Proceedings of ITC18, 2008

Clustered Data Storage for Multi-site Fusion Experiments

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

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

LABCOM data acquisition and management system has already provided full functions in both the local and remote participations for the LHD experiments. This study newly added the function to deal with experimental raw data not only in one experimental device but also from multiple distant sites. Its original distributed structure has enabled the multi-site modification with a minimum change mainly within the data location indexing database for clustered storage. However, the access permission and restriction for each site's data and users should be strictly implemented. The system has started its operation since 2008 under bilateral collaboration between LHD, QUEST, and GAMMA10 experiments, aiming to organize "Fusion Virtual Laboratory" in Japan.

Keywords: LABCOM, clustered storage, SINET3, LHD, QUEST, GAMMA10, Fusion Virtual Laboratory

1 Introduction

Remote participation technology is one of the most important fundamentals for modern fusion experiments [1, 2]. It is recently based on over 10 Gbps information highways, in which many Giga-bytes or sometimes Tera-bytes experimental data are shared by distributed collaborators.

On the other hand, the amount of experimental data which continuously keep growing (Fig. 1) often causes the operational staffs too heavy management burden. Such the increasing costs of data management will possibly be optimized by an intensive administration of the data storage system through the ultra-wideband networks. The Internet Data Center (IDC), which provides centralized monitoring and control for data resources, is a typical example to streamline the data management in commercial fields. This solution would be also required in physics research experiments.

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

SNET has been hosted by NIFS since 2001 fiscal year, at first for the LHD remote participation activities [5, 6]. From 2005 FY, the bilateral collaboration programs between NIFS and research centers of other universities has additionally come into operation. The most typical example of it is the All-Japan spherical tokamak (ST) research program [7] where 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 to be able to deal with multiple experiments and their data simultaneously. In the following sections, required specifications and applied implements are described with their effectiveness.

2 Objective: Access controls for multiple sites

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

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

When sharing the clustered storage volumes among different experimental sites, a clear distinction should be given to the access permission and restriction for data and users of each site. These access controls shall be implemented on the indexing database by adding a new "site" key to prior "diagnostic (data) name" and "shot number"

author's e-mail: nakanishi.hideya@lhd.nifs.ac.jp

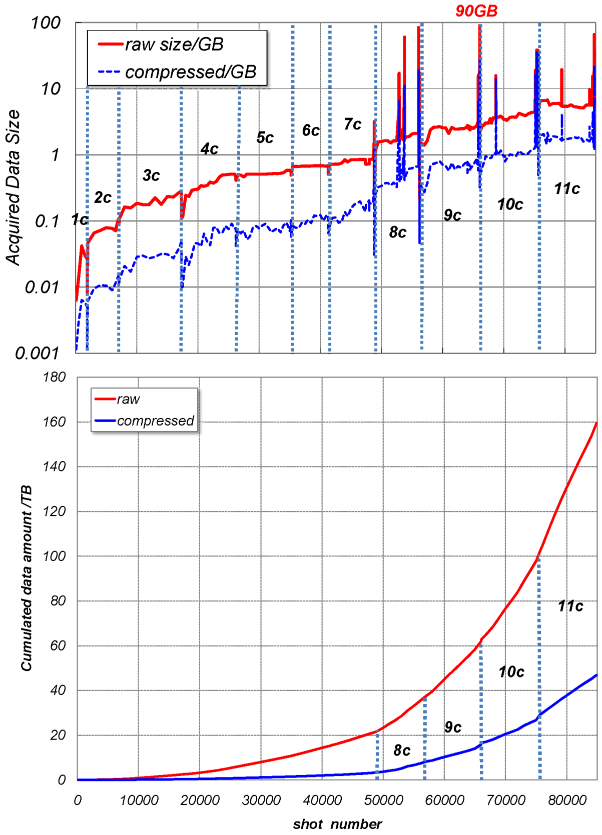

Fig. 1 Data growth in LHD, per-shot size (top) and cumulative amount (bottom): 1c~11c mean annual experimental campaigns of LHD. 90 GB/shot is the world record of acquired raw data in one plasma experimental discharge. All the acquired data are kept online to be accessible for every collaborators.

keys. The "site" should control both the diagnostic data and the user groups. Table 1 shows the essential part of this upgrade.

3 New LABCOM/X data acquisition and management system

R&D for LABCOM data acquisition and management system has been started since 1995, aiming for constructing a new plasma diagnostic data system for Large Helical Device (LHD) experiment in NIFS. As the first plasma was established in March 1998 [10], it has experienced ten years' annual campaigns until now.

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

The LABCOM system has originally a distributed structure in which data acquisition and storage elements are completely separated on fast network [12]. When wide-area networks (WAN) could be equivalent to local one (LAN) on its throughput, there is no logical difference between them. The multi-site modification was, therefore, realized with a minimum change mainly on the facilitator database which informs the data locations shown in Fig. 2.

Table 1 Related tables in "facilitator" database; components of main "shot" table (left) and contents of new "site" table (right): Bold-typed **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

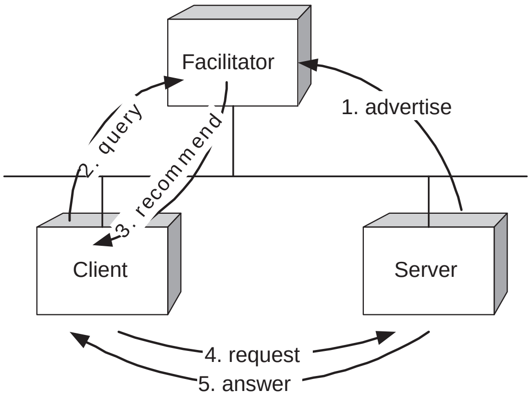

Fig. 2 Recommend-type facilitator model: The facilitator never mediate the requests but only recommend the appropriate server to send them [13]. It is suitable for the distributed data store and retrieval system which must transfer many binary large objects (BLOBs) without any bottlenecks.

The access restriction between multi-sites' data and user groups has been implemented by a combination of database's user account corresponding to site's user group and its access permission to registered IP addresses. It means that every stored data belong to their own site, and also the data retrieval computers are independently registered for each site.

The main "shot" table shown in Table 1 contains more than 14 million entries of the experimental data. By means of database's embedded acceleration of key indexing, however, a query search be answered in about 0.14 second.

The multiple sites' data handling system has started the operation since September 2008, under bilateral collaborations between LHD in NIFS, QUEST in RIAM, Kyushu Univ., and GAMMA10 in PRC, Univ. of Tsukuba. Due to such the topological evolution, we newly name the data system as "LABCOM/X". Fig. 3 show the schematic

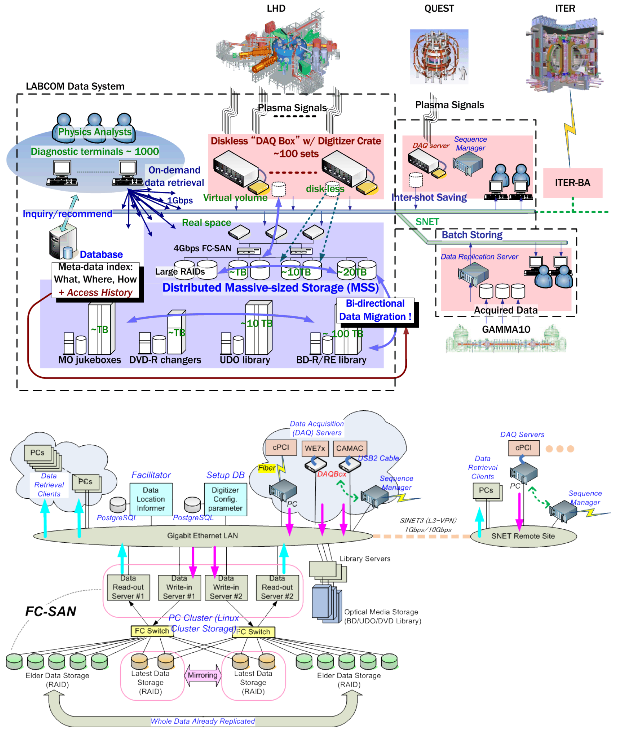

Fig. 3 Multi-site data acquisition and management system based on SNET; physical view (top) and logical scheme (bottom).

views.

4 GFS2 storage cluster and data replication

For a multi-sites' data repository, it will be quite essential that plural I/O servers should work redundantly and even in load-balancing. The cluster filesystem provides the synchronization mechanism of the content data among them. We use Red Hat Global File System (GFS) [14] and its version 2 (GFS2) afterward, whose I/O performance is almost the same as ordinary local ones like xfs or ext3 (Table 2).

Generally, cluster filesystems such as Sun's Lustre File System or IBM's General Parallel Filesystem (GPFS) provides better performance in writing huge data volumes by means of split I/O into many storage nodes. To keep the consistency among their distributed chunks, it usually needs at least one metadata server or service process. On the other hand, GFS never split a file into many or distributed chunks. It only provides a distributed file locking mechanism to synchronize the file appearance among the cluster node computers. So, it is also possible for us to use GFS volume as a local filesystem without any metadata server. This feature is quite advantageous when some GFS volumes are filled and changed to be read-only ones, as shown in Fig. 3.

Another possibility to make data replication was hardware or software mirroring. When using a number of huge disk arrays, however, it would be more disadvantageous due to the extremely long rebuilding time in recovery when inconsistency disorders happened.

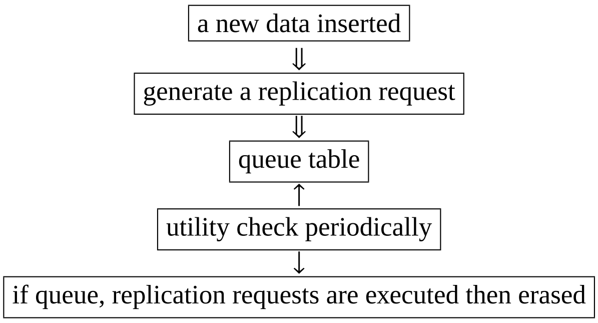
a new data inserted
⇓
generate a replication request
⇓
queue table
↑
utility check periodically
↓
if queue, replication requests are executed then erased

Fig. 4 Replication queuing algorithm between database table and the utility.

After above mentioned discussions, we have made a specific utility to replicate newly appeared data files which runs in cooperation with the facilitator database. Applied replication scheme is a simple combination of request queuing and their cyclic batch execution, as shown in Fig.4. It never check the equality between the source and destination volumes, only making the incremental copies of newly appeared files. Such the loosely tied data mirroring mechanism is rather preferable for flexible storage operations.

We have also changed the data migration scheme. Between the distant data acquisition (DAQ) servers and storage ones, we previously adopted the network filesystem (NFS) to share the cluster volume on local-area network, in other words, within a single LHD site. However, it could be less reliable for the distant data sharing because NFS was designed basically to be used on LAN. Moreover, over seventy NFS clients were constantly connected to the NFS server during the experimental sequences and occasionally caused overloads on server.

For the above reasons, we have abandoned NFS and applied the ftp-based method for it. As its client only establishes a network session during the file transfer and will be disconnected when completed, the server-side load efficiency has been much improved. It also has an advantage to be easily replaced by some higher-throughput parallel-session ftp, such as GridFTP [15], both for the future extension and far distant migration.

Table 2 Throughput difference between local (ext3, xfs) and cluster (gfs2) filesystems: These are the 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

5 Conclusion and remarks for future

The LABCOM storage cluster has proved the effectiveness for the use of multiple fusion experiments. It is on-demand extendable with FibreChannel storage area network (FC-SAN) and multiple disk volumes.

Considering about the preprogrammed sequential operation and the data granularity of fusion experiments, the simple network locking algorithm provided by GFS2 seems more preferable for us to fast split I/O cluster filesystems having a metadata server. Newly developed replication utility also provides a good flexibility for sustaining the data protection on it. The ftp-based new migration scheme has proved its reliability having no trouble during one year experiment on LHD.

We will further advance the "Fusion Virtual Laboratory" in Japan to demonstrate the next-generation and coming 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 Cyber Science Infrastructure (CSI) development project of the National Institute of Informatics (NII).

References

[1] J. How and V. Schmidt, Fusion Eng. Design **60**, pp. 449–457 (2002).

[2] J. Vega *et al.*, Rev. Sci. Instrum. **74**, pp. 1773–1777 (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. Design **81**, pp. 2051–2055 (2006).

[6] K. Tsuda, Y. Nagayama, T. Yamamoto, R. Horiuchi, S. Ishiguro and S. Takami, Fusion Eng. Design **83**, pp. 471–475 (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. Design **82**, pp. 1203–1209 (2007).

[9] M. Ohsuna, H. Nakanishi, S. Imazu, M. Kojima, M. Nonomura, M. Emoto, Y. Nagayama and H. Okumura, Fusion Eng. Design **81**, pp. 1753–1757 (2006).

[10] O. Motojima *et al.*, Phys. Plasmas **6**, pp. 1843–1850 (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**, pp. 171–177 (2006) [in Japanese].

[12] H. Nakanishi, M. Kojima and S. Hidekuma, Fusion Eng. Design **43**, pp. 293–300 (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).