Sharing lattice QCD data over a widely distributed file system

T Amagasa¹, S Aoki², Y Aoki³, T Aoyama³, T Doi³, K Fukumura², N Ishii³, K-I Ishikawa⁶, H Jitsumoto⁷, H Kamano⁵, Y Konno¹¹, H Matsufuru⁸, Y Mikami¹¹, K Miura³, M Sato¹, S Takeda⁹, O Tatebe¹, H Togawa⁷, A Ukawa¹⁰, N Ukita¹, Y Watanabe⁴, T Yamazaki¹ and T Yoshie¹

¹ Center for Computational Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8577, Japan
² Yukawa Institute for Theoretical Physics, Kyoto University, Kyoto, Kyoto 606-8502, Japan
³ Kobayashi-Maskawa Institute for the Origin of Particles and the Universe (KMI), Nagoya University, Nagoya, Aichi 464-8602, Japan
⁴ RIKEN Nishina Center, Wako, Saitama 351-0198, Japan
⁵ Research Center for Nuclear Physics, Osaka University, Ibaraki, Osaka 567-0047, Japan
⁶ Graduate School of Science, Hiroshima University, Higashi-Hiroshima, Hiroshima 739-8526, Japan
⁷ Information Technology Center, University of Tokyo, Bunkyo-ku, Tokyo 113-8658, Japan
⁸ Computing Research Center, KEK, Tsukuba, Ibaraki 305-0801, Japan
⁹ Institute of Physics, Kanazawa University, Kanazawa, Ishikawa 920-1192, Japan
¹⁰ RIKEN Advanced Institute for Computational Science, Kobe, Hyogo 650-0047, Japan
¹¹ Hitachi Solutions East Japan, Ltd., Kawasaki,Kanagawa 210-0007, Japan

E-mail: yoshie@ccs.tsukuba.ac.jp

Abstract. JLDG is a data-grid for the lattice QCD (LQCD) community in Japan. Several large research groups in Japan have been working on lattice QCD simulations using supercomputers distributed over distant sites. The JLDG provides such collaborations with an efficient method of data management and sharing.

File servers installed on 9 sites are connected to the NII SINET VPN and are bound into a single file system with the GFarm. The file system looks the same from any sites, so that users can do analyses on a supercomputer on a site, using data generated and stored in the JLDG at a different site.

We present a brief description of hardware and software of the JLDG, including a recently developed subsystem for cooperating with the HPCI shared storage, and report performance and statistics of the JLDG. As of April 2015, 15 research groups (61 users) store their daily research data of 4.7PB including replica and 68 million files in total. Number of publications for works which used the JLDG is 98. The large number of publications and recent rapid increase of disk usage convince us that the JLDG has grown up into a useful infrastructure for LQCD community in Japan.

1. Introduction
Numerical simulation of lattice QCD (LQCD) is the only non-perturbative method to understand hadron/nuclear physics from the first principles of strong interaction of quarks and gluons.
Fundamental data of simulations called “gauge configurations” require huge computer resources to be generated, and are valuable because they can be used to study various aspects of elementary particle physics. With this reason, it is usual in LQCD community to organize a large research collaboration which uses various supercomputers distributed over distant sites.

The Japan Lattice Data Grid (JLDG) [1] provides such collaborations with an efficient method of data management and sharing. Since the official start of operation in 2008, the JLDG has been improved in various ways. Among others, the following two have drastically improved usability of the JLDG. 1) Implementation of FUSE mount which enables users to mount the JLDG file system as a unix file system. Operation of FUSE mount started in 2011. 2) Development of a fast data copy subsystem between the JLDG and the HPCI Shared Storage, where the HPCI is a Japanese national project started in 2011 for constructing High Performance Computing Infrastructure. The subsystem was constructed in 2013 and started its operation in 2014.

In this presentation, we give the first overall report of the JLDG, because we have not reported details of the JLDG anywhere else yet, and recent usage of the JLDG convinces us that the JLDG is now working as an indispensable infrastructure for daily research of LQCD in Japan.

2. System overview

The JLDG consists of the main system and two subsystems as shown in Fig. 1.

![Figure 1. Schematic diagram of the main and two subsystems of the JLDG](image)

The main system provides users a grid-based file system. File servers installed on 9 main lattice QCD research sites (see Fig. 1 and author addresses of this report for official names of JLDG sites) are connected to the NII SINET [2] VPN called HEPnet-J/sc, operated with 1Gbps bandwidth, and are bound into a single partition file system with the GFarm [3], a grid-base file system software. Because the file system looks the same from any sites, users can do analyses
(measurement of physical quantities) on a supercomputer on a site, using data generated and stored in the JLDG at a different site. As of April 2015, 32 file servers provide users with 5.8PB disk space in total.

The ILDG (International Lattice Data Grid) project [4] shown in the upper right part of Fig. 1 is an activity to share valuable lattice QCD configurations worldwide. Collaborations make their configurations public via the ILDG, after they completed physics study of interest, so that other collaborations can study other physics with these data. The ILDG is a grid of 5 regional grids, each of them being constructed with different grid software and operated with different policy. The ILDG defines standard data format, markup language and interface to download configurations from one of regional grids. The JLDG has a subsystem which enables ILDG users to access to a public directory of the JLDG file system using the ILDG interface. In other words, the JLDG works as the Japanese grid of the ILDG.

The HPCI (High Performance Computing Infrastructure) [5] is a national project to provide computational platform realized by connecting major supercomputers in Japan. In addition to computational resources such as K computer, the HPCI system includes the HPCI Shared Storage (HPCI-SS), a gfarm-based file system, which provides research groups with file space. The HPCI covers a wide area of sciences. Several lattice QCD researches are selected as HPCI projects and use mainly the K computer. Fundamental data generated by these project are once stored in the HPCI-SS and then transferred to the JLDG for measurement of physical quantities using supercomputers connecting to the JLDG. A subsystem of the JLDG for cooperation with the HPCI-SS enables such users to transfer data between the two grid file systems in multiple streams. See the middle right part of Fig. 1.

3. JLDG main system
3.1. Hardware and user interface

![Schematic diagram of the JLDG system at a typical site](image)

Figure 2. Schematic diagram of the JLDG system at a typical site

We provide two user interfaces to access the JLDG file system, grid-ftp and JLDG FUSE mount. In order to explain the JLDG system and its design concept, we show in Fig. 2 a schematic diagram of hardware components at a typical JLDG site.
Three networks are involved in the system. One is a subnet of the VPN HEPnet-J/sc shown in the upper part of the figure. JLDG file servers are attached to the subnet and bound to a single partition file system together with file servers installed at other JLDG sites. A grid-ftp server is running on one of the JLDG servers, which enables grid-ftp access to files in the JLDG via a grid-ftp interface of the GFarm.

The VPN connects JLDG sites directly without passing through firewalls of participating universities or institutes, and therefore can be a backdoor to each site. In order to reduce potential security risks, we prepare a closed network which we call a buffer network (shown in the middle part of Fig. 2) between the VPN and the site LAN (in the lower part). JLDG file servers are attached to the buffer network in addition to the VPN, while grid-ftp clients are connected to both the buffer network and the site LAN. We setup the JLDG servers in a way that users can access the servers only through grid-ftp protocol via the buffer network. (In particular, we prohibit user login to JLDG servers.) We think that this framework makes it quite difficult for intruders to access the VPN.

In order to implement FUSE mount of the JLDG file system, FUSE clients have to communicate directory to all of the JLDG file servers via a GFarm data port. The L3 switch shown in the figure routes packets between the VPN and the site LAN. ACL of the switch restricts ports and target FUSE clients so that any unnecessary communications are rejected.

3.2. Software
JLDG software consists of seven packages.

The Gfarm developed at the AIST and the University of Tsukuba provides a grid-base global file system. It utilizes a grid middleware Globus toolkit [6]. The Uberftp package developed at the NCSA provides an interactive and non-interactive grid-ftp access to the JLDG file system. (The grid-ftp interface of the gfarm is included in the gfarm package and is installed in some of the JLDG servers.) A supplemental package Gfarm2fs installed in all JLDG servers and FUSE clients provides FUSE access to the JLDG file system.

Access control to files and directories of the JLDG file system has to be realized independently of individual JLDG sites. The gfarm supports an access control similar to that of Linux based on user and group. As usual, grid-certificate technology is employed to manage users. We operate a private CA using the Naregi-CA package developed by the NAREGI project [7]. A grid user certificate is issued to each user, and the certificate is mapped to one of gfarm users. In order to reduce management cost of user and group, we operate VOMS (virtual organization management system) using the VOMS package developed by the Open Science Grid project [8]. A management server periodically imports (once a day) the user/group database of the VOMS, updates the mapping of grid certificate and gfarm user, and distributes it to all JLDG servers.

In order to detect systematically important events of the JLDG system, such as failures of servers, we started to operate the Zabbix [9] monitoring system very recently.

3.3. Features of JLDG file system
Based on our experience of operating the JLDG over 7 years, we summarize some features of the JLDG file system, which we think important for operating a large scale file sharing system.

- **Easiness of adding file space:**
  In general, capacity of a file sharing system has to be increased according to user demand. The JLDG file space can be easily added by buying file servers and installing them as JLDG servers. Because the JLDG file system remains unchanged except file system size, users do not have to re-arrange their data.

- **Solidness against data loss or corruption:**
  The JLDG creates automatically three replicas (including original), when a user file is
uploaded to the JLDG file system. Because file replicas are created on three different servers, we can recover files even if two of servers are broken. File replication provides another merit: we can make two servers off-line for maintenance, while continuing operation. In order to detect data corruption due to e.g. a bug of file server firmware, a check-sum (md5sum) is calculated and compared when a file or replica is created. This is an efficient and low-cost method of detecting data corruption, and is implemented in the JLDG very recently. See ref. [10] for details.

- Flexible access control:
The JLDG supports an access control to files and directories based on user and group. In addition to the unix-like access control, permission to specified user/group can be set to files and directories, which enables us to respond precisely to user’s request.

- Fast file copy in multiple streams:
The JLDG supports a fast file copy in multiple streams, which can be useful and efficient when many files (such as all files in a directory) have to be copied. Throughput reaches network bandwidth for some cases. An example is given in sec. 4.3.

3.4. Operation policy and method
We list below our operation policy and method. Some of them may be quite different from those for other grid systems and therefore readers may be interested in.

- Any lattice QCD collaboration in Japan can use the JLDG without charge with no quota.
- One can store data of any type (QCD configurations, quark propagators · · ·), if these data should be shared within a collaboration.
- The JLDG team is organized and discusses everything of the JLDG. Members work as volunteers. The team consists of computer scientists who developed and improve the JLDG, representatives of JLDG sites and representatives of research collaborations.
- A part of administrative and maintenance work is outsourced.
- We have no JLDG specific budget. A part of physics or computer science budgets is devoted to the JLDG activity.

4. Cooperation subsystem with HPCI-SS
4.1. Purpose and solution
As mentioned in Introduction, the HPCI provides computing resources of supercomputers to some lattice QCD research projects. In addition, the HPCI provides them disk space as a part the HPCI-SS. Such research projects need to transfer their data between the HPCI-SS and the JLDG, because some of supercomputers they use for generating data connect only to the HPCI-SS and hence data can be stored only in the HPCI-SS, while supercomputers they want to do measurements on these data connect only to the JLDG. They also want to archive all fundamental data in the JLDG, before the HPCI project finishes.

The subsystem of cooperation with the HPCI-SS is a client machine which enables users to mount both the JLDG and the HPCI-SS file systems, and to copy files in multiple streams between the two file systems.

4.2. Authentication
There were almost no technical difficulty to develop such a client, because both file systems employ the common grid file system software gfarm, which supports FUSE mount and file copy in multiple streams.

The only issue resides in user authentication, or how to treat user grid certificate. Even if the JLDG accepts the HPCI user certificate for user authentication, problem still remains because
the gfarm can map only one grid certificate to a gfarm user (global name), and therefore a person has to have two gfarm user names (one associated with the JLDG certificate and one with the HPCI certificate), unless the user abandons the account associated with the JLDG certificate.

We have agreed to compromise this issue as follows.

- A user can have two gfarm user names associated with the JLDG and the HPCI certificates, respectively. In this case, the user has to care file access control, e.g. set read/write permission of a directory to the two gfarm user names.
- A user can use only one gfarm user name associated with the HPCI certificate. In this case, he/she has to copy proxy certificate manually to JLDG clients.

We summarize in Fig. 3 how users handle grid certificates for the cooperation subsystem.

![Flow diagram of grid (proxy) certificates for the cooperation subsystem with the HPCI-SS. Pictures of K computer and the HA-PACS are taken from [11, 12], respectively.]

Figure 3. Flow diagram of grid (proxy) certificates for the cooperation subsystem with the HPCI-SS. Pictures of K computer and the HA-PACS are taken from [11, 12], respectively.

4.3. Performance and a use case

File copy from the HPCI-SS to the JLDG is carried out with high performance. We plot in Fig. 4 throughput (MB/s) versus number of files copied, where file size is fixed to 1GB (typical size of files user transfers) and number of files per stream to 16. Throughput exceeds 80 MB/s when number of files is larger than 100, and reaches 90 MB/s for 128 files. Because the JLDG is operated with 1Gbps network, the achieved throughput corresponds to almost full use of the network bandwidth.

The cooperation subsystem with the HPCI-SS has already been used by several research collaborations. Fig. 5 shows data size of transferred files per day, plotted against date of the transfer job. We see more than 4TB data are transferred in several days.
Figure 4. Performance of file copy from the HPCI-SS to the JLDG via the cooperation subsystem with the HPCI-SS

Figure 5. Data size per day, transferred from the HPCI-SS to the JLDG by the most active research group

5. Statistics
Currently, 15 research groups which include 61 researches use the JLDG to share daily research data. Table 1 classifies groups and users according to organizations to which a representative of the research group belongs to, or a researcher belonged to when he/she joined the JLDG VO.

| Organization | # of groups | # of users |
|--------------|-------------|------------|
| Tsukuba      | 5           | 20         |
| KEK          | 2           | 5          |
| Osaka        | 2           | 9          |
| Nagoya       | 1           | 5          |
| Kyoto        | 1           | 0          |
| Riken        | 1           | 0          |
| Others       | 3           | 22         |

As Fig. 6 shows, data size stored in the JLDG increases very rapidly, and reached 4.7PB (including replica) in April 2015. Number of files are also increases rapidly to reach 68 million. Fig. 7 shows number of publications for works which used the JLDG.

6. Conclusions
The JLDG continues operation for almost seven years since 2008. Advantage of sharing research data on the widely distributed file system has been gradually recognized. As seen in sec. 5, disk usage increases rapidly since 2013. Number of publications for works which used the JLDG for data sharing also increases gradually and reached 98 in total (as of Sept. 2014).

These observations convince us that the JLDG has grown up to a useful infrastructure for lattice QCD community in Japan to manage and share data within research groups. The success of the JLDG means that strategies we employed for data sharing system match demand of our community.

We continue improving and enhancing the JLDG further, hoping that sharing lattice QCD data on a widely distributed file system leads to reformations of computing style. For example, if simulation program running on supercomputers can read/write data from/to the JLDG file system, data driven cooperative use of supercomputers at distant sites will be realized.
Acknowledgments
We are grateful to strong encouragement including several financial supports from both lattice QCD community and computer science community. It supported a close collaboration of lattice QCD physicists and computer scientists in the JLDG team, which we believe is an important element leading to the success of the JLDG.

JLDG is currently supported by “HPCI strategic program Field 5”. Cooperation subsystem with HPCI-SS was developed through the HPCI System Research Project (Project ID:hp120108).

References
[1] http://www.jldg.org/
[2] http://www.sinet.ad.jp/
[3] http://sourceforge.net/projects/gfarm/
[4] Davies C T H, Irving A C, Kenway R D and Maynard C M 2003 Proc. XXth Int. Symp. on Lattice Field Theory (LATTICE 2002) (Boston) Nucl. Phys. B (Proc. Suppl.) 119 (Amsterdam: North-Holland/Elsevier) p 225;
Beckett M G, Joo B, Maynard C M, Pleiter D, Tatebe O and Yoshie T 2011 Comput. Phys. Commun. 182 1208, and references therein.
[5] http://www.hpci-office.jp/
[6] http://toolkit.globus.org/toolkit/
[7] https://www.naregi.org/
[8] https://www.opensciencegrid.org/
[9] http://www.zabbix.com/
[10] Tatebe O and Yoshie T 2015 these proceedings
[11] http://www.aics.riken.jp/
[12] http://www.ccs.tsukuba.ac.jp/