Adaptive Data Migration Scheme with Facilitator Database and Multi-Tier Distributed Storage in LHD

NAKANISHI Hideya , OHSUNA Masaki, KOJIMA Mamoru, IMAZU Setsuo [b] , NONOMURA Miki, WATANABE Kenji [c] , MORIYA Masayoshi [c] , NAGAYAMA Yoshio, and KAWAHATA Kazuo

National Institute for Fusion Science, 322-6 Oroshi-cho, Toki 509-5292, Japan
[b] *Pretech Corp., 1-19-13 Kanayama-cho, Atsuta-ku, Nagoya 456-0002, Japan*
[c] *N.S.M. Inc., Softpia Japan WS24-305, 6-52-18 Imajuku, Ogaki, Gifu 503-0807, Japan*

Abstract

Recent "data explosion" induces the demand for high flexibility of storage extension and data migration. The data amount of LHD plasma diagnostics has grown 4.6 times bigger than that of three years before. Frequent migration or replication between plenty of distributed storage becomes mandatory, and thus increases the human operational costs. To reduce them computationally, a new adaptive migration scheme has been developed on LHD's multi-tier distributed storage. So-called the HSM (Hierarchical Storage Management) software usually adopts a low-level cache mechanism or simple watermarks for triggering the data stage-in and out between two storage devices. However, the new scheme can deal with a number of distributed storage by the facilitator database that manages the whole data locations with their access histories and retrieval priorities. Not only the inter-tier migration but also the intra-tier replication and moving are even manageable so that it can be a big help in extending or replacing storage equipment. The access history of each data object is also utilized to optimize the volume size of fast and costly RAID, in addition to a normal cache effect for frequently retrieved data. The new scheme has been verified its effectiveness so that LHD multi-tier distributed storage and other next-generation experiments can obtain such the flexible expandability.

Key words: LABCOM/X, LHD, HSM, multi-tier distributed storage, access history, intelligent migration

1. Introduction

Recent "data explosion" demands higher potential of distributed storage and mass data migration among them. Acquired data amount for each LHD plasma discharge has grown 4.6 times bigger than that of three years before (Fig. 1). In such circumstances, data migration or replication between a large number of distributed storage devices becomes much frequent, which increases the human operational and maintenance costs.

LHD already had about seventy data acquisitions (DAQ) in the tenth campaign [1]. Increased number of parallel DAQ units also makes maintenance burden heavier. Therefore, "more distributed acquisition and more centralised operations" become indispensable to cope with high-efficiency I/O throughputs and much enlarged data volume.

To reduce the related human burden by means of computer automation, this study has tried to develop a new intelligent data migration scheme on the LHD multi-tier distributed storage. As many storage devices are already used in the LHD data system, not only the inter-tier migration but also the intra-tier replication or moving should be managed by this new scheme so that it can help us in extending or replacing storage equipment.

This paper describes the detailed investigation for

Email address: nakanishi.hideya@lhd.nifs.ac.jp (NAKANISHI Hideya).

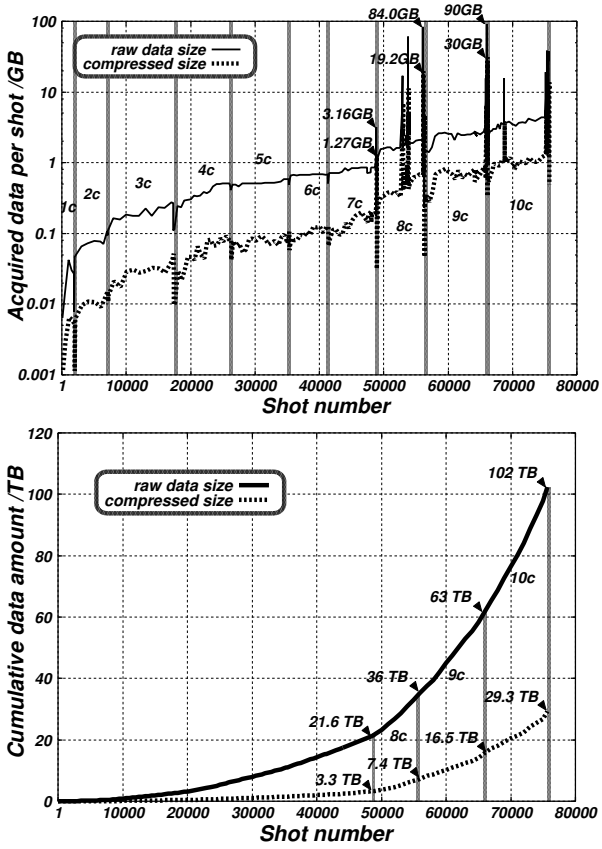

Fig. 1. Growth of shot-by-shot data acquired by LABCOM system (top) and its cumulative amount (bottom).: In the last tenth campaign, ordinary short-pulse experiments produced at maximum 4.67 GB/shot raw data, having about 170 shots everyday.

the requested specifications toward the new intelligent migration scheme first, and secondly its schematic advantages are compared to the usual hierarchical storage management (HSM) mechanisms. Evaluation for the potential of this new approach and further discussions will be given at the end.

2. Requirements for Intelligent Data Management

The LHD data acquisition and management system, namely *LABCOM system*, originally had a three-tier storage structure [2,3]. The first tier consisted of local hard-drives of DAQ computers, and the second tier was a cluster of paired RAID devices. In the bottom tier there were the massively-sized storage (MSS) which can contain many recordable media with the picking robotics inside.

As LHD has already experienced ten annual campaigns, we had to upgrade not only the capacities of storage devices but also their technologies one after another. At the end of tenth campaign, we have seven RAID pairs in the second-tier and three MO jukeboxes, four DVD-R/+R changers, one UDO library, and one BD-R/-RE library in backend (Fig. 2). Every optical recording media contain independent filesystems one by one in a standard "Universal Disk Format (UDF)" [4], which is commonly used in various optical discs. They were introduced because of their largest capacity among available products of those days. See the comparison in Table 1.

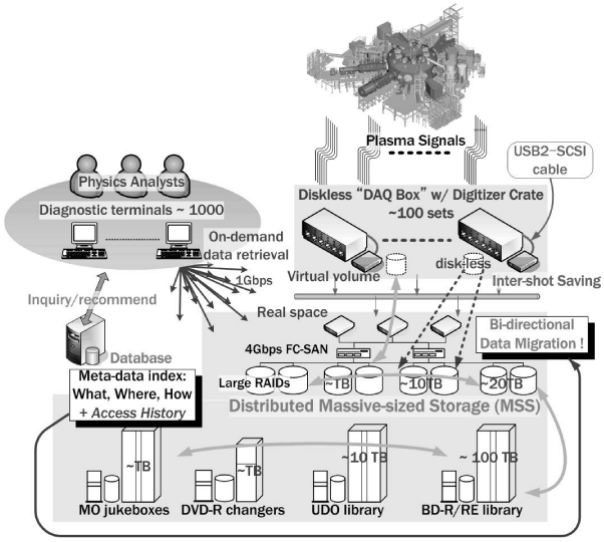

Fig. 2. New structure of LABCOM/X DAQ and multi-tier storage system: As the 1st-tier "DAQ Box" keeps raw data only on its volatile memory, they must be saved to the 2nd-tier RAID just after acquisition for each shot will be completed.

For the past ten years, several replacement of the obsolete equipment had to be done to obtain faster throughput and higher capacity. New automation scheme, therefore, should be helpful for human operators in extending or replacing them. In other words, it must manage not only the inter-tier data migration but also the intra-tier replication or moving as automatically as possible. This innovative project is named as *"LABCOM/X"* because the program will be the tenth revision corresponding to the tenth LHD campaign.

Table 1
History and comparison of optical storage media used in LHD. ISO-130 is 135×153×11 (mm). DVD±R cost is for double-sided one.

	MO	DVD±R	UDO	BD-R
operation	1997∼2005	2002∼	2005∼	2006∼
capacity (GB)	4.8	4.7 / 9.4	15 / 30	50
write (MB/s)	2.3 / 4.6	1.35×16	4 / 8	4.5×4
size (mm)	ISO-130	ϕ120×1.2	ISO-130	ϕ120×1.2
price (US$)	32.2	12.5	52.3	20.6

2.1. *Volatile First-Tier Storage*

In our previous study [5], we have successfully evaluated the diskless DAQ computer which is free from hard-drive disorders. Not having the local data accumulation on DAQ computers, we can omit the batch-processing migration from DAQ computers to RAIDs (Fig. 2), whose task ran every night after daily experiments ended.

Instead, acquired raw data will be temporarily written on so-called *RAMdisk* filesystem. As they are volatile entities on the limited size of RAMdisk, shot-by-shot migration to the RAID storage becomes mandatory. Older shot data should be deleted before the next shot, therefore, its migration function should be synchronously activated by the acquisition task.

Because our migration utility "MigrateOS" was a nightly scheduled batch process which pulled the data from DAQ PCs, we will need another utility for the new purpose.

2.2. *Multi-Purpose Migration Utility: "MigrateFS/X"*

New migration utility should push the archived file into the second tier storage between every experimental shots. Its behavior is synchronized by the completion message from the acquisition task. After successfully flushing the archived file through SMB/CIFS or NFS network filesystem, it will register new entry on the index database for the new persistent entity and then make the volatile one void.

However, traditional asynchronous migration is still indispensable because the real-time DAQ often has no time to execute the data compression in a steady-state experiment. New utility should do it with embedded zlib and JPEG-LS algorithms after the sequences end [6].

In addition to above mentioned inter-tier migration, the new utility should also manage the intra-tier data moving and replication. Especially in large-scale distributed storage, the alternation of storage devices would frequently take place (Fig. 3). In this case, it is only needed to replicate the old data tree into the new device by using this utility. It can synchronously update the data existence information in the index database by issuing an SQL transaction.

A widely adaptive intelligent management will be realized if a new computational scheme can make a decision to trigger migration. For the remote control of distributed storage, the related function should be awaken through a TCP socket communication. In the next sub-section, the details will be described.

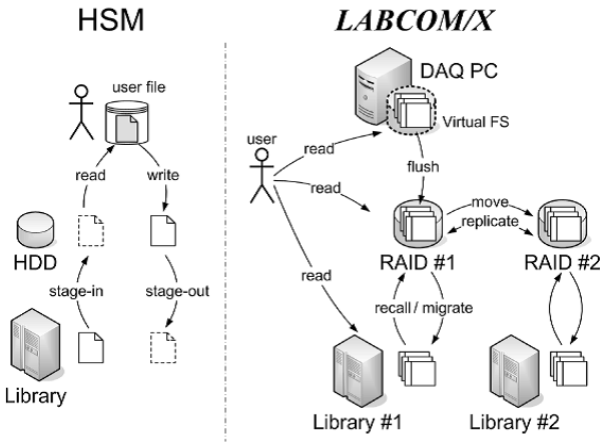

Fig. 3. Schematic comparison between usual HSM and LABCOM/X multi-tier migration: In HSM, the data entity is a unique existence in principle, however, LABCOM/X enables multiple replications to be scattered in distributed storage. Thus it can provide the intra-tier data moving and even disaster recovery functionalities.

2.3. *Intelligent Data Recall Using Access History*

A new computational scheme is necessary to decide when, what, and how to start migration. As all the data entities are registered in the index database, the most important function is to perceive whether each data object satisfies the necessary conditions to be migrated or not.

The judgement rules to trigger the migration can be considered as follows:

(i) Volatile entities should be migrated into the persistent storage at the first opportunity.

(ii) Data archives in the 2nd-tier storage will be migrated to the 3rd-tier incrementally after the experimental sequence completely stopped. It is almost the same as nightly scheduled backup.

(iii) As the 3rd-tier storage use removable recording media, such as 50 GB Blu-ray Disc, archived data files are reordered so that every data files having the same experimental number should be contained together into one media.

(iv) Archived objects in the 3rd-tier libraries would be recalled to make a replica in the 2nd-tier when some conditions below are satisfied:
  (a) Accumulated request count for the data object is above the definite threshold.
  (b) Time intervals between recent several requests are less than the pre-defined value.
  (c) Averaged time interval between every past requests is less than the definite threshold.

(v) Recalled replica would be erased from the 2nd-tier when any of above (iv) conditions are no

longer satisfied.

This intelligent recall mechanizm, a combination of access history and MigrateFS/X utility, can also optimize the volume size of fast and costly RAID. Different from usual cache mechanisms, properly tuned (iv)-(b), (c) thresholds can prevent a shot survey analysis from wiping out the 2nd-tier storage. This system does not have any backup process either. It only has a media replication inside the library.

3. HSM vs. New LABCOM/X Migration System

There are some commercial software so-called HSM system [7–9]. Some adopt rather low-level cache mechanism between fast hard-drive and slower library device. The others apply simple watermarks for triggering the data stage-in and stage-out between these two devices. The HSM software is generally suitable for the read dominant environment having many users.

On the other hand, massive-sized storage system (MSS) for physics experiments is designed primarily to deal with the rushing outputs of raw diagnostic data. Our new scheme, therefore, can deal with a number of distributed storage by introducing the facilitator database that manages the whole data locations with their access histories and retrieval priorities. Figure 3 shows the difference between HSM and our new migration scheme.

Such a write dominant usage reveals the HSM disadvantages (Table 2). Lengthy backup tasks certainly wipe out the cache region and make it of no effect.

From another viewpoint of disaster recovery, it might fall into the fatal situation if the data mapping information for every recording media has been broken or lost. All the recorded media are used as a part of the huge virtual volume whose internal usage is a complete "black box". So, we have no way to read a single media if the mapping information will be lost.

In our system, all media are independently readable because they are written in standard UDF. It also simplifies to make backup media. Due to this media portability, we never need to read all the media in recovering from possible media errors.

Table 2
Functional comparison between usual HSM software and LABCOM/X new migration system: RAID capacity shows the percentage of the MSS volume.

	HSM		LABCOM/X
behavior	cache-type	watermarks	history-driven
host	unique		multiple
device connection	direct-attached		distributed
sync. trigger	always	watermarks	conditioned
sync. behavior	write-back	flushing	file copy
concurrent I/O	always (cache)	yes (flushing)	avoidable
when read	always cached	always stage-in	manageable
RAID capacity	5∼10 %	10∼20 %	any
stream write speed	$<$ MSS write		∼RAID
cached contents	unspecified		selected
media portability	no		yes (UDF)

4. Summary and Future

By the inspection described in the previous section, our new scheme have been confirmed to be the most promising solution than other existing technology to manage the virtual volume expansion like HSM. By utilizing the flexible expandability, it could be applied not only for the LHD multi-tier distributed storage but also for other biggest fusion experiments and the next-generation projects.

This result shows very well that such the innovative approach can enable us to realize ten-times bigger DAQ system having one thousand DAQs for fusion plasma diagnostics. We aim to establish the next-generation technology to advance the outputs of this LHD study.

Acknowledgements

This work is performed with the support and under the auspices of the NIFS Collaborative Research Program; NIFS06(07)ULHH503, NIFS05KCHH004, and NIFS06(07)PLHH002. It is also supported by Softpia Japan's Collaborative Research project *"Development of Image Grabbing, Recording, and Fast Retrieval System for Medical Diagnostic Database Using Distributed Storage Technology"* in 2006 and 2007.

References

[1] S. Sudo, Y. Nagayama, M. Emoto, H. Nakanishi, H. Chikaraishi, S. Imazu, C. Iwata, Y. Kogi, M. Kojima, S. Komada, S. Kubo, R. Kumazawa, A. Mase, J. Miyazawa, T. Mutoh, Y. Nakamura, M. Nonomura, M. Ohsuna, K. Saito, R. Sakamoto, T. Seki, M. Shoji, K. Tsuda, M. Yoshida, LHD Team, Control, data acquisition and remote participation for steady-state operation in LHD, Fusion Eng. Design 81 (15-17) (2006) 1713–1721.

[2] Nakanishi H., Kojima M., Ohsuna M., Nonomura M., Imazu S. and Nagayama Y., Multi-Layer Distributed Storage of LHD Plasma Diagnostic Database, J. Plasma Fusion Res. SERIES 7 (2006) 361–364.

[3] H. Nakanishi, M. Emoto, M. Kojima, M. Ohsuna, S. Komada, LABCOM group, Object-oriented data handling and oodb operation of lhd mass data acquisition system, Fusion Eng. Design 48 (1-2) (2000) 135–142.

[4] Wikipedia, Universal Disk Format (UDF), http://en.wikipedia.org/wiki/Universal_Disk_Format (2007).

[5] H. Nakanishi, M. Ohsuna, M. Kojima, S. Imazu, M. Nonomura, Y. Nagayama, K. Kawahata, Portability improvement of LABCOM data acquisition system for the next-generation fusion experiments, Fusion Eng. Design 82 (5-14) (2007) 1203–1209.
[6] M. Ohsuna, H. Nakanishi, S. Imazu, M. Kojima, M. Nonomura, M. Emoto, Y. Nagayama, H. Okumura, Unification of ultra-wideband data acquisition and real-time monitoring in LHD steady-state experiments, Fusion Eng. Design 81 (15-17) (2006) 1753–1757.
[7] Wikipedia, Hierarchical storage management (HSM), http://en.wikipedia.org/wiki/Hierarchical_Storage_Management (2007).
[8] IBM, High Performance Storage System, http://www.hpss-collaboration.org/hpss/ (2007).
[9] Quantum, AMASS, http://www.quantum.com/Products/Software/AMASS/Index.aspx (2007).