Contribution ID: **123** Type: **oral presentation**

BioinfoGRIDBlast: a new approach at automatic database management, update and replication for Grid-enabled BLAST

Thursday 10 May 2007 14:20 (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).

The EGEE platform is most useful for research groups producing uneven computational workloads. In these cases the cost of joining the EGEE Grid is substantially lower of that of an owned cluster. This can provide benefits to many research groups such as bioinformatics research groups. BioinfoGRID EU project is already using the Grid for computationally intensive tasks. A dramatic performance gain is relatively easy to obtain for applications which can be trivially parallelized.

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.

BioinfoGRIDBlast is an effort for providing the user a Blast system over the Grid with automatic database management.
The core for BioinfoGRIDBlast provides workload slicing into small jobs, and jobs tracking and management over the Grid. On top of this the following functionalities have been implemented:
1) An updater engine maintains the BLAST biological databases on the Grid constantly updated. The updates are polled and fetched from the FTP sites of origin for each database.
2) Older versions of the biological databases are also kept available on the Grid,but instead of uploading each version in full BioinfoGRIDBlast implements a patch system based on xdelta to dramatically reduce storage costs.
3) A dynamic Replication Engine (RE) keeps track of the usage for each database. The RE then constantly adapts the number of replicas for each database, keeping it proportional to the database's recent usage. This keeps Blast responsiveness high while keeping storage costs bearable.

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)

We noticed a few tricky commands such as lcg-cp not exiting after download is completed. These can be worked around relatively easily once they are known. The efficiency of the queueing system might be improved: sometimes not the optimal queues are chosen by the Broker.

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 if a typical use case for a computationally intensive application which can leverage the Grid obtaining dramatic speedups compared to local execution. Blast is also very widely used in Bioinformatics research groups. A working, centrally maintained and optimized grid Blast installation can provide significant benefits for the bioinformatics community. Problems to be faced when porting Blast to the Grid platform are: Grid job management, databases updates management, availability of all versions of several databases, amount of replication for each of the databases (storage costs are significant). We addressed such problems in our BioinfoGRIDBlast. BioinfoGRIDBlast sports an automatic management system for keeping the databases constantly updated and to dynamically adjusts the amount of replication for each database. The amount of replication is based on the recent usage amount for each database; this constantly optimizes the balance of storage costs vs grid availability.

Primary author: Mr TROMBETTI, Gabriele (CNR-ITB)

Co-authors: Dr ORRO, Alessandro (CNR-ITB); Mr MERELLI, Ivan (CNR-ITB); Dr MILANESI, Luciano (CNR-ITB)

Presenter: Mr TROMBETTI, Gabriele (CNR-ITB)

Session Classification: Data Management

Track Classification: Data Management