Contribution ID: **37** Type: **Oral**

Analysis of Metagenomes on the EGEE Grid

Tuesday 12 February 2008 11:00 (20 minutes)

Metagenomic analysis requires several iterations of alignment and phylogenic classification steps. Source samples reach several millions of sequences. These sequences are compared to the eukaryotic species of the "Non-redundant" database.
The deployment process involves three stages: First, public databases are copied in relevant SEs to reduce the access time by increase the geographic replication of the data. Second, the available resources are tested through short test jobs that check the different operations. Finally, the experiment is performed.
The sequences of the source sample are split into different jobs. Each job is submitted through the RBs to the CEs that have been selected in the test phase. Jobs copy all the relevant databases from close SEs to the local storage, install locally the BLAST and clustalW software and execute the scripts. After the completion of the job, results are copied back through the SEs and GridFTP as a backup solution.

3. Impact

Metagenomic analysis is needed in the cases in which it is impossible to grow significant samples of isolated specimens. Many bacteria cannot survive alone, and require the interaction with other organisms. In such cases, the information of the DNA available belongs to different kinds of organisms.
Four experiments have been executed with up to 800K sequences. The environment has enabled to reach performances of more than 8.000 sequences per hour. The complete experiment was performed in 10 days. A standard PC would have taken 1.5 CPU years in optimal conditions and would not reach more than 66 seqs/hour).
However, the failure ratios of the jobs are high. In the largest case, 55% of the jobs were resubmitted. From the failing jobs, 49% end in the aborted state, 26% had problems accessing the catalogue, 7% fail using the wget command, 4% could not install the BLAST tool due to problems in the configuration of the compiler and 14% were cancelled due to its long duration.

URL for further information:

www.grycap.upv.es/bio

4. Conclusions / Future plans

An environment has been developed to fragment, automate and check the operations of Metagenomic analysis. It has been tuned-up considering the most efficient and reliable resources, the optimal job size, and the data transference and database reindexation overhead. The environment re-submits faulty jobs, detect endless tasks and ensure that the results are correctly retrieved.
New metagenomic studies are being completed, and the full processing chain is being enriched with more steps.

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

biomed, biocomputation, metagenomics, service challenges

1. Short overview

A Metagenome is a sample of several complete genomes of several living beings. The analysis of metagenomes is a key issue in biological research, but it is a computationally intensive task.
An environment has been developed and deployed on top of EGEE and has been successfully used for the

analysis of four metagenomes from digestive track, soil and sea bacteria consortia. The environment has enabled to complete a study which would have taken 1.5 CPU years in optimal conditions in 10 days.

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

Presenter: Dr BLANQUER, Ignacio (UPV)

Session Classification: Life Sciences

Track Classification: Application Porting and Deployment