Data Acquisition System for Steady State Experiments at Multi-Sites

H. Nakanishi 1), M. Emoto 1), Y. Nagayama 1), T. Yamamoto 1), S. Imazu 1), C. Iwata 1), M. Kojima 1), M. Nonomura 1), M. Ohsuna 1), M. Yoshida 1), M. Hasegawa 2), A. Higashijima 2), K. Nakamura 2), Y. Ono 3), M. Shoji 1), S. Urushidani 4), M. Yoshikawa 5), K. Kawahata 1)

1) National Institute for Fusion Science (NIFS), Toki, Japan
2) Research Institute for Applied Mechanics, Kyushu University, Kasuga, Japan
3) University of Tokyo, Tokyo, Japan
4) National Institute of Informatics (NII), Tokyo, Japan
5) Plasma Research Center, University of Tsukuba, Tsukuba, Japan

E-mail contact of main author: nakanisi@nifs.ac.jp

Abstract. A high-performance data acquisition system (LABCOM system) has been developed for steady state fusion experiments in Large Helical Device (LHD). The most important characteristics of this system are the 110 MB/s high-speed real-time data acquisition capability and also the scalability on its performance by using unlimited number of data acquisition (DAQ) units. It can also acquire experimental data from multiple remote sites through the 1 Gbps fusion-dedicated virtual private network (SNET) in Japan. In LHD steady-state experiments, the DAQ cluster has established the world record of acquired data amount of 90 GB/shot which almost reaches the ITER data estimate. Since all the DAQ, storage, and data clients of LABCOM system are distributed on the local area network (LAN), remote experimental data can be also acquired simply by extending the LAN to the wide-area SNET. The speed lowering problem in long-distance TCP/IP data transfer has been improved by using an optimized congestion control and packet pacing method. Japan–France and Japan–US network bandwidth tests have revealed that this method actually utilize 90% of ideal throughput in both cases. Toward the fusion goal, a common data access platform is indispensable so that detailed physics data can be easily compared between multiple large and small experiments. The demonstrated bilateral collaboration scheme will be analogous to that of ITER and the supporting machines.

1. Introduction

As the fusion plasmas have their own fluctuation frequencies between kHz and MHz, the diagnostic sampling rates must be always higher than them. In case of a spatial profile measurement, 50 or 100 channels are typically used so that their total data production will go easily up to 100 MB/s. Until recently, they were acquired after the end of short pulse plasma discharges. In steady state experiments, however, they must be acquired, processed, displayed, and also stored in real time (RT). Therefore, a high speed continuous data acquisition system has been developed in LHD [1].

Nowadays, fusion researches are being progressed by bilateral collaborations among a few large experiments and number of small experiments. Toward the fusion goal, a common data access platform is becoming more indispensable so that physicists can easily make detailed comparisons between multiple large and small experiments [2].

The data amount for fusion plasma diagnostics keeps growing in about 50% per year, i.e. 10 times in 5 years. This tendency is definitely observed in LHD (FIG. 1), and also similar phenomena can be seen in many fusion experiments in the world [3-4]. For the next generation fusion experiment, the new data access platform is expected to provide more scalable performance to cover 10 or 100 times data growth through the whole life of each experimental project.

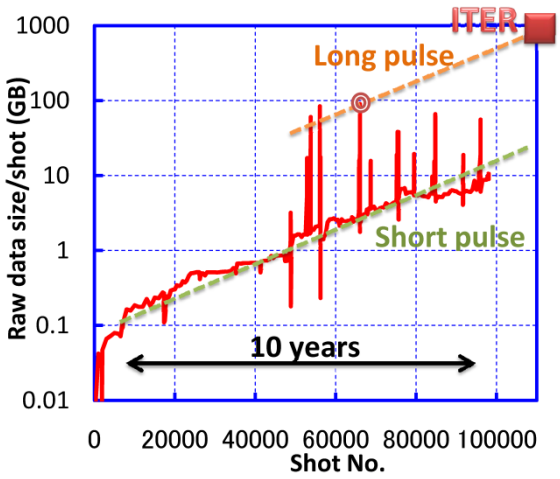

FIG. 1. Data growth in LHD: Each spike means the steady-state trial. A red double circle shows the world record of 90 GB.

2. LABCOM System

The LABCOM system primarily consists of a number of distributed DAQ computers, storage devices, and the index database which stores all the data locations. Its key objectives are; (1) RT DAQ with the same sampling rates as burst acquisition; (2) scalability of both the performance and quantity, even in the number of sites. LHD now uses 80 DAQs in parallel, and acquires 10.6 GB/shot raw data in short pulse experiments of 3 min. iteration. In other words, it has about 170 shots per day and produces roughly 1.8 tera-byte (TB) plasma diagnostic data in a day.

2.1. Real Time Data Acquisition

In steady state operations, each RT DAQ node can continue nonstop acquisition at the rate of maximum 110 MB/s. Of course, their data streams can be obtained, processed, and displayed by the RT monitoring clients. In order to reduce the client loads for receiving the massive data stream, each server has the functionality to thin out the outgoing stream into 1/N according to each client's request. The high-speed RT DAQ has another potential ability that it could unify the fast DAQ system for RT feedback controls and that for detailed physical diagnostics [5]. Until now, they are separately implemented by using different hardware and operating systems.

As for storing the endless data streams, the continuous 110 MB/s writing cannot be achieved by a single hard drive and therefore a stripe set of multiple hard drives, namely RAID-0, is adopted in every node. In order to store a long data stream having indefinite time duration, the LABCOM system has invented a new idea of "sub-shot" that contains 10 second definite time chunk out of the stream. The full length data over the plasma duration will be cut into numbers of 10 second sub-shots. It is because anyone could not refer the unclosed file until the end of long pulse discharge if the RT DAQ continued writing the data stream into a single file [6-7]. The inside diagram handling the data blocks is shown in FIG. 2. The DAQ cluster has established the world record of acquired data amount of 90 GB/shot in a long pulse experiment of LHD, which almost reaches the ITER data estimate of 100 or 1000 GB/shot (FIG. 1).

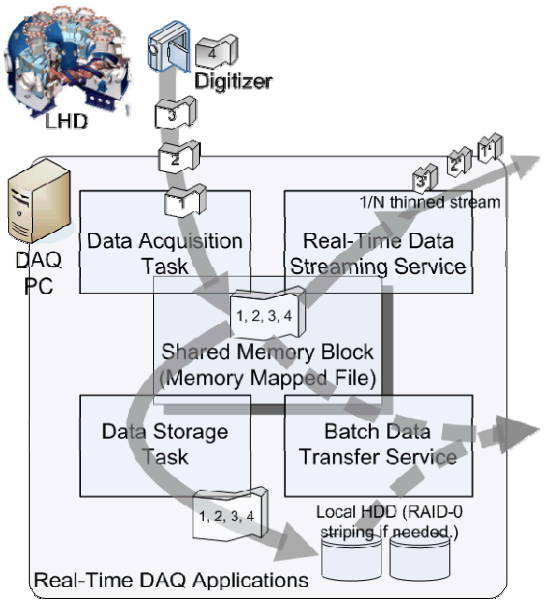

FIG. 2. Inside diagram of streaming DAQ: All the data blocks are handled on the volatile random access memory (RAM) [5]. The sub-shot data files will be newly made and closed every 10 seconds so that any data clients can refer them even if the RT DAQ is still continuing.

2.2. Data Store

As for the data archiving storage, the massively parallel processing (MPP) structure is important for scalable input/output (I/O) performance and also for data redundancy [8-9]. As MPP is the key technology in modern high-performance computing, it is also promising to cover the ITER data amount.

Hundreds of tera-byte compressed data are stored in the two-stage storage in LHD. The first one is a cluster of hard drive arrays for faster I/O, and the second consists of Blu-ray Disc (BD) libraries for data security [10]. The data accessing "Retrieve" utility always consults the data indexing database. This data mediation service by "Facilitator" is a distinguishing characteristic that manages peer-to-peer data handling among many client and server computers (see FIG. 3 (top)) [11].

As shown in FIG. 2, the first data entity appeared in DAQ unit is volatile on random access memory (RAM), it is indispensable to be stored into persistent storage like hard drive arrays. In order to avoid that there only exists one data entity so long, another copy must be replicated on the spot. To satisfy those requirements, the LABCOM data store once adopted the symmetric storage cluster on FibreChannel-based storage area network (FC-SAN) (FIG. 3). It provided enough fast data read/write speed through 4 Gbps FibreChannel switches and links, however, a unit trouble often caused the whole system disorder because of their tightly linked structure. In order to obtain higher availability and faster throughput with larger capacity, we newly adopted the network attached storage (NAS) system based on 10 gigabit Ethernet.

To replace the Red Hat Global File System (RH-GFS) previously used on FC-based storage cluster, we have adopted the "IznaStor" distributed Key-Value Store (KVS) file archiving system (see FIG. 3 (bottom)). It provides high availability and reliability due to their hot plug-and-play nodes, load balancing, automatic file replication, and the scalable I/O throughput

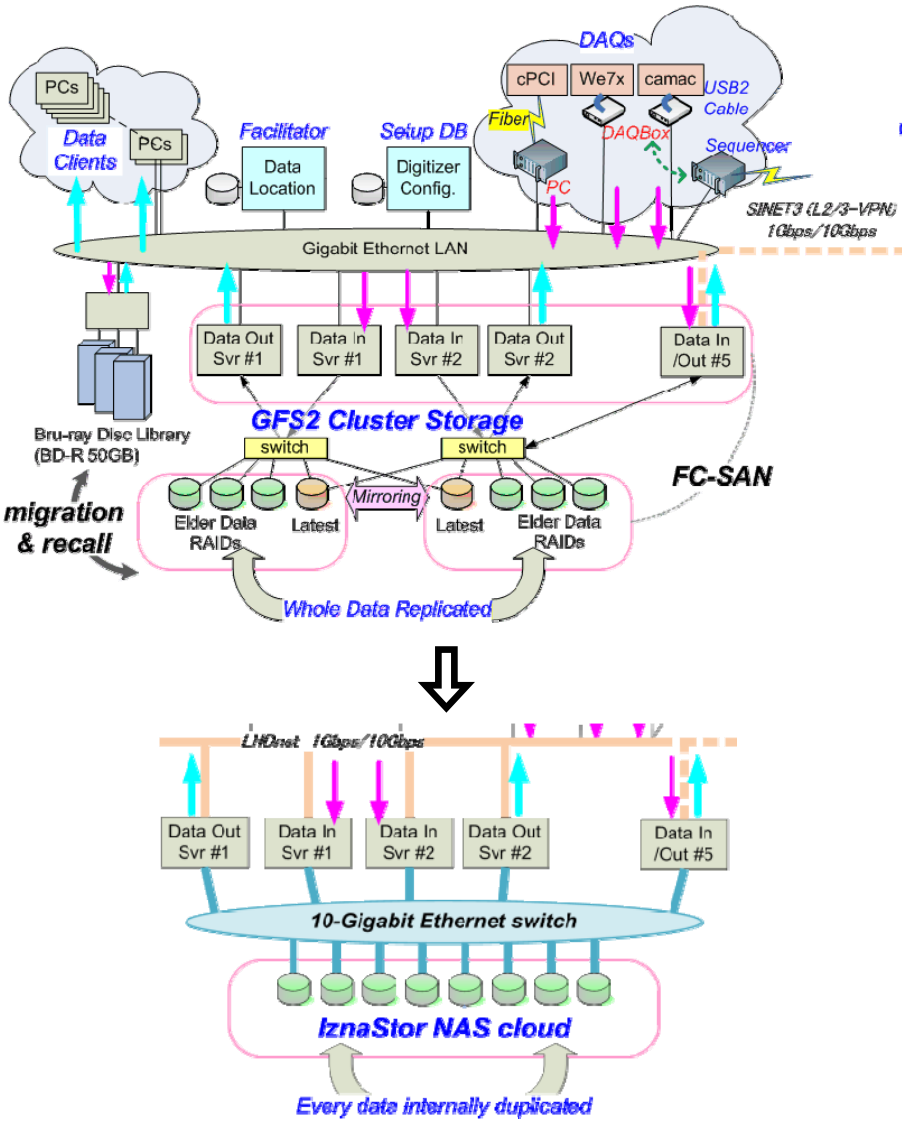

FIG. 3. Schematic view of the LABCOM DAQ and store system: The storage cluster had a mirrored structure on FibreChannel storage area network (FC-SAN) to store replicated files previously (left). They are replaced by distributed network attached storages (NAS) named "IznaStor" in 2010.

without any single point of failure (SPOF). The read/write performance enhancement is verified as shown in TABLE I.

Distributed KVS can be considered as a promising technology for the massive sized data archives such as large fusion experiments. Especially in LABCOM system, the data retrieval queries are made simply by the data (diagnostic) name and the shot number which can be easily translated into the "key" string of the KVS. Of course, the data stream is stored as "value" itself. It seems more suitable for the experimental data store to simply get/put the archived data from/to the storage, and provides much scalable I/O performance with high

TABLE I: PERFORMANCE COMPARISON BETWEEN OLD AND NEW STORAGE

Storage device/Filesystem	**I/O throughput**
4 Gbps FC-RAID/RH-GFS2	107 MB/s
10 Gbps IznaStor/FUSE isfs (virtual fs)	172 MB/s
10 Gbps IznaStor/KVS put/get (native)	403/620 MB/s

availability, instead.

3. Multi-site Data Access Platform

Major Japanese fusion laboratories and universities are connected each other through the layer-2 or layer-3 virtual private network (VPN) on SINET3 [12], namely SNET [13] (see FIG. 4). As every element of the LABCOM system are distributed on the local area network (LAN), the data of remote fusion devices can be acquired simply by extending LAN to the wide-area SNET. In addition to the LHD experiments at NIFS, the LABCOM system acquires data from two remote experiments: QUEST at Kyushu University [14], GAMMA 10 at the University of Tsukuba [15].

To realize a common data access platform from many DAQ servers to many data retrieval clients across multiple remote sites, the above mentioned "Facilitator" plays the most important role by mediating all the data access. The ownership of each site's data and the access privileges to the proper user group must be implemented as an extension to the LABCOM data system developed originally for the LHD experiment [16].

In order to provide the better presence of the remote experiment, the high-definition video streaming not only for displaying the plasma in real time but also for the view of the experiment control room has been operating since 2009. Modern digital cameras often produce high bandwidth data streams. For instance, a popular VGA (640x480) full colour camera outputs 70 or 80 MB/s continuously. Therefore, issues in multi-site data access can be thought as (1) network throughput, and (2) access control of data and users in multiple experiments.

3.1. TCP Long Fat Pipe Network (LFN) Problem

The speed lowering problem has been revealed more seriously in long-distance TCP/IP communications on SNET. For instance, the effective throughput over 1 000 km distance between LHD in NIFS and QUEST in Kyushu Univ. was only 60 Mbps on 1 Gbps SNET. The standard data transferring protocol TCP/IP has a well-known "long fat pipe network (LFN)" problem that deteriorates the effective throughputs of long-distance wideband

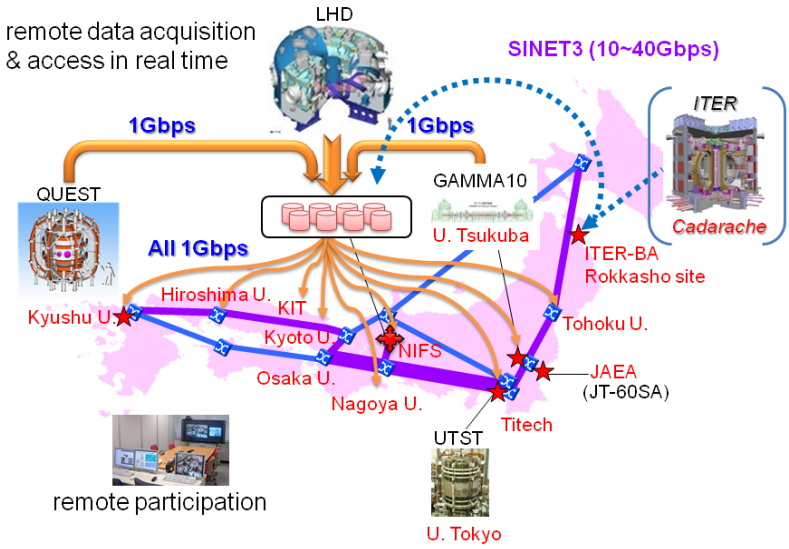

FIG. 4. Multi-site data platform on SNET: All the diagnostic data of each device are migrated into main storage in RT, and then delivered to every remote and local site.

communications [17]. It is usual, for instance, to get less than 1 Mbps speed through the 10 Gbps oversea shared link. For the recent two years, we practically did some LFN tests between Japan-US and Japan-France [18], in order to develop a reliable way for making the remote experiment more realistic.

By tuning the TCP related kernel parameters with the optimized congestion control method and also applying the packet pacing algorithm, we can drastically improve the network performances against the LFN problem. Our Japan-France bandwidth tests compared two passes of 10 Gbps international links; one is the shared lines of SINET3, GEANT2 and Renater, and the other is dedicated lines of APAN/JGN2plus and SURFnet. Under 4 Gbps rate limit, both have achieved a very stable throughput of 3.5 Gbps (see FIG. 5). It was achieved by a combination of the TCP parameter tuning and the inter-packet gap (IPG) control on Ethernet hardware driver. Also Japan-US tests proved 451 Mbps through the firewall having 460 Mbps forwarding performance (FIG. 6). This was done by using PSPacer packet pacing software instead of IPG hardware tuning. This verified acceleration method could be applied for the network accelerating gateway for each SNET collaboration site.

3.2. Access Control for Data Security

When used in multiple experimental sites, the data access control is indispensable for both user and data belonging to each site. However, the data security and the high performance

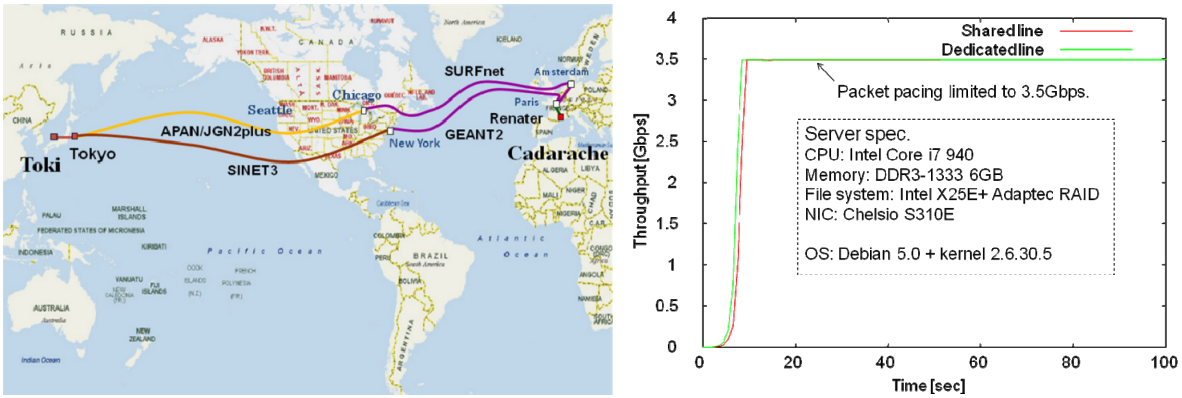

FIG. 5. Results of Japan-France bandwidth test: The fiber distance is about 15 000 km whose round-trip latency time is about 310 ms. Between 10 Gbps ports at both ends, the bandwidth was limited to 4 Gbps for the 10 Gbps shared backbone; SINET3, GEANT2, and Renater. The dedicated path through APAN/JGN2plus and SURFnet was also used.

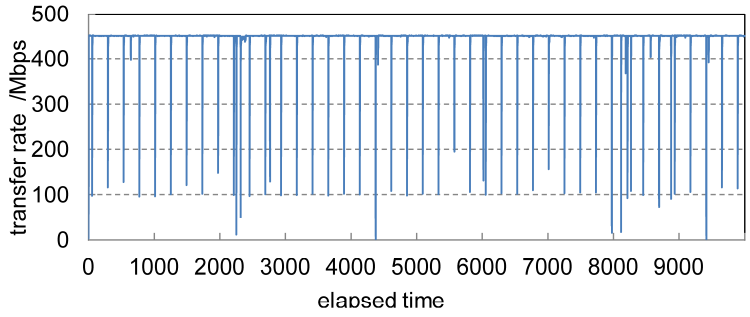

FIG. 6. Result of Japan-US bandwidth test between NIFS and GA, San Diego, CA: With the limited forwarding rate of 460 Mbps, very stable 451 Mbps throughput was obtained over 10 000 s. The comb-like drops are considered due to some periodical task awoken on the receiver computer.

accessibility are opposed ideas that are both essential to the multi-site data platform.

On this shared platform, any researchers can access every remote and local experiment with similar performance regardless of his/her whereabouts. In order not to spoil the network throughputs, a light-weight access control is implemented to give proper data privileges among multiple experimental groups: Each site's data belong to the site group and are accessible only by the client PCs which have been registered to the site group in advance. On the other hand, the backbone SNET provides enough network security because it is intrinsically a closed VPN by itself.

3.3. Remote Operation and Monitoring

In order to operate remote DAQs interactively and monitor their status in real time, a Web-based utility has been newly developed as a part of the LABCOM system. It is based on the so-called "Agent oriented technology" [19] in which the DAQ agent runs on every node to accept commands and respond the resulting status through IP multicast protocols.

We already use it on LHD, QUEST, and GAMMA 10 remote DAQ operations and found that such operation assisting GUIs are indispensable for controlling and monitoring many remote DAQs simultaneously. In addition to the DAQ related tasks and digitizers, remote power controller can be also controlled by this tool. It is quite helpful to make a fresh start on electricity in case of computer or hardware accidental malfunctions.

4. Conclusion

In the LABCOM system, multiple sites' experimental data are shared on SNET. Its performance is enough scalable to almost cover the ITER data amount [20]. The demonstrated bilateral collaboration scheme will be analogous to that of ITER and the supporting machines.

This data sharing platform is a strong infrastructure to allow cross validation not only between different fusion experiments, but also between modelling or simulation results and the experimental data when coupled with Grid or high-performance computing resources. By sharing innumerable experimental and computational results on the same data accessing platform, researchers may find new scaling laws more easily. Direct pattern search for specific plasma waveform or picture may also be enabled.

In conclusion, the modern high bandwidth network gradually gets rid of the distant gap between local and remote sites. The bilateral collaboration scheme of experimental resource sharing may also suggest the international system for future fusion researches.

References

[1] NAKANISHI, H., et al., "Design for real-time data acquisition based on streaming technology", Fusion Eng. Des. **56-57** (2001) 1011–1016.
[2] NAKANISHI, H., et al., "Portability Improvement of LABCOM Data Acquisition System for the Next-Generation Fusion Experiments", Fusion Eng. Des. **82** (2007) 1203.
[3] FARTHING, J.W., et al., "20 Years of Data Acquisition at JET", Proc. 4th IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research, San Diego, USA, July 21–23, 2003.
[4] HOW, J.A., et al., "Trends in Computing Systems for Large Fusion Experiments",

Fusion Eng. Des. **70** (2004) 115.
[5] OHSUNA, M., et al., "Unification of Ultra-Wideband Data Acquisition and Real-Time Monitoring in LHD Steady-State Experiments", Fusion Eng. Des. **81** (2006) 1753.
[6] NAKANISHI, H., et al., "Ultra-Wideband Real-Time Data Acquisition in Steady-State Experiments", J. Plasma Fusion Res. **82** (2006) 171 (in Japanese).
[7] NAKANISHI, H., et al., "Data Acquisition and Management System of LHD", Fusion Sci. Techno. **58** (2010) 445–457.
[8] KOJIMA, M., et al., "Object-Oriented Design for LHD Data Acquisition Using Client-Server Model", Fusion Eng. Des. **43** (1999) 433.
[9] NAKANISHI, H., "Development of Scalable and Distributed Data Acquisition and Storage Systems for Fusion Plasma Diagnostics", PhD Thesis, Graduate University for Advanced Studies, Hayama, Japan (2003) (in Japanese).
[10] NAKANISHI, H., et al., "Multi-Layer Distributed Storage of LHD Plasma Diagnostic Database", J. Plasma Fusion Res. Ser. **7** (2006) 361.
[11] NAKANISHI, H., et al., "Adaptive data migration scheme with facilitator database and multi-tier distributed storage in LHD", Fusion Eng. Des. **83** (2008) 397–401.
[12] NATIONAL INSTITUTE OF INFORMATICS, Science Information NETwork 3 SINET3 (2007), http://www.sinet.ad.jp/
[13] TSUDA, K., et al., "Virtual laboratory for fusion research in Japan", Fusion Eng. Des. **83** (2008) 471–475.
[14] ZUSHI, H., et al., "Study of Edge Turbulence from the Open to Closed Magnetic Field Configuration during the Current Ramp-up Phase in QUEST", this conference EXS/P2-22 (2010).
[15] YOSHIKAWA, M., "Fluctuation Suppression during the ECH induced Potential Formation in the Tandem Mirror GAMMA 10", this conference EXC/P8-21 (2010).
[16] NAKANISHI, H., et al., "Clustered Data Storage for Multi-Site Fusion Experiments", Plasma Fusion Res. **5** (2010) S1042.
[17] YOSHINO, T., et al., "Performance optimization of TCP/IP over 10 gigabit ethernet by precise instrumentation", Proc. 2008 ACM/IEEE conference on Supercomputing, No.11.
[18] YAMAMOTO, T., et al., "Configuration of the virtual laboratory for fusion researches in Japan", Fusion Eng. Des. **85** (2010) 637-640.
[19] ALKHATIB, G., RINE, D., "Agent Technologies and Web Engineering: Applications and Systems", Information Science Publishing, Hershey, Pennsylvania (2008).
[20] LISTER, J.B., et al., "The Status of the ITER CODAC Conceptual Design", Fusion Eng. Des. **83** (2008) 164.