Contribution ID: **67** Type: **Oral**

Estimating the Performance of BLAST runs on the EGEE Grid

Tuesday 13 April 2010 11:00 (15 minutes)

Estimating the response time of large experiments is a key issue for achieving an efficient load balancing and minimizing the failure rate. This requires having a good knowledge of both the application and the infrastructure. This work describes a set of experiments that have conducted to the definition of a performance model that can be used to estimate the response time of the selected resources and to adapt the partition of the load to the dynamic status of the resources.

Detailed analysis

The work has concentrated on two lines: First, the analysis and selection of the parameters that provide information about the performance of a resource, using the GLUE schema (SPECint, SPECfp, average and maximum queuing time among others). Second, the analysis of the factors affecting the performance of BLAST searches (reference database size, input sequence size, sequence length, operation parameters and resemblance of input and reference data). Therefore, several experiments have been completed fixed one or several of the variables. The experiments were repeated fixed the same conditions to reduce variability and increase significance. Input data has been based on the UniProtdatabase, introducing random modifications to control the execution conditions. Submission was performed explicitly selecting the CEs and measuring the information provided by the BDII. The publication delay is also considered as an error factor. Results were obtained in different resources at different times, and a performance model is being adjusted using this information.

Conclusions and Future Work

The work describes a model to estimate response time of BLAST runs in the EGEE grid. The work will be completed with experiments validating the model and introducing new parameters from other components, such as the Workload Manager and the LRMS. This will provide a more complete model to characterise the execution of BLAST jobs in the grid, and the improvement of scheduling policies. Dynamic adaptation of load is feasible, especially on pilot submission schemas, which are also considered.

Impact

The results obtained have revealed that database size and input data size have a direct linear impact on the response time, thus leading to a simple prediction model. Variability due to similarities on the input and reference data, as well on the request for larger or shorter output has a minor impact on the performance and can be ignored. Results on the effect of the performance capabilities of the results are on the way. Static values will be easier to fix, but dynamic parameters will need to be feed directly on the model.
Although there are other studies on the literature about the estimation of performance of BLAST, in the knowledge of the authors this is the first study on a production Grid infrastructure. The results will be relevant, not only for the users of the biomed community, but also for the developers of QoS schedulers. This is an important research line, especially considering the sustainability of Grid infrastructures, and the consideration of external providers.

Keywords

Performance estimation, BLAST, dynamic scheduling

URL for further information

www.grycap.upv.es

Authors: Mr CARRIÓN-COLLADO, Abel (UPV); Dr BLANQUER-ESPERT, Ignacio (UPV); Prof. HERNÁNDEZ-GARCÍA, Vicente (UPV)

Presenter: Dr BLANQUER-ESPERT, Ignacio (UPV)

Session Classification: Bioinformatics

Track Classification: Experiences from application porting and deployment