

UNICORE - An Uniform Platform for Chemistry on the Grid

Jarosław Pytliński, Łukasz Skorwider,
ICM Warsaw University, Pawińskiego 5a, 02-106 Warsaw, Poland
Valentina Huber
Research Center Jülich, D-52425 Jülich, Germany
Michał Wroński and Piotr Bała
Faculty of Mathematics and Computer Science, N. Copernicus University
Chopina 12/18, 87-100 Toruń, Poland,

1 Introduction

Research in the area quantum chemistry and molecular biology requires computer resources usually not available at the user workstation even though current desktop computers provide significant CPU speed, and become important computational platforms. However, nontrivial problems still require large memory, disk space and CPU time, available only at remote systems or sites.

Users sharing external as well as departmental or even local computers have to solve number of practical problems. They are faced with different site policies and practices such as different security, different user identification, different priorities or scheduling mechanisms and so on. Distributed computational resources cannot be effectively utilized using typical unix tools based on the remote login. Additionally current users adopt latest developments in the graphical user interfaces and simple do not want to use traditional, command line based tools.

Recent advances in computer technology, especially grid tools make them good candidate for development of user interfaces to computing programs and resources [1]. Computational grids enable sharing a wide variety of geographically distributed resources and allow selection and aggregation of distributed resources across multiple organizations for solving large scale computational and data intensive problems in science. Computational chemistry traditionally provides large number of important, nontrivial and computationally intensive problems. Currently various *ab initio* calculations as well as various molecular modeling techniques become standard tools for a large number of users such as chemists or biologists. One should note that most of them have no longstanding experience in large scale computing and expect intuitive tools which are easy to learn and use.

In order to attract users, grid interfaces must be easy to install, simple to use and can be used in distributed environment. Available web technology which can provide powerful user interface, cannot achieve high numerical efficiency required for most simulation problems. This leads us to solutions consisting on web interface for the user and traditional, usually batch type, numerical simulation engine.

Easy access to remote high performance resources is being developed significantly in last few years. However most of the development goes to the design and production of general tools which can be used in different situations and provide simple access to remote resources. The existing solutions like globus, [2] Legion, [3] LSF [4] address most important issues. Unfortunately, these tools require an advanced computer skills for installation and usage. Moreover, in order to provide required functionality user application must be modified, sometimes significantly. This is not possible for commercial or legacy codes which cannot be easily adopted by the user. Most grid solutions have limited user interface not addressing application specific issues.

Most of the quantum chemical and molecular biology applications have no dedicated user interfaces, however some graphical interfaces exist. Some expensive commercial graphical tools are available as well as academic software but even sophisticated tools such as GaussView [5], Tripos [6] or Cerius [7] are able to run user applications only on local workstations and do not take advantage of grid environment. These and many other applications can be used as aid for input preparation, but in most cases job must be prepared and submitted manually by the user.

Another approach is presented by WebMo [8] which is web based submission system for quantum chemistry codes such as Gaussian [9], Gamess [10] and Mopac [11]. This tool is limited to the local batch systems and has no grid capabilities. The web submission to the geographically distributed systems is possible within BioCore [12] which is web interface to the molecular dynamics code NAMD [13]. Currently this system is limited to the particular MD code and single visualization package (VMD). NPACI Gamess Portal presents analogous approach for quantum mechanical code.

2 Grid tools

UNICORE is uniform interface to the computer resources which allows users to prepare, submit and control application specific jobs and file transfers [14]. Jobs to be run on the same system can be grouped in job-groups. The user specifies target system for job group as well as resource requirements for CPU time, number of CPUs amount of memory and disk space for job. Jobs can have complicated structure with various job subgroups which can be run at different systems. Subgroup jobs and file transfers can be ordered with user defined dependencies. The user input is mapped to the target system specific commands and options by the UNICORE infrastructure. Compared to other tools UNICORE has wider functionality, is more flexible and allows for much easier integration of the user interface with external quantum chemistry or biomolecular applications.

The UNICORE architecture is based, like other grid middleware, on the three tier model. It consists of user, server and target system tier. The user tier consists of the graphical user interface, the UNICORE client, written as Java application. It offers the functions to prepare and control jobs and to set up and maintain the user's security environment. The UNICORE job can be build from multiple parts which can be executed asynchronously or dependently on different systems at different UNICORE sites. Within the UNICORE environment user has a comfortable way to use distributed computing resources without having to learn about site or system specifics. Details of the UNICORE architecture can be found elsewhere [14].

3 Computational chemistry applications

The UNICORE software provides general framework for running user applications. On the other hand there exists wide range of applications which are commonly used in quantum chemistry and molecular biology. For the advanced user it is possible to develop scripts for running application within UNICORE, but this takes time and requires considerable experience. In the simplest approach one can prepare script job specifying program which should be run. This approach depends on the details of the program installation and execution on the target system and cannot be used without modifications if the target system changes.

In order to overcome this disadvantage the biomolecular software is introduced to the UNICORE through the Incarnation DataBase (IDB) entries which describe details of the local installation. IDB allows for definition of the environment variables for the script executed by the user. In this way we

can define variables required for particular application or define path to the program. These entries must be customized for each target system and can be used by any user.

The UNICORE execution model allows to include shell commands in the IDB entries. This allows to prepare any environment for execution shell. Knowledge of defined variables is required for preparing UNICORE job, but allows to run the same script without modifications on any target machine.

The UNICORE infrastructure allows also for more sophisticated registration of the applications available on the target machine through `SOFTWARE_RESOURCE` entries in the IDB. The program name and program version together with shell commands required to set up execution environment are stored in the dedicated section of the IDB file. The UNICORE client has build-in capabilities to check software resource entries on the particular target systems and user gets informed if the chosen program is not available. This mechanism however cannot be used with the simple user script jobs and is accessible only when user job is prepared with the dedicated plugin to the UNICORE client.

Taking advantage of the environment variables defined in the IDB at target systems we have developed UNICORE Abstract Job Objects (AJO) for most popular biomolecular applications including Gaussian98 [9], Gromos96 [15], Amber [16] and Charmm [17]. The AJOs can be easily modified by the user in the part including application input, exactly as it is done for batch jobs. However prepared AJOs can be run without any modifications using UNICORE middleware on any target system.

UNICORE script job still requires knowledge of the input files for the biomolecular applications. In most cases it is prepared with standard text editor and requires significant experience from the user. Any mistake in the file format results an error and extends time in which results will be obtained.

4 UNICORE plugins

We have used UNICORE as framework in development dedicated user interface to biomolecular applications. The plugin mechanism available in the UNICORE client become very attractive and efficient mechanism for integration of applications with grid middleware. Plugin is written in Java and is loaded to the UNICORE client during start, or on request. Once it is available in the client, in addition to the standard futures such as preparation of the script job, user gets access to the menu which allows preparation application specific jobs.

4.1 Gaussian Plugin for UNICORE Client

We have developed the UNICORE plugin for one of the most popular quantum chemical application Gaussian98. The user can set up type of *ab initio* calculations and add options specific to the particular type of simulations. Plugin allows only for options which do not conflict with the chosen simulation type. The theory level, basis set, total charge and multiplicity can be set. All options are presented with the defaults which correspond to the simple calculations.

In a separate window user adds atoms coordinates required for the calculations. Currently the most popular Cartesian coordinates (XYZ format) can be used together with Z-matrix format, specific to the *ab initio* calculations. The user can input coordinates in dedicated window or load from the file. The most popular formats are supported. Read coordinates can be easily modified by the user.

Once all parameters are set user can generate valid Gaussian98 input with pressing proper button in the plugin. The text input in the application specific format is presented in the client window. Because input file is generated automatically, the user gets advantage of the correct syntax and proper

combination of the options. Any further modifications to the input simply can be performed by the change of the chosen parameter at any time during input preparation.

An advanced user can edit generated input and introduce changes he wants based on his own experience with Gaussian98. At any time he can return to the automatically generated input by pressing generate input button. The job file can be saved and used in the future. Gaussian plugin is also able to import valid Gaussian98 input prepared by the user, either by hand using any text editor or by an external application.

Once input is ready user can specify target system and resources required for jobs using standard UNICORE client facilities. For example user can check job status, monitor execution and retrieve output to the local workstation. All these functions can be performed from the UNICORE client with a single login during client startup. User can monitor job status and retrieve output to any computer connected to the network with the UNICORE client installed, in particular other than one used for job submission.

4.2 CPMD Plugin

The CPMD plugin, developed at the Research Center Jülich, provides the preparation and setting up options for Car-Parrinello Molecular Dynamics calculations [18]. The CPMD code is a plane wave/pseudopotential implementation of Density Functional Theory, particularly designed for *ab initio* molecular dynamics [19]. The plugin allows to specify the full set of configuration parameters for CPMD simulation, like creation of the input file, specification of the CPMD software, selection of the pseudopotential library, needed data sets and the disposition of the result files, etc.

The preparation of the CPMD input file, which has a complex format, requires knowledge of many keywords to describe atom coordinates, molecular symmetry, geometry optimization, electronic properties, etc. The graphical interface *CPMD Wizard* assist users and generates these parameters automatically to the text area, where it can be modified by experienced users by hand, e.g. they can add new keywords. It checks the input data and their interdependencies and prevents users from entering incorrect values while requesting to set mandatory parameters. Once generated input can be stored in the XML format to be reused and modified through the wizard in the future. It is also possible to load existing local or remote plain files to the text area and to modify them here [20].

In addition, the plugin checks the availability of the software on sites and allows job submission only to sites, where the CPMD program is installed. Furthermore the site administrators can specify in the Incarnation DataBase several CPMD versions with appropriated environments, e.g. different paths to binaries, and the plugin displays available versions in a choice box to be selected by a user.

The CPMD code needs also the specification of the directory containing pseudopotential files for calculation. Some but not all sites provide default pseudopotential libraries. The plugin checks the availability of the library on the target system and provides it as a default. Users can specify also their own library. The files from the local directory will be automatically transferred to the remote system.

4.3 Amber plugin

The experience with the development of plugin for quantum chemical application was used for development of the user interface to the Amber 6.0, one of the most popular molecular dynamics code. Developed user interface allows for preparing input for MD simulations in various modes: constant energy, constant temperature or constant pressure. In each mode user is able to specify most important parameters of the simulations such as time step, total simulation time, initial time and others.

Because MD job requires input files with initial coordinates, the topology, parameters and possibly others. User has to specify names of proper files using dedicated window. These files are included in job build by plugin and will be automatically transferred to the target system. In the same way user defines names of the files which will be created and should be transferred back to the user workstation. Target system together with required resources are specified by the user using UNICORE client which is also used for job submission and control.

4.4 PDB plugin

The PDB plugin allows for access to the remote databases from the UNICORE client. It allows user to prepare query to the PDB database [21], widely used crystallographic database. The query is submitted to the PDB web interface and result is returned to the UNICORE client. The search results can be browsed, visualized or saved for later processing. The PDB plugin introduces remote database interface to the UNICORE Client and provides user with uniform access to the various, not only computational, resources.

5 BioGRID

The UNICORE software was used to establish European computational grid - EUROGRID [22]. BioGRID is application oriented grid which adopts EUROGRID infrastructure to the specific area, namely molecular biology and quantum chemistry [23].

An integrated part of BioGRID is information deployment. User, especially one without large experience, requires information on the application he can use to solve particular problem, as well as information about available software and hardware resources. Some of this information can be obtained directly from the UNICORE client, however tools addressing resource information are still under development. We have decided to supply user with required information using web technology. The BioGRID Web site [23] has been started and contains all information on resources available together with description of the programs and links to the most important related material. This web site gives opportunity to download UNICORE client and provides user with detailed description of the installation and configuration process. User can also download plugins and can find extended support information.

Acknowledgements This work is supported by European Commission under IST grant 20247. The software was used using EUROGRID facilities at ICM Warsaw University (Poland), Forschungszentrum Jülich (Germany), University of Manchester - CSAR (UK), IDRIS (France) and University of Bergen - Parallab (Norway).

References

1. C. Kesselman I. Foster, editor. *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan Kaufman Publishers, USA, 1999.
2. I. Foster and C. Kesselman. Globus: A metacomputing infrastructure toolkit. *Int. J. Scientific Applications*, 11(2):115–128, 1997.
3. Legion. University of Virginia. Charlottesville. VA USA [Http://legion.virginia.edu](http://legion.virginia.edu).
4. LSF. Platform Computing. <http://www.platform.com>.
5. GaussView. Gaussian Inc., Pittsburgh PA. USA. 2000.

6. Tripos. Tripos Inc. St. Louis. USA.
7. Cerius². Accelrys, San Diego. USA.
8. Webmo. <http://www.webmo.net>.
9. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, P. Salvador, J. J. Dannenberg, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople. Gaussian 98. 2001.
10. M. W. Schmidt, K. K. Baldridge, J. A. Boatz, M. S. Gordon, S. T. Elbert, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery. General atomic and molecular electronic structure system. *J. Comput. Chem.*, 14:1347–1363, 1993.
11. J. J. P. Stewart, L. P. Davis, and L. W. Burggraf. Semi-empirical calculations of molecular trajectories: method and applications to some simple molecular systems. *J. Comp. Chem.*, 8(8):117–23, 1987.
12. M. Bhandarkar, G. Budescu, W. F. Humphrey, J. A. Izaguirre, S. Izrailev, L. V. Kal, D. Kosztin, F. Molnar, J. C. Phillips, and K. Schulten. Biocore: A collaboratory for structural biology. In A. G. Bruzzone, A. Uchrmacher, and E. H. Page, editors, *Proceedings of the SCS International Conference on Web-Based Modeling and Simulation*. San Francisco, California, 1999.
13. L. Kal, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. Namd2: Greater scalability for parallel molecular dynamics. *J. Comp. Phys.*, 151:283–312, 1999.
14. UNICORE. Pallas. Germany. <http://www.unicore.org>.
15. W. Van Gunsteren and H. J. C. Berendsen. *GROMOS (Groningen Molecular Simulation Computer Program Package)*. Biomos, Laboratory of Physical Chemistry, ETH Zentrum, Zurich, 1996.
16. P. Kollman. *AMBER (Assisted Model Building with Energy Refinement)*. University of California, San Francisco, USA, 2001.
17. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.
18. CPMD Plugin. <http://www.fz-juelich.de/unicore-test/cpmd>.
19. CPMD consortium. <http://www.cpmid.org>.
20. V. Huber. Supporting car-parrinello molecular dynamics with unicore. In *Computational Science ICCS - San Francisco, USA, Proceedings, Pt. 1*, volume 2073, pages 560–567. Springer Verlag, 2001.
21. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235-242, 2000.
22. EUROGRID. <http://www.eurogrid.org>.
23. BioGRID. <http://biogrid.icm.edu.pl>.