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 capability of storage distribution and dynamic migration among them. The data volume 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 data 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 priority. Not only the inter-tier migration but also the intra-tier replication or data moving is 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 mechanism for frequently retrieved data. This new scheme has been verified its effectiveness so that LHD multi-tier distributed storage and other next-generation experiments can apply such further 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 dynamic 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 the circumstances, the data migration procedures between multi-tier distributed storages need much closer attention. Operation for data migration or replication between increased number of distributed storage devices becomes much frequent, and thus increases the human operational and maintenance costs.

LHD already had about seventy data acquisitions (DAQ) in the 10th campaign [1]. Increasing number of parallel DAQ units have also made the operational and maintaining burden quite heavier. Therefore, "more distributed acquisition and more centralised operations" should be indispensable to cope with both 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 developed a new intelligent data migration scheme on the LHD

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

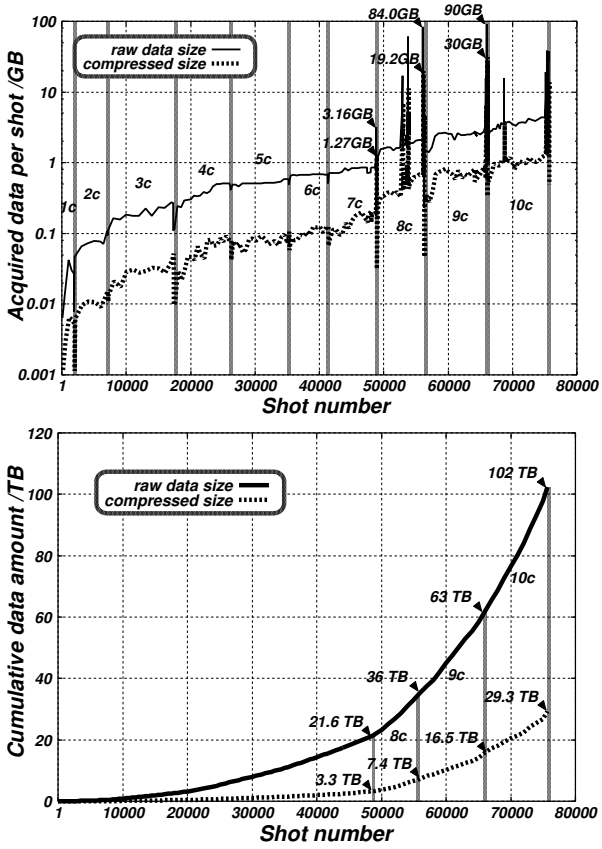

Fig. 1. Growth of shot-by-shot data size acquired by LABCOM system (top) and its cumulative amount (bottom).: In the last 10th campaign, ordinary short-pulse experiments produced at maximum 4.67 GB/shot raw data, having about 170 shots everyday.

multi-tier distributed storage. As there are already many storage devices, old and new RAIDs and libraries, in LHD data system, not only the inter-tier migration but also the intra-tier replication or data moving should be managed by this new scheme so that it can be helpful in extending or replacing storage equipment.

This paper describes the detailed investigation for the requested specifications toward the new intelligent migration scheme first, and secondly its schematic advantages comparing to the usual hierarchical storage management (HSM) mechanizms. Evaluation for this new approach's potential and further discussions will be given at last.

2. Requirements for Intelligent Data Management

Originally, the LHD data acquisition and management system, namely *LABCOM system*, had three-tier storage structure [2,3]. First tier consisted of local disks in DAQ computers, and second tier was a cluster of paired RAID devices. The backend tier was the massively-sized storage (MSS) which can contain many recordable medias with media changing robotics inside.

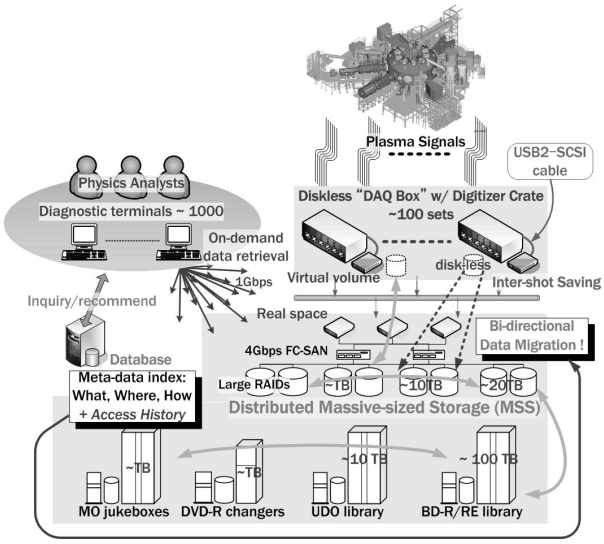

Fig. 2. New structure of LABCOM/X DAQ and multi-tier storage system: As the 1st-tier "DAQ Box" keeps raw data only on its volatile memory, they must be saved to the 2nd-tier RAID just after acquisition for each shot will be completed.

As LHD has already experienced ten annual campaigns, not only the capacities of storage devices but also their technologies themselves have changed one after another. At the end of 10th campaign, we have seven RAID pairs in the second-tier storage and three MO jukeboxes, four DVD-R/+R changers, one UDO library, and one BD-R/-RE library in backend (Fig. 2). Of course, we already experienced some replacement of obsolete devices by up-to-date ones of faster throughput and higher capacity.

New automation scheme, therefore, should be helpful for human operators in extending or replacing storage equipment. In other words, it must manage not only the inter-tier data migration but also the intra-tier replication or moving.

We have named this innovative project as *"LABCOM/X"* because it corresponds to its tenth revision toward the 10th and later LHD campaigns.

2.1. *Making the First-Tier Volatile*

In our previous study [4], we have successfully made the DAQ frontend computers diskless for reducing the opportunity to recover their disorder. By giving up data accumulation on their local hard-drive, we can reduce the batch-processing migration tasks from double action in three tiers to a single one between two tiers (Fig. 2).

It can save our human cost and time for daily post-experimental works.

On the other hand, acquired raw data are written temporarily into files and directories 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-processing task with data pulling action, we needed another utility suitable for this new purpose.

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

New migration utility should push data archived files into the second or other storage tier. Its behaviors are synchronized by the TCP socket messages sent primarily from the acquisition task. After successfully flushing the archive files through SMB/CIFS or NFS network filesystem, it will register new entry on the index database for the new data entity and then erase old ones.

Compatibility for the batch-processing, namely asynchronous, migration function in case of steady-state operation should be also sustained. Real-time DAQ often has no choice but to output non-compressed raw data, and then this utility should do that with embedded zlib and JPEG-LS algorithms in making archive files afterward [5].

In addition to above mentioned inter-tier migrations, this new utility should manage the intra-tier data moving or replication. Especially in the large-scale distributed storage, the alternation of storage device generations would occur frequently (Fig. 3). In most cases, just an evacuation after the whole data replication is necessary under the condition of data existence information in indexing database being synchronously modified. This utility proceeds it by simultaneously communicating in SQL messages. If whole the procedures should be done by manual operation, they may cause heavy human burden.

To enable the automatic management of starting migration, related control functions should be awaken remotely through the TCP socket communication. As described in the next sub-section, a widely adaptive intelligent storage management would be realized if an computational scheme to make a decision to trigger migration could send the socket message there.

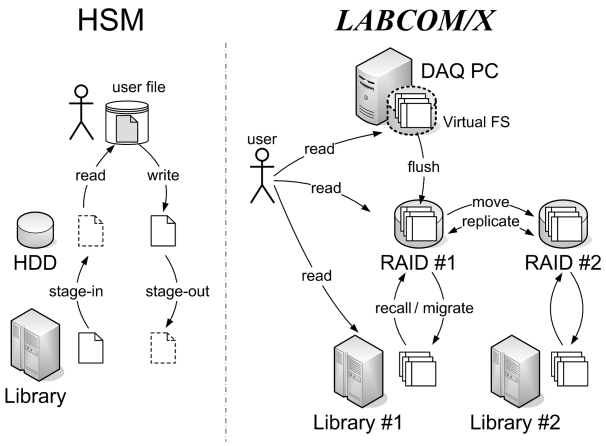

Fig. 3. Schematic comparison between usual HSM and LABCOM/X multi-tier migration systems: In HSM, the data entity is a unique existence in principle, however, LABCOM/X enables multiple replications to be scattered in distributed storage. Thus it can provide the intra-tier data moving and even disaster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

2.3. *Intelligent Data Recall based on 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 index meta-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 for each data object can be considered as follows:

(i) Volatile entities should be forced to migrate into persistent storage at the first opportunity.

(ii) Data archives in the 2nd-tier storage, i.e. RAID, 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 stored in the 3rd-tier libraries would be recalled to make a replica in the 2nd-tier when some conditions below are satisfied:
  (a) Accumulated reference count for the data object is above the pre-defined threshold.
  (b) Time interval between the most recent two references is shorter than pre-defined time.
  (c) Averaged time interval between every past references is less than the definite time.

(v) Once recalled replica should be erased from the 2nd-tier when any of above conditions are no

Table 1
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 independence	no		yes (UDF)

longer satisfied.

This intelligent recall mechanizm, a combination of access history and MigrateFS/X utility, can also optimize the necessary volume of fast and costly RAID as a normal cache mechanism for frequently referred data.

3. HSM vs. New LABCOM/X Migration System

There are some commercial software, so-called HSM system [6–8]. Some of them adopts 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. Stored data volume continues drastic growth recently. Our new scheme, therefore, can deal with a number of distributed storage by the facilitator database that manages the whole data locations with their access histories and retrieval priority. Figure 3 shows the difference between HSM and our new migration system.

Such the write dominant usage reveals the HSM disadvantages very well (Table 1). Lengthy backup always wipe out the cache area every night, and make the advantage of automatic cache mechanizm void.

From another viewpoint of disaster recovery, it could fall into the fatal situation if the data mapping information which manages every recording media by its own format has been broken or lost. All the contained recordable media are used as a part of the huge virtual volume, and thus its inside is so-called a “blackbox” so that we have no way to read a single media without mapping information.

Of course, all the media in our system are readable independently because they are written in standard UDF. It also help us to make backup media so easily. Due to this media portability, we never need to read the whole area in case of recovery from some media error.

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. It will be expected to be applied not only for the LHD multi-tier distributed storage but also for other today’s biggest fusion experiments and the next-generation projects by utilizing its further expandability.

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 plasma diagnostics. We aim to establish the next-generation technology to advance the outputs of this LHD’s 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] H. Nakanishi, M. Ohsuna, M. Kojima, S. Imazu, N. Kuroda, Y. Nagayama, K. Kawahata, Portability improvement of LABCOM data acquisition system for the next-generation fusion experiments, (to be published in Fusion Eng. Design).

[5] 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.
[6] Wikipedia, Hierarchical storage management, http://en.wikipedia.org/wiki/Hierarchical_Storage_Management (2007).
[7] IBM, High Performance Storage System, http://www.hpss-collaboration.org/hpss/ (2007).
[8] Quantum, AMASS, http://www.quantum.com/Products/Software/AMASS/Index.aspx (2007).