Large-Capacity, High-Reliability Grid Storage: iStorage HS Series (HYDRAstor)

KAWANABE Masazumi, YOSHIMURA Shigeru, UTAKA Junya
YOSHIOKA Hiroshi, MIZUMACHI Hiroaki, KATO Mitsugu

Abstract

With the rapid increase in corporate data that must be stored long-term, such as the backup of management information and the archiving of e-mails with customers, the need for safe, easy storage of big data has been rising higher than ever. HYDRAstor is a grid storage system that meets these needs. Adopting a revolutionary grid architecture to achieve high performance, scalability and reliability as well as operation/management labor saving, it is suitable for the storage of big data. This paper describes the outline and features of the technologies used in HYDRAstor and introduces actual cases in which it is used.

Keywords

big data, grid storage, deduplication, scalability, virtualization, high reliability

1. Introduction

NEC is the sole Japanese vendor providing a scale-out storage platform, "HYDRAstor," that can store big data such as corporate data, mail archives and image data with high reliability and extensibility.

HYDRAstor can extend performance to the industry-top level. A famous American commentator, W. Curtis Preston, who is the executive editor of TechTarget's Storage Media Group, introduced HYDRAstor as the fastest deduplication storage. For storage capacity, it can be expanded from the minimum configuration of 71 TB (terabytes) to the maximum configuration of 6.3 PB (petabytes) (logical capacity with 20x compression).

2. Outline

HYDRAstor is a large-capacity, high-reliability storage platform adopting grid architecture. It has the following features that make it suitable for the storage of big data:

(1) Dynamic expansion and auto-optimization of performance and capacity

Grid architecture allows performance and capacity to be scaled out dynamically by adding nodes. Capacity is scalable with respect to user data, which is increasing every day.

Data distributed over multiple nodes is virtualized and recognized as being located in a single storage pool, which can be handled as a large-capacity data "warehouse."

(2) Cost reduction with efficient data storage

HYDRAstor's unique deduplication technology, DataRedux, can improve data storage efficiency dramatically and reduce actual physical disk capacity with respect to the amount of stored data. This means that a large amount of data can be stored at a low cost.

(3) Even higher reliability than RAID

HYDRAstor distributes data over multiple nodes using "Distributed Resilient Data." The addition of two to six parity fragments can be specified. When three parity fragments are added, the original data can be recovered even when three data blocks are lost simultaneously. This reliability is higher than that of RAID6, which is generally known to withstand the simultaneous failure of up to two HDDs. It is therefore possible to store a large amount of data with high reliability.

(4) Long-term data storage through node replacement

HYDRAstor's Dynamic Topology feature enables the replacement of an old node with a new node while maintaining the existing data, without the need for data migration. By scheduling periodic node replacement and replacing old nodes with new nodes sequentially, it is possible to store a large amount of data for a long time

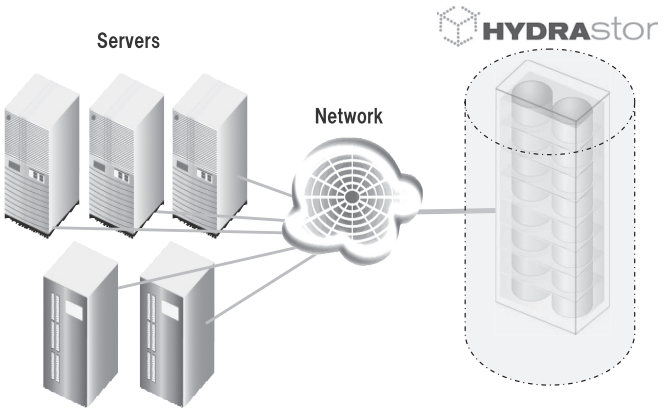

Fig. 1 HYDRAstor system configuration.

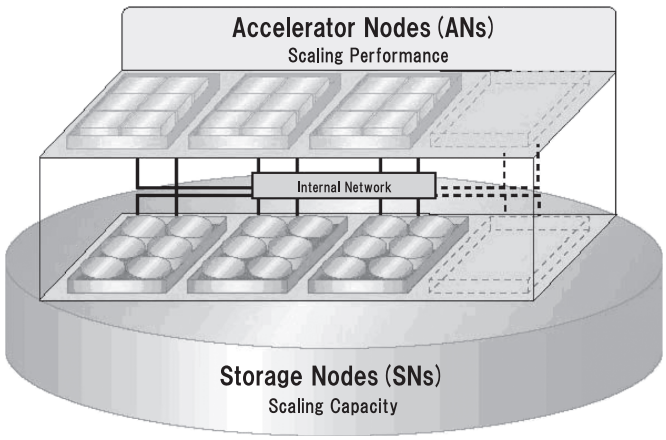

Fig. 2 Grid storage architecture.

without migrating data.

For data access protocols, HYDRAstor supports NFS (Network File System), CIFS (Common Internet File System) and OST (OpenStorage Technology). These are widely diffused network file sharing protocols, so HYDRAstor can be used as a network sharing server by a wide range of platforms including UNIX, Linux and Windows. **Fig. 1** shows the configuration of a system using HYDRAstor.

In section 3, we will discuss specific details of these features.

3. Features of HYDRAstor

3.1 Dynamic Expansion and Auto–optimization of Performance and Capacity

HYDRAstor is composed of two kinds of nodes: accelerator nodes, which process data requests, and storage nodes, which store actual data blocks (**Fig. 2**). The addition of accelerator and storage nodes makes it possible to dynamically extend performance and capacity (**Fig. 3**).

The maximum data write rate is 750 MB/s per node and can be increased linearly as far as the environmental conditions of the external network permit. Capacity can be extended from 71 TB to 6.3 PB (logical capacity with 20x compression). Nodes can be added as required without considering data storage locations. Furthermore, auto-configuration optimization (Dynamic Topology) technology causes these additional nodes to be recognized automatically by the system and places the data in distributed locations autonomously in the optimum

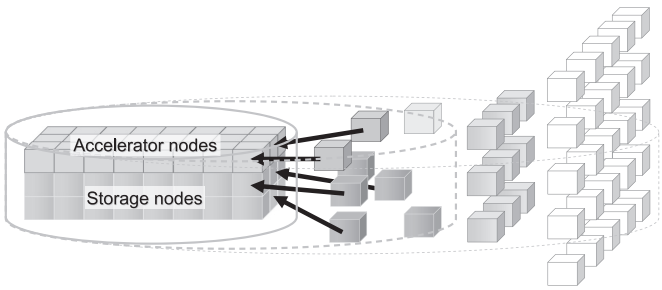

Fig. 3 Auto-configuration optimization (Dynamic topology).

configuration so that no bottleneck is produced.

This Dynamic Topology technology facilitates the following operation management tasks that had previously been extremely complex, thereby drastically reducing management costs:

- Capacity scaling following increases in stored data quantity
- Performance scaling following increases in data transfer quantity
- Improvement of performance bottlenecks
- Node replacement in case of fault

3.2 Cost Reduction with Efficient Data Storage

DataRedux, a deduplication technology unique to HYDRAstor, checks for duplication to avoid rewriting data that has already been written to the storage. This greatly improves data storage efficiency and provides the system with high performance and high cost efficiency.

DataRedux separates the written data into variable lengths intellectually so as to detect duplication with existing data above this maximum length. This technique makes it possible to detect data duplication that is undetectable using the fixed-length data division technique (**Fig. 4**).

This deduplication technology drastically reduces the amount of data transferred to the disk and therefore the required physical disk capacity with respect to the stored data. This means that the daily data writing operations to the disk can be done more quickly and at a lower cost. When this technology is applied to remote replication, further compression of transferred data becomes possible. Therefore, the amount of data transferred to remote sites is greatly reduced so that remote replication with a low-speed circuit using a narrowband frequency can be implemented.

3.3 Even Higher Reliability than RAID

With the deduplication technology described in the previous subsection, a single data block is shared by multiple items of data. In such a system, the accidental loss of a single data block affects all the data referencing that data block, sometimes extending to a very broad range. To prevent this problem, HYDRAstor uses redundant distribution of data (Distributed Resilient Data technology) to achieve higher reliability than the traditional RAID (Redundant Array of Independent Disks) system. HYDRAstor improves reliability by fragmenting the data block to be saved, adding redundancy codes and storing the data fragments by distributing them to multiple storage nodes.

Fig. 5 shows a case in which the original data block is split into nine fragments, with three redundancy codes added to them. In this example, data fragments 1 to 12 are distributed over four storage nodes. In this case, the original data can be

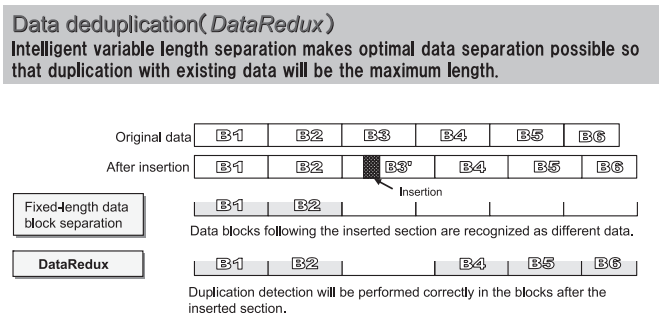

Fig. 4 DataRedux.

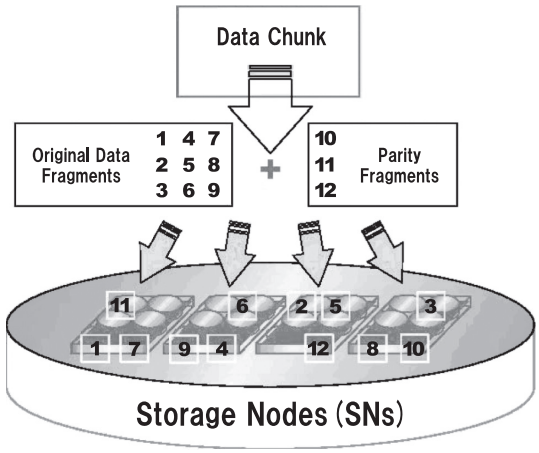

Fig. 5 Distributed Resilient Data.

recovered even if three of the twelve fragments are lost simultaneously. The reliability in this case is higher than with RAID6, which is generally known to withstand the simultaneous failure of up to two HDDs. The level of redundancy can be set freely according to the importance of the stored data, so the administrator can build and manage the system flexibly.

Should HYDRAstor experience a failure, it detects the failure location automatically and executes reconfiguration processing in the background, which means that the administrator does not need to perform the troublesome management tasks that are usually required. This reconfiguration is processed by multiple storage nodes with sufficient processing capabilities, without overhead that would hinder other processing operations being executed.

Fig. 6 shows how fragments 2, 5 and 12, lost by a failure, are immediately detected and automatically reconfigured into other storage nodes.

Distributed Resilient Data technology achieves reliability that exceeds that of existing disk storage products and reduces management costs related to storage faults.

3.4 Long–term Data Storage through Node Replacement

The Dynamic Topology of HYDRAstor enables the replacement of an old node with a new node while maintaining existing data, without the need for batch data migration.

- When a new node is added to HYDRAstor, data is relocated from other nodes to the new node.
- When an old node is deleted from the configuration, the data stored in it is automatically relocated to the

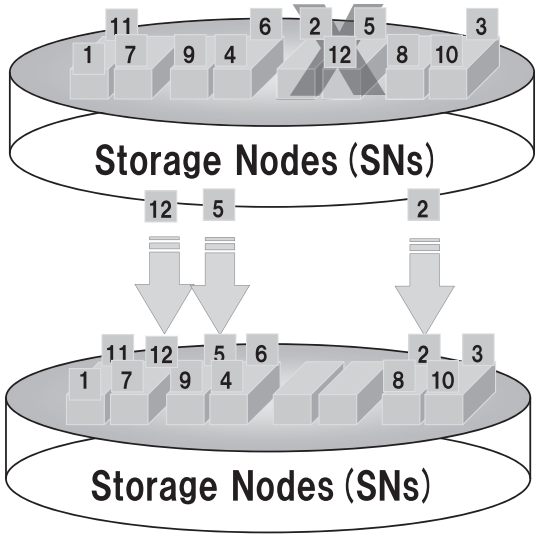

Fig. 6 Autonomous data recovery.

remaining nodes so that the overall system is well-balanced.

By scheduling periodic node replacement and replacing old nodes with new nodes sequentially, it is possible to gradually replace the system hardware with new hardware, allowing the system to store a large amount of data for a long time without migrating data.

4. Use Cases

In this section, we will consider the case study of a customer who has introduced HYDRAstor for handling large amounts of video and image data. Before this introduction, the customer could not save this content, which was increasing every day, in their servers, so instead stored them in a warehouse in the form of tapes. This method hindered the effective utilization of past content because retrieving the desired content took a long time (**Fig. 7**).

We solved this problem and enabled quick video distribution/editing by saving large-capacity content temporarily in HYDRAstor and transferring it through the network when necessary (**Fig. 8**). In this system, we also use tape devices for the storage of content that has not been referenced for a long time.

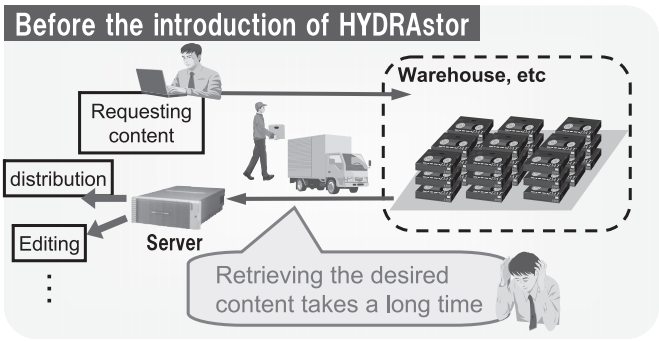

Fig. 7 Before the introduction of HYDRAstor.

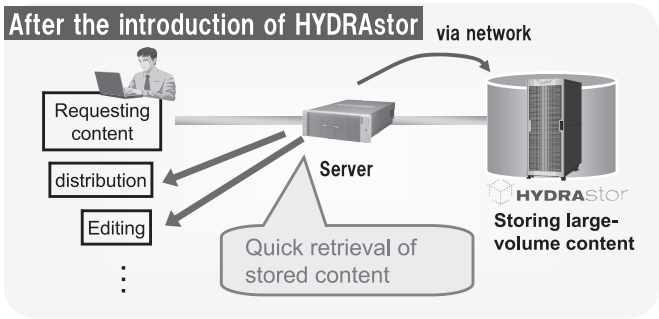

Fig. 8 After the introduction of HYDRAstor.

5. Conclusion

Meeting improvements in the performance of base servers and increases in the capacity of HDDs, HYDRAstor will continue to evolve as an advanced big data storage platform.

*UNIX is a registered trademark of The Open Group in the U.S. and other countries.
*Linux is a registered trademark of Linux Torvalds in the U.S. and other countries.
*Windows is a registered trademark of Microsoft Corporation in the U.S. and other countries.

Authors' Profiles

KAWANABE Masazumi
Manager
1st IT Software Division
IT Software Operations Unit

YOSHIMURA Shigeru
Manager
1st IT Software Division
IT Software Operations Unit

UTAKA Junya
Manager
1st IT Software Division
IT Software Operations Unit

YOSHIOKA Hiroshi
Senior Manager
1st IT Software Division
IT Software Operations Unit

MIZUMACHI Hiroaki
Executive Expert
1st IT Software Division
IT Software Operations Unit

KATO Mitsugu
Assistant General Manager
1st IT Software Division
IT Software Operations Unit

The details about this paper can be seen at the following.

Related URL:

http://www.necam.com/HYDRAstor/

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.7 No.2 Big Data

Vol.7 No.2
September, 2012

Special Issue TOP