Clustered Data Storage for Multi-Site Fusion Experiments

Hideya NAKANISHI, Mamoru KOJIMA, Masaki OHSUNA, Setsuo IMAZU, Miki NONOMURA, Takashi YAMAMOTO, Masahiko EMOTO, Yoshio NAGAYAMA, Kazuo KAWAHATA, LHD exp. group, Makoto HASEGAWA[1)], Aki HIGASHIJIMA[1)], Kazuo NAKAMURA[1)] and Masayuki YOSHIKAWA[2)]

National Institute for Fusion Science, 322-6 Oroshi-cho, Toki 509-5292, Japan
[1)]*RIAM, Kyushu University, 6-1 Kasuga-kouen, Kasuga 816-8580, Japan*
[2)]*PRC, University of Tsukuba, Tsukuba 305-8577, Japan*

(Received 7 January 2009 / Accepted 3 August 2009)

The LABCOM data acquisition and management system already provides full functionality to both local and remote participants in Large Helical Device (LHD) experiments. This study newly added a function for dealing with raw experimental data not only from one experimental device but also from multiple distant sites. Its original distributed structure has enabled the multi-site modification to be made with a minimum of change, mainly within the data location indexing database for clustered storage. However, access permissions and restrictions for each site's data and users should be strictly implemented. The system began operation in 2008 under bilateral collaborations between the LHD, QUEST, and GAMMA10 experiments, with the goal of organizing a Fusion Virtual Laboratory in Japan.

Keywords: LABCOM, clustered storage, SINET3, LHD, QUEST, GAMMA10, multi-site, access control, Fusion Virtual Laboratory (FVL)

DOI: 10.1585/pfr.5.S1042

1. Introduction

Remote participation technology is fundamental for modern fusion experiments [1, 2]. It is currently based on information highways with capacities of more than 10 Gbps, in which many gigabytes or sometimes terabytes of experimental data are shared by distributed collaborators.

On the other hand, the amount of experimental data, which continue to grow (Fig. 1), often causes too heavy a management burden for operational staff. The increasing cost of data management may be optimized by intensive administration of data storage systems through ultra-wideband networks. The Internet Data Center (IDC), which provides centralized monitoring and control for data resources, is a typical example intended to streamline data management in commercial fields. This solution would also be required in physics research experiments.

SINET3 is a Japanese academic information highway operated by the National Institute of Informatics (NII) which has a 10 or 40 Gbps backbone [3]. It also serves Layer-2 or Layer-3 IP virtual private network (VPN), SNET, exclusively for the fusion research community [4]. It is intrinsically equipped with both wide bandwidth and high security.

SNET has been hosted by the National Institute for Fusion Science (NIFS) since the 2001 fiscal year (FY), initially for Large Helical Device (LHD) remote participation activities [5,6]. Starting in FY 2005, bilateral collaboration programs between NIFS and the research centers of other universities have additionally come into operation. The most typical example is the All-Japan Spherical Tokamak (ST) research program [7], in which remote data acquisition can be realized between its new experimental device QUEST and LHD's data repository.

In this study, we have modified the LHD data acquisition and management system so it can deal with multiple experiments and their data simultaneously. In the following sections, the required specifications and applied implementations are described along with their effectiveness.

2. Access Control for Multiple Sites

As mentioned in the previous section, one of the most important objectives of this study is to build easily extendable data storage with centralized management. The LHD data repository already possesses multiple disk volumes and a FibreChannel-based storage area network (FC-SAN) built by yearly increases in capacity. FC-SAN is the *de facto* standard for massive storage shared for various uses.

The LABCOM data system can already provide full functionality to both local and remote participants in LHD fusion experiments [8, 9]. In this study, however, we must add a new function to deal with raw data acquired not only from one experimental device but also from multiple distant sites.

author's e-mail: nakanisi@nifs.ac.jp

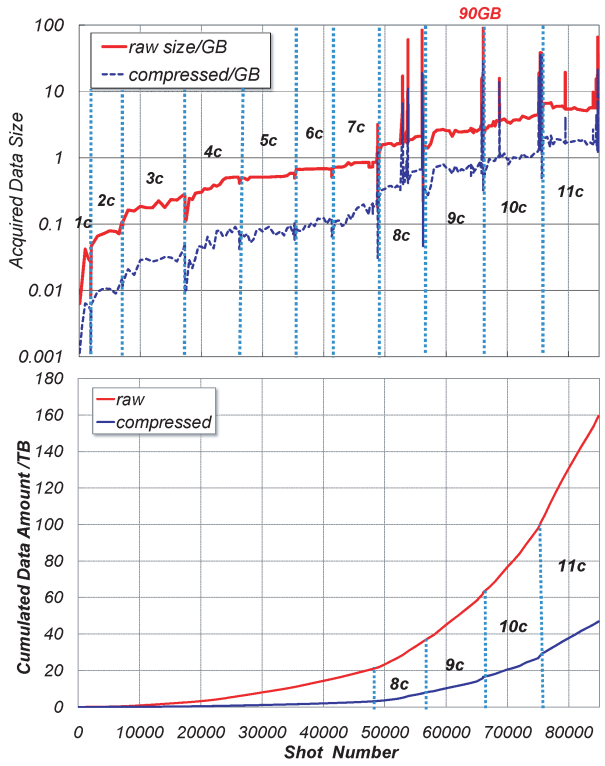

Fig. 1 Data growth in LHD, per-shot size (top) and cumulative amount (bottom): 1c-11c represent annual experimental LHD campaigns. 90 GB/shot is the world record for raw data acquired in one experimental plasma discharge. All acquired data are kept online to be accessible to every collaborator.

Table 1 Related tables in "facilitator" database; components of main "shot" table (left) and contents of new "site" table (right): Bold-faced **shot#, diag#,** and **site#** are the primary keys.

Column	Modifiers
shot	not null
subshot	not null
diag_id	not null
host_id	not null
media_id	not null
regist_no	not null
note_id	not null
site_id	not 0 default 1

site_id	site_name
1	lhd
2	quest
3	gamma10

When sharing clustered storage volumes among different experimental sites, a clear distinction should be made between access permissions and restrictions on the data and users of each site. These access controls will be implemented in the indexing database by adding a new "site" key to existing "diagnostic (data) name" and "shot number" keys. Table 1 shows the essential part of this upgrade. Here, the "site" key should control both diagnostic data and user groups.

Access control among multiple sites' data and user groups has been implemented by a combination of database application accounts dedicated to each site group and access permissions to registered IP addresses. Each stored data set belongs to its own site, and also data retrieval computers are independently registered for each site. The site name must be unique across multiple experimental sites; however, the same diagnostic name and shot number can be used across multiple sites. When a research collaborator joins multiple site groups, he/she can register his/her host to all the sites to access data from them. See Fig. 2.

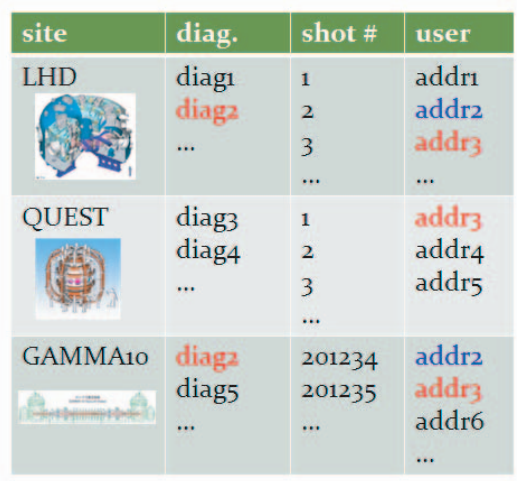

Fig. 2 Relationship between data retrieval query keys and permitted user addresses for each site.

3. New LABCOM/X Data Acquisition and Management System

R&D for the LABCOM data acquisition and management system began in 1995, with the goal of constructing a new plasma diagnostic data system for the LHD experiment in the NIFS. As the first plasma was established in March 1998 [10], it has now experienced ten years of annual campaigns.

One of the LHD's most remarkable achievements was to establish a new world record for the amount of diagnostic data acquired in one fusion plasma discharge. This was achieved by means of a new ultra-wideband real-time data acquisition technology whose maximum performance is up to 160 MB/s for each digitizer front-end [11].

The LABCOM system originally had a distributed structure in which data acquisition and storage elements are completely separated on a fast network [12]. When wide-area networks (WANs) can be equivalent to local-area networks (LANs) in throughput, there is no logical difference between them. The multi-site modification was, therefore, realized with a minimum of change, mainly in the facilitator database that informs the data locations, as shown in Fig. 3.

The main "shot" table in Table 1 contains more than

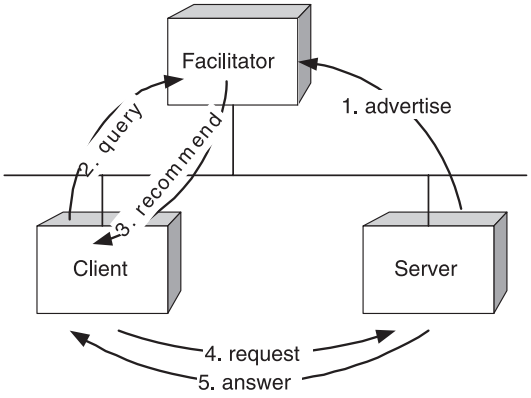

Fig. 3 Recommend-type facilitator model: The facilitator never mediates requests but only recommends the appropriate server to which to send them [13]. It is suitable for a distributed data storage and retrieval system that must transfer many binary large objects (BLOBs) without bottlenecks.

14 million entries of experimental data. By means of the database's embedded acceleration of key indexing, however, a query search can be answered in about 0.14 second.

Our multi-site data system began operating in September 2008 under bilateral collaborations between the LHD at the NIFS, QUEST at the Reseach Institute for Applied Mechanics (RIAM), Kyushu University, and GAMMA10 at the Plasma Research Center (PRC), University of Tsukuba. It applies a new framework, the Fusion Virtual Laboratory, in which users can access data regardless of their whereabouts. This activity makes up SNET, which is based on a closed VPN on the Japanese academic Internet backbone SINET3, as described in Sec. 1.

Due to this topological evolution, we renamed the system LABCOM/X; its present structure is shown in Fig. 4. Four pairs of replicated 24 TB RAIDs make up the clustered storage in which the 2-way FibreChannel switching fabrics provide a redundant SAN. Among them, only one pair of RAIDs is synchronized in real time for storing newly acquired data files; the others preserve older files. The gateway I/O servers to the SAN are also redundant and provide load balancing; currently, two are used for write-in and two for read-out. Data-producing data acquisition (DAQ) servers and retrieval clients can access the I/O servers equivalently from any place on the LHD LAN or SNET WAN.

4. GFS2 Storage Cluster and Data Replication

For a multi-site data repository, it is quite essential that plural I/O servers work redundantly and even provide load-balancing. The cluster filesystem provides a mechanism for synchronizing content data among them. We used Red Hat Global File System (GFS) [14] and later adopted its version 2 (GFS2), whose I/O performance is almost the same as that of ordinary local systems such as xfs or ext3 (Table 2).

Table 2 Throughput difference between local (ext3, xfs) and cluster (GFS2) filesystems: Results from 100 MB write tests of `dd if = /dev/zero of = outfile bs = 1024 count = 102400` and `count = 1048576`.

filesystem	I/O rate (100 MB)	I/O rate (1 GB)
ext3	0.635 s 165 MB/s	8.63 s 124 MB/s
xfs	0.811 s 129 MB/s	8.53 s 126 MB/s
GFS2	0.869 s 121 MB/s	6.68 s 161 MB/s

Generally, cluster filesystems such as Sun's Lustre File System or IBM's General Parallel File System (GPFS) provide better performance in writing huge data volumes by means of splitting I/O into many storage nodes. To maintain consistency among their distributed chunks, they usually need at least one metadata server or service process. On the other hand, GFS never splits a file into many or distributed chunks. It provides only a distributed file-locking mechanism to synchronize the file's appearance among cluster node computers. Thus, it is also possible for us to use a GFS volume as a local filesystem without a metadata server. This feature is quite advantageous when some GFS volumes are filled and changed to read-only, as shown in Fig. 4.

To realize the embedded data replication scheme, some possibilities that apply hardware or software mirroring can be considered. When using a number of huge disk arrays, however, these would be disadvantageous due to the extremely long rebuilding time in recovery when inconsistencies have occurred.

Therefore, we have made a specific utility to replicate newly appearing data files; it runs in cooperation with the facilitator database. The applied replication scheme is a simple combination of request queuing and cyclic batch execution, as shown in Fig. 5. It never checks the equivalence of the source and destination volumes, only making incremental copies of newly appearing files. Such a loosely tied data mirroring mechanism is rather preferable for flexible storage operations.

We have also changed the data migration scheme. Between the remote DAQ servers and the storage servers, we previously adopted the Network File System (NFS) to share the cluster volume on a LAN, in other words, within a single LHD site. However, it could be less reliable for remote data sharing because NFS was designed to be used on LANs. Moreover, more than 70 NFS clients were constantly connected to the NFS server during the experimental sequences and occasionally caused overloads on the server.

For the above reasons, we have abandoned the NFS in favor of applying the FTP-based method on it. As FTP clients establish a network session only during file transfer and are disconnected when the transfer is complete, the server-side load efficiency has been much improved. FTP

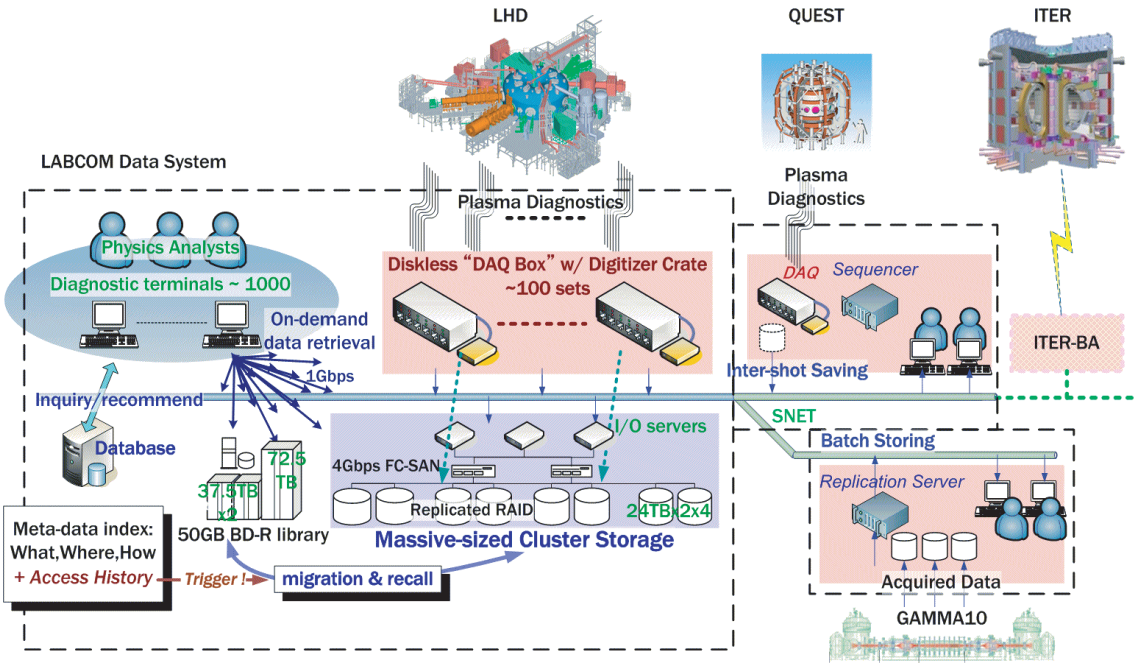

Fig. 4 LABCOM/X multi-site data system based on SNET: The database at the left end performs the *facilitator* function, and user diagnostic terminals are the data clients. DAQ and I/O servers at both the LHD and SNET sites correspond to the distributed servers in the *facilitator* model.

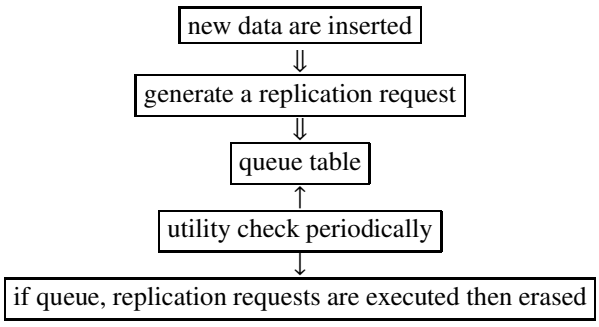

Fig. 5 Replication queuing algorithm between the database table and the utility.

also has the advantage of being easily replaced by some higher-throughput parallel-session FTP, such as GridFTP [15], both for future extension and far distant migration.

5. Conclusion and Discussion

The LABCOM storage cluster has proven its effectiveness for use in multiple fusion experiments. It is on-demand extensible with an FC-SAN and multiple disk volumes.

As mentioned in Sec. 4, parallel filesystems such as Sun Lustre FS or IBM GPFS are beneficial to I/O-critical applications such as high-performance computing (HPC) clusters, digital media production or archiving centers, and broadcasting stations. Red Hat GFS has no throughput improvement by means of parallel I/O; however, it is easier to have symmetrical replication volumes for data redundancy instead. Taking the preprogrammed sequential operation and data granularity of fusion experiments into account, the distributed file-locking mechanism of GFS2 is simple but suitable for our replicated storage.

A newly developed replication utility also provides good flexibility for sustaining data protection. The new FTP-based migration scheme has proven its reliability without any trouble during one year of operation at LHD and remote sites. We will further advance the Fusion Virtual Laboratory in Japan to demonstrate next-generation and forthcoming ITER and ITER-BA multi-site experiments.

Acknowledgments

This work is performed with the support and under the auspices of the NIFS Collaborative Research Program: NIFS08ULHH503, KUTR031, KUTR035 and NIFS07-KUGM021. It is also supported by the Cyber Science Infrastructure (CSI) development project of the National Institute of Informatics (NII).

[1] J. How and V. Schmidt, Fusion Eng. Des. **60**, 449 (2002).
[2] J. Vega *et al.*, Rev. Sci. Instrum. **74**, 1773 (2003).
[3] NII, *SINET3* http://www.sinet.ad.jp/ (2007).
[4] *SNET* http://snet.nifs.ac.jp/ (2006) [in Japanese].
[5] M. Emoto, T. Yamamoto, S. Komada and Y. Nagayama, Fusion Eng. Des. **81**, 2051 (2006).
[6] K. Tsuda, Y. Nagayama, T. Yamamoto, R. Horiuchi, S. Ishiguro and S. Takami, Fusion Eng. Des. **83**, 471 (2008).
[7] *All-Japan ST Research Program* http://www.nifs.ac.jp/kenkyo/icr/st.html (2008) [in Japanese].

[8] H. Nakanishi, M. Ohsuna, M. Kojima, S. Imazu, M. Nonomura, Y. Nagayama and K. Kawahata, Fusion Eng. Des. **82**, 1203 (2007).
[9] M. Ohsuna, H. Nakanishi, S. Imazu, M. Kojima, M. Nonomura, M. Emoto, Y. Nagayama and H. Okumura, Fusion Eng. Des. **81**, 1753 (2006).
[10] O. Motojima *et al.*, Phys. Plasmas **6**, 1843 (1999).
[11] H. Nakanishi, M. Ohsuna, M. Kojima, S. Imazu, M. Nonomura, M. Emoto, H. Okumura, Y. Nagayama, K. Kawahata and LHD exp. group, J. Plasma Fusion Res. **82**, 171 (2006) [in Japanese].
[12] H. Nakanishi, M. Kojima and S. Hidekuma, Fusion Eng. Des. **43**, 293 (1999).
[13] K. Kawagome *et al.*, *Distributed Objects Computing* (Kyoritsu Publishing, Tokyo, Japan, 1999) [in Japanese].
[14] Red Hat, Inc., *Global File System* http://www.redhat.com/docs/manuals/enterprise/RHEL-5-manual/Global_File_System/ (2007).
[15] *GridFTP* http://www.globus.org/toolkit/data/gridftp/ (2008).