The PortalGIGA Project

Fabrício A. B. Silva[1], Walfredo Cirne[2], Hermes Senger[1], Eduardo R. Hruschka[1], Cléver R. G. de Farias[1], Pedro Pascutti[3], Carla Osthoff[4], Laurent Dardenne[4], Reynaldo Novaes[5]

[1]Universidade Católica de Santos (UNISANTOS) - Programa de Mestrado em Informática - Santos – SP – Brazil
{fabricio,senger,erh,cleverfarias}@unisantos.br

[2]Universidade Federal de Campina Grande (UFCG) – Departamento de Sistemas e Computação - Campina Grande – PB – Brazil
walfredo@dsc.ufcg.edu.br

[3]Universidade Federal do Rio de Janeiro (UFRJ) – Instituto de Biofísica Carlos Chagas Filho (IBCCF) - Rio de Janeiro – RJ - Brazil
pascutti@biof.ufrj.br

[4]Laboratório Nacional de Computação Científica (LNCC) – Petrópolis – RJ – Brazil
{osthoff, dardenne}@lncc.br

[5]HP Brasil R&D – Porto Alegre – RS – Brazil
reynaldo.novaes@hp.com

Abstract. *This paper describes the PortalGIGA project, whose main objective is to create a Portal to allow the seamless execution of Bag-of-Tasks applications on the Grid. The target applications of the PortalGIGA project are data mining, black-oil reservoir simulation and molecular dynamics simulation. This paper presents the overall architecture of the portal and gives a general overview of the target applications.*

1. Introduction

A computational grid, or simply grid for short, provides access to heterogeneous resources in a geographically distributed area, allowing the integration of these heterogeneous and distributed resources into a unified computer resource. Computational grids are usually built on top of specially designed middleware platforms, the so-called grid platforms. Grid platforms enable the sharing, selection and aggregation of a variety of resources including supercomputers, servers, workstations, storage systems, data sources and specialized devices that are geographically distributed and owned by different organizations [Baker et al. 2002].

Among the most suitable applications for running on a grid are the Bag-of-Tasks (BoT) applications, which are parallel applications whose tasks are independent of each other. Examples of BoT applications include Monte Carlo simulations, massive searches (such

as key breaking), parameter-sweep applications, image manipulation, and several data mining algorithms.

The main objective of the PortalGIGA project is the seamless execution of three classes of applications on a Grid Platform: data mining algorithms, black-oil reservoir simulation and molecular dynamics simulation. The Grid infrastructure needed will be built over the middleware MyGrid [Cirne et al. 2003], which is especially designed for running BoT applications on a grid, and the GIGA Network [GIGA 2005]. The GIGA Network is a 10 Gbps network interconnecting several research centers in Brazil. The GIGA Network and the PortalGIGA project are funded by RNP, which is the agency that manages the Brazilian National Research Network.

2. Target Applications

This section presents the target applications of the PortalGIGA project: data mining, black oil reservoir simulation and molecular dynamics simulation. The section also exemplifies how these applications can be executed on a grid as BoT applications.

2.1. Data mining

Some recent works suggest that grids are natural platforms for developing high performing data mining services [Canataro and Talia 2003, Orlando et al. 2002]. More specifically, [Orlando et al. 2002] describes the application of two data mining algorithms (DCP and K-means) in the Knowledge Grid proposed by Canataro [Canataro and Talia 2003]. The DCP Algorithm enhances the popular Apriori Algorithm [Agrawal 1996], which is an algorithm for mining association rules, whereas the K-Means [Kaufman and Rousseeuw 1990] is a popular clustering algorithm. However, other DM techniques can take advantage of a grid infrastructure. In general, several data-mining applications can be classified as parameter-sweep, BoT applications [Silva et al. 2004]. Since those applications are composed of independent tasks, they are suitable for execution in a grid environment, where machine heterogeneity and significant network delays are common.

In the PortalGIGA project, we will make available a specialized Portal for data analysts, which will be able to execute efficiently BoT data mining applications on the grid. We also intend to give support to other steps of the process of Knowledge Discovery in Databases [Fayyad et al. 1996], such as data preprocessing/transformation and support for interpretation and evaluation of results.

2.2. Reservoir Simulation

A reservoir simulator is a sophisticated computer program used to predict the future performance of a reservoir based on its current state and past performance. It is the main computational tool used by petroleum engineers to explore methods for increasing the ultimate recovery of hydrocarbons from a reservoir. The simulator solves the system of partial differential equations describing multiphase fluid flow (oil, water, gas) in a porous reservoir rock.

Simulation of large reservoirs or entire fields containing millions of grid blocks and as many as a thousand wells entails solution of a large set of differential equations. The computer resources required to solve the large set of equations governing multiphase flow in the reservoir can grow rapidly depending on the reservoir size, number of grid blocks and the type of steeping scheme used on the model.

The mechanics of the simulations can be described as follows: the reservoir is first divided into segments using X,Y,Z axes. Rock and fluid properties are then assigned to each block to describe the reservoir system. Computations are carried out for all phases in each block at discrete time steps. The results usually consist of production volumes and rates, pressure and saturation distributions, material balance errors, and other process specific information provided at selected time steps.

In the PortalGIGA project we intend to make available a portal that will be able to execute on the grid several simulations of a given reservoir, each simulation with a different configuration of production/injection wells. In the first version of the Portal we will use BOAST [Chang et al. 1992] as black-oil reservoir simulation program. BOAST is a multi-phase, three-dimensional black-oil simulator developed by the US Department of Energy. The BOAST program simulates isothermal, Darcy flow in three dimensions. It assumes reservoir fluids can be described by three fluid phases (oil, gas, and water) of constant composition with physical properties that depend on pressure only. These reservoir fluid approximations are acceptable for a large percentage of the world's oil and gas reservoirs.

2.3. Molecular Dynamics Simulation

Currently researchers at IBCCF (UFRJ) and LNCC, both members of the PortalGIGA project, are studying HIV-1's protease and a family of her inhibitors through molecular dynamics techniques.

The inhibition of HIV's protease has been reaching success on clinical treatments. However, the fast emergence of drug resistance by the HIV virus is a difficult problem to be resolved. Different HIV's resistant mutations have been appearing according to the region of the planet. For example, in southern Brazil there are resistant mutants that have different amino acid sequences, when compared to HIV variants found in other parts of the country. Moreover, a World Health Organization's forecast states that around ninety percent of the people infected by the virus, at the beginning of this century, will belong to the third world. Therefore, it is necessary to increase both theoretical and experimental efforts that enable the development of new drugs against regional mutants.

In the PortalGIGA project a researcher will be able to run concurrently several simulations of inhibitor/protease interactions on a computational grid. It will be generated several mutated variants of HIV and protease inhibitors, and each pair will be executed on an available node of the grid.

3. Architecture of the Portal

Figure 1 shows the basic architecture of the PortalGIGA Platform.

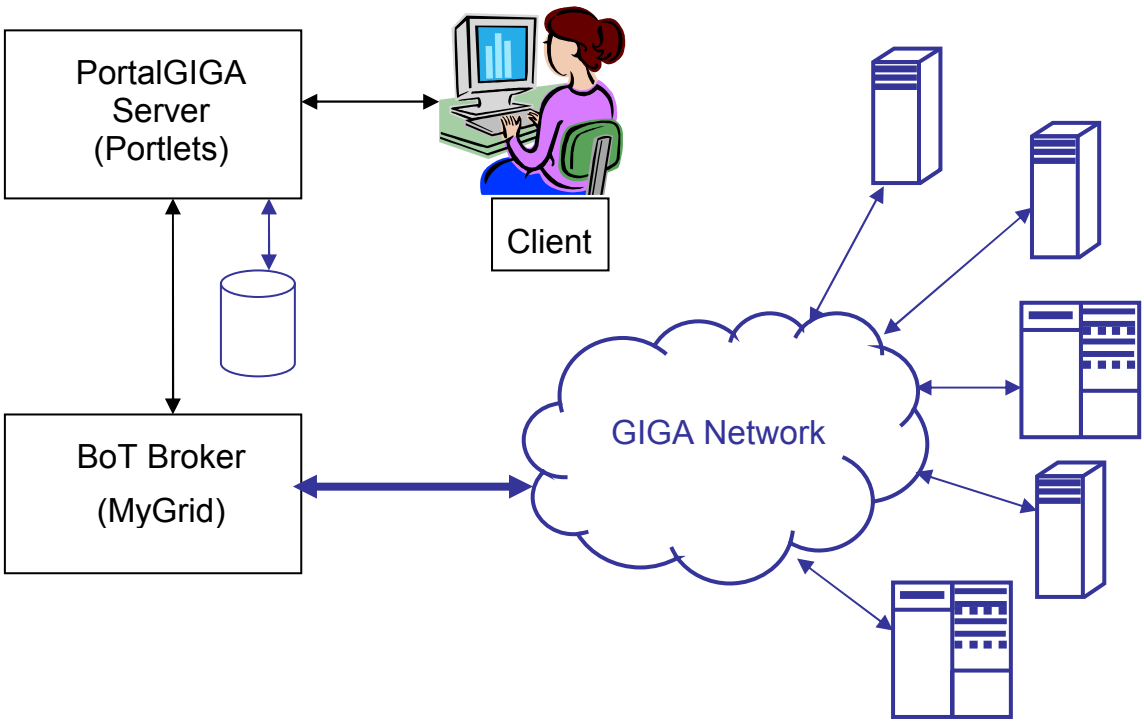

Figure 1 – Architecture of the PortalGIGA project

The Portal has two main components: the PortalGIGA server and the BoT scheduler (Mygrid). The PortalGIGA server will be built using Java Portlet technology [Java 2005] and will be made available through a Jetspeed server [Jetspeed 2005]. A Portlet is a web component written in Java, which processes requests and generates dynamic content. For each target application of the PortalGIGA project, there will be a set of dedicated portlets that will be responsible to generate/access input data, to execute the BoT application and to receive, evaluate and store the results.

As the BoT application broker we will use the MyGrid middleware. The MyGrid platform [Cirne et al. 2003] was especially conceived to support the execution of BoT applications, which constitute a class of parallel applications that can be partitioned in several independent tasks. Usually, these tasks have an infrequent need for communication.

The main benefits of MyGrid are twofold: minimal installation effort and ease of use. Most grid platforms (for an example see [Foster 1997]) can only be installed and configured by system administrators. Moreover, installation procedures are usually manually repeated in a considerable number of machines. MyGrid enables regular users to create their own grid to run applications on whatever resources they have access to, without the need for these users to get involved into grid details and administrative procedures for heterogeneous platforms.

Since MyGrid focuses on BoT applications, its working environment consists of a small set of services to enable its users to manipulate their files on the grid. Consequently, neither previous software installation nor shared file system are needed on machines. A user is required to install and configure MyGrid only on one machine, which is called home machine. Interactions with other machines are supported by the Grid Machine Interface (GMI). The GMI provides a minimal set of services that must be available in a machine so it can be used as a machine for grid computing, the so-called grid machine. These services consist of: (1) remote executing on a grid machine; (2) termination of a running task; and (3) file transfers between the home and grid machines.

4. Conclusion

This paper described the general architecture and the target applications of the PortalGIGA project, which is a portal conceived to permit the seamless execution of three types of BoT applications on a grid using the GIGA network. The PortalGIGA project has started on December 2004 and has an expected duration of 12 months. Five research institutions participate on the project, and we expect that the resulting portal will be useful for several research groups in the areas of data mining, black-oil reservoir simulation and molecular dynamics simulation.

5. Acknowledgements

We would like to thank RNP for the financial support.

References

Agrawal, R. et al. (1996) "Fast Discovery of Association Rules*"* In: Advances in Knowledge Discovery and Data Mining, Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Editors, MIT Press, pp. 307-328.

Baker, M., Buyya, R., Laforenza, D. (2002) "Grids and Grid Technologies for Wide-area Distributed Computing", Software - Practice and Experience, v. 32, pp. 1437-1466, John Wiley and Sons ltd, England.

Canataro, M., Talia, D. (2003) "The Knowledge Grid", Communications of the ACM, v.46, n.1.

Chang, M.M., Sarathi, P., Heemstra, R. J., Cheng A.M., Pautz, J. F., (1992) "User's Guide and Documentation Manual For BOAST-VHS for the PC"**,** Topical Report NIPER-542

Cirne, W., Paranhos, D., Costa, L., Santos-Neto, E., Brasileiro, F., Sauvé, J., Oshtoff, C., Silva, F., Silveira, C. (2003) "Running Bag-of-Tasks Applications on Computational Grids: The MyGrid Approach". Proceedings of the 2003 International Conference on Parallel Processing.

Fayyad, U. M., Shapiro, G. P., Smyth, P. (1996) "From Data Mining to Knowledge Discovery : An Overview". In: Advances in Knowledge Discovery and Data Mining, Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Editors, MIT Press, pp. 1-37.

Foster, I., Kesselman, C. (1997) “Globus: A Metacomputing Infrastructure Toolkit”. International Journal of Supercomputing Applications, 11(2):115-128.

Giga Network web site (2005) “http://www.projetogiga.org.br/”

Java Portlet Specification (2005) “http://www.jcp.org/aboutJava/communityprocess/final/jsr168/”

Jetspeed web page (2005) “http://portals.apache.org/jetspeed-1/”

Kaufman, L., Rousseeuw, P. J., (1990) “Finding Groups in Data: an Introduction to Cluster Analysis”, Wiley Series in Probability and Mathematical Statistics.

Orlando, S., Palmerini, P., Perego, R., Silvestri, F., (2002) “Scheduling High Performance Data Mining Tasks on a Data Grid Environment”, Proceedings of Int. Conf. Euro-Par 2002, 27-30 August 2002, Paderborn, Germany, LNCS 2400 - Springer-Verlag - Pag. 375-384.

Silva, F.A.B., Carvalho, S., Senger, H., Hruschka, E.R., Farias, C.R.G., (2004) “Running Data Mining Applications on the Grid: A Bag-of-Tasks Approach”, Proceedings of the 2004 International Conference on Computational Science and Its Applications, LNCS 3044 - Springer-Verlag - Pag. 168-177.