My WorkSphere: Integrated and Transparent Access to Gfarm Computational Data Grid through GridSphere Portal with Metascheduler CSF4

Zhaohui Ding[1], *Yuan Luo*[1], *Xiaohui Wei*[1]

zhaohui.ding@email.jlu.edu.cn; pp.jordan@email.jlu.edu.cn; weixh@jlu.edu.cn

College of Computer Science & Technology, Jilin University,

Changchun, Jilin, China 130012[1]

Chris Misleh[2], *Wilfred W. Li*[2], *Peter W. Arzberger*[2]

wilfred@sdsc.edu, parzberg@ucsd.edu

National Biomedical Computation Resource, University of California, San Diego,

La Jolla, CA, USA, 92093[2]

Osamu Tatabe[3]

tatebe@cs.tsukuba.ac.jp

Department of Computer Science, University of Tsukuba, Tsukuba, Japan[3]

Abstract: In order to deliver cyberinfrastructure to the general scientific and biomedical research community, transparent access and ease of use is of critical importance. Applications in systematic modeling of biological processes across scales of time and length demand more and more sophisticated algorithms and larger and longer simulations. The increased level of sophistication requires that cyberinfrastructure developers either work closely with the applications scientists, or develop middleware that flattens the learning curve for these scientists to use the grid willingly and transparently. The need to develop generalized and reusable components for cyberinfrastructure developers and the desire for customized solutions by application scientists create a dilemma that may take time to resolve. One reasonable approach is to let each do what they know best, and bridge the gap through innovative research and technology development.

In this approach, we adopt technology that enables applications to execute in the grid environment without modifications, and without knowledge of specific computational resources being utilized. This separation, of scientific application development and the subsequent use of the grid, means that the cost of entry to the grid is minimal. We have previously demonstrated that the Gfarm file system may be used as a computational data grid, in which preexisting applications may be moved to the where the data is located to take advantage of local file system input and output [1]. To fully support Grid Security Infrastructure (GSI) authentication and utilize the local batch schedulers on different clusters that are part of Gfarm file system, we have demonstrated the use of Community Scheduler Framework (CSF4) with LSF (Load Sharing Facility), and SGE (Sun Grid Engine) as batch

schedulers [2]. Here we report the latest advances in the use of Gfarm as a computational data grid, with CSF4 as the metascheduler, through a GridSphere portal based environment, termed My WorkSphere. All the components utilized are open source and available freely.

The GridSphere portal framework [3] enables the reuse of functionalities exposed as portlets, java technology based web components, but portable for different application scenarios. In this work environment termed My WorkSphere, a user can apply for access to compute resources through GAMA (Grid Account Management Architecture) portlet, which automates the user X509 certificate creation, signing, and proxy management [4]. A user who has never used the grid can start using the grid through the portal environment. The same X509 certificate may also be used directly, if desired, since it's a properly signed certificate by the organization running the GAMA server.

We have developed a prototype CSF4 portlet that allows users to submit jobs to satisfied resources through a web browser. A CSF4 server can forward the job request, along with the user credentials, to different heterogeneous clusters using the GRAM (Grid Resource Allocation & Management) Protocol. CSF4 uses the Java Commodity Grid Kit and GT4 (Globus Toolkit 4) delegation service to support full delegation of user proxies to both Pre-WS GRAM and WS-GRAM. The jobs are submitted to different clusters with local batch schedulers, such as SGE, dynamically based on metadata provided by MDS (Monitoring and Discovery System) or through FCFS (first come first serve)/round robin scheduling algorithm. CSF4 is capable of working with different local schedulers, like LSF, PBS (Portable Batch System), SGE and Condor, and via both GT4 WS GRAM and GT2 GRAM (Pre-WS GRAM).

For simple access to data and application deployment, Gfarm-Fuse (File System in User Space) is used to enable a familiar Unix environment in which compilation, installation of software is only required once per platform, and the filesystem is transparently mounted and unmounted without user intervention. This also allows clusters without public IP addresses for all the compute nodes to mount Gfarm via NFS (Network File System), for a performance penalty.

We have compiled and deployed into Gfarm commonly used bioinformatics software such as MEME, and BLAST, and other computational mathematics, biology and chemistry applications, such as FFTW, APBS, and AutoDock. We are currently conducting scalability and stability experiments using these applications, and also exploring the integration of CSF4 into other web service toolkits such as Opal [5,6] to enable different clients to take advantage such a transparent computational data grid infrastructure.

References

1. Li WW, Arzberger PW, Yeo CL, Ang L, Tatebe O, et al. Proteome Analysis using iGAP in Gfarm. *; 2005; Grid Asia 2005, Singapore. World Scientific Press.

2. Wei X, Li WW, Tatebe O, Xu G, Hu L, et al. (2005) Integrating Local Job Scheduler - LSF™ with Gfarm™. Lecture Notes In Computer Science 3758: 197.

3. Gridsphere (2004) GridSphere. http://www.gridsphere.org.

4. Bhatia K, Chandra S, Mueller K. GAMA: Grid Account Management Architecture. *; 2006; Melbourne, Australia. pp. In Press.

5. Krishnan S, Stearn B, Bhatia K, Greenberg JP, Li WW, et al. (2005) Web services-based data integration for life science computations. SDSC Technical Report. TR-2005-2002 p.

6. Li WW, Krishnan S, Mueller K, Ichikawa K, Date S, et al. Building cyberinfrastructure for bioinformatics using service oriented architecture; 2006; Singapore. pp. 39-46.