Proceedings of the 5th IEEE International Workshop on

Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications

IDAQ

WEST — Intelligent Data Acquisition and Advanced Computing Systems — EAST

IDAACS'2009

ACS

IDAACS'2009

The crossing point of Intelligent Data Acquisition & Advanced Computing Systems and East & West Scientists

RENDE (Cosenza), ITALY

September 21-23, 2009

ORGANIZED BY

RESEARCH INSTITUTE OF INTELLIGENT COMPUTER SYSTEMS,
TERNOPIL NATIONAL ECONOMIC UNIVERSITY, UKRAINE

IN COOPERATION WITH

DEPARTMENT OF ELECTRONICS, INFORMATICS AND SYSTEMS,
UNIVERSITY OF CALABRIA, ITALY

SPONSORED BY

IEEE Ukraine Section I&M/CI Joint Societies Chapter

Ternopil National Economic University, Ukraine

Department of Electronics, Informatics and Systems, University of Calabria, Italy

IEEE Ukraine Section, IEEE Instrumentation & Measurement Society, IEEE Italy Section, IEEE Region 8

IEEE Catalog Number: CFP09803-CDR
ISBN: 978-1-4244-4882-1
Library of Congress: 2009905888

Technical support: IDAACS'09 OrgCom, Phone: +380352-43-6038 Fax:+380352-436354
e-mail: orgcom@idaacs.net

IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
21-23 September 2009, Rende (Cosenza), Italy

MolDynGrid Virtual Laboratory as a part of Ukrainian Academic Grid infrastructure

A.O. Salnikov [2], I.A. Sliusar [2], O.O. Sudakov [1,2], O.V. Savytskyi [1], A.I. Kornelyuk [1]
[1] Institute of Molecular Biology and Genetics, National Academy of Sciences of Ukraine, 150, Acad.Zabolotny Str., Kyiv, Ukraine, kornelyuk@imbg.org.ua, http://imbg.org.ua
[2] Information and Computer Center, National Taras Shevchenko University of Kyiv, Glushkova prosp., 2, Kyiv, Ukraine, cluster@cluster.kiev.ua, http://grid.org.ua

***Abstract* - MolDynGrid virtual laboratory has been established for collaborative research in computational biology and bioinformatics that requires high processing power and huge storage space. This research involves molecular dynamics (MD) simulations of biological macromolecules, such as proteins, nucleic acids and their complexes. Also, a comparison of MD simulation results with experimental NMR spectroscopy data is proposed.**

The main aim of MolDynGrid virtual laboratory is to provide an efficient infrastructure for automation of MD job processing in Grid. It will assist research workers to use Grid for their needs without having them to be familiar with highly complicated internals of the Grid.

***Keywords* - Grid, virtual laboratory, molecular dynamics, GROMACS, web-portal**

I. INTRODUCTION

EGEE (ENABLING GRIDS FOR E-SCIENCE) project has been carrying out in Europe since April, 2004. Now within the framework of EGEE the whole European Grid infrastructure is operated [1]. Biology and medicine are among the main directions, chosen in Europe for developing and implementing of Grid technologies. First of all, it concerns creating databases of hereditary diseases patients. The BioinfoGRID project combines bioinformatics services and applications for molecular biology users with the Grid infrastructure created by the EGEE [2]. Biomedical Grids are created for drawing up databases of various clinics with the purpose of creating a virtual hospital [3]. Grid-medicine is IT-infrastructure containing a specialized computer service, adapted for solving problems of processing biomedical data. Accordingly, Grid-medicine resources are computer resources, like specialized bases of medical data, medical devices and complexes [3].

The key concept of Grid technologies is a virtual organization – a group of people, working in various research organizations, but having a common aim and ability to share their resources. Some examples of the created virtual laboratories and projects in area of Grid - medicine are considered:

- medical diagnostic images processing;
- modeling a patient's body for choosing treatment tactics and surgical intervention;
- Grid technologies in pharmacy;
- Grid in genomic medicine;
- virtual biomedical universities and e-learning.

As a part of the program for informatization of the National Academy of Sciences, the Ukrainian Academic Grid (UAG) has been implemented and is functioning since 2005. In 2007, under the initiative of the Ministry of Education and Science of Ukraine, National Grid infrastructures for support of scientific researches and educations were created in Ukraine.

The development of the first Grid segment in Ukraine, and in perspective a national-wide Grid network, give an opportunity of successful integration into the international scientific projects which are carried out in research centers in Europe and other parts of the world [3].

In the post-genomic era, much attention is focused on the 3D structures and conformational flexibility of proteins as the novel targets for drug design. Standard molecular dynamics simulations of biomolecular systems at full atomic resolution are typically limited to the nanosecond timescale and therefore need more powerful hardware and software systems.

II. VIRTUAL LABORATORY CONCEPTS

MolDynGrid virtual laboratory was established in 2008 for interdisciplinary researches in computational structural biology and bioinformatics, which required a high processing power and immense storage space [4]. In particular, this involves molecular dynamics simulations of biological macromolecules such as proteins, nucleic acids and their complexes. It is also important to compare MD simulations results with experimental NMR spectroscopy data on proteins dynamics. MolDynGrid virtual laboratory was established as a part of Ukrainian Academic Grid infrastructure.

Laboratory's main aim is to provide an efficient utilization of Grid infrastructure for automation of MD job processing. It will assist biologists in using Grid for their needs without requiring them to be familiar with highly complicated IT internals.

Characteristics of resources needed for molecular dynamics calculations fit well with concepts of the Grid. Historically, Grid has grown on top of the Internet, which is now available world-wide. The Internet itself began as

information sharing tool for use by limited number of research organizations. Grid comes one step next in resource sharing – not only information can be shared, but also processing power, software, maybe some devices and, of course, data storage capacities. Another fundamental idea of the Grid is that resources are provided for groups of people with the same interests - Virtual Organizations (VO).

III. Distributed File Storage System

In the beginning of the Ukrainian Academic Grid infrastructure, only processing power and software sharing had been implemented. This was done by means of Nordugrid ARC Middleware. Implementing data storage space sharing was the next step that was required for successful deployment of MD calculations in the Grid. So, we have implemented the storage infrastructure between clusters that support the MolDynGrid VO [5].

Implementation was directed by users requirements. The users wanted to access storage resources from a single point and didn't care which particular storage element holds their data. An index is required to accomplish this task. When some data is going to be stored, a process of finding needed space should not require user intervention, still providing a single entry point to the user. File access policies should be the same no matter what particular storage element is used. User's personal certificate DN and VO membership are used to identify him on the Grid. So, access permissions must be set on the basis of this attributes. Of course, comfortable user interface should be available for operations with data independently of computational job management.

The first standard solution we had used is Replica Location Service provided by Globus Toolkit. It implements a data location index which acts as an entry point for automated storage operations. We had also chosen ARC's Smart Storage Element service instead of GridFTP as local storage agent. It supports setting access rights for each individual file and standard secure HTTP protocols. Comfortable user interface was implemented ourselves as a plug-in module for the Midnight Commander, a full-featured file manager for UNIX systems. Distributed storage operation scheme is shown at Figure 1.

Development of user interface which provides access to distributed storage began with implementation of "stels" – a shell wrapper for ARC file listing tool which transforms flat line file listing into a directory-tree like output. We simply parse all slashes and consider a part of name between slashes as a directory name, than we build file tree and show it the way like "ls" for local filesystem. But using command line still is not comfortable for users, especially for biology scientists that do not have experience working with Linux shell. So, we developed extended filesystem plugin for MC based on our "stels" wrapper.

There are several advantages of choosing MC extfs plugin as UI implementation:

- Easy to install – all that you need is to copy our script to plugins directory of MC and add plugin name to the list;
- After installation every user may use the plugin by creating an empty file with name in format <URL of RLS server> dot rls;
- Just pressing Enter on this file will provide an access to RLS, representing content as a directory tree ready to navigate in familiar way. You are free to use standard combinations like F5 to copy or F8 to remove files [6].

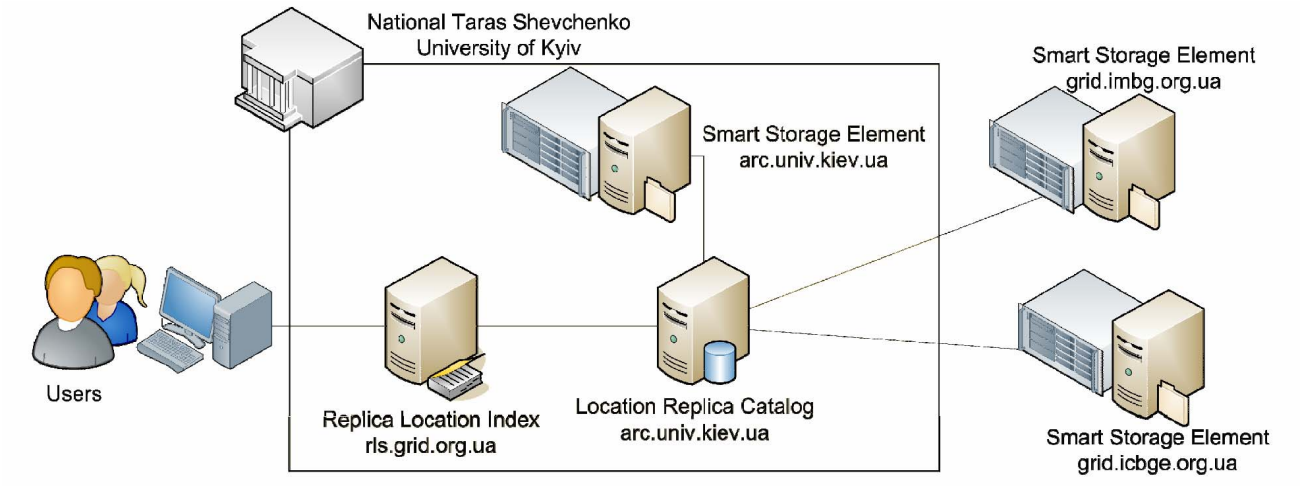

Fig. 1. Distributed storage system.

IV. MolDynSub_CLI

Several interfaces have been created on the way to the aim of automated MD computation job management, from the command line tools to full-functioning web-portal. Our researches are based on MD simulation GROMACS package [7]. Full cycle of computation consists of several steps: Ion generation, Energy minimization, Position Restraints and

Molecular Dynamics. Each step require user to run a chain of programs that GROMACS package provides. We want to utilize Grid computing power for these calculations, therefore Grid job description is also required. To accomplish this, an automation script "moldynsub_CLI" providing step by step GROMACS cycle execution in Grid had been written. All the parameters needed for different GROMACS applications and general Grid job description are gathered into one common file "params", which is used by the automation script. The script operates the following algorithm: it calls Grid tool for user authorization if proxy certificate was not found or expired, generates default parameter description file if it does not exist, submits full-cycle GROMACS MD computation to Grid if all required files exist in the current directory, or submits MD computation from checkpoint if checkpoint file found.

V. Virtual Laboratory Web-portal

Automation script and Midnight Commander plug-in were widely used in MolDynGrid virtual laboratory while we came one step next in providing MD automation interfaces. Creating of web-based interfaces that eliminate usage of command line at cluster and provide access to virtual laboratory resources was the aim of web-portal development. In order to start working with MolDynGrid, the scientist just needs to open web-browser with personal certificate installed and have his credentials already delegated to MyProxy server. Next step is the navigation to the URL: http://www.moldyngrid.org. The web-page of MolDynGrid portal is depicted on Figure 2.

The portal is functioning on web-server with Apache backend running on Fedora Linux. All the modules use PHP CGI, without any Java applets. This is much faster solution and does not require installing Java JRE and application server like Tomcat on the server like existing web-portal solution based on GridSphere. Grid middleware UI required on web server.

Portal divided into several blocks:

- Informational block – contains documentation about portal and several virtual lab tools usage instructions;
- Computation block – provides interface to MDP parameters customization and job submission;
- Trajectory database – provide interface to previously calculated trajectories located on distributed storage.

Informational block contents are available to all guests directly from the home page. Access to other blocks providing interfaces to laboratory resources, is restricted. To gain access you need to have a valid certificate signed by one of the Grid CA and to be recognized as a member of MolDynGrid VO [8]. Authorization procedure is controlled by "Authorization Gateway". Certificate validation implemented on Apache level, requesting client certificate from browser. Certificate DN used by CGI script to verify VO membership.

Computational block is composed of two parts: parameter specification interface and job submission script. All users have an ability to save different profiles of parameters for job submission, modify their values and also specify job and trajectory names, mailing options and other grid parameters. Submission script is based on moldynsub_CLI tool developed earlier. To submit a job, portal needs a user delegation, which is requested from MyProxy service.

Trajectory database was created using MySQL backend. Interface to the database is available from restricted portal area. This interface provides information about available trajectories and properties (parameters of simulation) for each trajectory. When a new job is submitted from the Computational block, a new record in the database is created.

Fig. 2. MolDynGrid virtual laboratory web-page.

All this components, interoperating together, works as an single entity – MolDynGrid virtual laboratory portal, that make MD simulations easy in the Grid.

VI. First Application Results

At this time, MD simulations were performed for human immunodeficiency virus protease (HIV-1 protease), mammalian and eubacterial tyrosyl-tRNA synthetases and their isolated functional N- and C-terminal domains, transfer RNAs and specific TyrRS–tRNA complexes [4,8].

Computed MD trajectories are located on distributed file storage system in Grid and are accessible for the future analysis. The resources of MolDynGrid are available free of charge for registered members upon observance of usage rules. Information about MolDynGrid is available from the web-site http://www.moldyngrid.org.

At the moment the Ukrainian Grid infrastructure (http://grid.org.ua/monitors/) used for calculations consists of 18 clusters with about 1500 CPUs available for Grid. MolDynGrid usually utilizes computing elements (CE) of 5 clusters and storage elements (SE) of 2 clusters that correspond to 900 CPUs and 3 TBytes of disk space, respectively.

VII. Conclusions

The MolDynGrid virtual laboratory was established and developed for molecular dynamics simulations in Grid and provides powerful facilities for solving of resource consuming problems of biomolecular modeling. The efficiency of MolDynGrid was proved by application to different problems. Virtual laboratory is used by biologists that are not familiar with low and middle level of Grid functionality.

Future development of MolDynGrid will include the implementation of Analytical block with interfaces to the most important algorithms and integration with gLite middleware [9].

References

[1] F. Gagliardi, B. Jones, F. Grey, M.E. Bégin, M. Heikkurinen, "Building an infrastructure for scientific Grid computing: status and goals of the EGEE project," *Philos Transact A Math Phys Eng Sci*. 2005 363(1833), pp. 1729-1742.

[2] L. Milanesi, I. Merelli, "High performance Grid based implementation for genomics and protein analysis," *Stud Health Technol Inform*. 2006;120, pp. 374-380.

[3] A.I. Kornelyuk, O.P. Mintser, "Modern computational Grid-technologies and their applications in medical researches," *Medical Informatics and Engineering*, Vol. 1, № 1, (2008), pp. 23-29.

[4] MolDynGrid Vitual Laboratory, https://www.moldyngrid.org.

[5] M. Zynovyev, S. Svistunov, O. Sudakov, Yu. Boyko, "Ukrainian Grid Infrastructure: Practical Experience," *Proceedings of 4-th IEEE Workshop IDAACS 2007*, September 6-8, 2007, Dortmund, Germany, pp. 165-169.

[6] A.O. Salnikov, I.A. Sliusar, O.O. Sudakov, Yu.V. Boyko, A.I. Kornelyuk, "Implementing the File Storage System in the Ukrainian Academic Grid Infrastructure," *Abstracts of 21th International CODATA Conference "Scientific Information for Society-from Today to the Future* (October 5-8, 2008, Kyiv), p. 31.

[7] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J. Berendsen, "GROMACS: fast, flexible and free," *J. Comput. Chem.*, 2005, 26, pp. 1701-1718.

[8] A.O. Salnikov, I.A. Sliusar, O.O. Sudakov, Yu.V. Boyko, "Ukrainian GRID-Segment authorization system," *Proceedings, III International Conference "Electronics and Applied Physics",* October, 25-27, 2007, Kyiv, Ukraine, p. 163.

[9] gLite 3.1 home page, 2008, http://www.glite.org.