Multi-Layer Distributed Storage of LHD Plasma Diagnostic Database

NAKANISHI Hideya, KOJIMA Mamoru, OHSUNA Masaki, NONOMURA Miki, IMAZU Setsuo, and NAGAYAMA Yoshio
National Institute for Fusion Science, Oroshi-cho 322-6, Toki 509-5292, Japan
e-mail: nakanisi@nifs.ac.jp

Abstract

At the end of LHD experimental campaign in 2003, the amount of whole plasma diagnostics raw data had reached 3.16 GB in a long-pulse experiment. This is a new world record in fusion plasma experiments, far beyond the previous value of 1.5 GB/shot. The total size of the LHD diagnostic data is about 21.6 TB for the whole six years of experiments, and it continues to grow at an increasing rate. The LHD diagnostic database and storage system, *i.e.* the LABCOM system, has a completely distributed architecture to be sufficiently flexible and easily expandable to maintain integrity of the total amount of data. It has three categories of the storage layer: OODBMS volumes in data acquisition servers, RAID servers, and mass storage systems, such as MO jukeboxes and DVD-R changers. These are equally accessible through the network. By data migration between them, they can be considered a virtual OODB extension area. Their data contents have been listed in a "facilitator" PostgreSQL RDBMS, which now contains about 6.2 million entries, and informs the optimized priority to clients requesting data. Using the "glib" compression for all of the binary data and applying the three-tier application model for the OODB data transfer/retrieval, an optimized OODB read-out rate of 1.7 MB/s and effective client access speed of 3∼25 MB/s have been achieved. As a result, the LABCOM data system has succeeded in combination of the use of RDBMS, OODBMS, RAID, and MSS to enable a virtual and always expandable storage volume, simultaneously with rapid data access.

Keywords:
data acquisition, mass storage system, OODBMS, RAID, DVD changer, LHD, LABCOM

1. Introduction

In the 7th LHD campaign of 2003∼2004, quasi-steady-state plasma experiments whose longest duration was ∼756 s were performed successfully. The primary LHD data acquisition system, named the LABCOM system, then established a new world record for acquisition data amount of 3.16 GB in one discharge. This was far beyond the previous

record of about 1.5 GB/shot by JET [1]. Even in the short-pulse operation, which usually repeats about 150 shots per day, the whole acquisition data amount has been over 1 GB/shot. Figure 1 shows the growth curve by shot number.

On the other hand, the LHD diagnostics have over 40 kinds of plasma measurements with up to 2000 signal channels in total. A considerable number of these require fast data acquisition even in steady-state experiments. The greater part of the new world record was acquired by such fast sampling real-time digitizers, which provide quite different capabilities from the conventional CAMAC digitizers. To realize fast real-time data acquisition, we have performed R&D for new digitizer systems [2]. In the 7th campaign, we have begun to operate the NI PXI/CompactPCI and Yokogawa WE7000 digitizers, which can achieve 80 MB/s and 2.2 MB/s continuous data acquisition, respectively. A PXI frame grabber can also deal with 16 MB/s video stream for measurements using high-resolution CCD cameras. As their cost-performance ratio is quite reasonable in comparison to CAMAC, their utilization is becoming widespread in LHD. In the 7th and 8th campaigns, we had ten and four new WE7000 and PXI installations, respectively, with only one new CAMAC installation.

This technological shift to new digitizers has brought about an explosion in output data quantity. The intense increase in amount of diagnostic data inevitably leads to larger storage volume requirements every year. As shown in Fig. 2, the total size of the LHD diagnostic data for the previous six years is about 21.6 TB, and it continues to grow at an increasing rate. Therefore, the data storage system must be sufficiently flexible and easily expandable to allow maintenance of the whole data integrity. However, large capacity and rapid read/write performance are conflicting properties in a mass storage system. For enormous databases, it is quite difficult to maintain good responsiveness without highly sophisticated tuning and optimization.

Here, we describe the realization of the LABCOM data system and discuss its achieved performance.

2. Data Acquisition and Database

The database and storage system for LHD raw data has three categories of storage layers. The first is the 50∼250 GB local disk arrays for each data acquisition computer. Acquired raw data will be compressed by "zlib" and then stored in the virtual volume, which is provided by the local object-oriented DBMS. OODBMS was adopted because of the seamless connection between the volatile data objects in C++ applications and their persistent instances in OODB space [3].

The parts of the OODB client/server system, however, intrinsically share so much

information with each other that their communications often require excessive network bandwidth. Therefore, we first adopted "glib" compression of all the binary data to improve the apparent read/write speed. The three-tier application model for the OODB data was also applied for the transfer/retrieval programs. Thus, an optimized OODB readout rate of 1.7 MB/s and effective client access speed of 3∼25 MB/s have been achieved.

Even though the OODB virtual space can contain many binary large objects (BLOB) inside, DBMS usually has less functionality to directly manage TB∼PB huge virtual volumes. However, media library equipment, such as magnetic tape (MT) libraries or DVD changers, are often used for mass storage systems. In a similar way, hierarchical storage management (HSM) systems will be used, which will enable a huge virtual file system.

HSM is a well-established method, which provides automatic stage-in/stage-out file migration between a definite logical file system and its front-end cache area. When OODBMS volumes are held in files, however, their sizes can easily reach as large as 4 GB. Such large file operations will cause longer time lags for any HSM to complete the stage-in/-out processes. On the other hand, the granularity of plasma diagnostic data is usually kS∼MS/channel, which is much smaller than popular storage media, such as 200 GB MT cartridges, and 4.7 GB DVD-R. Therefore, the data access patterns will be almost random. Based on examination of HSM with the MT library, we concluded that randomly accessible media, such as MO and DVD, are more appropriate for fusion experimental data [4].

Due to this mismatch between OODB and HSM, we have developed a new OODB volume extension mechanism by translating their BLOBs into files and directories of the file system as explained in the next section.

3. Multi-Layer Mass Storage System

As the plasma diagnostics raw data usually consist of multiple channels of lengthy time series signals, its occupied volume in data storage becomes much larger than usual relational databases in other fields, even if they have similar numbers of record entries.

The number of LHD data entries can be estimated from the total shot number multiplied by the diagnostic varieties and the backup replications. At present, the system contains about 6.2 million entries, and the primary part of 3.4 million entries is information for distributed data locations. To promptly return a query result, a fast index search of the relational database management system (RDBMS) will usually be applied. Millions of record entries occupy a few GB of RDBMS volume. Plasma raw data, therefore, should

be stored independently outside the database, to prevent any deceleration of its index searching.

The LABCOM data storage system, therefore, has applied a completely distributed architecture based on fast network. It realizes data redundancy, fail-safe capability, and even load-balancing function by means of replication pairs of every storage server, which are equally accessible through the network. All of their contents are listed in a "facilitator" PostgreSQL RDBMS, and informed to any data retrieval clients on demand. Figure 3 shows a schematic view of this system.

Storage servers in the latter two layers consist of files and directories in the file system, not in the OODBMS volume. To enable seamless extension from the three-tier model of OODB, the same application server program runs in all of them, and accesses the file system instead for data retrieval. In addition, by means of the data migration mechanism from OODB to file system, they can be logically considered as an OODB extension area.

The second layer consists of multiple sets of huge redundant disk array (RAID) servers, to provide fast data retrieval to clients. The third has a few sets of so-called mass storage systems (MSS). For the first four campaigns, three sets of 1.2 TB magneto-optical (MO) disk jukeboxes were applied. Subsequently, 1.8 TB or 3.3 TB DVD-R changers were adopted until 2004. Figure 4 shows the storage structure. The numbers of running servers in each layer are 40, 4, and 5, respectively.

Table 1 shows the cost comparison between the two kinds of third layer storage equipment. The recording media only account for a small part of the total storage cost, and the most expensive devices are libraries or changers with virtual volume management software. Even though the prices of HDDs and their arrays (RAID) always decrease rapidly, this hardly affects the per-byte cost as long as we continue to use or reinforce the same equipment. With application of next-generation DVD storage media, such as Blu-ray Disc or HD DVD, the per-byte cost may again decrease markedly.

4. Results and Discussion

Data retrieval speed to the clients is the most important property to evaluate a database and storage system. Figure 5 shows the speed differences between each kind of storage server. Note that the multi-channel diagnostic data were stored in one file per shot in 2nd and 3rd layer storage. As designed, the 2nd layer RAIDs have been shown to consistently provide a comfortable speed.

From OODBMS, the apparent speed of 31.5 MB raw data retrieval was 2.1 MB/s, while the real I/O rate was 0.8 MB/s. The acceleration ratio was almost threefold, which was achieved by the data compression ratio. The difference between the 1st and 2nd retrieval

can be considered due to the internal cache mechanism. In general, internal OODBMS operations involve heavy address translating calculations between persistent object images and volatile memory instances. Therefore, the data retrieval speed would be considerably improved with application of more powerful PCs. Roughly 2- or 3-fold increases in speed can be obtained easily by using ∼GHz Pentium 4 PCs, where the bottlenecks of data retrieval may exist just in the transaction overheads of both HDD readout and TCP/IP telecommunications.

The preprocessing delays in 3rd layer storage can be easily understood as the robot moving time to pick up and make the MO or DVD media ready. They cannot respond quickly in random data access, whereas they could provide vast online archive spaces instead.

This also provides insight into how it should be possible to optimize the facilitator's recommendation priority for data retrieval requests; new or often referred data must exist in RAID servers as soon or for as long as possible, while aged data, which will be referred to less, can be stored only in the 3rd layer. Here, the time to search indexes in the facilitator RDBMS can always be negligible (less than 1 s) as compared to the whole elapsed time.

We conclude that the LABCOM database and storage system has succeeded in combination of RDBMS, OODBMS, RAID, and MSS to realize a virtual and always expandable storage volume. It simultaneously enables rapid data retrieval with some optimization and acceleration mechanisms.

References

[1] J. W. Farthing, Proc. 4th IAEA TM of Control, Data Acquisition and Remote Participation for Fusion Research, San Diego, 21-23 July 2003.

[2] H. Nakanishi *et al.*, Fusion Eng. Design **56-57** 1011 (2001).

[3] H. Nakanishi *et al.*, Fusion Eng. Design **48**, 135 (2000).

[4] H. Nakanishi, PhD Thesis, The Graduate University for Advanced Studies, Hayama, Japan, 2003.

Table 1: Cost comparison of LHD mass storage systems. The 1st generation MO jukebox is about 20 times more expensive than the 2nd generation DVD changers. Prices include recording media and management softwares.

Equipment	Media	Unit Price	Cost (/JPY)
HP SureStore 1200ex	4.8GB MO	20 M JPY	17.5 M/TB
Pioneer DRM-7000	4.7GB DVD-R	3 M JPY	0.95 M/TB

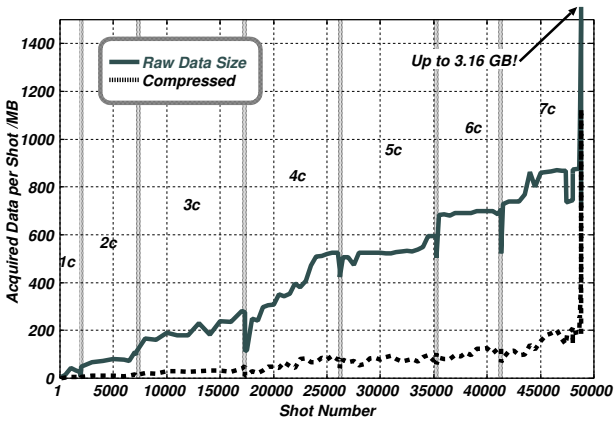

Figure 1: By-shot data growth in LABCOM data acquisition system.

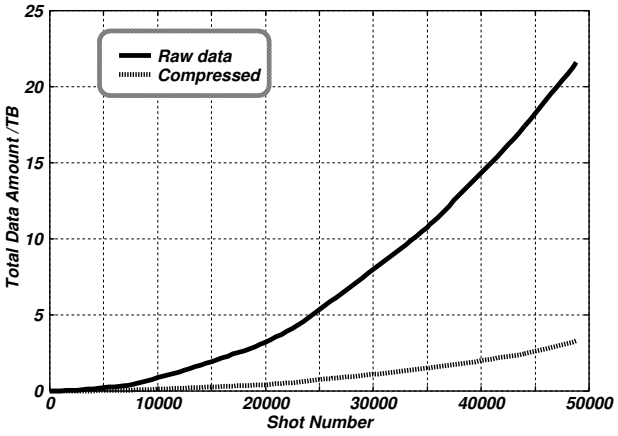

Figure 2: By-shot data growth in LABCOM data archives.

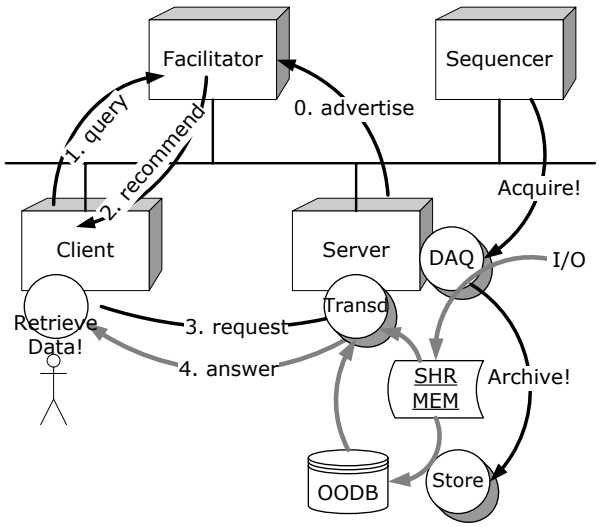

Figure 3: Cooperation between data servers and the facilitator. The data clients never refer to the OODB directly, but their requests are sent to and answered from the application server "Transd", which can access the file system instead when running on the 2nd or 3rd layer storage servers.

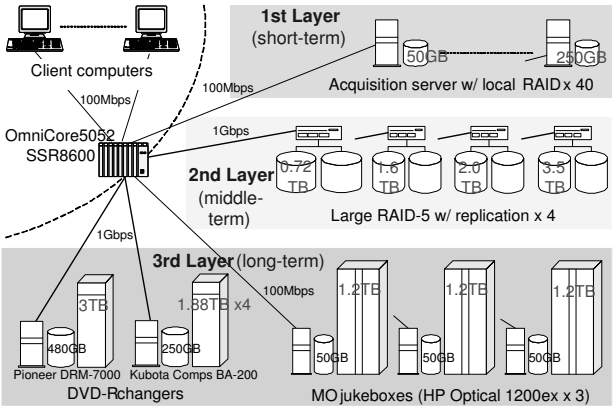

Figure 4: Multi-layer structure of LABCOM data storage.

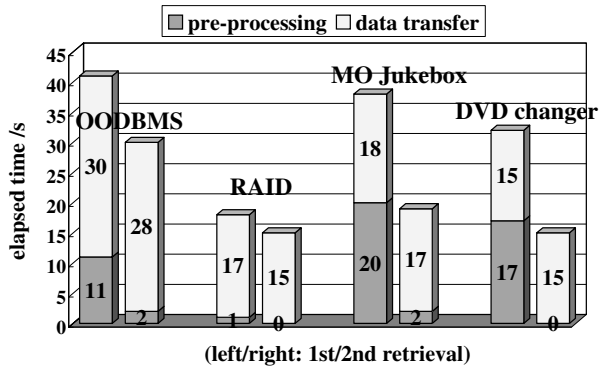

Figure 5: Data retrieval speeds from different kinds of storage. Elapsed times are for retrieving the same 12 MB of compressed data (raw size, 31.5 MB) of 126-channel Hα measurement. The client PC has dual 450 MHz Pentium-III processors with 512 MB memory and 100 Mbps Fast Ethernet port, while OODBMS servers run on dual 200 MHz Pentium-Pro machines.