

Architecture of a Grid-Enabled Research Platform with Location-Transparency for Bioinformatics

Yoshiyuki Kido¹

kido@ais.cmc.osaka-u.ac.jp

Shoji Hatano³

qin@im.ac.cn

Susumu Date²

date@ais.cmc.osaka-u.ac.jp

Juncai Ma³

ma@im.ac.cn

Shingo Takeda²

stakeda@ais.cmc.osaka-u.ac.jp

Shinji Shimojo⁴

shimojo@cmc.osaka-u.ac.jp

Hideo Matsuda²

matsuda@ist.osaka-u.ac.jp

¹ R&D Department, Bio Science Division, Mitsui Knowledge Industry Co., Ltd., Harmony Tower 21st Floor, 1-32-2 Honcho, Nakano-ku, Tokyo 164-8721, Japan

² Graduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka, Suita-shi, Osaka 565-0871, Japan

³ Information Network Center, Institute of Microbiology, Chinese Academy of Sciences Bei-1-tiao 13-hao, Zhongguancun, Haidian, Beijing 100080, China

⁴ Cybermedia Center, Osaka University, 5-1 Mihogaoka, Ibaraki-shi, Osaka 567-0047, Japan

Abstract

The recent advance in information technologies has brought about the borderlessness in every field of both science and business. The borderlessness has increasingly made activities in interdisciplinary field more important. This current situation produces a strong demand that people want to establish a virtual group, organization and society for their business and scientific purposes irrespective of the actual structure formed by organizations. Remarkably, bio sciences require a research platform that satisfies such demand for further development. In this paper, we present a research platform for bioinformatics in detail. The prominent feature of the research platform is the use of Grid and its location transparency, which means that bio scientists and researchers are able to utilize a large amount of computational power for their analysis and to access data of their interest without being aware of where data and computational resources are located. The usefulness and feasibility of the architecture of the research platform is shown as well as future issues to achieve toward the final goal of our research in this paper.

Keywords: grid, grid portal, lifescience, bioinformatics, location-transparency

1 Introduction

Recently, both scientific and business world have been transformed to a borderless world. The development of information technologies in this decade has provided people with a method of integrating computational resources distributed on remote sites into a single system. It causes the situation that data, information and even knowledge exchanged over physical organizational border have increased in amount. Sometimes, more data is exchanged in a cross-organizational way than within an organization. Furthermore, the surplus of computational power can be shared by multiple alliance sites. This situation produces a strong demand that people want to establish a virtual group, organization and community for their business and research purposes, irrespective of the actual structure formed by organizations.

In particular, bio sciences strongly require a research platform that satisfies the demand above for further advancement. In-silico drug design research explains this movement well. In-silico drug design is one of the most attractive and fascinating research areas in bio sciences because it is expected to dramatically slash the cost required for creating a drug. It, however, requires the flexible and seamless integration of bio-related databases that hold protein, chemical, genome sequence, drug data report and so forth and high-performance computers for the processing of large volume of data. In other words, a research platform that allows scientists and researchers to freely and securely access a variety of computational resources on a virtual organization without having being aware of the location of it is essential for feature drug design. Importantly, these resources are, in general, distributed and administrated by multiple organizations and therefore are so difficult to integrate because of the difference of security policy and the heterogeneity of computational resources.

Emerging Grid is expected to satisfy the demand described above in the era of “borderless”. It offers a set of technologies for the sharing and coordinated use of diverse resources in dynamic, distributed “virtual organizations” (VOs) [5]. Examples of such technologies include security, communication, and resource monitoring. With these technologies, scientists and researchers are able to realize a computational environment spanned over multiple organizations such as universities, scientific institutions and enterprise organizations. In reality, many scientific projects in the world has adopted the Grid to realize a system on a virtual organization composed of multiple organizations. NEESgrid [10] is just an example of such scientific projects.

Today, the Grid and related technologies have been standardized to a suite of technologies inspired by the fusion of web services and traditional HPC-based Grid technologies through GGF (Global Grid Forum), where standardization issues are discussed by researchers, engineers and scientists (<http://www.gridforum.org/>). As a result of the intensive activities in these days, the Grid-related technologies are becoming mature, meaning that the stability and reliability of the technologies are high. However, the current Grid technologies have a serious problem in user convenience. In other words, the Grid technologies require the detailed knowledge and experience on it for practical and effective use in spite of the mature of the Grid technologies. Moreover, scientists and researchers who have little knowledge on the Grid have much difficulty in benefiting from the Grid.

In this paper, we show a Grid-enabled research platform with location transparency that allows scientists and researchers to securely access data of their interest and benefit from high-performance computational resources without being aware of the location of data and computational resources. More specifically, a research platform on which scientists and researchers can perform Gridified BLAST [2], and ClustalW [12] deployed on Osaka University for molecular sequences stored in bio-resource databases that will be provided by CAS (Chinese Academy of Sciences), on CAS without being aware of the location of data and program code. Importantly, the work summarized in this paper is positioned preliminary work of our research. The final goal of our research is to realize a research platform through which scientists establish a virtual group, organization and community for their research purposes, irrespective of the actual structure formed by organizations.

This paper is organized as follows. Section 2 presents two key technologies for building a Grid-enabled research platform with location transparency for bioinformatics, that is, GUIDE, which is a Grid Portal, and GSI-SFS, which is a secure file system on the Grid. After that, Section 3 details a Grid-enabled research platform connecting CAS and Osaka University. In Section 4, future issues and directions toward the final goal in this research is discussed. Finally, Section 5 concludes this paper.

2 Background and Key Technologies

In this section, we first describe the background and motivation of the collaboration performed by CAS and Osaka University. After that, two key technologies that take a role of much importance as building blocks of the research platform built over the network between CAS and Osaka University are detailed. First, a Grid Portal software named GUIDE is presented and subsequently GSI-SFS is

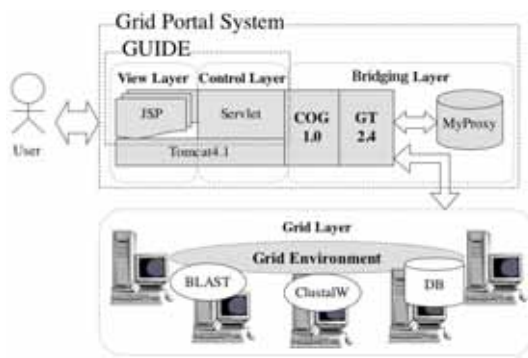


Figure 1: GUIDE architecture.



Figure 2: Snapshots of GUIDE.

briefly explained. These software have been developed in the BioGrid Project promoted by Osaka University and are available from the project web site; (<http://www.biogrid.jp/>).

2.1 Background and Motivation

China is becoming one of the most attractive countries for bio scientists and researchers. The reason can be explained from the fact that China has so many kinds of species. In practice, China has 1,244 fish species and 3,862 bird species [14], and these numbers are higher in comparison with other countries. With increased expectation and hope, the SDB (Scientific Database) Project has started to establish databases for inventory of biological species and specimens by CAS. This project involves 32 institutions, 12 of which are biological research institutes. In this project, Grid technologies have been adopted as building blocks for integrating and federating databases distributed in China. Today, a variety of preliminary research has been planned and implemented through the collaboration with domestic and international research partners.

The work reported in this paper is one of such collaborative researches by two bio-related Grid projects promoted by the Information Network Center, CAS and the Cybermedia Center, Osaka University. The goal of this preliminary collaboration toward the SDB project is to testify the usefulness and validity of the Grid technologies for building a research platform for bioinformatics. More specifically, this collaborative research has been performed to investigate whether it is possible or not to build a Grid-enabled research platform that allows scientists and researchers to easily utilize the computational power and access data of interest without having be aware of data and program code location.

2.2 GUIDE (Grid User Interface to Distributed Environment)

In spite of the recent mature of the Grid technologies, the Grid and related technologies are, in reality, inconvenient and therefore difficult for scientists and researchers to use without detailed knowledge and experience on them. This situation prevents scientists from taking advantage of the benefits brought by the Grid. It may cause the loss of the opportunities of dramatic advance in science and technology. Notably, this kind of situation is remarkable in bio scientific situation. It means that bio scientists have much difficulty in reducing computational time for bio scientific analysis application with the Grid technologies although the possibilities and benefits brought by the Grid fascinate bio scientists and researchers. For the point of view described above, we have developed a Grid portal system, that is, GUIDE for bio sciences. GUIDE provides an intuitive set of web interfaces to Gridified BLAST and ClustalW. These two programs have been adopted because of the prevalence and essentiality in bio

scientific research. BLAST requires a large number of computational power dynamically aggregated by the Grid. ClustalW aligns multiple nucleic or protein sequences simultaneously in order to elucidate their relatedness as well as their evolutionary origins. These two tools are frequently utilized by scientists so that ClustalW is performed by being fed the search result of BLAST. Figure 1 illustrates the overview of GUIDE architecture. GUIDE system consists of three layers; View Layer, Control Layer, and Bridging Layer.

View Layer provides an intuitive web interfaces for computationally-accelerated BLAST and ClustalW. Technically, this layer is built with JSP (Java Server Pages) on top of Tomcat, a servlet engine. This layer helps system user concentrate on their analytic work without being aware of the complexity in the Grid. Snapshots of GUIDE are shown in Fig. 2. Control Layer has a central role of importance in GUIDE. This layer is composed of multiple servlet programs that interconnect View Layer and Bridging Layer. The main task of this layer is to process data from and to user and then exchange the data after the translation so that it is interoperable and accepted by the receiver. In Bridging Layer, offered is the functionality of bridging the border between traditional web technologies and the Grid. To the end, CoG (Commodity of the Grid) 1.0 [13] has been adopted. CoG is a suite of Java APIs for Globus toolkit, which is a de facto standard implementation of the Grid concept (<http://www.globus.org/>). Servlet programs in Control Layer interact with Grid Bridging Layer via CoG APIs to utilize the Grid technologies.

Take note that MyProxy [9] has been adopted to take advantage of GSI (Grid Security Infrastructure) [3] based on PKIX (Public Key Infrastructure X.509), which delivers a single sign-on functionality detailed in Section 3.1. MyProxy holds X.509 certificates which system users register in advance. With MyProxy, system users are able to access user certificates necessary for GSI authentication from anywhere on the Internet. More detailed security mechanism in GUIDE is clarified in Section 3. Through the orchestra of web technologies and Grid technologies, GUIDE produces the synergy of allowing scientists to intuitively benefit from the complex Grid technologies.

2.3 GSI-SFS (Grid Security Infrastructure - Self-Certifying File System)

Scientists and researchers demand on a secure and easy-to-use data sharing and exchanging mechanism for their collaboration on the Internet. Advanced bio sciences require a diversity of bio-related data to be effectively treated. Seamless access to data of scientists' interest, which are geographically distributed on the Internet, takes a role of importance for today's bio scientific research. This fact simultaneously means that the success of advanced bio sciences after the paradigm shift highly depends on the establishment of efficient and effective research infrastructure, allowing bio researchers and scientists to seamlessly access data of their interest in an on-demand manner. In addition to that, robust data security is essential for bio sciences because bio-related data is inherently confidential and privacy-rich. Although the current Grid technologies offer to us a few promising capabilities pertaining to data operation, such as third-party data transfer [1], their capabilities are so difficult to use for scientists without detailed knowledge on the Grid. This lack of convenient data access methods prevent bio researchers and scientists who really want to benefit from the Grid from utilizing the Grid. For the reason described, a secure file system has been developed. This file system has been designed so that it achieves the following five requirements for a data access method on the Grid.

1. Single disk image: Scientists and researchers want to access data of interest as if the data were located on a local disk. Even the use of URI (Universal Resource Identifier) is not preferable for scientists.
2. Data confidentiality: In order to protect bio data to be inherently protected, strong data security is essential. Data in transit on the network should be protected by strong cryptography.
3. Exclusiveness: Access information to a certain data, rather than the data itself, must be protected.

4. On-demand remote data access: Scientists want to access to data of interest. The reason can be explained partly from the point of risk management view.
5. User convenience: Scientist never wants to type his or her password whenever he/she accesses data located in the remote computer.

To satisfy these five requirements, the secure file system named GSI-SFS has been built with two existent promising technologies; GSI and SFS (Self-certifying File System) [8]. The former is a service shipped in Globus toolkit and allows users to freely utilize computational resources under their delegated authorization after single sign-on to the Grid. The latter is a secure file system being developed at MIT and allows each user to hold his/her own single disk image on a wide-area network. Furthermore, the data in transit are protected with strong cryptography in SFS.

Taking the use of SFS on the Grid into consideration, SFS is considered to be useful and easy-to-use file system since it has achieved four requirements listed except for user convenience requirement. However, a serious key management issue arises in case that SFS is used for data sharing on the Grid. This fact means that as the number of servers on the Grid increases, key management increasingly becomes cumbersome and error-prone. SFS, therefore, does not suite well on the Grid where hundreds of computers are expected to be connected.

GSI has the ability to solve this key management problem to allow scientists to access data of interest seamlessly after typing their password just one time. GSI is a core security service provided by Globus toolkit, and offers a single sign-on functionality that takes advantage of PKIX and public key cryptosystem. This prominent functionality is realized based on the delegation capability in GSI, which will be explained in Section 3.1. GSI takes full advantage of the delegation capabilities based on public key cryptosystem to produce the synergy of a single sign-on functionality. With this functionality, the user is able to conveniently access a diversity of heterogeneous computers without typing his password many times.

Figure 3 diagrams the architecture of GSI-SFS, which is a secure file system built based on GSI. GSI-SFS is composed of two modules; SFS module and GSI authentication module. SFS module delivers the original promising functionalities of SFS. These functionalities are realized in pure SFS part that mediates the communication between NFS server and client. SFS client has a functionality called certprog with which system administrator can register a program used for user authentication in advance. In GSI-SFS, GSI authentication module developed by the BioGrid project has been configured to use this certprog functionality. GSI authentication module performs authentication between a GSI-SFS user and a GSI-SFS server. After that, it passes the information necessary for the provision of original SFS functionality delivered via GSI-SFS server from SFS server. With this GSI-SFS, a user accesses data of their interest on multiple computers without typing a pair of user ID and password corresponding to the computer to access. In addition to that, as the authentication is performed between each user and SFS server, the access information is managed for each user. It means that GSI-SFS has achieved the listed five requirements by solving the complexity problem in key management in SFS without losing the salient features in original SFS.

3 A Grid-Enabled Research Platform over the Network Connecting China and Japan

3.1 Architecture

The research platform we have developed is composed of GUIDE module and GSI-SFS modules. Two security systems of these modules have been succeeded to be integrated into the platform. Before going the architecture of the research platform into detail, we will illustrate the mechanism of single sign-on functionality that has a core role for the integration of GSI-SFS and GUIDE modules. After that, we will explain how the research platform produces a synergy of location transparency as well as

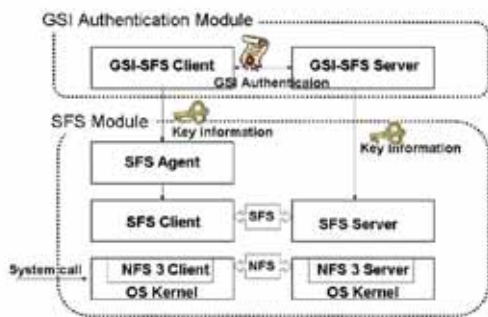


Figure 3: GSI-SFS architecture.

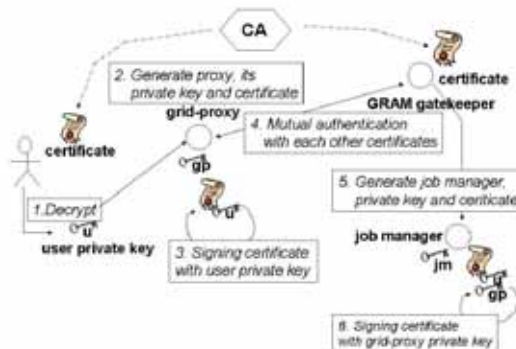


Figure 4: Delegation capability.

how the integrated security system functions. Figure 4 diagrams a delegation capability in GSI. This figure explains how a user is authenticated by a remote process when the user attempts to submit a job to a remote computer through GRAM [4] capability. In GSI each of computer and user has a X.509 certificate that verifies that the owner of the public key is the user. The certificate is also confirmed to be issued by Certificate Authority (CA), a trusted third-party entity. Technically, these certificates are digitally-signed with CA's private key, meaning that the validity of the certificates is confirmed only with CA's public key.

First, a user decrypts her encrypted private key stored in a local computer with her pass phrase and then generates an authentication proxy called grid-proxy and a proxy certificate signed with the user's private key. This grid-proxy has the responsibility of authenticating the user instead of the user. In case that the user submits a job to a remote computer, mutual authentication between GRAM gatekeeper and the grid-proxy takes place. GRAM gatekeeper identifies the grid-proxy by verifying the certificate of grid-proxy with the user's public key and vice versa. Subsequently after the mutual authentication, a job manager which manages a process for completing user submitted job and the certificate are generated in the same way as the interaction between the user and the gatekeeper. In a way described above, user authentication is delegated and propagated from entity to entity. Figure 5 illustrates how the two systems are integrated and how they work in the research platform. Importantly, the user can perform their analysis work such as BLAST and ClustalW in the exactly same fashion as the use of traditional web technologies, since the action described below takes place internally in the platform.

In signing in the research platform, (1) a scientist needs to submit his pair of user ID and password via web interface to GUIDE. This information is secured with SSL (Secure Socket Layer). (2) After receiving them, GUIDE retrieves the certificate of the scientist from MyProxy online repository and then (3) generates grid-proxy for the scientist. (4) When the scientist submits a BLAST service request to the platform, (5) the job manager produces a set of processes on a cluster system after mutual authentication illustrated in Fig. 5. Importantly, a grid proxy for process A on computer A is generated. (6) If process A accesses the database on the remote computer B, (7) then GSI-SFS authentication client automatically perform mutual authentication with gsi-sfs authentication server using the grid proxy for the job manager. After that, (8) database on the remote computer B is automatically mounted to the file system on computer A. (9) Process A automatically accesses data of interest as if the database on computer B were on a local disk. Most importantly, with this internal mechanism, the scientist can utilize computational power aggregated by GRAM and access data of interest without location of them.

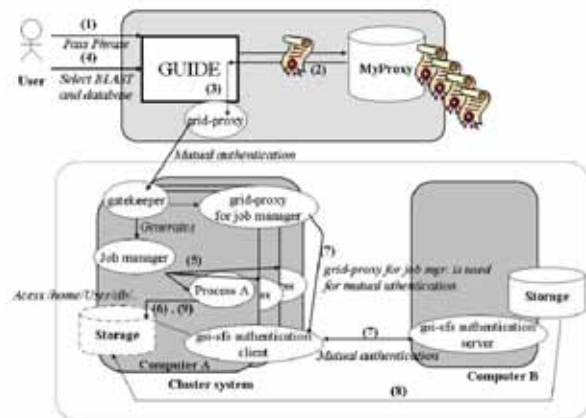


Figure 5: Architecture of grid-enabled research platform.

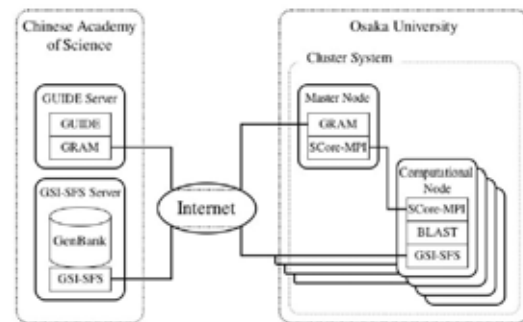


Figure 6: Software composition.

3.2 Implementation

In Osaka University, a cluster system composed of 79 nodes was used in the research platform. Except for one node usually as for master server, each of 78 nodes has two Pentium III processors running at 1.4 GHz. Ordinarily, the master server is used as a login server to this cluster system and has the responsibility of load management in the cluster system through SCore [6], a scheduling middleware. In CAS, two computers were used. Each of these computers has two Pentium 4 processors running at 2.0 GHz. A computer has hold GenBank public database, which is provided at NCBI web site; (<http://www.ncbi.nlm.nih.gov/GenBank/>).

In the research platform, the computational power by the cluster system in Osaka University and CAS's GenBank database, which plans to be replaced by CAS's proprietary database, were integrated with Grid technologies. Figure 6 overviews the composition of software in the research platform. We installed Globus toolkit 2.4 and GSI-SFS client into all computational nodes of the cluster system in Osaka University and the two computers in CAS so that the functionality of GSI-SFS server and GRAM are available among these computers. SCore, which is a scheduling middleware, and SCore-MPI were adopted to take advantage of computational power of the cluster system to decrease the computational time of BLAST. GUIDE system was deployed in a computer in CAS. On this research platform bio scientists and researchers can utilize computational power brought by the cluster system in Osaka University using data in CAS without being aware of data and code location.

4 Evaluation and Discussion

We described the architecture of the research platform which we developed in the preliminary work toward the final goal of our research described in Section 1. Technically, this architecture of this research platform is feasible to allow scientists and researchers to make use of computational resources and access valuable data in a location-transparent manner. In practical, however, there are three major future issues to be achieved.

First, performance issue needs be discussed. The research platform built over the network between CAS and Osaka University has serious problems in network connectivity. At the time of building this research platform, the network connecting CAS and Osaka University was so narrow. The effective bandwidth of the network was approximately 1 Mbps. In addition to that, the latency of the network

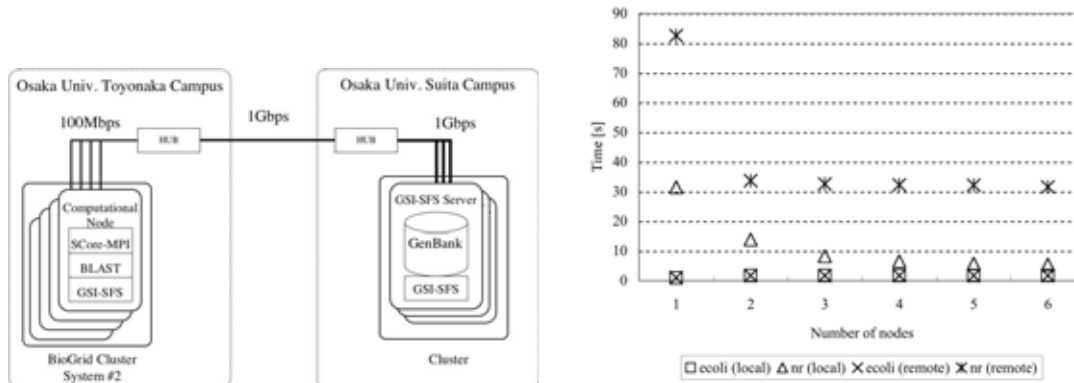


Figure 7: Performance test environment. Figure 8: The results of performance test.

was considerably large as the traffic passed through more than 20 routers. The actual RTT (Round Trip Time) was more than approximately 400 msec on average.

These network characteristics resulted in the long computation time of BLAST. In the research platform, BLAST loads approximately 500MB data from CAS’s database and then perform computation. This process took approximately one hour in comparison with approximately one minute on a node in the cluster system located in Osaka University. This is a so serious problem from the aspect of practical use, although it solves location transparency to achieve user-convenience. This problem pertaining to computation time is considered to cause from the throughput bottleneck arisen from high latency characteristics, since the performance of SFS is limited by that of TCP adopted in NFS, on which SFS is built. In order to confirm that the performance improves under a low-latency network, we performed a performance test in a campus network. Figure 7 and 8 shows the performance test environment and the result of the performance test, respectively. The performance test was performed on the research environment built on LAN in Osaka University, which was composed of two cluster systems connected on a fast and low-latency network. The effective RTT was 0.023 msec on average and the effective bandwidth was 93.56 Mbps. The research environment built for the test had the same architecture as the one built over the network between CAS and Osaka University.

In the performance test, the computational time required for BLAST was measured. First, we measured the time when database is located on the same node as the node performing computation by changing the number of nodes used for BLAST computation. After that, the computational time when database is located on other node than the one performing computation was measured. The graph indicates that the computational time was reduced to three times longer computation time in comparison with the one with a single node which locally holds database as well as that the computation time gradually decreased with the increase of computational nodes. The term “nr” and “ecoli” are the name of database contained in GenBank. This result illustrates that we can achieve more practical and effective computational time even on the long-distance network when we can take advantage of such network in near future. Furthermore, we consider that performance tuning of NFS will improve the situation. Also, we expect that the advancement of network technologies in these days alleviates this problem to some degree.

Next, security issues must be also discussed. Data security is an important issue inevitable for bio sciences. As described in Section 2.3, bio sciences are expected to treat more important and privacy-rich data and document. This situation is illustrated from the recent keyword in bio sciences such as personalized medicine and tele-medicine. Such advanced bio sciences in near future requires robust protection of data. For robust data protection, an authorization mechanism is essential as well as authentication mechanism. These two mechanisms needs be integrated so that the information

necessary for authentication and authorization, such as user permission and user attributes, is seamlessly exchanged to offer access control capability between two mechanisms. In the research platform presented in this paper, no authorization mechanism has been provided, while GSI-based authentication mechanism has been provided. For this problem, we would like to integrate an authorization mechanism most suitable for our research purpose. Examples of such mechanisms include Community Authorization Service [11] and PRIMA [7]. These mechanisms allow us to authorize the right permission to system users. The common characteristics of these authorization mechanisms are that these mechanisms have the capability of building authorization infrastructure based on VO. Also, we also consider the use of PERMIS (Privilege and Role Management Infrastructure Standards Validation), which is based on PMI (Privilege Management Infrastructure) using attribute certificates conforming to X.509 standard (<http://www.permis.org/>).

Finally, we need to discuss the scheduling problem of computational workload. In this preliminary research, we connected two sites to realize a Grid-enabled research platform. Therefore, we assumed that the computation is performed in Osaka University and data of interest is located in CAS. Taking the general situation that the Grid-enabled research platform is required into consideration, however, data and computation should take place from the aspect of load balancing and performance maximization. For example, assume that multiple cluster systems and databases are integrated for the research platform. In case that there are multiple sites holding the identical public database such as PDB and GenBank, preferable is that the database close in a network sense to the cluster system performing the analytic computation is selected. Also, the cluster system with low utilization must be utilized. Such a scheduling mechanism that offers the functionalities described above is required for our research platform. We plan to undertake this problem in near future. We consider that there are many issues to achieve other than these three major issues. However, notice that this paper reports the result of the preliminary work in our research, meaning that the work reported in this paper is just one step toward our final goal. Most important in this work is that the architecture that realizes a Grid-enabled research platform with location transparency is shown to be technically feasible.

5 Conclusion

This paper described the architecture of a research platform built over the actual network between China and Japan. The Grid technologies have been adopted as a building block for the research platform. Particularly, in this paper, we showed how two Grid-enabled technologies developed by BioGrid project promoted by Osaka University, that is, GSI-SFS and GUIDE were technically integrated to produce the synergy of location transparency concerning program code and data. On the other hand, this paper also shows that the current research platform has three major future issues to achieve for practical use.

We consider that the importance of this work can be explained from the fact that bio scientists and researchers come to be able to make use of computational resources and data resources without the location of resources on the research platform, although it still have serious problems to solve as described in Section 4. This is just a result of preliminary work toward our final goal of realizing a research platform through which scientists establish a virtual group, organization and community for their research purposes, irrespective of the actual structure formed by organizations. We would like to continue the research toward this final goal from now on.

Acknowledgments

This work was partly supported by the IT program of the Ministry of Education, Culture, Science and Technology of Japan (MEXT). We appreciate APAN (Asia Pacific Advanced Network) and PRAGMA (Pacific Rim Applications and Grid Middleware Assembly) research community for the provision of

this collaboration framework.

References

- [1] Allcock, W., GridFTP protocol specification, *Global Grid Forum Recommendation GFD*, 20, 2003.
- [2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.*, 25(17):3389–3402, 1997.
- [3] Buttler, R., Engert, D., Foster, I., Kesselman, C., Tuecke, S., Volmer, J., and Welch, V., A national authentication infrastructure, *Computer*, 33(12):60–66, 2000.
- [4] Czajkowski, K., Foster, I., Karonis, N., Kesselman, C., Martin, S., Smith, W., and Tuecke, S., A resource management architecture for metacomputing systems, *Proc. of IPPS/SPDP'98 Workshop on Job Scheduling Strategies for Parallel Processing*, 62–82, 1998.
- [5] Foster, I., Kesselman, C., Nick, J., and Tuecke, S., Grid services for distributed system integration, *Computer*, 35(6):37–46, 2002.
- [6] Ishikawa, Y., Tezuka, H., Hori, A., Sumimoto, S., Takahashi, T., O'Carroll, F., and Harada, H., RWC PC cluster II and score cluster system software - high performance linux cluster, *Proc. of 5th Annual Linux Expo*, 55–62, 1999.
- [7] Lorch, M., Adams, D., Kafura, D., Koneni, M., Rathi, A., and Shah, S., The PRIMA system for privilege management, authorization and enforcement in grid environments, *Proc. 4th International Workshop on Grid computing*, 109–116, 2003.
- [8] Mazieres, D., *Self-Certifying File System*, Doctoral dissertation, Massachusetts Institute of Technology, 2000.
- [9] Novotny, J., Tuecke, S., and Welch, V., An online credential repository for the Grid: MyProxy, *Proc. of the Tenth International Symposium on High Performance Distributed Computing (HPDC-10)*, 104–111, 2001.
- [10] Pearlman, L., Kesselman, C., Gullapalli, S., Spencer Jr., B.F., Futrelle, J., Ricker, K., Foster, I., Hubbard, P., and Severance, C., Distributed hybrid earthquake engineering experiments: Experiences with aground-shaking grid application, *Proc. of the 13th IEEE Symposium on High Performance Distributed Computing*, 14–23, 2002.
- [11] Pearlman, L., Welch, V., Foster, I., Kesselman, C., and Tuecke, S., A community authorization service for group collaboration, *Proc. of the IEEE third International Workshop on Policies for Distributed Systems and Networks*, 50–59, 2002.
- [12] Thompson, J.D., Higgins, D.G., and Gibson T.J., CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22(22):4673–4680, 1994.
- [13] von Laszewski, G., Foster, I., Gawor, J., and Lane, P., A java commodity grid toolkit, *Concurrency: Practice and Experience*, 13:643–662, 2001.
- [14] World Conservation Monitoring Centre, *Global diversity: Status of the Earth's Living Resources*, Chapman & Hall, 1992.