§28. Light-weight Data Archiving Structure by Using Distributed Filesystem 'GlusterFS' in LHD

Nakanishi, H., Ohsuna, M., Kojima, M., Nonomura, M., Imazu, S. (Pretech Corp.),
LABCOM Group,
Emoto, M., Nagayama, Y., Kawahata, K., Ida, K.

The LHD data acquisition and archiving system continued using cloud storage software named '*IznaStor/dSS*' for the past three years [1,2,3]. By using it, a new world record of acquired data amount was established from 90 GB to 328.5 GB for a 30-minute long pulse plasma sustainment in 2012 (Fig. 1).

The cloud storage provided us a lot of advantageous features for LHD's data operation. Hot plug and play nodes are very easy to be scaled out without any service stop. Actually, we hot plugged a new node during the 15th annual campaign and successfully increased the capacity on the fly. Internal auto-replication also reduces the operator's burdens for data preservation, in addition to the load balancing capability among all the member nodes.

In recovering from an accidental stop of one or several nodes, however, the member nodes occasionally showed unstable behaviors like the so-called "split-brain" or "amnesia" state. Once those phenomena happened, the recovery process took a very long time with a heavy cpu load continuing for at least several days and sometimes more than a week. In some cases, it never came to a successful end.

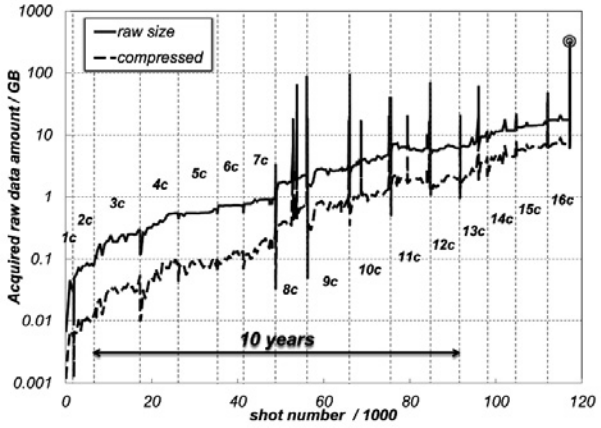

Fig. 1. Annual growth of acquired data in LHD: Compressed size shows the usage on storage capacity. Some upper spikes mean the long-pulse experiments, and the highest one signed by double circle is the world record of 328.5 GB/pulse.

Such an unstable situation might probably result from the mutual consultation mechanism among symmetrically equal nodes. Such a symmetric structure is suited for scaling out the system with a great number of active nodes; however, it is not necessarily an optimal solution for a project oriented private cloud storage consisting of a designated number of nodes. In latter case, one or several node failure could be so serious because it may have rather a big capacity comparing to the total storage size.

Considering the above mentioned problem on RAID -based cloud storage, we have planned to innovate a new middleware providing more simplified storage mechanism. We found the candidate software, *"OpenStack/Swift"* and *"GlusteFS"*, and made some verification tests on them by using the real mass of LHD data for more than half a year. Table **I** shows the functional comparison between them.

Consequently, we have selected *GlusteFS* to replace the present *IznaStor*. Compared to cloud storage, it only implements very limited structures; raid-0 striped, raid-0 like distributed, raid-1 like replicated, and their combination (Fig. 2). Because of the system simplicity, it can provide abilities of light weight read/write access and easy maintenance in case of malfunctions. In *GlusterFS*, we can manage the storage volumes as standard filesystems like XFS at the servers.

Since the LABCOM data system is implemented to be independent of the storage structure, it is easy to plug off the old *IznaStor* and on the new *GlusteFS*. The effective I/O speed is also confirmed to be on the same level as estimated from raw performance of disks. This achievement contributes not only for the next LHD campaign but also could be informative to implement the next generation fusion experimental data system such as ITER CODAC and its data archiving system.

1) H. Nakanishi, *et al.*: Fusion Sci. Technol. 58 (2010) 445.
2) H. Nakanishi, *et al.*: Nucl. Fusion 51 (2011) 113014.
3) H. Nakanishi, *et al.*: Fusion Eng. Des. 87 (2012) 2189.

Table I. Comparison of the functional differences

	OpenStack Swift	*GlusterFS*
redundancy	object replica	file/brick replica
load balancing	partition	brick (=dir)
fail-over	object, partition, zone	file, brick, volume
rebuild	manual rebalance	manual rebalance
protocols	http rest (put/get)	FUSE, nfs, rest
gateway	proxy	(FUSE client)
other features	static ring info.	stateless
r/w speed	—	450/250–320 MB/s

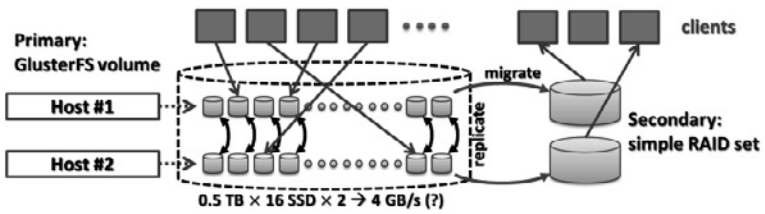

Fig. 2. Schematic view of the combination use of SSD-based *GlusterFS* "distributed replicated" volume and the normal RAIDs of HDDs. The data migration from the primary to the secondary storage will be made very soon but with some delay, i.e. asynchronously.