8 thousand millions of independent runs respectively.

Authors: Dr BARTOLI, Lisa (Biocomputing Group, University of Bologna); Dr CAROTA, Luciana (INFN-CNAF-Bologna, Italy); Dr MONTANUCCI, Ludovica (Biocomputing Group, University of Bologna); Dr MARTELLI, Pier Luigi (Biocomputing Group, University of Bologna); Dr FARISELLI, Piero (Biocomputing Group, University of Bologna); Prof. CASADIO, Rita (Biocomputing Group, University of Bologna)

Co-authors: Dr DONVITO, Giacinto (INFN-Bari,Italy); Prof. MAGGI, Giorgio (INFN-Bari,Italy)

Presenter: Dr CAROTA, Luciana (INFN-CNAF-Bologna, Italy)

Session Classification: Poster and Demo Session

Track Classification: Poster session