The iStorage HS Series Features the Superior Data Compression and High-Speed Transmission Capabilities that are Essential Functions of Big Data Storage

HIRONAGA Ryuji, KAWANABE Masazumi

Abstract

The Big Data era has arrived and the data volume that ICT systems now have to handle has exploded. At such a time, storage systems are required to deal competently with sudden data volume increases. HYDRAstor is a scale-out storage system that meets these needs. Adopting a revolutionary grid architecture it achieves high performance, scalability and reliability as well as operation/management labor saving and thereby guarantees a world leading performance. This paper describes the features adopted by HYDRAstor and introduces new functions that meet the NEC Cloud IaaS requirements.

Keywords

object storage, bandwidth limit, big data, deduplication, scale-out, high reliability

1. Introduction

NEC is the Japanese provider of "HYDRAstor," the scale-out storage platform for backup archives that can store big data; such as corporate data, mail archives and image data, with high reliability. It also offers extensibility according to increased data volumes.

HYDRAstor can extend performance to the upper levels of industry and is suitable for the storage of big data. It features;

(1) Efficient data storage via a deduplication technology
(2) Dynamic expansion and auto-optimization of performance and capacity
(3) Even higher reliability than RAID
(4) Long-term data storage through node replacement

By taking the opportunity to mount new functions, HYDRAstor now supports the object storage interface and the bandwidth limit function, in order to meet the requirements of the NEC cloud computing platform service, NEC Cloud IaaS.

Details of the features and new functions of HYDRAstor are described below.

2. Features of HYDRAstor

2.1 Efficient Data Storage via a Deduplication Technology

HYDRAstor's unique deduplication technology, DataRedux, together with its physical compression technology can improve data storage efficiency dramatically and reduce the actual physical disk capacity with respect to the amount of stored data. DataRedux, a deduplication technology checks for duplication in order to avoid rewriting data that has already been written to the storage. The rest of the non-duplicated data is processed for storage via physical compression. By optimally using resources over the entire nodes in the system, the duplication data checking and writing process is conducted with the greatest speed.

DataRedux separates the written data into variable lengths logically, so as to detect duplication with existing data above a maximum length. This technique makes it possible to detect any data duplication that is undetectable using the fixed-length data division technique. **Fig. 1** shows that DataRedux enables the correct detection of block duplication following an inserted section, even when a file is stored that has had data inserted in the original file.

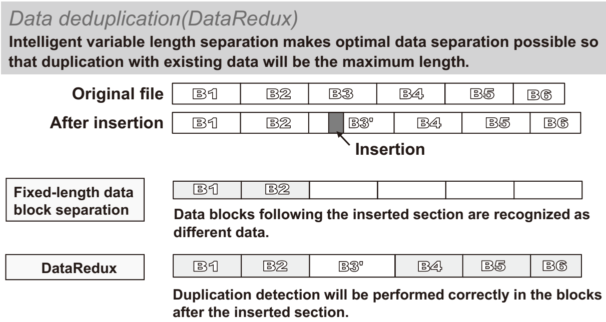

Fig. 1 DataRedux.

This deduplication technology reduces the amount of data transferred to the disk significantly and therefore the required physical disk capacity with respect to the stored data. This means that 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 may be implemented via a low-speed circuit using a narrowband frequency.

2.2 Dynamic Expansion and Auto-optimization of Performance and Capacity

The unique grid architecture employed by HYDRAstor allows performance and capacity to be scaled out dynamically by adding nodes. Data distributed over multiple nodes is virtualized and recognized as being located in a single storage pool, so that provisioning resources is not required.

HYDRAstor is composed of three kinds of nodes: accelerator nodes that process data requests, storage nodes that store actual data blocks, and hybrid nodes that possess functions both of accelerator nodes and of storage nodes. The addition of accelerator, storage and hybrid nodes makes it possible to dynamically extend both performance and capacity.

The performance is 40 TB/h per node (when the OpenStorage high speed deduplication option is employed) and it may be increased linearly as far as the environmental condition of the external network permits. Capacity can be extended from 156 TB to 37.4 PB (logical capacity with 20x compression). Nodes may be added as required without considering the data storage locations. Furthermore, auto-configuration optimization technology (Dynamic Topology) causes these additional nodes to be recognized automatically by the existing system and places the data in distributed locations autonomously in the optimum configuration; so that capacity, performance and reliability are all maximized.

The Dynamic Topology technology facilitates the following operation management tasks that had previously been extremely complex and significantly reduces the management costs:

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

2.3 Even Higher Reliability than RAID

With the deduplication technology described in Section 2.1, 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, which sometimes extends over 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. The Distributed Resilient Data technology 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. 2 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 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 such a system flexibly.

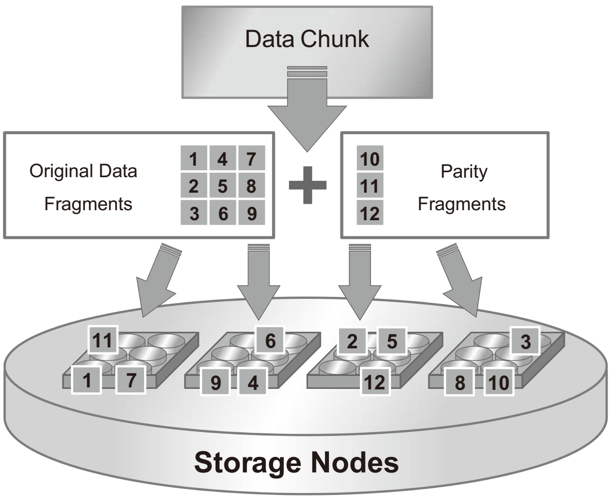

Fig. 2 Distributed resilient data.

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. HDD/nodes available within the system are used for data recovery, therefore, it is not necessary to have spare resources such as a hot-swap HDD for RAID, etc. Moreover, when a fault occurs to an HDD, only logically usable data in the faulted HDD is recovered and stored to several HDDs. This process can be implemented from several to several 10s times faster compared to that of the RAID technology. This reconfiguration is also processed by multiple storage nodes with sufficient processing capabilities, without overheads that would hinder the execution of other processing operations.

Fig. 3 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 the management costs that are related to storage faults.

2.4 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 data migration when updating system hardware.

When a new node is added to HYDRAstor, data stored in other nodes is automatically relocated to the new node. Then, when an old node is deleted from the configuration, the data stored in it is automatically relocated to the remaining nodes so that the overall system is well-balanced.

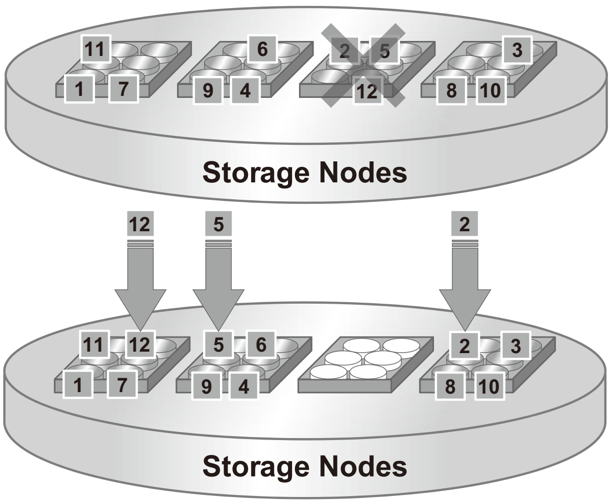

Fig. 3 Autonomous data recovery.

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. Thereby, allowing the system to store a large amount of data for a long time without migrating any of it.

3. New Functions of HYDRAstor

HYDRAstor supports several data access protocols: NFS, CIFS and OST (OpenStorage Technology - an exclusive protocol for Symantec NetBackup). In addition to these, it also supports an interface for the object storage architecture to meet the requirements of NEC Cloud IaaS.

The object storage interface uses REST API, an internet-oriented protocol that is compliant with HTTP protocol for Web access. Therefore, the object storage interface can be used for a wide range of platforms including UNIX, Linux and Windows. On the other hand, it is not suitable for storing data for which it is necessary to rewrite partially, to be accessed randomly or to be often updated. It is suitable for storing data such as video data, backup files, archive data, etc. which are not expected to be updated once stored.

HYDRAstor supports interfaces that include Amazon Simple Storage Service (Amazon S3) API, a de facto standard of an object storage interface among public cloud computing, and another interface that is compatible with the Swift API used as an object storage protocol among OpenStack environments.

NEC Cloud IaaS employs these interfaces supported by HYDRAstor to provide object storage services that are compatible with Amazon S3.

The object storage service provided by NEC Cloud IaaS allows simultaneous access by users of multiple tenants. However, if a large volume of accesses from a certain tenant is performed, the data response speed to other tenants may be degraded.

In order to resolve such an issue, HYDRAstor supports the bandwidth limit function. This function designates the upper limit of read/write bandwidth for each tenant and suppresses accesses that exceed these limits (**Fig. 4**).

With this function, a large volume of accesses carried out

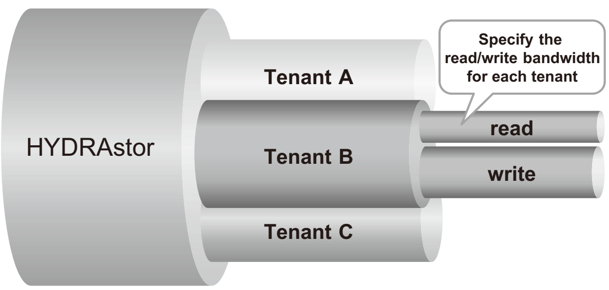

Fig. 4 Bandwidth limits.

by a certain tenant can be suppressed and influence on a user's access from other tenants can be reduced.

4. Conclusion

The present paper introduces HYDRAstor, which newly mounts object storage interfaces and a bandwidth limit function in order to implement an object storage service via NEC Cloud Iaas. In the coming FY2015, further functional reinforcement is expected, such as improved Amazon S3 compatibility, etc. We will continue our R&D to advance HYDRAstor, aiming for it to become the optimal storage platform for big data.

* NetBackup is a trademark or registered trademark of Symantec Corporation and/or its affiliates in the U.S. and other countries.
* UNIX is a registered trademark of The Open Group in the U.S. and other countries.
* Linux is a registered trademark of Linus Torvalds in the U.S. and other countries.
* Windows is a registered trademark of Microsoft Corporation in the U.S. and other countries.
* Amazon S3, an Amazon Simple Storage Service, is a trademark of Amazon.com, Inc. and/or its affiliates in the U.S. and other countries.
* OpenStack is a registered trademark or trademark of OpenStack Foundation.

Authors' Profiles

HIRONAGA Ryuji
Manager
Storage Systems Department
IT Platform Division

KAWANABE Masazumi
Manager
Storage Systems Department
IT Platform Division

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

Related URL:

HYDRAstor - Scale-out Grid Storage Platform
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.9 No.2 Special Issue on Future Cloud Platforms for ICT Systems

Vol.9 No.2
June, 2015

Special Issue TOP

NEC Information