Huawei OceanStor Dorado V6 All-Flash Storage Systems

Technical White Paper

Issue	**1.7**
Date	**2019-08-01**

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://e.huawei.com

Contents

1 Executive Summary

Huawei OceanStor Dorado V6 all-flash storage systems are designed for enterprises' mission-critical services. They use FlashLink® dedicated to flash media to achieve 0.2 ms stable latency. The gateway-free HyperMetro feature provides an end-to-end active-active data center solution, which can smoothly evolve to the geo-redundant disaster recovery (DR) solution to achieve 99.9999% solution-level reliability. Inline deduplication and compression maximize the available capacity and reduce the operating expense (OPEX). OceanStor Dorado V6 meets the requirements of enterprise applications such as databases, virtual desktop infrastructure (VDI), virtual server infrastructure (VSI), and file sharing, helping the financial, manufacturing, and carrier industries evolve smoothly to all-flash storage.

This document describes and highlights the unique advantages of OceanStor Dorado V6 in terms of its product positioning, hardware and software architecture, and features.

2 Overview

2.1 OceanStor Dorado V6 Family

OceanStor Dorado V6 includes Dorado3000 V6 SAS，Dorado5000 V6 SAS，Dorado5000 V6 NVMe，Dorado6000 V6 SAS，Dorado6000 V6 NVMe，Dorado8000 V6 SAS，Dorado8000 V6 NVMe，Dorado18000 V6 SAS，and Dorado18000 V6 NVMe.

Figure 2-1 Dorado3000 V6 SAS

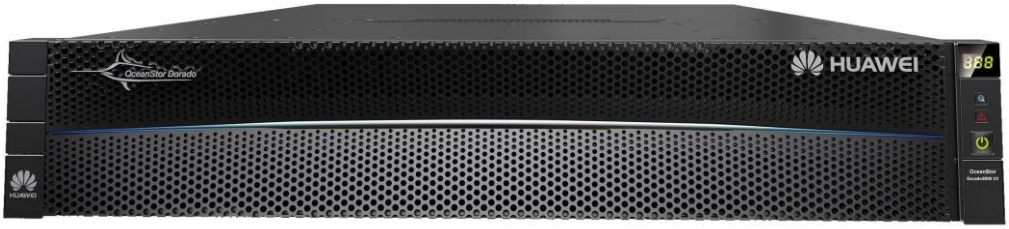


Figure 2-2 OceanStor Dorado5000/6000 V6(SAS)

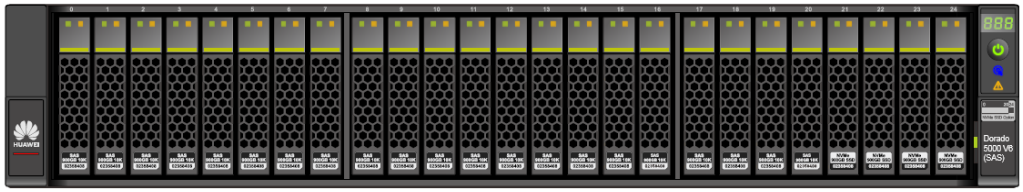


Figure 2-3 OceanStor Dorado5000/6000 V6(NVMe)

Figure 2-4 OceanStor Dorado8000/18000 V6(SAS)

Figure 2-5 OceanStor Dorado8000/18000 V6(NVMe)

3 System Architecture

3.1 Concepts

3.1.1 Controller Enclosure

The OceanStor Dorado5000 V6 supports 2U controller enclosures. The controller enclosure contains 2 storage controllers that process all storage service logic. It provides core functions such as host access, device management, and data services. A controller enclosure consists of a system subrack, controllers, interface modules, power modules, BBUs, management modules and SAS/NVMe SSDs.

OceanStor Dorado5000V6(NVME) controller enclosure contains 36 NVME disk slots and OceanStor Dorado5000 V6(SAS) controller enclosure contains 25 SAS disk slots.

Figure 3-1 Front view of OceanStor Dorado5000 V6(NVMe)

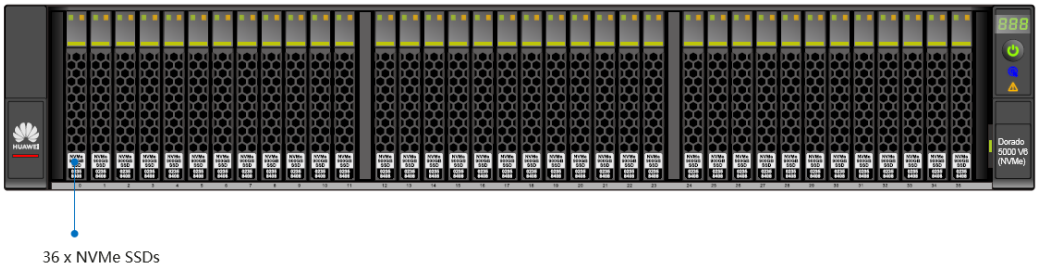


Figure 3-2 Back view of OceanStor Dorado5000 V6(NVMe)

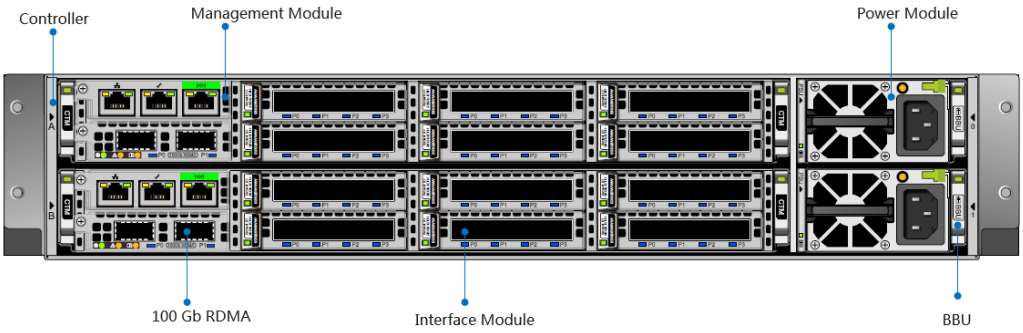


Figure 3-3 Front view of OceanStor Dorado5000 V6(SAS)

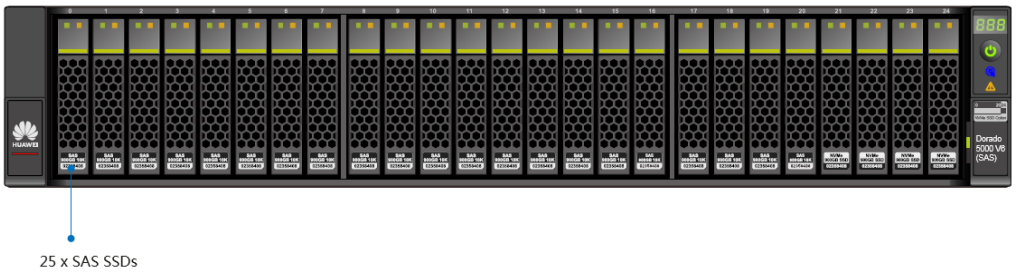
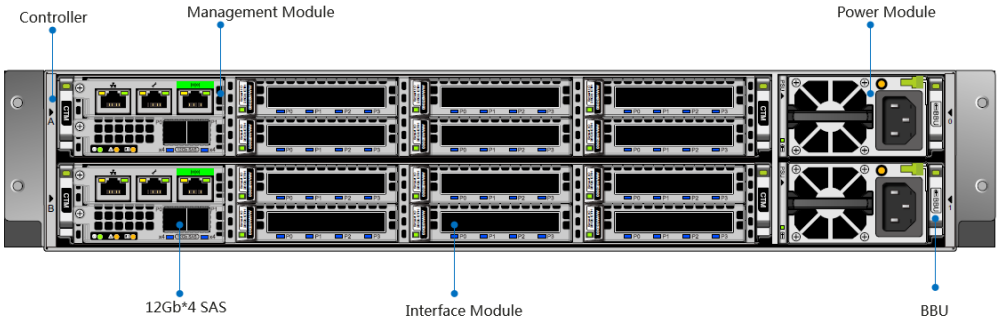


Figure 3-4 Back view of OceanStor Dorado5000 V6(SAS)

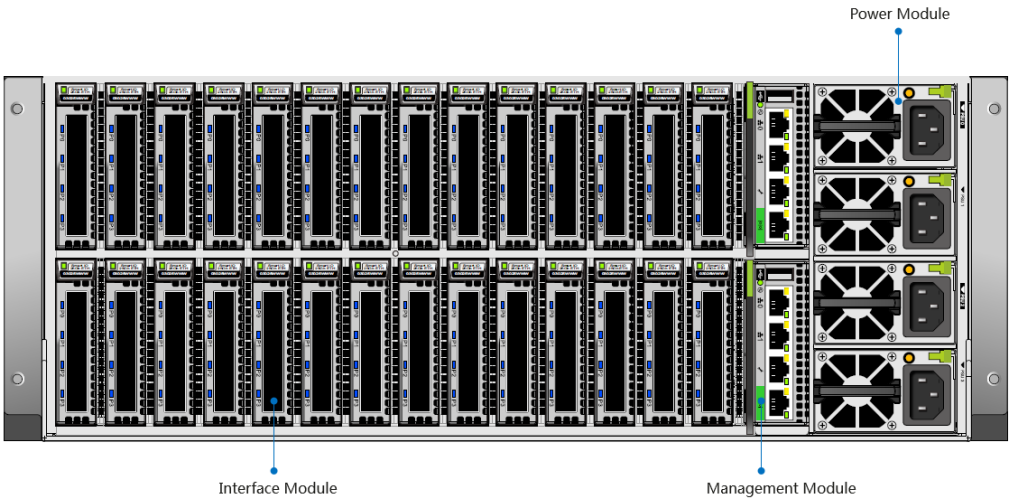


The OceanStor Dorado8000/18000 V6 supports 4U controller enclosures. The controller enclosure contains 4 storage controllers that process all storage service logic. It provides core

functions such as host access, device management, and data services. A controller enclosure consists of a system subrack, controllers, interface modules, power modules, BBUs, and management modules.

Figure 3-5 Front view of OceanStor Dorado8000/18000 V6(NVMe/SAS)

Figure 3-6 Back view of OceanStor Dorado8000/18000 V6(NVMe/SAS)



3.1.2 Controller

An OceanStor Dorado V6 controller is a computing module consisting of the CPU, memory, and main board. It processes storage services, receives configuration and management commands, saves configuration data, connects to disk enclosures, and stores critical data onto coffer disks.

Each controller enclosure has two or four controllers. Every two controllers form a pair for high availability. If a single controller fails, the other controller takes over the storage services to guarantee service continuity. The front-end I/O modules on the controllers provide host access ports. The port types include 8 Gbit/s, 16 Gbit/s, or 32 Gbit/s Fibre Channel, 100GE, 40GE, 25GE, and 10GE.

3.1.3 Disk Enclosure

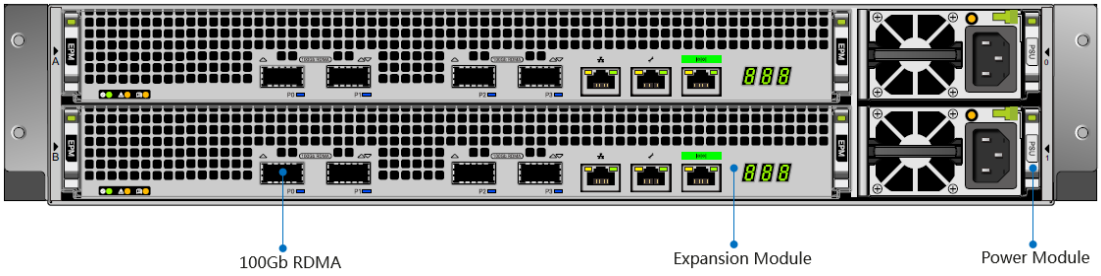
SSD Enclosure

NVMe SSD Enclosure

A 2U NVMe SSD enclosure of OceanStor Dorado V6 houses 36 x NVMe SSDs. It consists of a system subrack, expansion modules, power modules, and SSDs. A NVMe SSD enclosure provides 2 expansion modules, and each expansion module provides 4 x 100Gb RDMA ports for scale-up.

Figure 3-7 Front view of NVMe SSD enclosure

Figure 3-8 Back view of NVMe SSD enclosure



SAS SSD Enclosure

A 2U SAS SSD enclosure of OceanStor Dorado V6 houses 25 x SAS SSDs. It consists of a system subrack, expansion modules, power modules, and SSDs. A SAS SSD enclosure provides 2 expansion modules, and each expansion module provides 2 x 12Gb*4 ports for scale-up.

Figure 3-9 Front view of SAS SSD enclosure

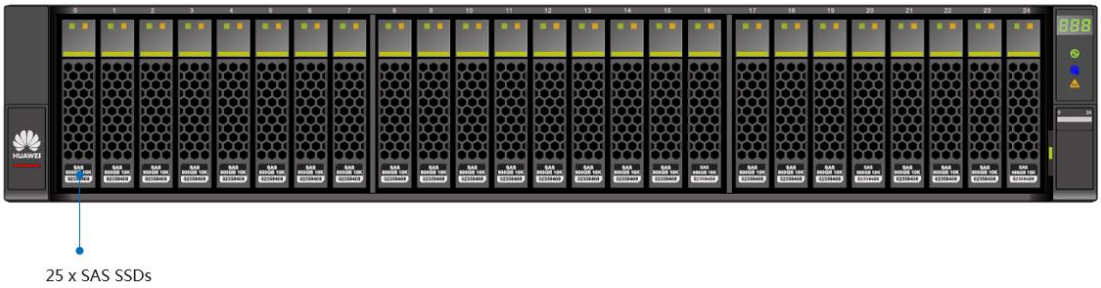
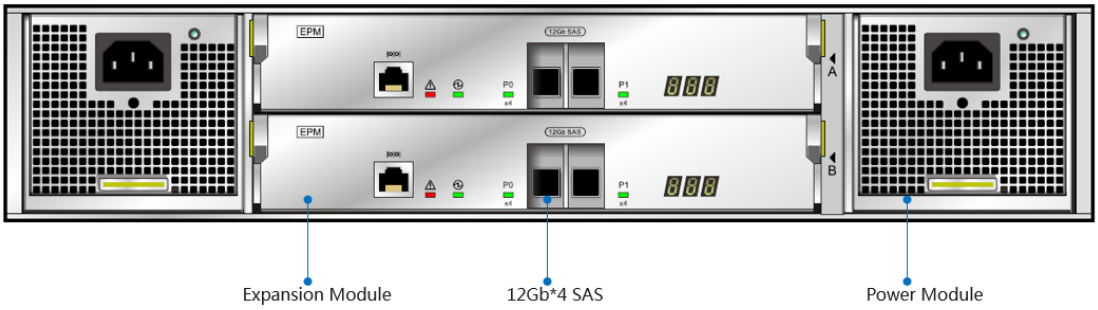


Figure 3-10 Back view of SAS SSD enclosure



3.1.4 Disk Domain

A disk domain consists of multiple disks. RAID groups select member disks from a disk domain. OceanStor Dorado V6 can have one or more disk domains and supports disk domains across controller enclosures.Each disk domain can have SSDs of two different capacities.

Figure 3-11 Disk domain across controller enclosures

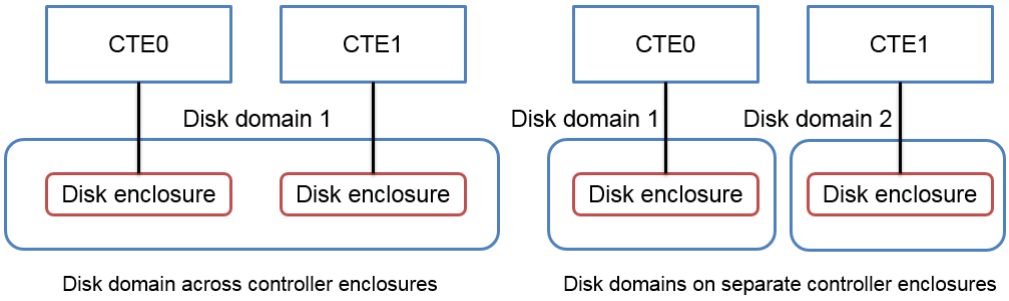


Figure 3-11 shows a dual-controller system. You can create a disk domain that contains all disks in the system or create a separate disk domain for each controller enclosure.

When creating a disk domain, you must specify the hot spare policy and encryption type.

You can choose to use a high or low hot spare policy, or not to use any one. The policy can be changed online.

- When you use a high hot spare policy, the disk domain reserves great hot spare space for data reconstruction in the event of a disk failure. The hot spare space increases non-linearly with the number of disks.
- When you use a low hot spare policy, which is the default setting, the disk domain reserves a small amount of hot spare space (enough for the data on at least one disk) for data reconstruction in the event of a disk failure. The hot spare space increases non-linearly with the number of disks.
- If you do not use a hot spare policy, the system will not reserve hot spare space.

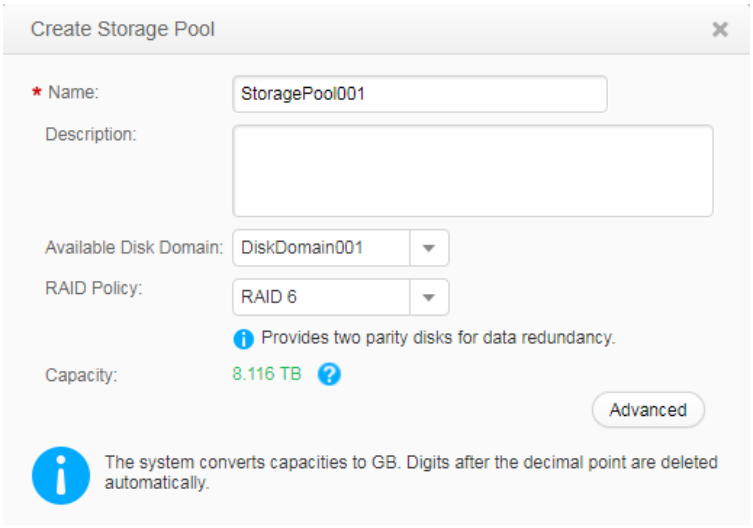
3.1.5 Storage Pool

Storage pools, which are containers of storage resources, are created in disk domains. The storage resources used by application servers are all from storage pools. Each disk domain can have only one storage pool.

You must specify the RAID level when creating a storage pool. By default, a storage pool has all the available capacity of the selected disk domain.

By default, a storage pool uses RAID 6, which meets the reliability requirements in most scenarios while providing high performance and capacity utilization. When the capacity of a single disk is large (for example, 8 TB), reconstruction of a single disk will take a long time, which reduces reliability. In this case, RAID-TP can be used for higher reliability.

Figure 3-12 Creating a storage pool



3.1.6 RAID

OceanStor Dorado V6 uses a Huawei proprietary algorithm, Erase-Code (EC), to implement RAID5, RAID 6, RAID-TP, RAID 10*. RAID-TP is able to tolerate three faulty disks, providing high system reliability.

NOTE

If you require the specifications marked by *, contact Huawei sales personnel.

OceanStor Dorado V6 uses the RAID 2.0+ block-level virtualization technology to implement RAID. With this technology:

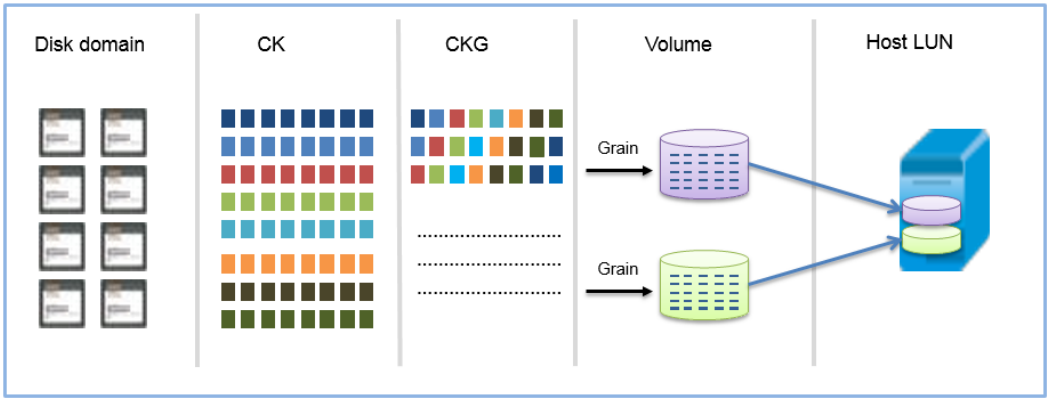
- Multiple SSDs form a disk domain.
- Each SSD is divided into fixed-size chunks (typically 4 MB per chunk) to facilitate logical space management.
- Chunks from different SSDs constitute a chunk group (CKG) based on the customer-configured RAID level.

Chunk groups support three redundancy configurations:

- RAID 5 uses the EC-1 algorithm and generates one copy of parity data for each stripe.
- RAID 6 uses the EC-2 algorithm and generates two copies of parity data for each stripe.
- RAID-TP uses the EC-3 algorithm and generates three copies of parity data for each stripe.

A chunk group is further divided to smaller-granularity (typically, 8 KB) grains, which are the smallest unit for data writes. OceanStor Dorado V6 adopts full-stripe write to avoid extra overhead generated in traditional RAID mechanisms. Figure 3-13 shows RAID mapping on OceanStor Dorado V6.

Figure 3-13 RAID mapping on OceanStor Dorado V6



OceanStor Dorado V6 uses EC to support more member disks in a RAID group, improving space utilization.

Table 3-1 Space utilization of RAID groups using EC

RAID Level	Number of Member Disks Recommended by EC	Space Utilization	Number of Member Disks Recommended by the Traditional Algorithm	Space Utilization
RAID 5	24+1	96%	7+1	87.5%
RAID 6	23+2	92%	14+2	87.5%
RAID-TP	22+3	86.9%	Not supported	NA

If a disk is faulty or is removed for a long time, the chunks on this disk are reconstructed. The detailed procedure is as follows:

1. The disk becomes faulty and the chunks on it become unavailable.
2. The RAID level degrades for the chunk groups that contain the affected chunks.
3. The system allocates idle chunks from the storage pool for data reconstruction.
4. Based on the RAID level of the storage pool, the system uses the normal data columns and parity data to restore the damaged data blocks and writes them to the idle chunks.

Because the faulty chucks are distributed to multiple chunk groups, all of the affected chunk groups start reconstruction at the same time. In addition, the new chunks are from multiple disks. This enables all disks in the disk domain to participate in reconstruction, fully utilizing the I/O capability of all disks to improve the data reconstruction speed and shorten data recovery time.

OceanStor Dorado V6 uses both common and dynamic RAID reconstruction methods to prevent RAID level downgrade and ensure system reliability in various scenarios.

- Common reconstruction

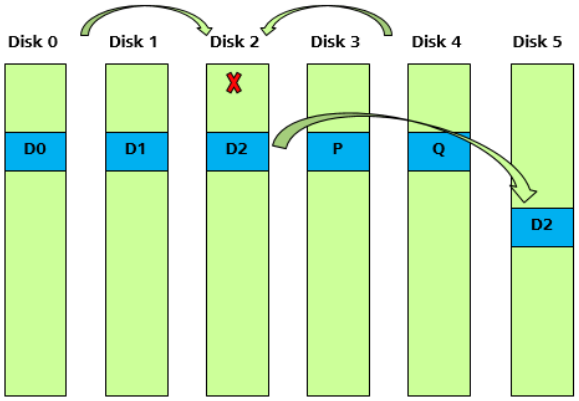
  A RAID group has M+N members (M indicates data columns and N indicates parity columns). When the system has faulty disks, common reconstruction is triggered if the

number of normal member disks in the disk domain is still greater than or equal to M+N. During reconstruction, the system uses idle chunks to replace the faulty ones in the chunk groups and restores data to the new chunks. The RAID level remains M+N.

In Figure 3-14, D0, D1, D2, P, and Q form a chunk group. If disk 2 fails, a new chunk D2_new on disk 5 is used to replace D2 on disk 2. In this way, D0, D1, D2_new, P, and Q form a new chunk group and the system restores the data of D2 to D2_new.

After common reconstruction is complete, the number of RAID member disks remains unchanged, maintaining the original redundancy level.

Figure 3-14 Common reconstruction



- Dynamic reconstruction

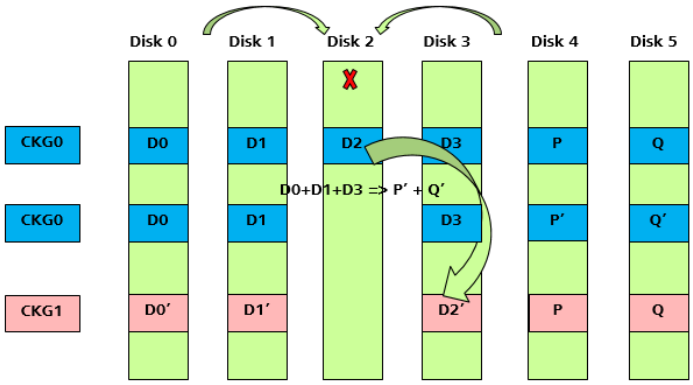
  If the number of member disks in the disk domain is fewer than M+N, the system reduces the number of data columns (M) and retains the number of parity columns (N) during reconstruction. This method retains the RAID level by reducing the number of data columns, ensuring system reliability.

  During the reconstruction, the data on the faulty chunk is migrated to a new chunk group. If the system only has M+N-1 available disks, the RAID level for the new chunk group is (M-1)+N. The remaining normal chunks (M-1) and parity columns P and Q form a new chunk group and the system calculates new parity columns P' and Q'.

  In Figure 3-15, there are six disks (4+2). If disk 2 fails, data D2 in CKG0 is written to the new CKG1 as new data (D2') and the RAID level is 3+2. D0, D1, and D3 form a new 3+2 CKG0 with new parity columns P' and Q'.

  After the reconstruction is complete, the number of member disks in the RAID group is decreased, but the RAID redundancy level remains unchanged.

Figure 3-15 Dynamic reconstruction



The number of RAID members is automatically adjusted by the system based on the number of disks in a disk domain. Factors such as capacity utilization, reliability, and reconstruction speed are considered. Table 3-2 describes the relationship between the disks in a disk domain and RAID members.

Table 3-2 Number of disks and RAID members

Number of Disks in a Disk Domain (X)	Number of RAID Members	Hot Spare Space Under the High Policy
8 to 12	X-1	Equals to the capacity of 1 disk.
13 to 25	X-2	Equals to the capacity of 2 disks.
26 or 27	X-3	Equals to the capacity of 3 disks.
> 27	25	Greater than or equal to the capacity of 3 disks.

The number of RAID members (M+N) complies with the following rules:

1. If the number of faulty disks in a disk domain is less than or equal to the number of disks in the hot spare space, the system does not trigger dynamic reconstruction.
2. A high capacity utilization should be guaranteed.
3. M+N should not exceed 25.

When the number of disks is less than 13, the hot spare space equals to the capacity of one disk and M+N is X-1. This ensures the highest possible capacity utilization.

When a disk domain has 13 to 25 disks, the hot spare space equals to the capacity of two disks and M+N is X-2. This setting is to avoid dynamic reconstruction when multiple disks fail.

When a disk domain has 26 or 27 disks, the hot spare space equals to the capacity of three disks and M+N is X-3. Dynamic reconstruction will not be triggered if up to three disks fail (at different time).

When the number of disks is greater than 27, the maximum value of M+N will be 25. This ensures a high capacity utilization while limiting read amplification caused by reconstruction. For example, if a disk in a 30+2 RAID group becomes faulty, the system must read the chunks

from 30 disks to reconstruct each chunk in the affected chunk groups, resulting in great read amplification. To avoid this, the system limits M+N to 25.

When new disks are added to the system to expand capacity, the value of M+N increases with the number of disks. All new data (including data generated by garbage collection) will be written using the new RAID level, while the RAID level for the existing data remains unchanged. For example, a disk domain has 15 disks and uses RAID 6; M+N is 11+2. If the customer expands the domain to 25 disks, new data will be written to the new 21+2 chunk groups, while the existing data is still in the original 11+2 chunk groups. When garbage collection starts, the system will move the valid chunks in the original 11+2 chunk groups to the 21+2 chunk groups and then reclaim the original chunk groups.

OceanStor Dorado V6 has the following advantages in terms of data redundancy and recovery:

- Fast reconstruction

  All disks in the disk domain participate in reconstruction. Test results show that OceanStor Dorado V6 takes only 30 minutes to reconstruct 1 TB of data (when there is no new data written to the system), whereas traditional RAID takes more than 2 hours.
- Multiple RAID levels available

  OceanStor Dorado V6 supports RAID 5, RAID 6, and RAID-TP. You can choose the RAID level that meets your needs. RAID-TP allows three faulty disks and provides the highest reliability for mission-critical services.
- Intelligent selection of RAID member disks

  If a disk has a persistent fault, the system can intelligently reduce the number of member disks in the RAID group and use dynamic reconstruction to write new data with the original RAID level instead of a lower level, avoiding reduction in data reliability.
- Appending mechanism to ensure data consistency

  OceanStor Dorado V6 uses appending in full-stripe writes. This avoids data inconsistency in traditional RAID caused by write holes.

3.2 Hardware Architecture

3.2.1 Product Models

The OceanStor Dorado V6 series products include Dorado3000 V6 SAS，Dorado5000 V6 SAS，Dorado5000 V6 NVMe，Dorado6000 V6 SAS，Dorado6000 V6 NVMe，Dorado8000 V6 SAS，Dorado8000 V6 NVMe，Dorado18000 V6 SAS，and Dorado18000 V6 NVMe. All devices are compatible with standard 19 inches cabinet.

Table 3-3 OceanStor Dorado V6 product models

Model	Controller Enclosure	Number of Controllers per Enclosure	Disk Type
Dorado3000 V6	2 U enclosure with integrated disks	2	SAS

Model	Controller Enclosure	Number of Controllers per Enclosure	Disk Type
Dorado5000 V6	2 U enclosure with integrated disks	2	NVMe or SAS
Dorado6000 V6	2 U independent enclosure without disks	2	NVMe or SAS
Dorado 8000 V6	4 U independent enclosure without disks	2 or 4	NVMe or SAS
Dorado18000 V6	4 U independent enclosure without disks	2 or 4	NVMe or SAS

3.2.2 Huawei-Developed SSDs

OceanStor Dorado V6 uses Huawei-developed SSDs (HSSDs, 1 DWPD) to maximize system performance. HSSDs work perfectly with storage software to provide an optimal experience across various service scenarios.

An SSD consists of a control unit and a storage unit (mainly flash memory chips). The control unit contains an SSD controller, host interface, and dynamic random access memory (DRAM) module. The storage unit contains only NAND flash chips.

Blocks and pages are the basic units for reading and writing data in the NAND flash.

- A block is the smallest erasure unit and generally consists of multiple pages.
- A page is the smallest programming and read unit. Its size is usually 4 KB, 8 KB, or 16 KB.

Operations on NAND flash include erase, program, and read. The program and read operations are implemented at the page level, while the erase operations are implemented at the block level. Before writing a page, the system must erase the entire block where the page resides. Therefore, the system must migrate the valid data in the block to a new storage space before erasing it. This process is called garbage collection (GC). SSDs can only tolerate a limited number of program/erase (P/E) cycles. If a block on an SSD experiences more P/E cycles than others, it will wear out more quickly. To ensure reliability and performance, HSSDs leverage the following advanced technologies.

3.2.2.1 Wear Leveling

The SSD controller uses software algorithms to monitor and balance the P/E cycles on blocks in the NAND flash. This prevents over-used blocks from failing and extends the service life of the NAND flash.

HSSDs support both dynamic and static wear leveling. Dynamic wear leveling enables the SSD to write data preferentially to less-worn blocks to balance P/E cycles. Static wear leveling allows the SSD to periodically detect blocks with fewer P/E cycles and reclaim their data, ensuring that blocks storing cold data can participate in wear leveling.

3.2.2.2 Bad Block Management

Unqualified blocks may occur when the NAND flash is manufactured or used, which are labeled as bad blocks. HSSDs identify bad blocks according to the P/E cycles, error type, and error frequency of the NAND flash. If a bad block exists, the SSD recovers the data on the

bad block by using the Exclusive-OR (XOR) redundancy check data between the NAND flash memories, and saves it to a new block. Within the lifecycle of an HSSD, about 1.5% of blocks may become bad blocks. HSSDs have reserved space to replace these bad blocks, ensuring sufficient available capacity and user data security.

3.2.2.3 Data Redundancy Protection

HSSDs use multiple redundancy check methods to protect user data from bit flipping, manipulation, or loss. Error correction code (ECC) and cyclic redundancy check (CRC) are used in the DRAM of the SSDs to prevent data changes or manipulation; low-density parity check (LDPC) and CRC are used in the NAND flash to prevent data loss caused by NAND flash errors; XOR redundancy is used between NAND flash memories to prevent data loss caused by flash failures.

Figure 3-16 Data redundancy check

LDPC uses linear codes defined by the check matrix to check and correct errors. When data is written to pages on the NAND flash, the system calculates the LDPC verification information and writes it to the pages with the user data. When data is read from the pages, LDPC verifies and corrects the data.

HSSDs house a built-in XOR engine to implement redundancy protection between flash chips. If a flash chip becomes faulty (page failure, block failure, die failure, or full chip failure), redundancy check data is used to recover the data on the faulty blocks, preventing data loss.

3.2.2.4 Background Inspection

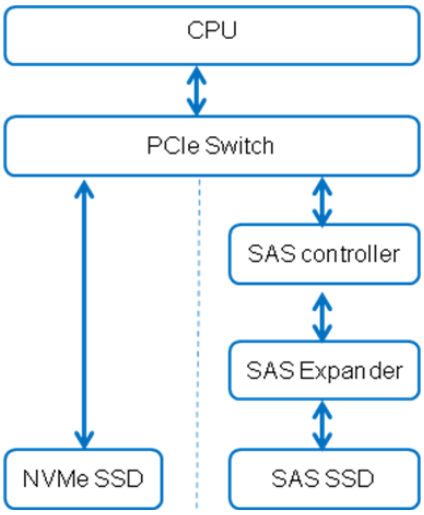
If data is stored in NAND flash for a long term, data errors may occur due to read interference, write interference, or random failures. HSSDs periodically read data from the NAND flash, check for bit changes, and write data with bit changes to new pages. This process detects and handles risks in advance, which effectively prevents data loss and improves data security and reliability.

3.2.2.5 Support for SAS and NVMe

Huawei HSSDs support both SAS and NVMe ports. NVMe is a more light-weighted protocol than SAS. Its software stack does not have a SCSI layer, reducing the number of protocol interactions. In addition, NVMe does not require a SAS controller or SAS expander on the hardware transmission path. The NVMe SSD directly connects to the CPU via the PCIe bus to

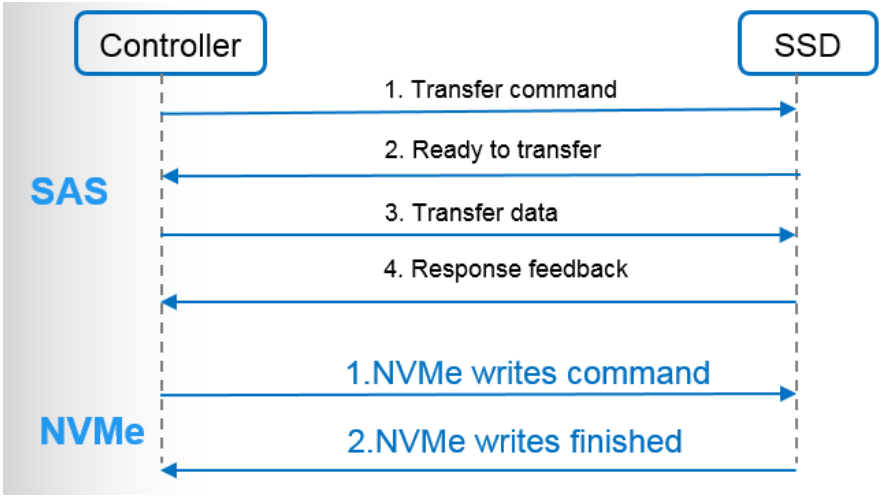
achieve lower latency. In addition, NVMe supports a larger concurrency and queue depth (64k queues, each queue with a depth of 64k), fully exploiting SSD performance. The NVMe HSSDs provide dual ports and are hot swappable, improving system performance, reliability, and maintainability.

Figure 3-17 Transmission paths of NVMe and SAS SSDs



NVMe SSDs reduce the number of interactions in a write request from 4 (in a SAS protocol) to 2.

Figure 3-18 SAS and NVMe protocol interactions



3.2.3 Huawei-Developed Chips

OceanStor Dorado V6 uses Huawei-developed chips, including SSD controller chips, front-end interface chips (SmartIO chips), and baseboard management controller (BMC) chips.

- SSD controller chip

HSSDs use new-generation enterprise-class controllers, which provide SAS 3.0 x2 and PCIe 3.0 x4 ports in compliance with industry standards. The controller features high performance and low power consumption. The controllers use enhanced ECC and built-in RAID technologies to extend the SSD service life to meet enterprise-level reliability requirements. In addition, this 28 nm chip supports the latest DDR4, 12 Gbit/s SAS, and 8 Gbit/s PCIe rates as well as Flash Translation Layer (FTL) hardware acceleration to provide stable performance at a low latency for enterprise applications.

- SmartIO chip

  Hi182x (IOC) is the first Huawei-developed storage interface chip. It integrates multiple interface protocols such as 8 Gbit/s, 16 Gbit/s, or 32 Gbit/s Fibre Channel, 100GE, 40GE, 25GE, and 10GE to achieve excellent performance, high interface density, and flexible configuration. The SmartIO chip can work as both initiator and target mode.

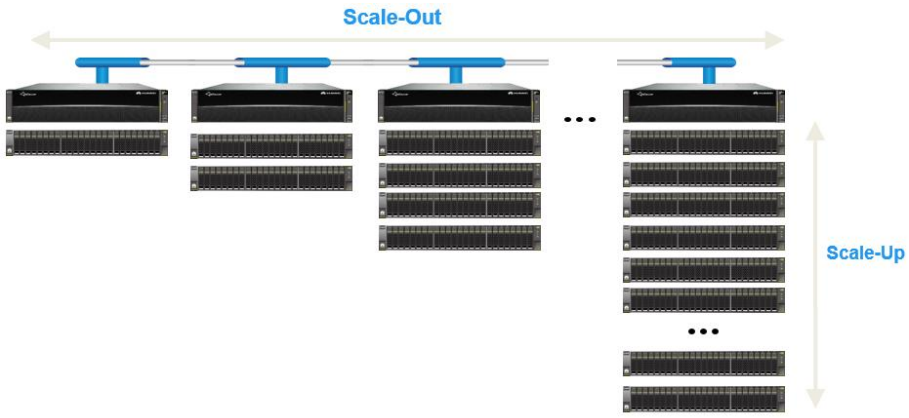
- BMC chip

  Hi1710 is a BMC chip. It consists of the A9 CPU, 8051 co-processor, sensor circuits, control circuits, and interface circuits. It supports the Intelligent Platform Management Interface (IPMI), which monitors and controls the hardware components of the storage system, including system power control, controller monitoring, interface module monitoring, power supply and BBU management, and fan monitoring.

3.2.4 Hardware Scalability

OceanStor Dorado V6 supports both scale-up and scale-out.

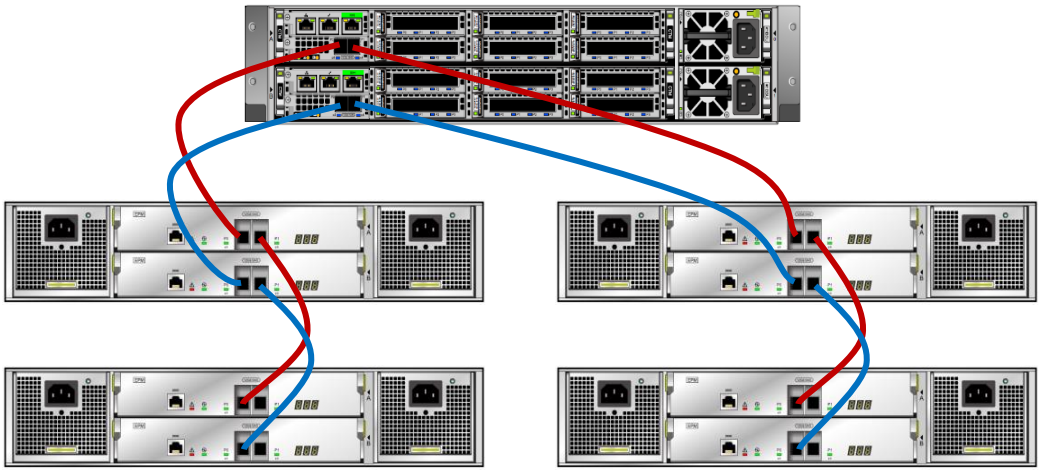
Figure 3-19 Scale-out and scale-up



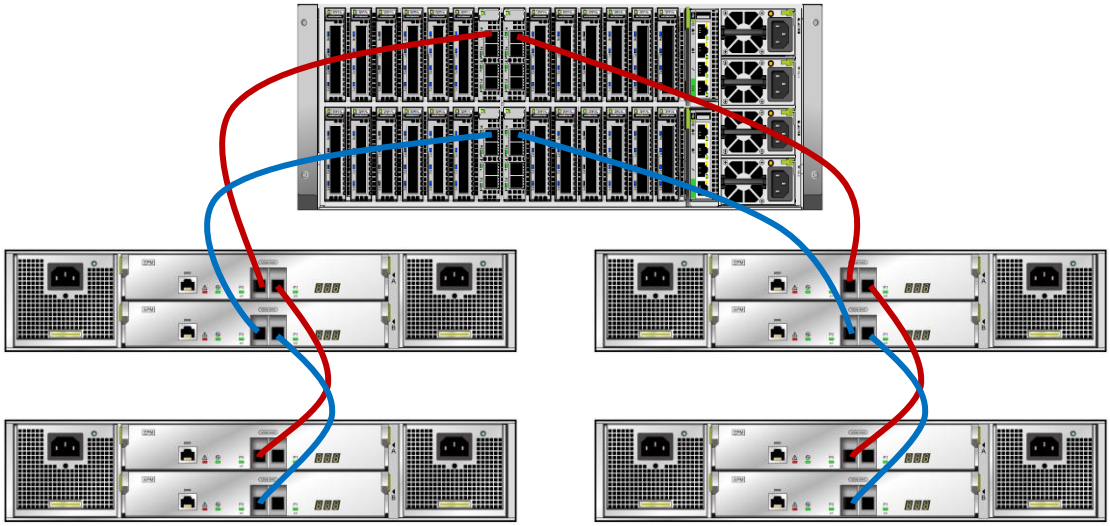
Scale-up

OceanStor Dorado5000/6000/8000/18000 V6(SAS) supports SSDs scale-up by SAS SSD enclosures through the redundant SAS 3.0 links.

Dorado5000/6000 V6(SAS) Scale-up

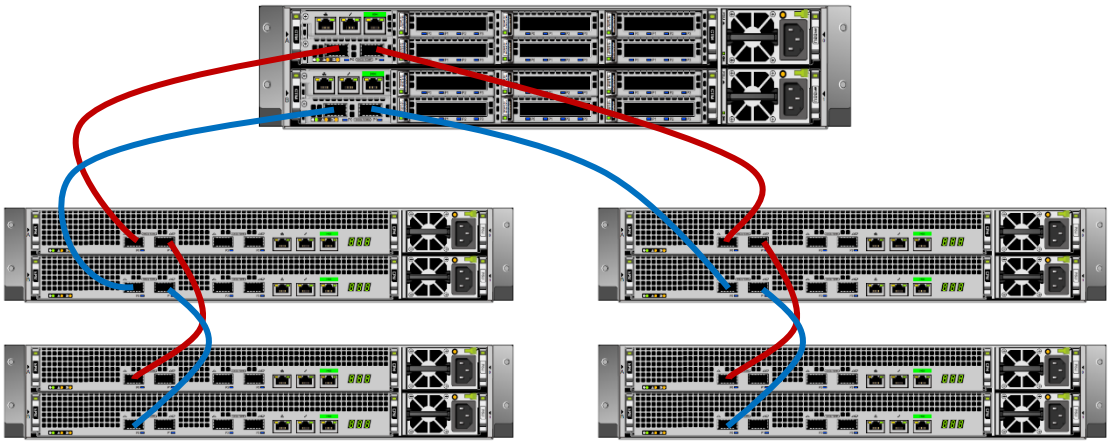


Dorado8000/18000 V6(SAS) Scale-up

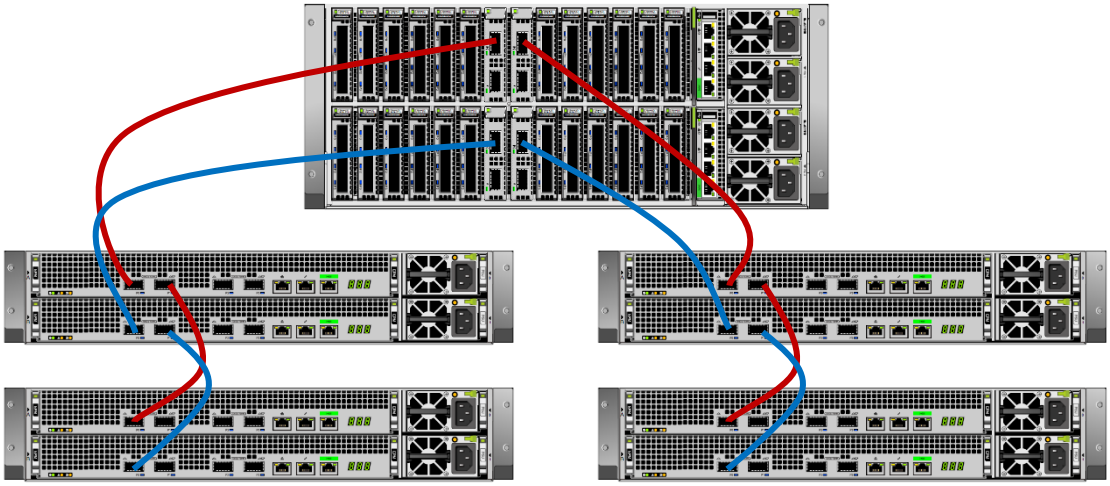


OceanStor Dorado5000/6000/8000/18000 V6(NVMe) supports SSDs scale-up by NVMe SSD enclosures through the redundant 100Gb RDMA links.

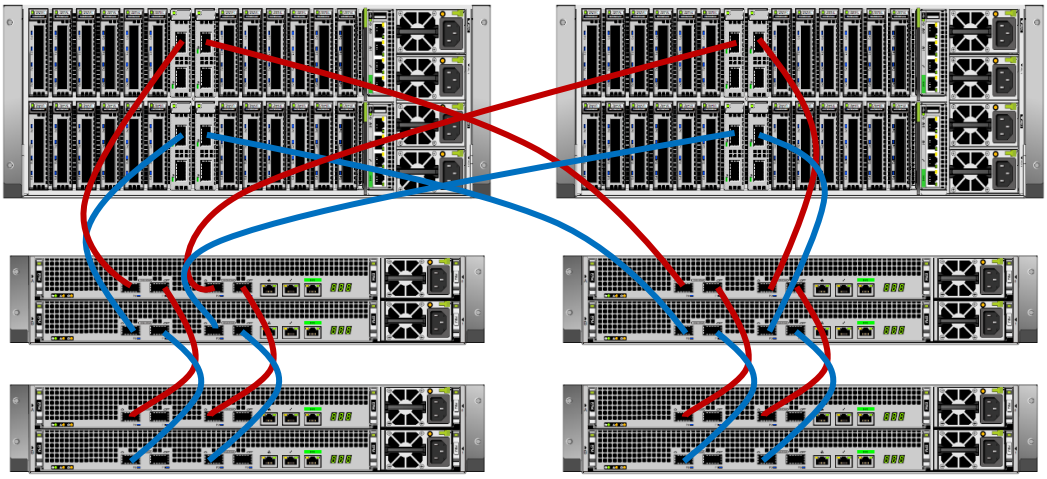
Dorado5000/6000 V6(NVMe) Scale-up



Dorado8000/18000 V6(NVMe) Scale-up



OceanStor Dorado8000/18000 V6(NVMe) supports NVMe SSD enclosures shared scale-up to 2 engines for higher reliability.

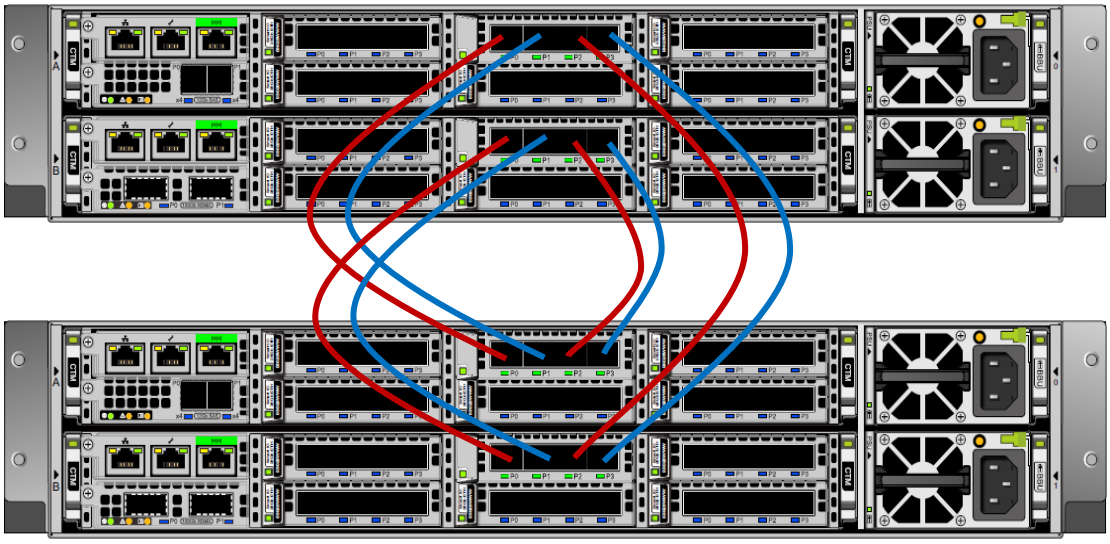


Scale-out

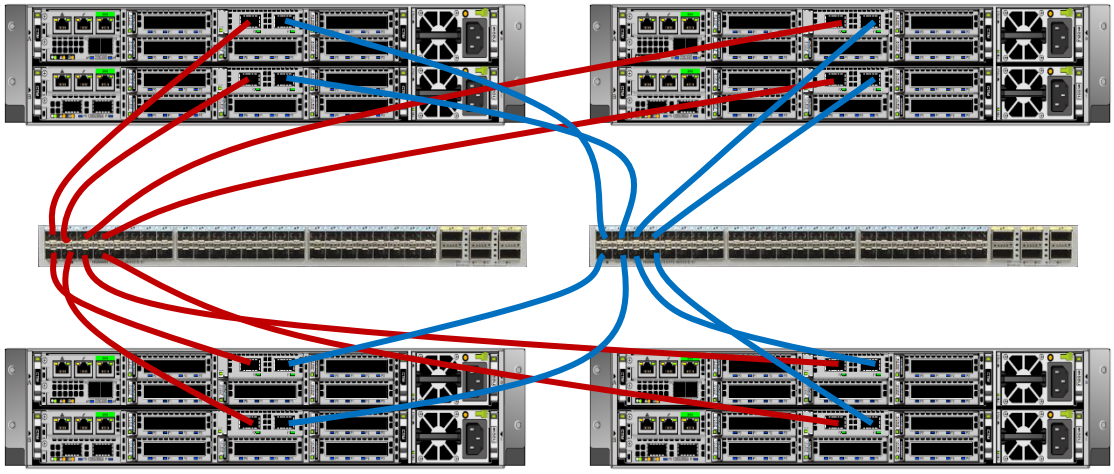
OceanStor Dorado5000/6000/8000 V6 supports scale-out up to 16 controllers interconnected by redundant 25/100 Gb RDMA links. OceanStor Dorado18000 V6 supports scaleout to 32 controllers. The 2 engines scale-out of Dorado5000/6000/8000 V6 is direct interconnected by 25/100 Gb RDMA links. The more than 2 engines scale-out of Dorado5000/8000 V6 is interconnected by redundant 100Gb RDMA switches.

The following figures show details of the network connections.

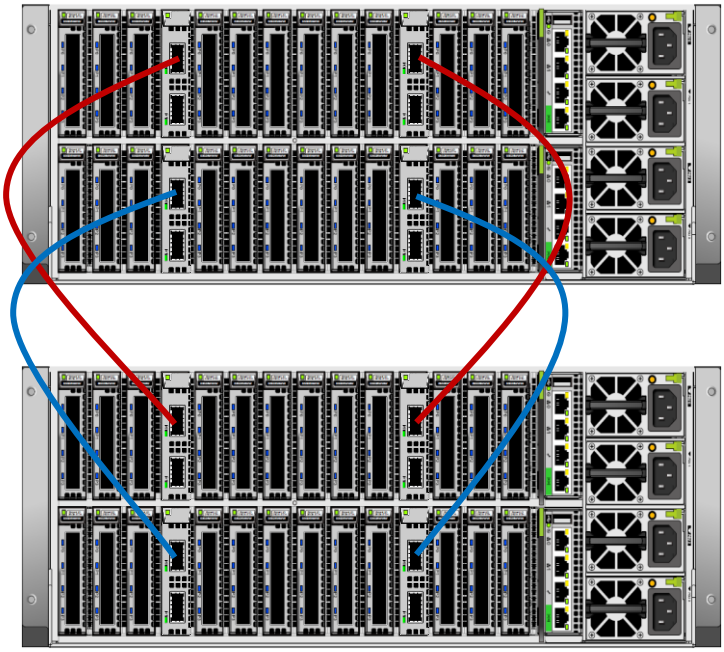
Dorado5000(NVMe) 4 controllers scale-out



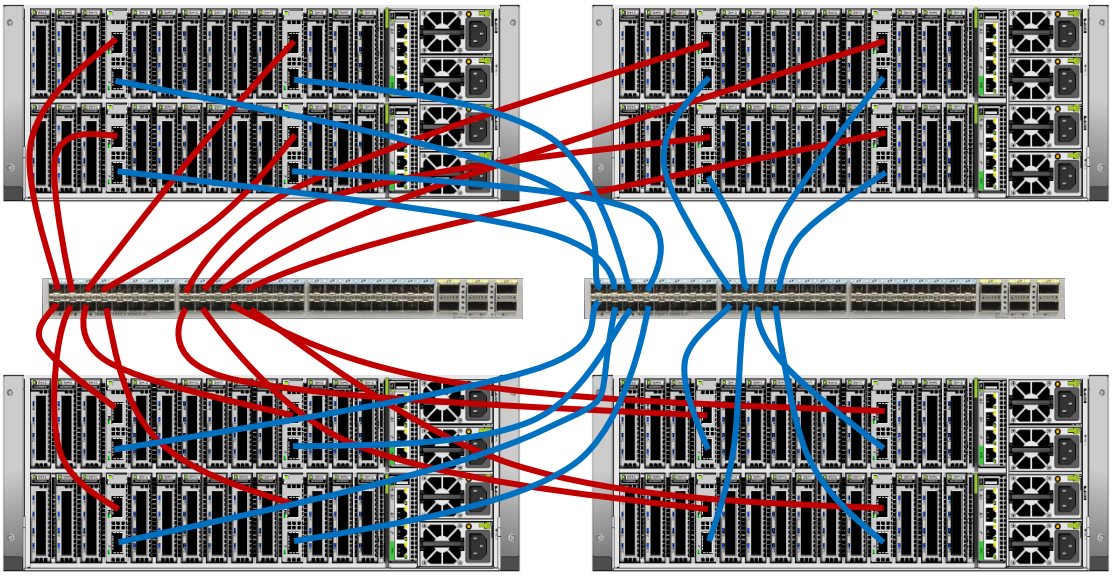
Dorado5000(NVMe) 8 controllers scale-out



Dorado8000(NVMe) 8 controllers scale-out



Dorado8000(NVMe) 16 controllers scale-out



3.2.5 Hardware Architecture Highlights

- Outstanding performance

  The hardware features end-to-end high-speed architecture, PCIe 3.0 buses, SAS 3.0 or PCIe 3.0 x 4 disk ports, and 8 Gbit/s, 16 Gbit/s, or 32 Gbit/s Fibre Channel, 100GE, 40GE, 25GE, or 10GE front-end ports. Huawei-developed NVMe SSDs contribute to high system performance at a low latency.

- Stable and reliable

  Tens of thousands of sets of these systems on live networks have consistently demonstrated the hardware maturity and fully redundant architecture. Stable and reliable PCIe hot swap technology allows online maintenance and replacement of NVMe SSDs.

  Storage system components are 1+1 redundancy and work in active-active mode. Normally, the two redundant components are working simultaneously and share loads. If one component fails or is offline, the other one takes over all loads without affecting ongoing services.

- Efficient

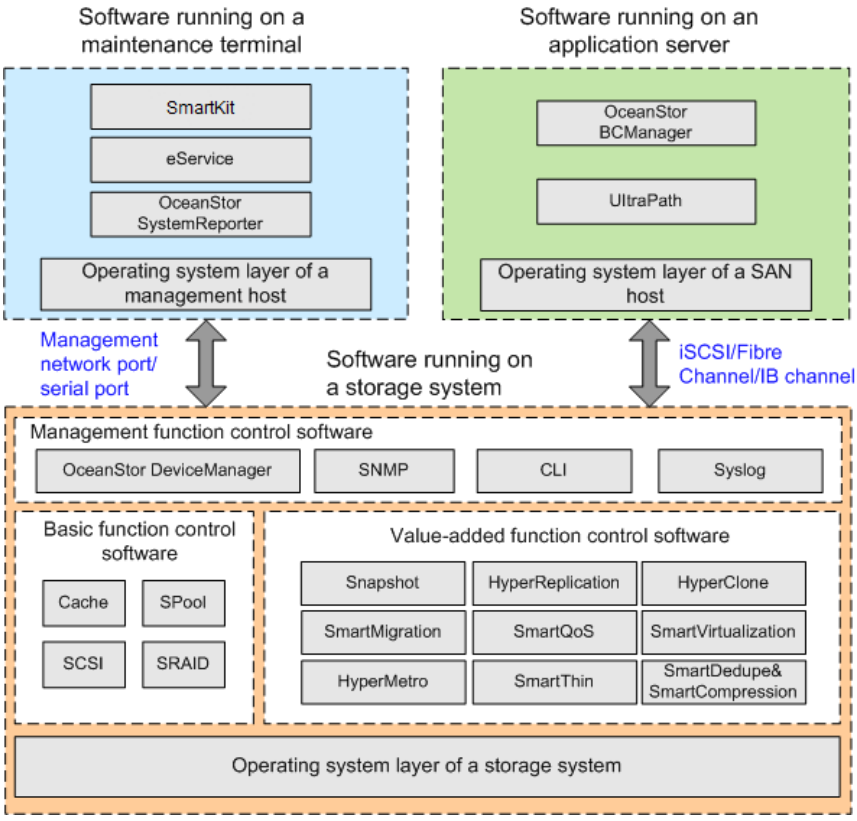
  OceanStor Dorado V6 supports both scale-out and scale-up, and its controllers and disks can be expanded online. Its I/O modules use a modular design and are hot swappable. Both its front-end and back-end ports can be configured on demand.

  The Turbo Module technology enables hot swap of controllers, fans, power supplies, interface modules, BBUs, and disk modules. All these modules can be operated online.

3.3 Software Architecture

OceanStor Dorado V6 uses a version of the OceanStor OS that has been designed specifically for SSDs and employs FlashLink® and comprehensive value-added features to provide excellent performance, robust reliability, and high efficiency.

Figure 3-20 Software architecture of OceanStor Dorado V6



The software architecture of the storage controller mainly consists of the cluster & management plane and service plane.

- The cluster & management plane provides a basic environment to run the system, controls multi-controller scale-out, and manages alarms, performance, and user operations.
- The service plane schedules storage service I/Os, permits data scale-out, and implements controller software-related functions provided by FlashLink®, such as deduplication and compression, redirect-on-write (ROW) full-stripe write, hot and cold data separation, garbage collection, global wear leveling, and anti-wear leveling.

3.3.1 FlashLink

FlashLink® associates storage controllers with SSDs by using a series of technologies for flash media, ensuring both reliability and performance of flash storage. The key technologies of FlashLink® include hot and cold data separation, end-to-end I/O priority, ROW full stripe write, global garbage collection, global wear leveling, and anti-wear leveling. These techniques resolve problems such as performance jitter caused by write amplification and garbage collection, and ensure a steady low latency and high IOPS of OceanStor Dorado V6.

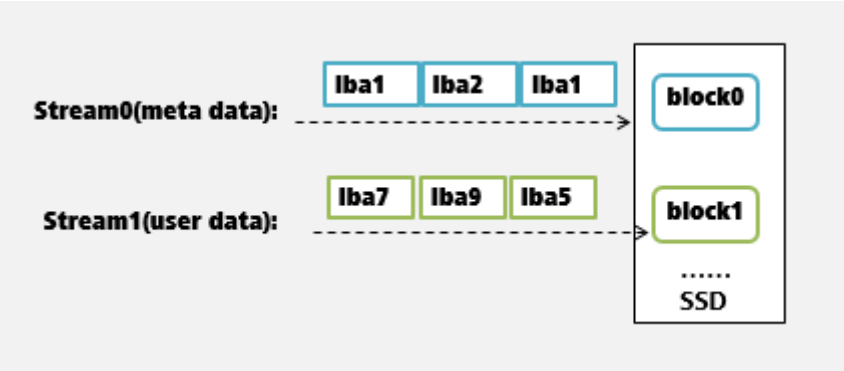
3.3.1.1 Hot and Cold Data Separation

During garbage collection, an SSD must migrate the valid data in the blocks that are to be reclaimed to a new storage space, and then erase the entire blocks to release their space. If all

the data in a block is invalid, the SSD can directly erase the whole block without migrating data.

Data in the storage system is classified into hot and cold data by change frequency. For example, metadata (hot) is updated more frequently and is more likely to cause garbage than user data (cold). FlashLink® adds labels to data with different change frequencies (user data and metadata) in the controller software, sends the data to SSDs, and writes the data to dedicated blocks to separate hot and cold data. In this way, there is a high probability that all data in a block is invalid, reducing the amount of data migration for garbage collection, and improving SSD performance and reliability.

Figure 3-21 Hot and cold data separation (1)



In Figure 3-22, the red and gray blocks represent metadata and user data, respectively.

If metadata and user data are stored in the same blocks, the blocks may still contain a large amount of valid user data after all the metadata becomes garbage, because metadata changes more frequently than user data. When the system erases these blocks, it must migrate the valid user data to new blocks, reducing garbage collection efficiency and system performance.

If metadata and user data are stored in different blocks, the system only needs to migrate a small amount of data before erasing the metadata blocks. This significantly improves the garbage collection efficiency.

Figure 3-22 Hot and cold data separation (2)

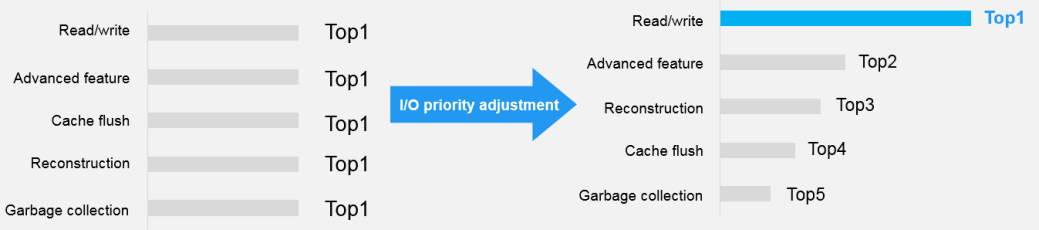
3.3.1.2 End-to-End I/O Priority

To ensure stable latency for specific types of I/Os, OceanStor Dorado V6 controllers label each I/O with a priority according to its type. This allows the system to schedule CPU and other resources and queue I/Os by priority, offering an end-to-end I/O-priority-based latency guarantee. Specifically, upon reception of multiple I/Os, SSDs check their priorities and process higher-priority I/Os first.

OceanStor Dorado V6 classifies I/Os into five types and assigns their priorities in descending order: read/write I/Os, advanced feature I/Os, reconstruction I/Os, cache flush I/Os, and

garbage collection I/Os. Control based on I/O priorities allows OceanStor Dorado V6 to achieve optimal internal and external I/O response.

Figure 3-23 End-to-end I/O priority

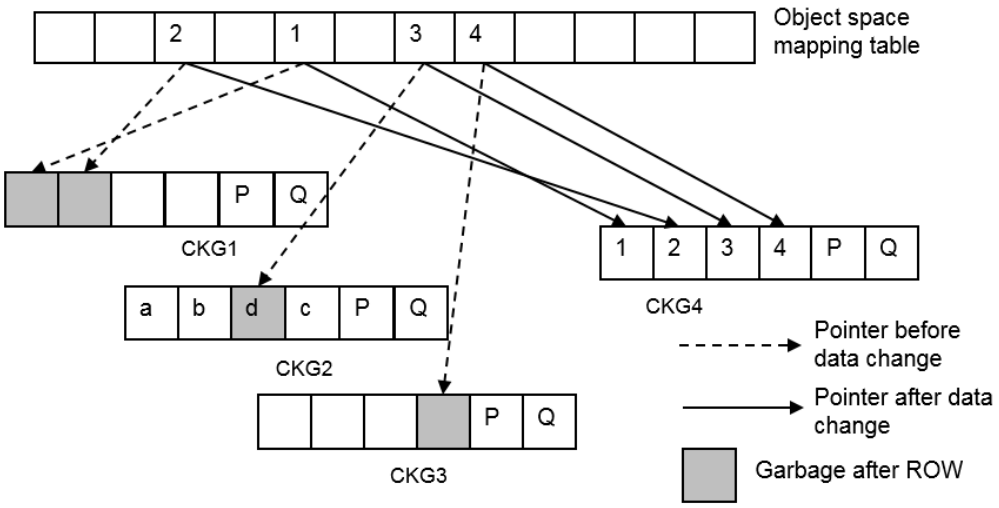


On the left side in the preceding figure, various I/Os have the same priority and contend for resources. After I/O priority adjustment, system resources are allocated by I/O priority.

3.3.1.3 ROW Full-Stripe Write

OceanStor Dorado V6 uses ROW full-stripe write, which writes all new data to new blocks instead of overwriting existing blocks. This greatly reduces the overhead on controller CPUs and read/write loads on SSDs in a write process, improving system performance in various RAID levels.

Figure 3-24 ROW full-stripe write



In Figure 3-24, the system uses RAID 6 (4+2) and writes new data blocks 1, 2, 3, and 4 to modify existing data.

In traditional overwrite mode, the system must modify every chunk group where these blocks reside. For example, when writing data block 3 to CKG2, the system must first read the original data block d and the parity data P and Q. Then it calculates new parity data P 'and Q', and writes P', Q', and data block 3 to CKG2. In ROW full-stripe write, the system uses the data blocks 1, 2, 3, and 4 to calculate P and Q and writes them to a new chunk group. Then it modifies the logical block addressing (LBA) pointer to point to the new chunk group. This process does not need to read any existing data.

Typically, RAID 5 uses 22D+1P, RAID 6 uses 21D+2P, and RAID-TP uses 20D+3P, where D indicates data columns and P indicates parity columns. Table 3-4 compares write amplification on OceanStor Dorado V6 using these RAID levels.

Table 3-4 Amplification in ROW-based full-strip write

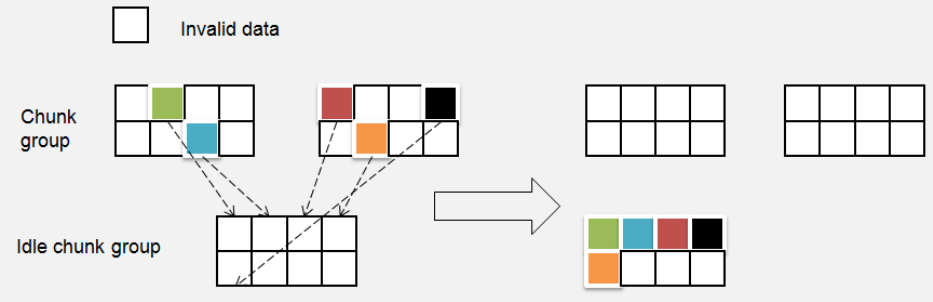
	Write Amplification of Random Small I/Os	Read Amplification of Random Small I/Os	Write Amplification of Sequential I/Os
RAID 5 (24D+1P)	1.04 (25/24)	0	1.04
RAID 6 (23D+2P)	1.08 (25/23)	0	1.08
RAID-TP (22D+3P)	1.13(25/22)	0	1.13

The performance differences between RAID 5 and RAID 6, and between RAID 6 and RAID-TP are only about 5%.

3.3.1.4 Global Garbage Collection

OceanStor Dorado V6 uses global garbage collection to reclaim the space occupied by invalid data blocks after ROW full-stripe write. Garbage collection is triggered when the ratio of garbage reaches a specified threshold. During garbage collection, the system migrates the valid data in the target chunk group to a new chunk group. Then the system reclaims all chunks in the target chunk group to release its space. At the same time, the system issues the **unmap** or **deallocate** command to SSDs to mark the data in the corresponding LBA area as invalid. The SSDs then reclaim the space. The garbage collection process is initiated by storage controllers and takes effect on all SSDs.

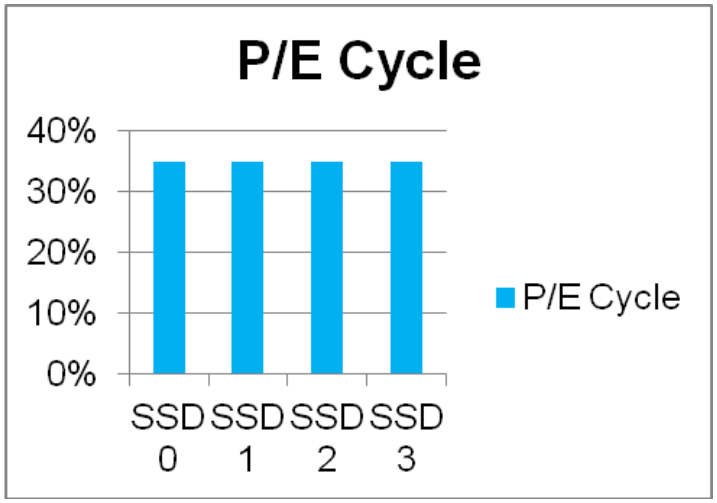
Figure 3-25 Global garbage collection



3.3.1.5 Global Wear Leveling and Anti-Wear Leveling

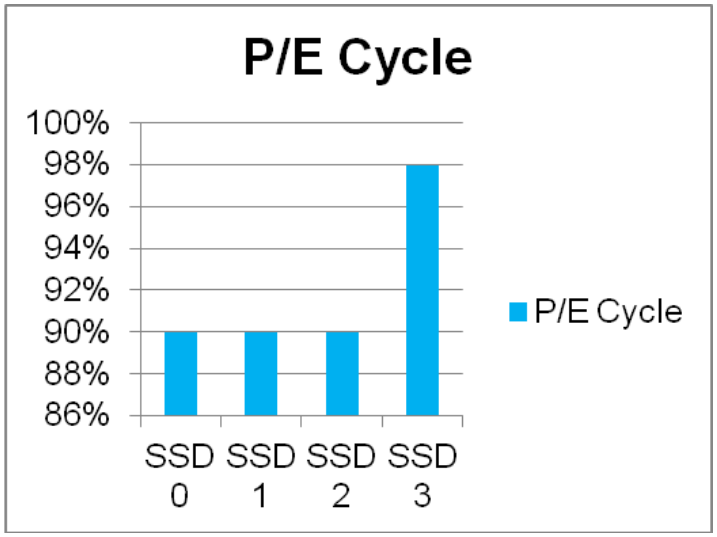
Different from HDDs, SSDs can only withstand a limited number of read and write operations. Therefore, an all-flash storage system requires load balancing between disks to prevent overly-used disks from failing. FlashLink® uses controller software and disk drivers to regularly query the disk wearing level from the SSD controller.

Figure 3-26 Global wear leveling



However, if SSDs are approaching the end of their life, for example, the wearing level exceeds 80%, multiple SSDs may fail simultaneously and data may be lost if global wear leveling is still used. In this case, the system enables anti-global wear leveling to avoid simultaneous failures. The system selects the most severely worn SSD and writes data onto it as long as it has idle space. This reduces that SSD's life faster than others, and you are prompted to replace it sooner, avoiding simultaneous failures.

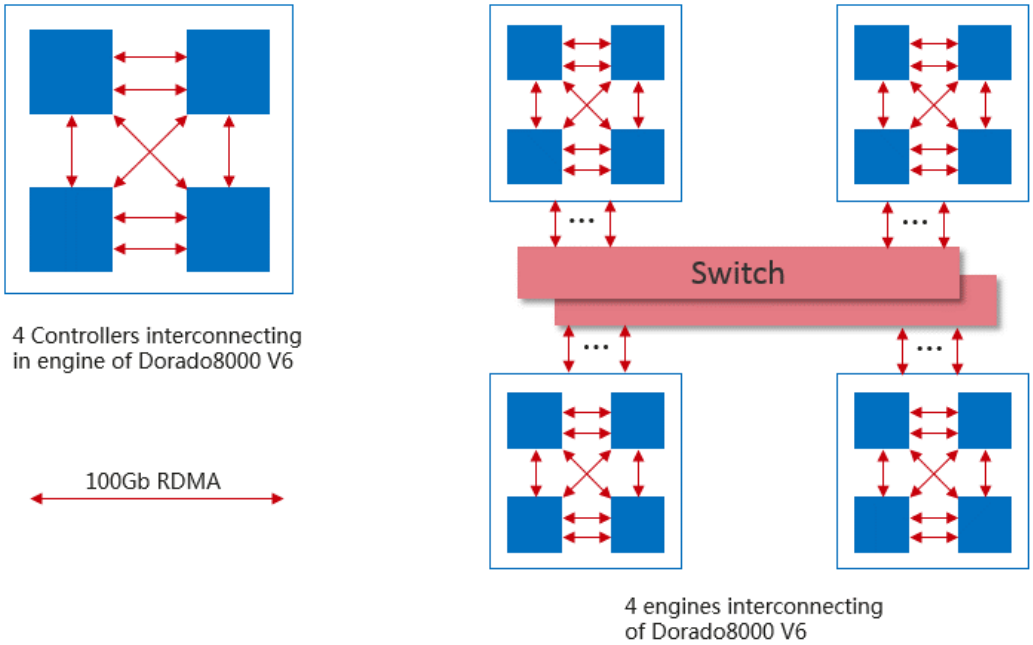
Figure 3-27 Global anti-wear leveling



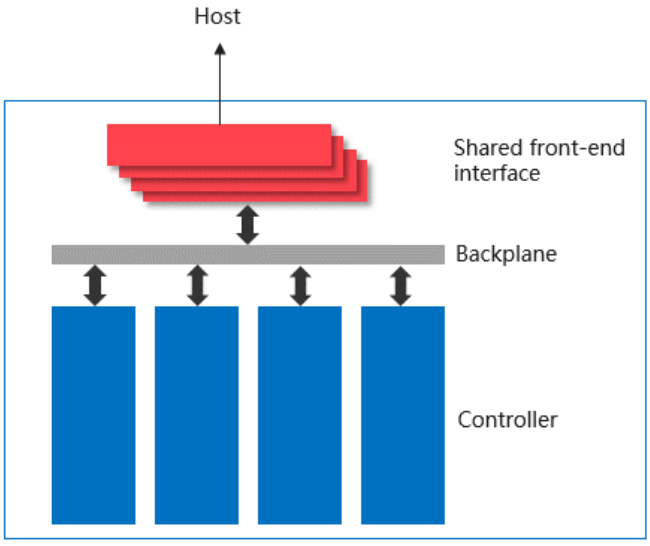
3.3.2 SmartMatrix

OceanStor Dorado V6 series uses the Smart Matrix multi-controller architecture. Controller enclosures can be scaled out to achieve linear increase in performance and capacity. Controllers in one controller enclosure(engine) are interconnected using onboard 100Gb RDMA links. Multiple controller enclosure are interconnected by 25/100 Gb RDMA links directly or via RDMA switches.

SmartMatrix Multi-Controller architecture

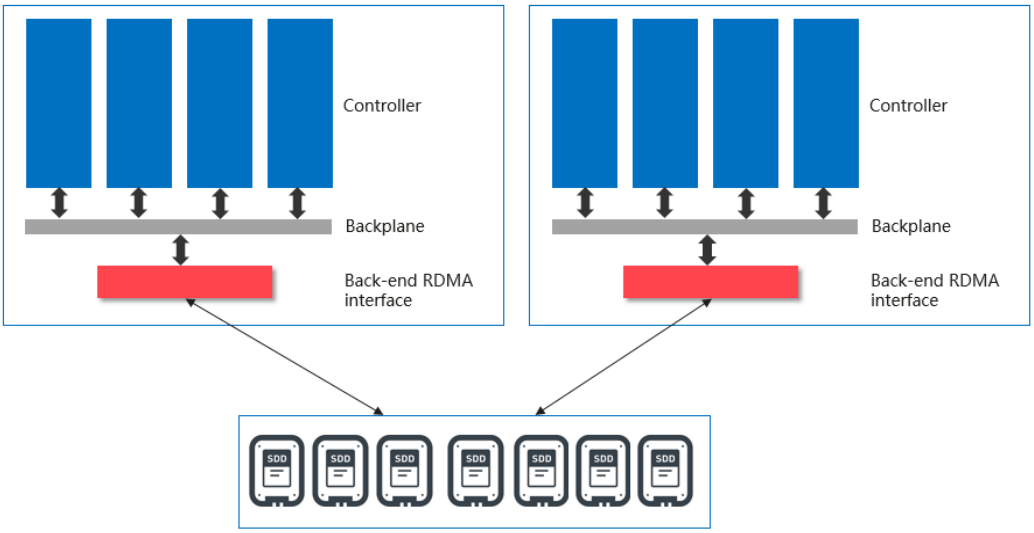


Dorado8000 V6 supports shared front-end interfaces fully connected to all 4 controllers via backplane in a controller enclosure(engine).Base on the architecture of shared front-end interface, Dorado8000 V6 supports continuous front-end link-up if one or more controllers fail. That means there is no impact including host link event when controller fails.



Shared front-end interface connecting to
4 controllers in Dorado8000 V6 engine

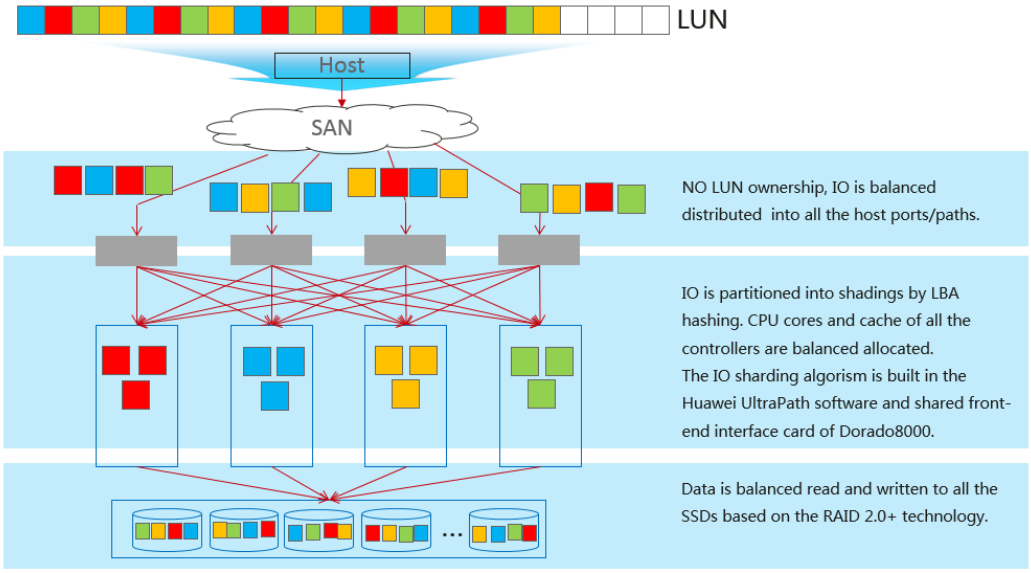
Dorado8000 V6(NVMe) supports cross-engine shared backend connecting which means SSD enclosures are connected to back-end interfaces of 8 controllers(2 engines) via 100Gb RDMA. Base on the cross-engine shared backend architecture, Dorado8000 V6(NVMe) provides the highest reliability in the industry which supports NO business interruption when multiple controllers fail(up to 7 controllers of 8) and one engine fails.



SSD enclosures shared connected to back-end of 8 controllers(2 engines)

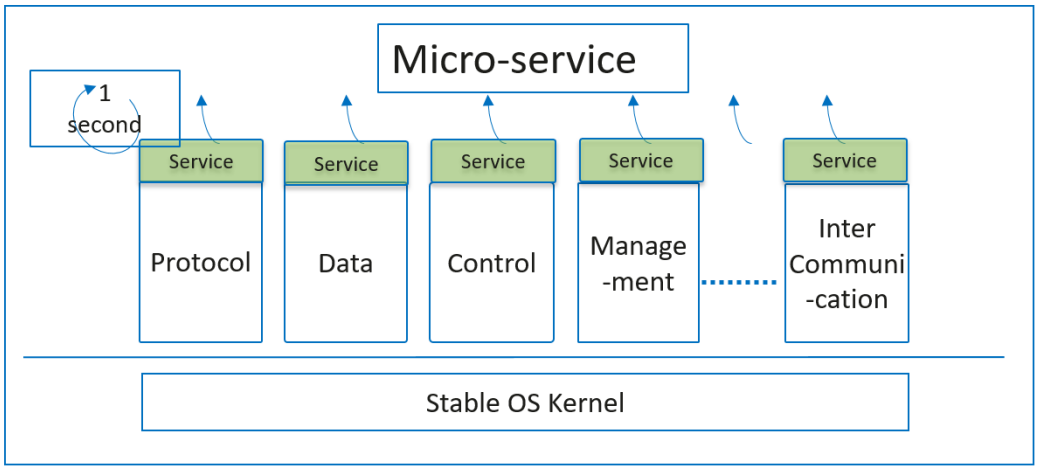
3.3.3 All balanced Active-Active architecture

Dorado5000/6000/8000 V6 is designed as an distributed all balanced active-active architecture, where all the system resources including host links, interfaces, controllers, CPU cores, cache, SSDs, are balanced allocated for the business processing. There is no ownership controller for a LUN, where the data of a LUN is evenly fragmented into 64MB shardings by LUN LBA hashing. The data shardings of a LUN are processed by all the controllers, where the CPU cores and cache resources can be leveraged evenly and globally. Base on the RAID2.0+ technology, the data after global deduplication and compression is evenly distributed into all the SSDs, where the SSDs undertake the same workload and wear leveling, to provide the same performance and lifetime.



3.3.4 In-Place Upgrading (NDU)

Dorado V6 supports 94% components in user mode, upgrading in 1s, no host connection loss by switching on global sharing frontend card, 6% components in steady kernel, upgrading with rebooting in minutes.



3.3.5 Value-added Features

OceanStor Dorado V6 provides powerful value-added features as the following, all the features can be enabled and share all the storage system resources like cpu and cache:

- Smart series software includes SmartDedupe, SmartCompression, SmartThin, SmartVirtualization, and SmartMigration, which improve storage efficiency and reduce user TCO.

- Hyper series software includes HyperSnap, HyperClone, HyperReplication, HyperMetro, HyperVault, and HyperLock, which provide disaster recovery and data backup.
- Cloud series software includes CloudBackup, which construct cost-effective cloud DR centers to reduce the OPEX.

3.3.6 Software Architecture Highlights

- Excellent performance

  FlashLink® realizes efficient I/O scheduling, providing high performance and low system latency.
- Stable and reliable

  Innovative RAID algorithms, value-added features, and multi-level reliability solutions ensure 99.9999% reliability and 24/7 stable service system operation.
- Efficient

  Multiple efficiency-improving features, such as heterogeneous virtualization and inline deduplication and compression, protect customers' investments.

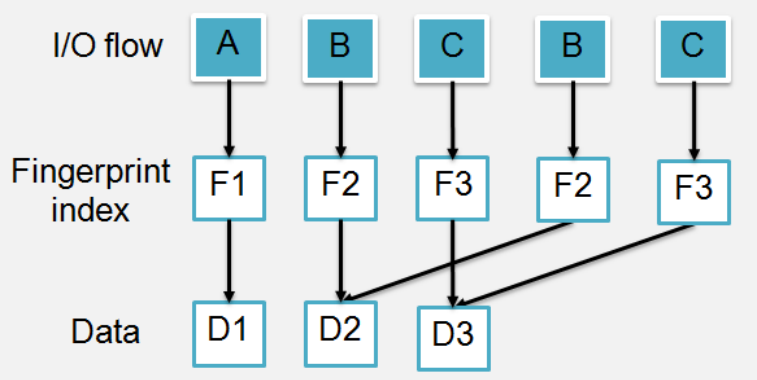
4 Smart Series Features

4.1 SmartDedupe (Inline Deduplication)

SmartDedupe allows OceanStor Dorado V6 to delete duplicate data online before writing data to flash media. The deduplication process is as follows:

The storage system divides the new data into blocks based on the deduplication granularity. Then for each block, the system calculates its fingerprint and compares it with the existing fingerprints. If the same fingerprint is found, the system reads the data corresponding to the fingerprint and compares that saved data to the new data block, byte by byte. If they are the same, the system increases the reference count of the fingerprint and does not write the new data block to the SSDs. If the fingerprint is not found or byte-by-byte comparison is not passed, the system writes the new data block to SSDs and records the mapping between the fingerprint and storage location.

Figure 4-1 Working principle of deduplication



SmartDedupe on OceanStor Dorado V6 has the following highlights:

- OceanStor Dorado V6 supports 4 KB and 8 KB deduplication granularities. You can enable or disable SmartDedupe on particular LUNs.

  The deduplication ratio depends on the application scenarios and user data contents. For applications that provide a high deduplication ratio (for example, VDI), it is recommended that you enable SmartDedupe and use 8 KB deduplication granularity to save space. In scenarios where the deduplication ratio is low, such as for databases, you can disable SmartDedupe to improve performance.
- OceanStor Dorado V6 supports byte-by-byte comparison to ensure data reliability.
- OceanStor Dorado V6 can identify zero data, which occupies no storage space.

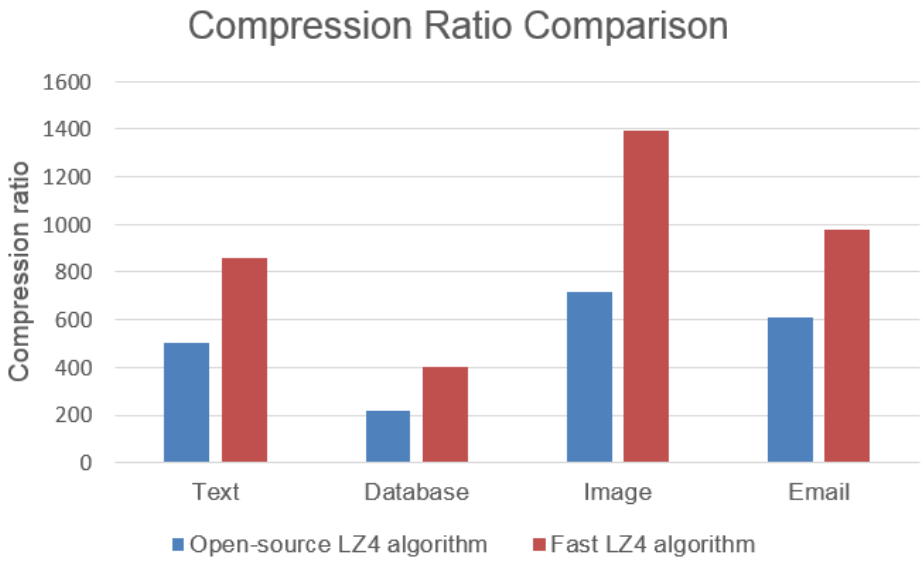
  When an application is reading data, zero is returned if a mapping relationship does not exist between LBA and the fingerprints. When an application writes zero data blocks, an internal zero page that requires no storage space is used to replace the zero data, improving the space utilization and system performance.

4.2 SmartCompression (Inline Compression)

SmartCompression compresses data online before writing data to flash media. In addition, compression is performed after deduplication, ensuring that no duplicate data is compressed and improving compression efficiency. SmartCompression reduces the amount of data written to SSDs and minimizes write amplification, improving the longevity of flash arrays.

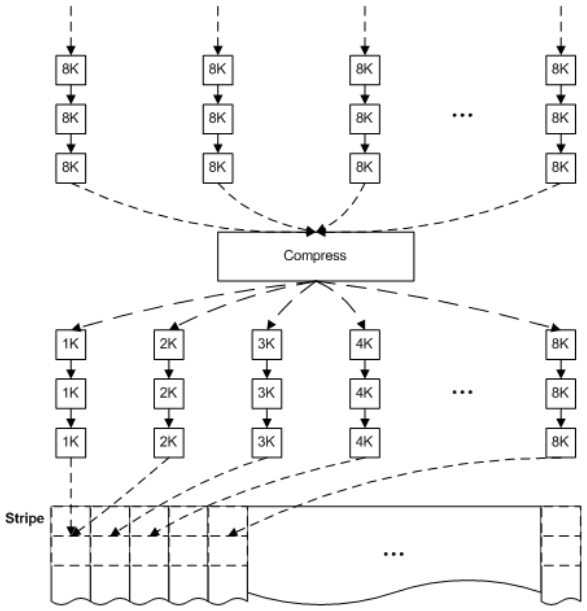
The compression algorithm is a compute-intensive program. Inline compression consumes significant CPU resources, affecting end-to-end performance of the system. Open-source compression algorithms that feature high performance and low compression ratio are commonly used in the industry, for example, LZ4, LZO, and Snappy. OceanStor Dorado V6 uses the Fast LZX, LZ4, and ZTSD algorithms, which are improvements of the open-source LZX, LZ4, and ZTSD compression algorithms and double the compression efficiency without decreasing the compression ratio.

Figure 4-2 Comparison between open-source and Fast LZ4 algorithms



The size of data blocks to be compressed can be 4 KB, 8 KB, 16 KB, and 32 KB. The compressed data is aligned by byte, which improves the compression efficiency and saves the storage space for compressed data. In the following figure, 8 KB data blocks are compressed, converged into full stripes, and then written to disks.

Figure 4-3 Working principle of compression



The compression ratio of OceanStor Dorado V6 also depends on user data. You can enable or disable SmartCompression for each specific LUN. In applications that require high performance, you can disable this function.

4.3 SmartThin (Intelligent Thin Provisioning)

OceanStor Dorado V6 supports thin provisioning, which enables the storage system to allocate storage resources on demand. SmartThin does not allocate all capacity in advance, but presents a virtual storage capacity larger than the physical storage capacity. This allows you to see a larger storage capacity than the actual storage capacity. When you begin to use the storage, SmartThin provides only the required space. If the storage space is about to use up, SmartThin triggers storage resource pool expansion to add more space. The expansion process is transparent to users and causes no system downtime. When the percentage of the LUN's used capacity to its total capacity exceeds Alarm Threshold (%), a threshold alarm will be created and sent to the application server. The threshold can be set from 50 to 99. OceanStor Droado V6 supports to change thin lun space transparently with no system downtime.

Application Scenarios

- SmartThin can help core service systems that have demanding requirements on business continuity, such as bank transaction systems, to expand system capacity non-disruptively without interrupting ongoing services.
- For services where the growth of application system data is hard to evaluate accurately, such as email services and web disk services, SmartThin can assist with on-demand physical space allocation, preventing wasted space.
- For mixed services that have diverse storage requirements, such as carrier services, SmartThin can assist with physical space contention, achieving optimized space allocation.

4.4 SmartQoS (Intelligent Quality of Service Control)

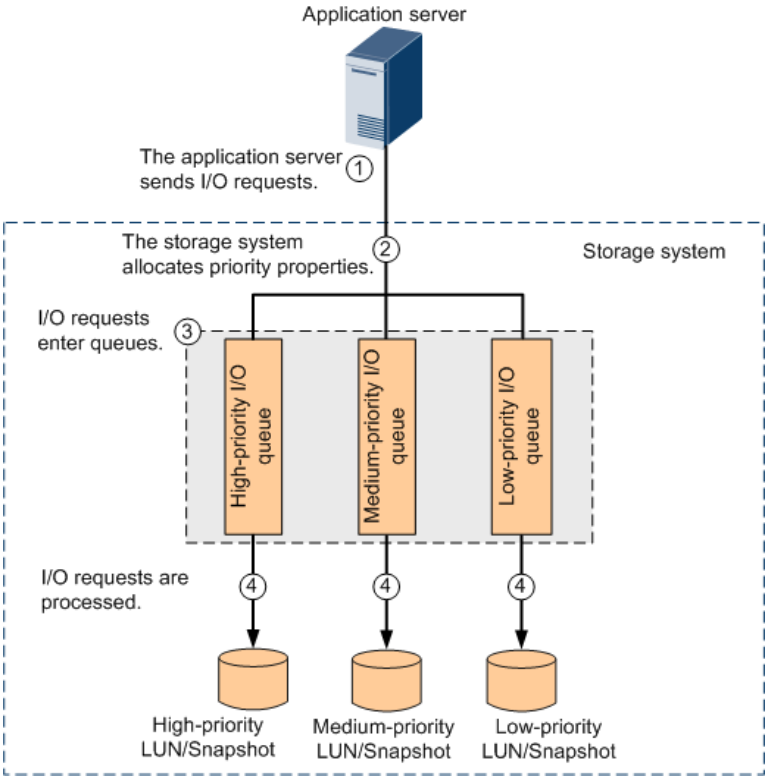
SmartQoS dynamically allocates storage system resources to meet the performance objectives of applications. You can set upper limits on IOPS or bandwidth for specific applications. Based on the upper limits, SmartQoS can accurately limit performance of these applications, preventing them from contending for storage resources with critical applications.

SmartQoS uses LUN- or snapshot-specific I/O priority scheduling and the I/O traffic control to guarantee the service quality.

I/O Priority Scheduling

This schedules resources based on applications' priorities, prioritizing applications with higher priorities in resource allocation to ensure their SLAs when storage resources are insufficient. You can configure an application as high, medium, or low priority.

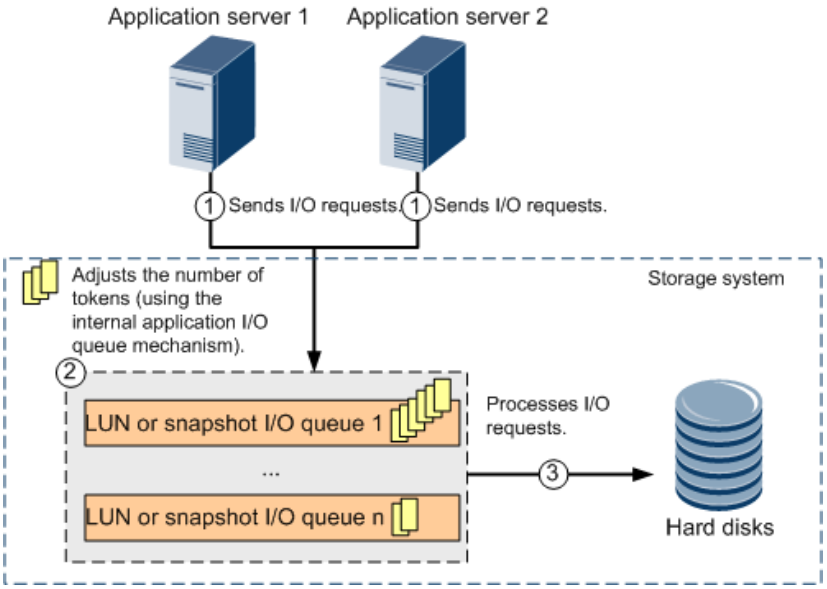
Figure 4-4 I/O priority scheduling process



I/O Traffic Control

This limits traffic of some applications by limiting their IOPS or bandwidth, thereby preventing these applications from affecting other applications. I/O traffic control is implemented based on hierarchical management, objective distribution, and traffic control management.

Figure 4-5 Managing LUN or snapshot I/O queues



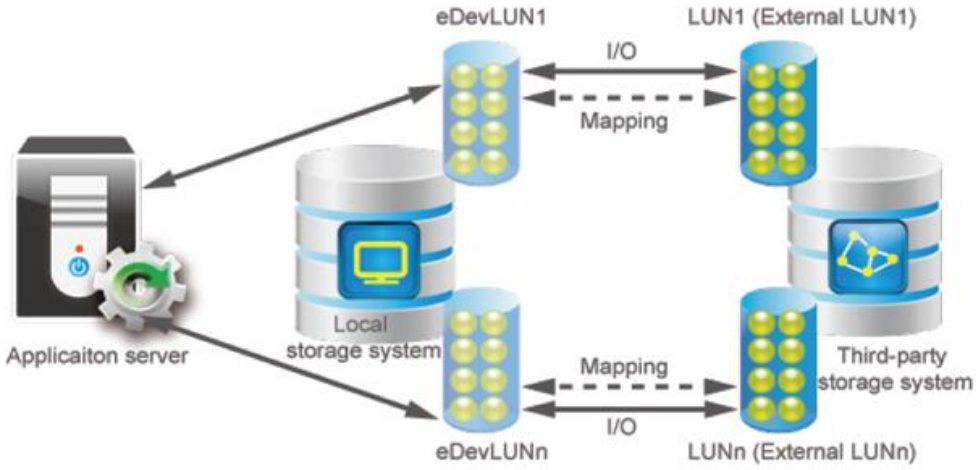
4.5 SmartVirtualization (Heterogeneous Virtualization)

OceanStor Dorado V6 uses SmartVirtualization to take over heterogeneous storage systems (including other Huawei storage systems and third-party storage systems), protecting customer investments. SmartVirtualization conceals the software and hardware differences between the local and heterogeneous storage systems, allowing the local system to use and manage the heterogeneous storage resources as its local resources. In addition, SmartVirtualization can work with SmartMigration to migrate data from heterogeneous storage systems online, facilitating device replacement.

The working principles of SmartVirtualization are as follows:

SmartVirtualization maps the heterogeneous storage system to the local storage system, which then uses external device LUNs (eDevLUNs) to take over and manage the heterogeneous resources. eDevLUNs consist of metadata volumes and data volumes. The metadata volumes manage the data storage locations of eDevLUNs and use physical space provided by the local storage system. The data volumes are logical presentations of external LUNs and use physical space provided by the heterogeneous storage system. An eDevLUN on the local storage system matches an external LUN on the heterogeneous storage system. Application servers access data on the external LUNs via the eDevLUNs.

Figure 4-6 Heterogeneous storage virtualization



SmartVirtualization uses LUN masquerading to set the WWNs and Host LUN IDs of eDevLUNs on OceanStor Dorado V6 to the same values as those on heterogeneous storage system. After data migration is complete, the host's multipathing software switches over the LUNs online without interrupting services.

Application Scenarios

- Heterogeneous array takeover

  As customers build data centers over time, the storage arrays they use may come from different vendors. Storage administrators can leverage SmartVirtualization to manage and configure existing devices, protecting investments.

- Heterogeneous data migration

  The customer may need to replace storage systems whose warranty periods are about to expire or whose performance does not meet service requirements. SmartVirtualization and SmartMigration can migrate customer data to OceanStor Dorado V6 online without interrupting host services.

4.6 SmartMigration (Intelligent Data Migration)

OceanStor Dorado V6 provides intelligent data migration based on LUNs. Data on a source LUN can be completely migrated to a target LUN without interrupting ongoing services. SmartMigration also supports data migration between a Huawei storage system and a compatible heterogeneous storage system.

When the system receives new data during migration, it writes the new data to both the source and target LUNs simultaneously and records data change logs (DCLs) to ensure data consistency. After the migration is complete, the source and target LUNs exchange information to allow the target LUN to take over services.

SmartMigration is implemented in two stages:

1. Data synchronization
   a. Before migration, you must configure the source and target LUNs.

b. When migration starts, the source LUN replicates data to the target LUN.

c. During migration, the host can still access the source LUN. When the host writes data to the source LUN, the system records the DCL.

d. The system writes the incoming data to both the source and target LUNs.

- If writing to both LUNs is successful, the system clears the record in the DCL.
- If writing to the target LUN fails, the storage system identifies the data that failed to be synchronized according to the DCL and then copies the data to the target LUN. After the data is copied, the storage system returns a write success to the host.
- If writing to the source LUN fails, the system returns a write failure to notify the host to re-send the data. Upon reception, the system only writes the data to the source LUN.

2. LUN information exchange

After data replication is complete, host I/Os are suspended temporarily. The source and target LUN exchanges information as follows:

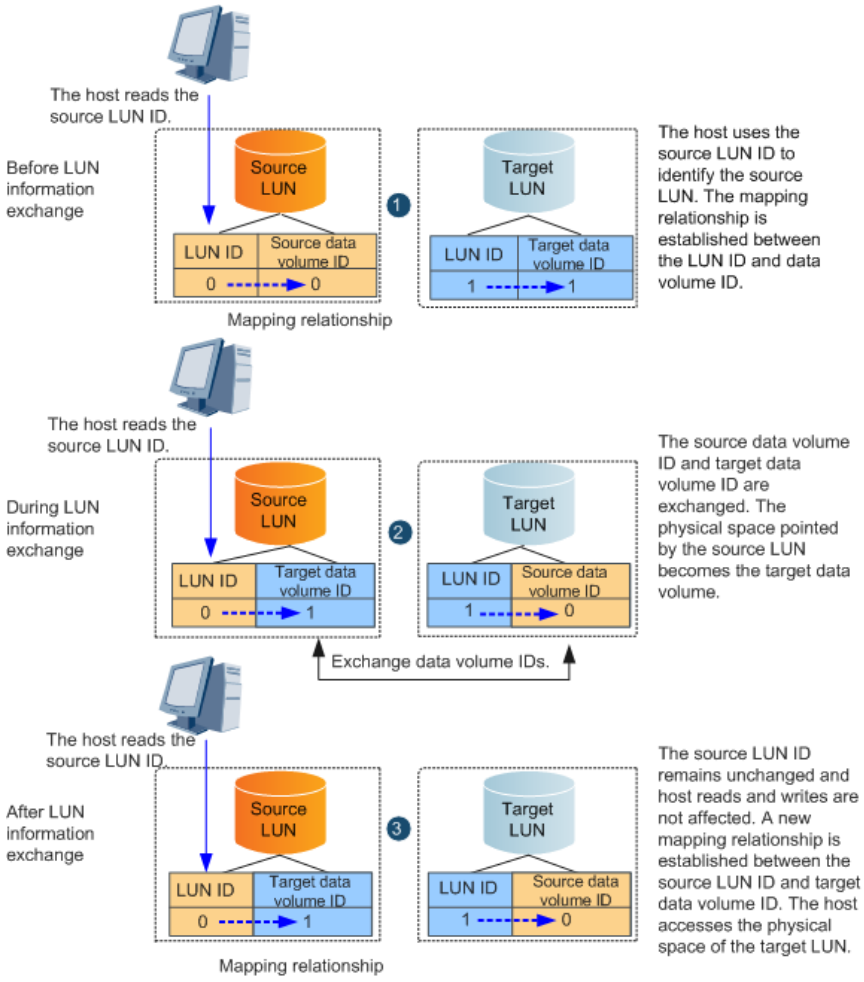
a. Before LUN information is exchanged, the host uses the source LUN ID to identify the source LUN. Because of the mapping relationship between the source LUN ID and the source data volume ID used to identify physical space, the host can read the physical space information about the source LUN. The mapping relationship also exists between the target LUN ID and target data volume ID.

b. In LUN information exchange, the source and target LUN IDs remain unchanged but the data volume IDs of the source and target LUNs are exchanged. This creates a new mapping relationship between the source LUN ID and target data volume ID.

c. After the exchange, the host can still identify the source LUN using the source LUN ID but reads physical space information about the target LUN due to the new mapping relationship.

LUN information exchange is completed instantaneously, which does not interrupt services.

Figure 4-7 LUN information exchange



Application Scenarios

- Storage system upgrade with SmartVirtualization

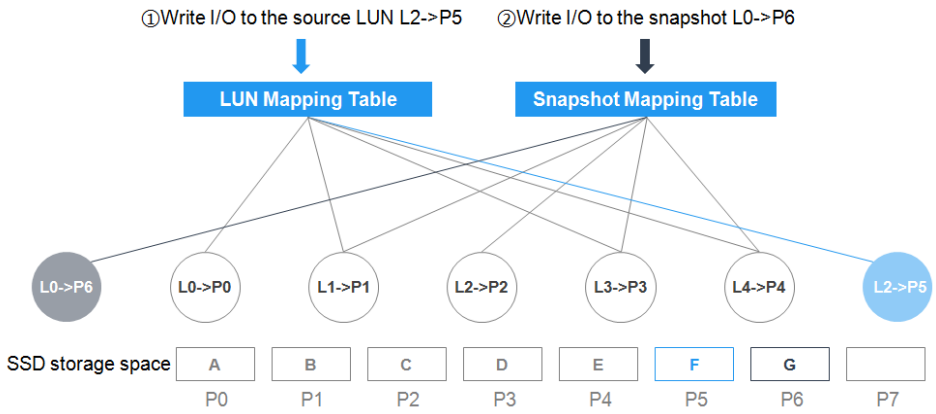
  SmartMigration works with SmartVirtualization to migrate data from legacy storage systems (from Huawei or other vendors) to new Huawei storage systems to improve service performance and data reliability.
- Data migration for capacity, performance, and reliability adjustments

5 Hyper Series Features

5.1 HyperSnap (Snapshot)

OceanStor Dorado V6 implements lossless snapshot using ROW. When snapshot data is changed, OceanStor Dorado V6 writes new data to new locations and does not need to copy the old data, reducing system I/O overhead.

Figure 5-1 ROW snapshot principle



In Figure 5-1, both the source LUN and snapshot use a mapping table to access the physical space. The original data in the source LUN is **ABCDE** and is saved in sequence in the physical space. The metadata of the snapshot is null. All read requests to the snapshot are redirected to the source LUN.

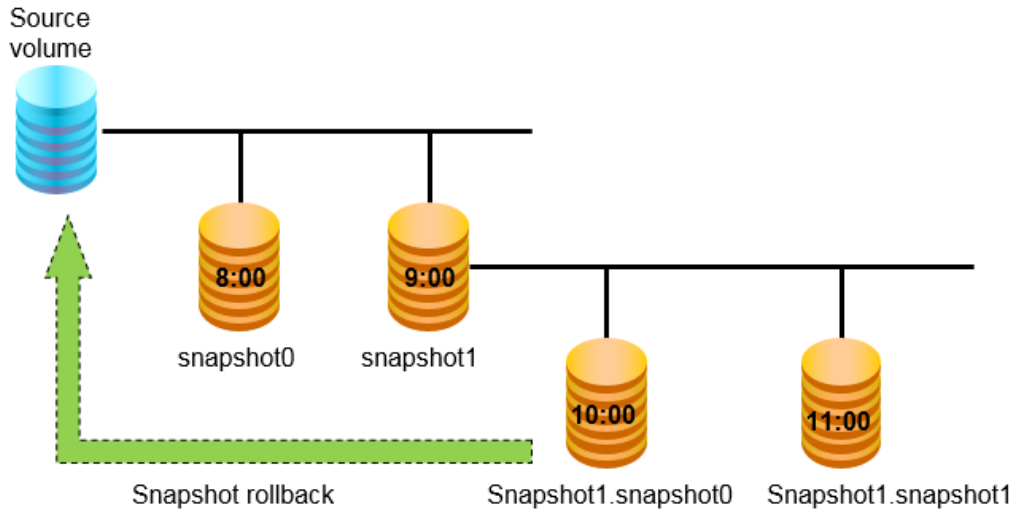
- When the source LUN receives a write request that changes **C** to **F**, the new data is written into a new physical space **P5** instead of being overwritten in **P2**. In the mapping metadata of the source LUN, the system changes **L2->P2** to **L2->P5**.

- If the snapshot must be modified, for example, **A** corresponding to **L0** must be changed to **G**, the system first writes **G** to **P6** and then changes **L0->P0** in the snapshot mapping table to **L0->P6**. Data in the source LUN is changed to **ABFDE** and data in the snapshot is changed to **GBCDE**.

HyperSnap implements writable snapshots by default. All snapshots are readable and writable, and support snapshot copies and cascading snapshots. You can create a read-only copy of a snapshot at a specific point in time, or leverage snapshot cascading to create child snapshots for a parent snapshot. For multi-level cascading snapshots that share a source volume, they can roll back each other and the source volume regardless of their cascading levels. This is called cross-level rollback.

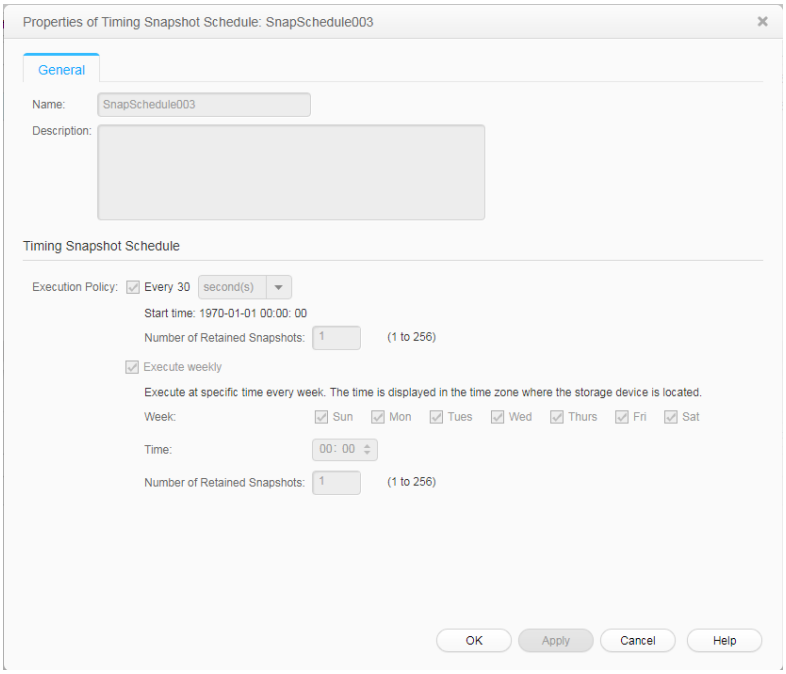
In Figure 5-2, **Snapshot1** is created for the source volume at 9:00, and **Snapshot1.snapshot0** is a cascading snapshot of **Snapshot1** at 10:00. The system can roll back the source volume using **Snapshot1.snapshot0** or **Snapshot1**, or roll back **Snapshot1** using **Snapshot1.snapshot0**.

Figure 5-2 Cascading snapshot and cross-level rollback



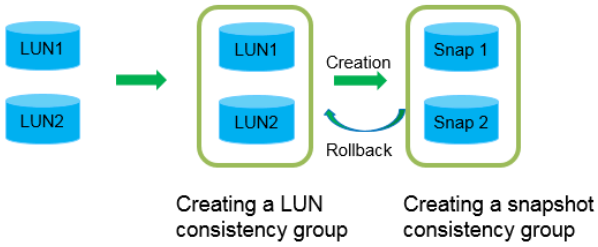
HyperSnap supports timed snapshots, which can be triggered weekly, daily, or at a custom interval (with a minimum interval of 30 seconds). The system supports multiple schedules and each schedule can have multiple source LUNs. Snapshots of the sources LUNs that share a schedule are in the same consistency group.

Figure 5-3 Configuring timing snapshot



HyperSnap supports snapshot consistency groups. For LUNs that are dependent on each other, you can create a snapshot consistency group for these LUNs to ensure data consistency. For example, the data files, configuration files, and logs of an Oracle database are usually saved on different LUNs. Snapshots for these LUNs must be created at the same time to guarantee that the snapshot data is consistent in time.

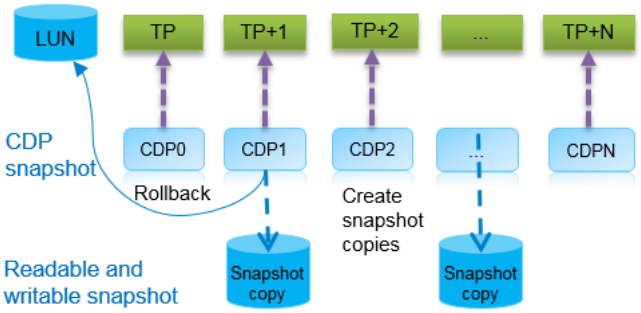
Figure 5-4 Snapshot consistency group



5.2 HyperCDP (Continuous Data Protection)

HyperCDP allows OceanStor Dorado V6 to generate high-density snapshots, which are also called HyperCDP objects. The minimum interval of HyperCDP objects is 10 seconds, which ensures continuous data protection and reduces the RPO. HyperCDP is based on the lossless snapshot technology (multi-time-point and ROW). Each HyperCDP object matches a time point of the source LUN. Dorado V6 supports HyperCDP schedules to meet customers' backup requirements.

Figure 5-5 HyperCDP snapshot principles



Technical Highlights:

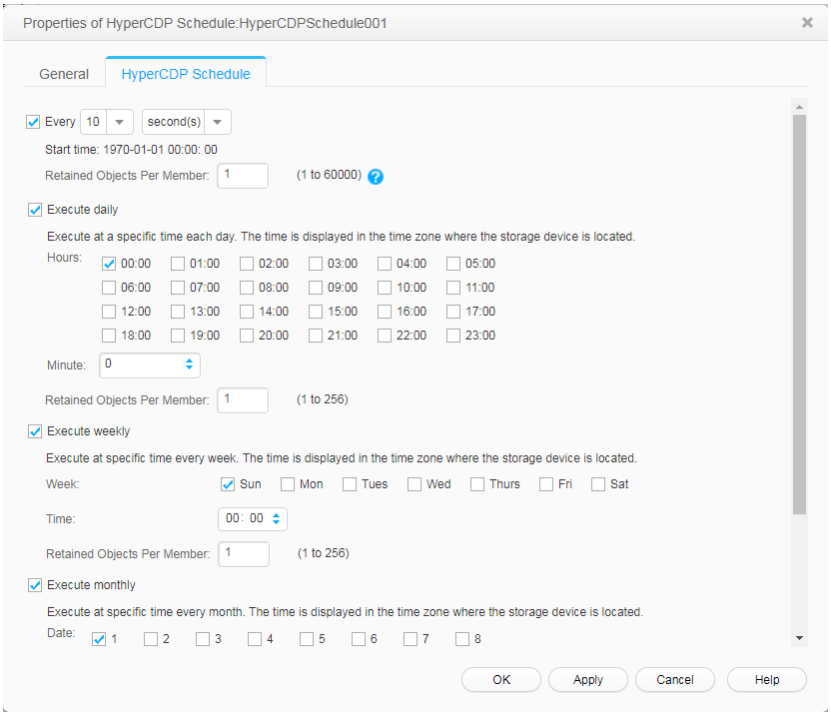
- Continuous protection, lossless performance

  HyperCDP provides data protection at an interval of seconds, with zero impact on performance and small space occupation.

- Support for scheduled tasks

  You can specify HyperCDP schedules by day, week, month, or specific interval, meeting different backup requirements.

Figure 5-6 HyperCDP schedule



- Intensive and persistent data protection

  A single LUN supports 60,000 HyperCDP objects. The minimum interval is 10 seconds. At this setting, continuous protection can be achieved for data within a week.

- Support for consistency groups

In database applications, the data, configuration files, and logs are usually saved on different LUNs. The HyperCDP consistency group ensures data consistency between these LUNs during restoration.

- HyperCDP duplicate for reads and writes

  Hosts cannot read or write HyperCDP objects directly. To read a HyperCDP object, you can create a duplicate for it and then map the duplicate to the host. The duplicate has the same data as the source HyperCDP object and can be read and written by the host. In addition, the duplicate can be rebuilt by a HyperCDP object at any time point to obtain the data at that time.

There are some restrictions when HyperCDP is used with other features of OceanStor Dorado V6.

Table 5-1 Restrictions of HyperCDP used with other features

Feature	Restriction
HyperSnap	• Source LUNs of HyperSnap can be used as the source LUNs of HyperCDP, but snapshot LUNs of HyperSnap cannot be used as the source LUNs of HyperCDP. • HyperCDP objects cannot be used as the source LUNs of HyperSnap.
HyperMetro	• Member LUNs of HyperMetro can be used as the source LUNs of HyperCDP, but HyperCDP objects cannot be used as member LUNs of HyperMetro. • HyperCDP rollback cannot be performed during HyperMetro synchronization.
HyperReplication	• The primary and secondary LUNs of HyperReplication can be used as the source LUNs of HyperCDP, but HyperCDP objects cannot be used as the primary or secondary LUNs of HyperReplication. • HyperCDP rollback cannot be performed during HyperReplication synchronization.
SmartMigration	Source LUNs of HyperCDP and HyperCDP objects cannot be used as the source or target LUNs of SmartMigration.
HyperClone	Source LUNs of HyperClone can be used as the source LUNs of HyperCDP. Before clone LUNs are split, they cannot be used as the source LUNs of HyperCDP.
SmartVirtualization	Heterogeneous LUNs cannot be used as the source LUNs of HyperCDP.

5.3 HyperCopy (Copy)

OceanStor Dorado V6 supports HyperCopy, which allows the system to create a complete physical copy of the source LUN's data on the target LUN. The source and target LUNs that form a HyperCopy pair must have the same capacity. The target LUN can either be empty or

have existing data. If the target LUN has data, the data will be overwritten by the source LUN during synchronization. After the HyperCopy pair is created, you can synchronize data. During the synchronization, the target LUN can be read and written immediately. HyperCopy supports consistency groups, incremental synchronization, incremental restoration, providing full backup for source LUNs. HyperCopy allows data copy between controllers, but does not support copy between different arrays.

HyperCopy is typically applied to:

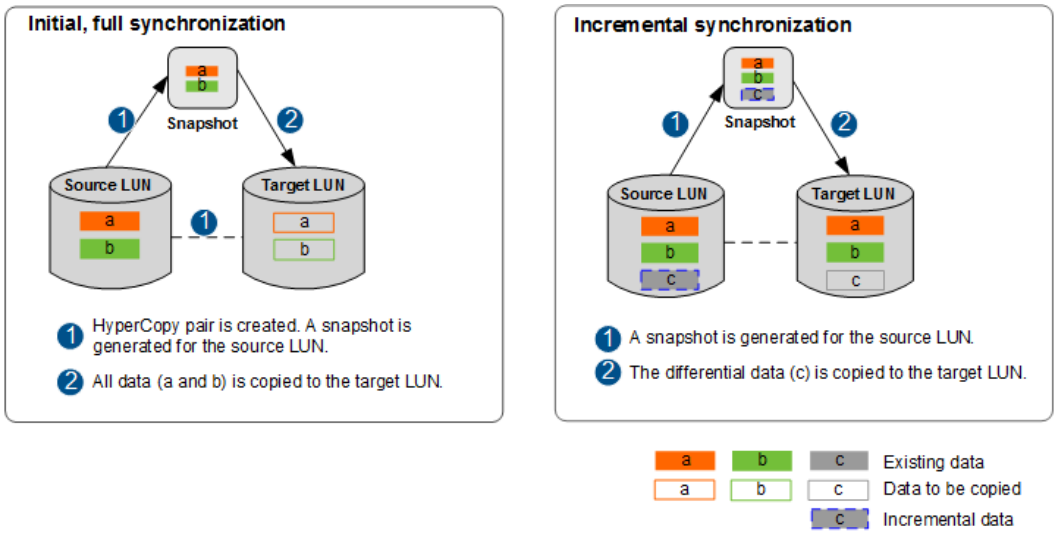
- Data backup and restoration
- Data analysis
- Data reproduction

Data Synchronization After HyperCopy Is Created

When data synchronization starts, the system generates an instant snapshot for the source LUN, and then synchronizes the snapshot data to the target LUN. Any subsequent write operations are recorded in a differential table. When synchronization is performed again, the system compares the data of the source and target LUNs, and only synchronizes the differential data to the target LUN. The data written to the target LUN between the two synchronizations will be overwritten. To retain the existing data on the target LUN, you can create a snapshot for it before synchronization.

The following figure illustrates the synchronization principle.

Figure 5-7 Data synchronization from the source LUN to the target LUN

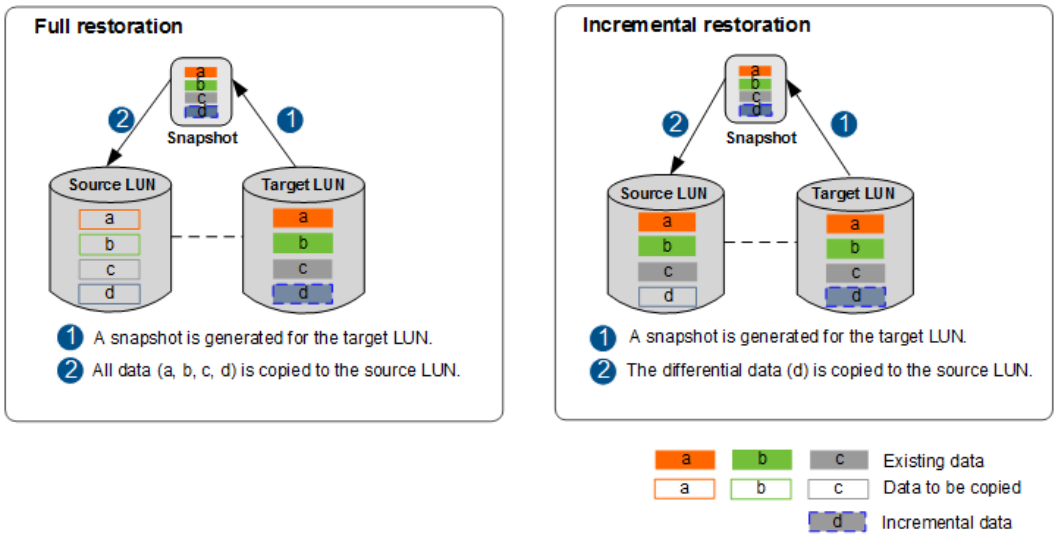


Restoration

If the source LUN is damaged, data on the target LUN can be restored to the source LUN. Restoration also supports full and incremental data synchronization. When restoration starts, the system generates a snapshot for the target LUN and synchronizes the snapshot data to the source LUN. For incremental restoration, the system compares the data of the source and target LUNs, and only synchronizes the differential data.

The following figure illustrates the restoration principle.

Figure 5-8 Restoration from the target LUN to the source LUN



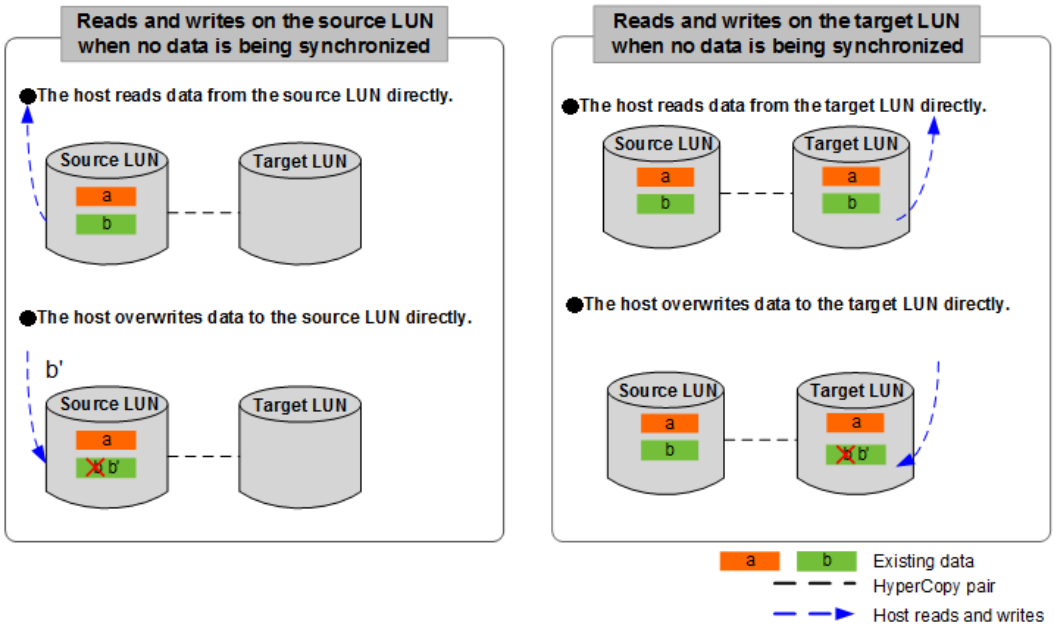
Immediate Read and Write

Read and write I/Os are processed in different ways when HyperCopy is or is not synchronizing data.

- When HyperCopy is not synchronizing data:

  The host reads and writes the source or target LUN directly.

Figure 5-9 Reads and writes when HyperCopy is not synchronizing data



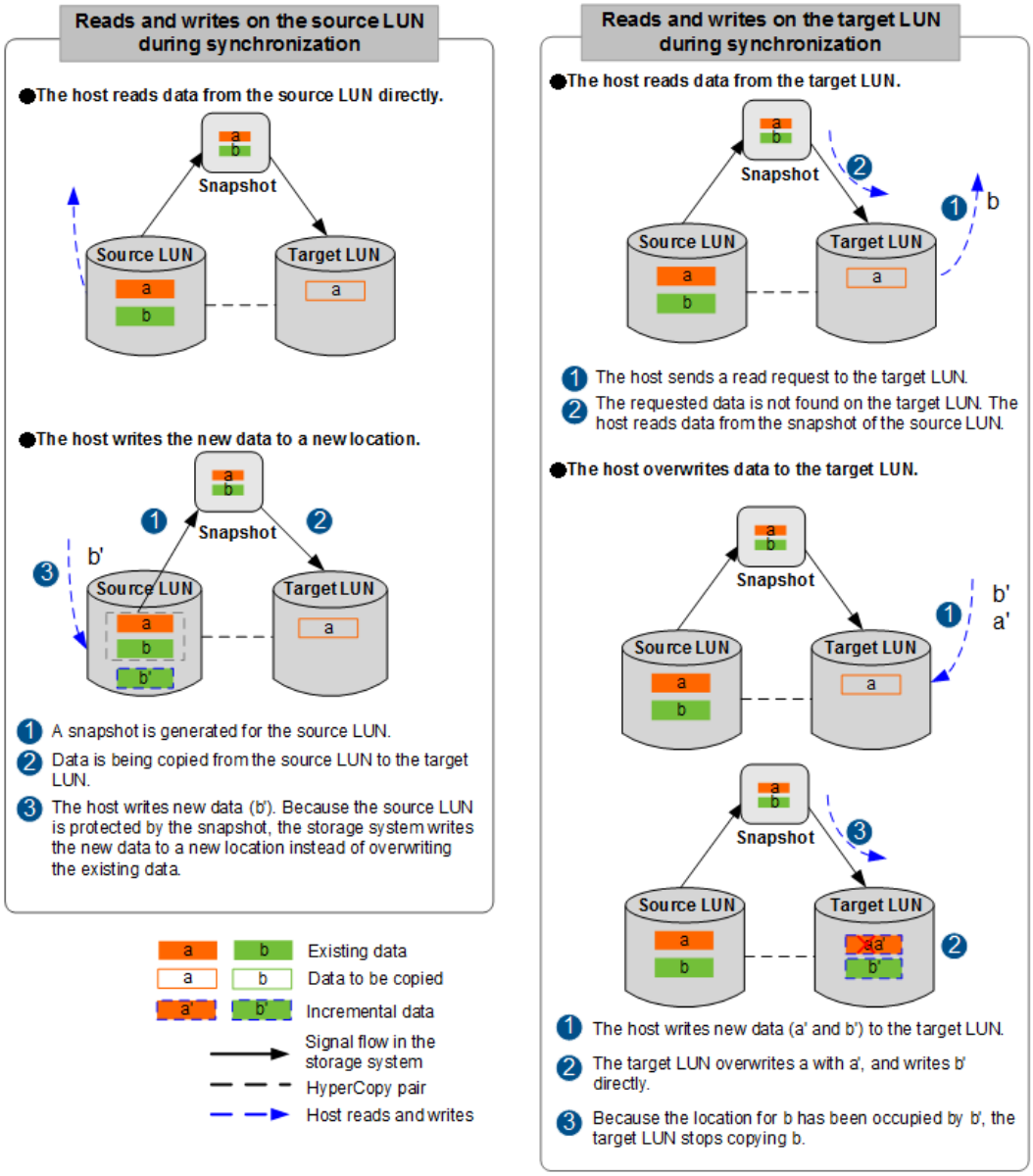
- When HyperCopy is synchronizing data:

  The host reads and writes the source LUN directly.

For read operations on the target LUN, if the requested data is hit on the target LUN (the data has been synchronized), the host reads the data from the target LUN. If the requested data is not hit on the target LUN (the data has not been synchronized), the host reads the data from the snapshot of the source LUN.

For write operations on the target LUN, if a data block has been synchronized before the new data is written, the system overwrites this block. If a data block has not been synchronized, the system writes the new data to this block and stops synchronizing the source LUN's data to it. This ensures that the target LUN can be read and written before the synchronization is complete.

Figure 5-10 Reads and writes when HyperCopy is synchronizing data



Consistency Group

You can add multiple HyperCopy pairs to a consistency group. When you synchronize or restore a consistency group, data on all member LUNs is always at a consistent point in time, ensuring data integrity and availability.

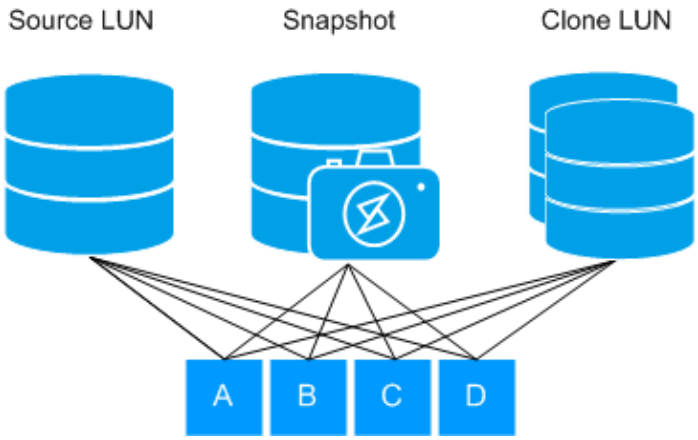
5.4 HyperClone (Clone)

HyperClone generates a complete physical data copy of the source LUN or snapshot, which can be used for development and testing without affecting the source LUN or snapshot.

After a clone LUN is created, it immediately shares the data with the source LUN and can be mapped to hosts for data access. You can split the clone LUN to stop data sharing with the source LUN and obtain a full physical copy of data. Hosts can read and write the clone LUN non-disruptively during and after the splitting. You can also cancel the splitting before it is complete to reclaim the storage space occupied by the physical copy and retain data sharing between the source and clone LUNs.

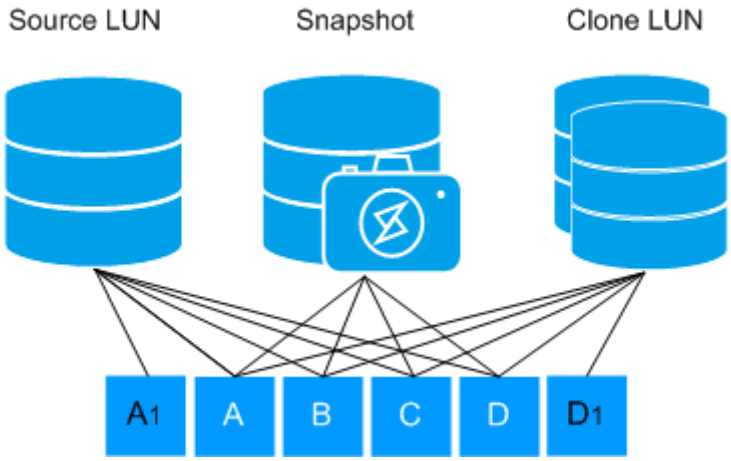
HyperClone is implemented based on snapshots. When a clone LUN is created, the system creates a readable and writable snapshot of the source LUN. The source and clone LUNs share data. When an application server reads data from the clone LUN, it actually reads the source LUN's data.

Figure 5-11 Clone LUN's data before data changes



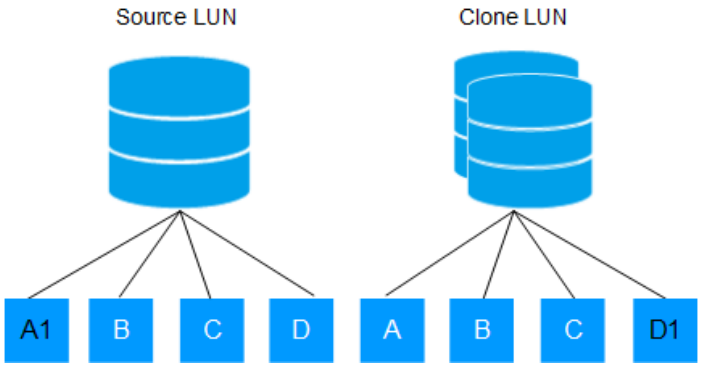
When an application server writes new data to the source or clone LUN, the storage system leverages ROW, which allocates a new storage space for the new data instead of overwriting the data in the existing storage space. As shown in Figure 5-12, when the application server attempts to modify data block A in the source LUN, the storage pool allocates a new block (A1) to store the new data, and retains the original block A. Similarly, when the application server attempts to modify block D in the clone LUN, the storage pool allocates a new block (D1) to store the new data, and retains the original block D.

Figure 5-12 Clone LUN's data after data changes



When a clone LUN is split, the storage system copies the data that the clone LUN shares with the source LUN to new data blocks, and retains the new data that has been written to the clone LUN. After splitting, the association between the source and clone LUNs is canceled and the clone LUN becomes an independent physical copy.

Figure 5-13 Clone LUN after splitting



OceanStor Dorado V6 supports consistent clones. For LUNs that are dependent on each other, for example, LUNs that save the data files and logs of a database, you can create clones for these LUNs' snapshots that were activated simultaneously to ensure data consistency between the clones.

Both HyperClone and HyperCopy can create a complete copy of data. The following table compares their similarities and differences.

Table 5-2 Comparison between HyperClone and HyperCopy

Item	HyperClone	HyperCopy
Copy type	Clone LUN	Copy relationship between the source and target LUNs
Immediate availability	Yes	Yes

Item	HyperClone	HyperCopy
Synchronization mode	No synchronization	Full and incremental synchronization and restoration
Consistency group	Not supported To ensure consistency of clones, you must create clones for consistently activated snapshots of the source LUNs.	Supported
Scope	Clones cannot be created between different controller pairs or storage pools.	Data copy can be performed between different controller pairs or storage pools.

5.5 HyperReplication (Remote Replication)

5.5.1 HyperReplication/S for Block (Synchronous Remote Replication)

OceanStor Dorado V6 supports synchronous remote replication between storage systems. HyperReplication/S writes each host's write I/O to both the primary and secondary LUNs concurrently and returns a write success acknowledgement to the host after the data is successfully written to both LUNs.

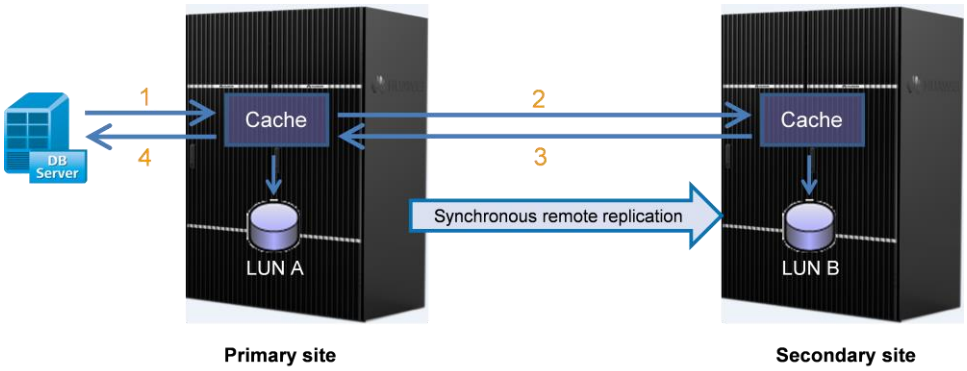
The general principles are as follows:

1. After a remote replication relationship is established between the primary and secondary LUNs, an initial synchronization is implemented to replicate all data from the primary LUN to the secondary LUN.
2. If the primary LUN receives a write request from the host during initial synchronization, the new data is written to both the primary and secondary LUNs.
3. After initial synchronization, data on the primary LUN is the same as that on the secondary LUN.

The following shows how I/Os are processed in synchronous remote replication.

1. The primary site receives a write request from the host. HyperReplication logs the address information instead of the data content.
2. The data of the write request is written to both the primary and secondary LUNs. If the LUNs use write-back, the data will be written to the cache.
3. HyperReplication waits for the write results of the primary and secondary LUNs. If the data has been successfully written to the primary and secondary LUNs, HyperReplication deletes the log. Otherwise, HyperReplication retains the log, and the data block enters the interrupted state and will be replicated in the next synchronization.
4. HyperReplication returns the write result of the primary LUN to the host.

Figure 5-14 I/O processing in synchronous remote replication



Technical Highlights

- Zero data loss

  HyperReplication/S updates data in the primary and secondary LUNs simultaneously, ensuring zero RPO.

- Split mode

  HyperReplication/S supports split mode, where write requests of production hosts go only to the primary LUN, and the difference between the primary and secondary LUNs is recorded by the differential log. If you want to resume data consistency between the primary and secondary LUNs, you can manually start synchronization, during which data blocks marked as differential in the differential log are copied from the primary LUN to the secondary LUN. The I/O processing is similar to the initial synchronization. This mode meets user requirements such as temporary link maintenance, network bandwidth expansion, and saving data at a certain time on the secondary LUN.

- Quick response and recovery

  HyperReplication/S immediately enters the **Interrupted** state in case of a system fault such as a link down failure or I/O error due to faults of the primary or secondary LUN. In the **Interrupted** state, I/Os are processed similarly to in split mode. That is, data is written only to the primary LUN and the data difference is recorded. If the primary LUN fails, it cannot receive I/O requests from the production host. After the fault is rectified, the HyperReplication/S pair is recovered based on the specified recovery policy. If the policy is automatic recovery, the pair automatically enters the **Synchronizing** state and incremental data is copied to the secondary LUN. If the policy is manual recovery, the pair enters the **To Be Recovered** state and must be manually synchronized. Incremental synchronization greatly reduces the recovery time of HyperReplication/S.

- Writable secondary LUN

When the secondary LUN is split or disconnected, you can cancel the write protection for the secondary LUN to receive data from the host.

The write protection for the secondary LUN can be canceled only when the following two conditions are met:

- The remote replication pair is in the split or interrupted state.
- Data on the secondary LUN is consistent with that on the primary LUN (when data on the secondary LUN is inconsistent, the data is unavailable, and the secondary LUN cannot be set to writable).

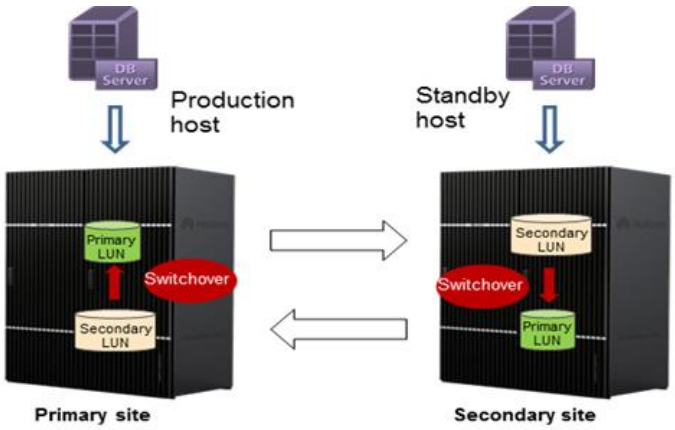
This function is used in the following scenarios:

- You want to use the data on the secondary LUN for analysis and mining without affecting services on the primary LUN.
- The production storage system at the primary site is faulty but the secondary site fails to take over services due to a primary/secondary switchover failure or communication failure.

OceanStor Dorado V6 can record the difference between the primary and secondary LUNs after host data is written to the secondary LUN. After the production storage system at the primary site recovers, you can perform incremental synchronization to quickly switch services back.

- Primary/secondary switchover

  A primary/secondary switchover is the process where the primary and secondary LUNs in a remote replication pair exchange roles.



  Primary/secondary switchover depends on the secondary LUN' data status, which can be:

  - Consistent: Data on the secondary LUN is a duplicate of the primary LUN's data at the time when the last synchronization was performed. In this state, the secondary LUN's data is available but not necessarily the same as the current data on the primary LUN.
  - Inconsistent: Data on the secondary LUN is not a duplicate of the primary LUN's data at the time when the last synchronization was performed and, therefore, is unavailable.

  In the preceding figure, the primary LUN at the primary site becomes the new secondary LUN after the switchover, and the secondary LUN at the secondary site becomes the new primary LUN. After the new primary LUN is mapped to the standby hosts at the secondary site (this can be performed in advance), the standby hosts can take over services and issue new I/O requests to the new primary LUN. A primary/secondary switchover can be performed only when data on the secondary LUN is consistent with that on the primary LUN. Incremental synchronization is performed after a primary/secondary switchover.

Note the following:

- When the pair is in the normal state, a primary/secondary switchover can be performed.
- In the split state, a primary/secondary switchover can be performed only when the secondary LUN is set to writable.

- Consistency group

  Medium- and large-size databases' data, logs, and modification information are stored on different LUNs. If data on one of these LUNs is unavailable, data on the other LUNs is also invalid. Consistency between multiple remote replication pairs must be considered when remote disaster recovery solutions are implemented on these LUNs. HyperReplication/S uses consistency groups to maintain the same synchronization pace between multiple remote replication pairs.

  A consistency group is a collection of multiple remote replication pairs that ensures data consistency when a host writes data to multiple LUNs on a single storage system. After data is written to a consistency group at the primary site, all data in the consistency group is simultaneously copied to the secondary LUNs to ensure data integrity and availability at the secondary site.

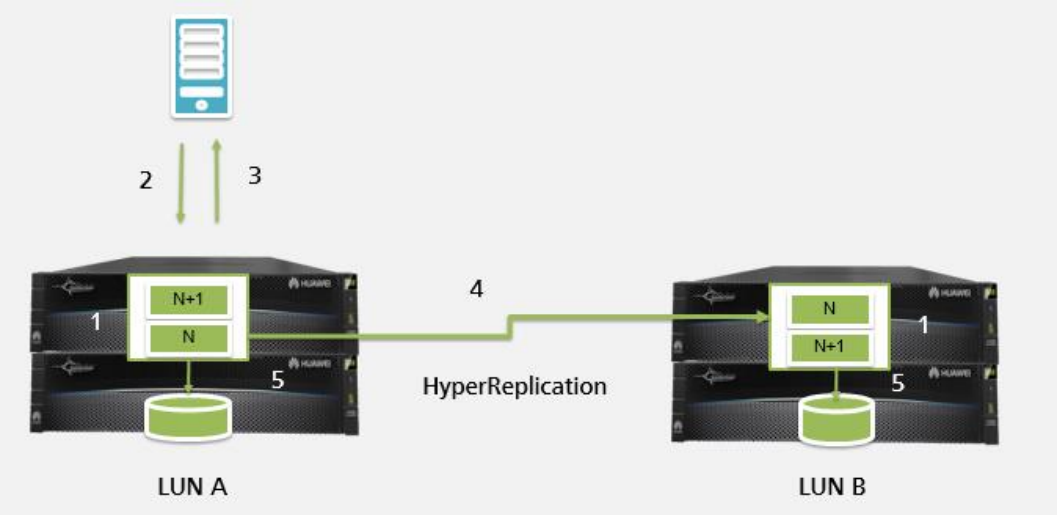
  HyperReplication/S allows you to add multiple remote replication pairs to a consistency group. When you set writable secondary LUNs for a consistency group or perform splitting, synchronization, or primary/secondary switchover, the operation applies to all members in the consistency group. If a link fault occurs, all member pairs are interrupted simultaneously. After the fault is rectified, data synchronization is performed again to ensure availability of the data on the secondary storage system.

5.5.2 HyperReplication/A for Block (Asynchronous Remote Replication)

OceanStor Dorado V6 supports asynchronous remote replication. After an asynchronous remote replication pair is established between a primary LUN at the primary site and a secondary LUN at the secondary site, initial synchronization is implemented. After the initial synchronization, the data status of the secondary LUN becomes **Synchronized** or **Consistent**. Then, I/Os are processed as follows:

1. The primary LUN receives a write request from a production host.
2. After data is written to the primary LUN, a write completion response is immediately returned to the host.
3. Incremental data is automatically synchronized from the primary LUN to the secondary LUN at the user-defined interval, which ranges from 3 seconds to 1,440 minutes. (If the synchronization type is **Manual**, you must trigger synchronization manually.) Before synchronization starts, a snapshot is generated for the primary and secondary LUNs separately. The snapshot of the primary LUN ensures that the data read from the primary LUN during the synchronization remains unchanged. The snapshot of the secondary LUN backs up the secondary LUN's data in case an exception during synchronization causes the data to become unavailable.
4. During the synchronization, data is read from the snapshot of the primary LUN and copied to the secondary LUN. After the synchronization is complete, the snapshots of the primary and secondary LUNs are deleted, and the next synchronization period starts.

Figure 5-15 Working principle of asynchronous remote replication



5.5.3 Technical Highlights

- Data compression

  The replication protocol supports Fiber Channel and Ethernet, each replication pair needs 2 channels. Both Fibre Channel and IP links support data compression by using the LZ4 algorithm, which can be enabled or disabled as required. Data compression reduces the bandwidth required by asynchronous remote replication. In the testing of an Oracle OLTP application with 100 Mbit/s bandwidth, data compression saves half of the bandwidth.

- Quick response to host requests

  After a host writes data to the primary LUN at the primary site, the primary site immediately returns a write success to the host before the data is written to the secondary LUN. In addition, data is synchronized in the background, which does not affect access to the primary LUN. HyperReplication/A does not synchronize incremental data from the primary LUN to the secondary LUN in real time. Therefore, the amount of data loss depends on the synchronization interval (ranging from 3 seconds to 1440 minutes; 30 seconds by default), which can be specified based on site requirements.

- Splitting, switchover of primary and secondary LUNs, and rapid fault recovery

  HyperReplication/A supports splitting, synchronization, primary/secondary switchover, and recovery after disconnection.

- Consistency group

  Consistency groups apply to databases. Multiple LUNs, such as log LUNs and data LUNs, can be added to a consistency group so that data on these LUNs is from a consistent time in the case of periodic synchronization or fault. This facilitates data recovery at the application layer.

- Interoperability with Huawei OceanStor converged storage systems

  Developed on the OceanStor OS unified storage software platform, OceanStor Dorado V6 is compatible with the replication protocols of all Huawei OceanStor converged storage products. Remote replication can be created among different types of products to construct a highly flexible disaster recovery solution.

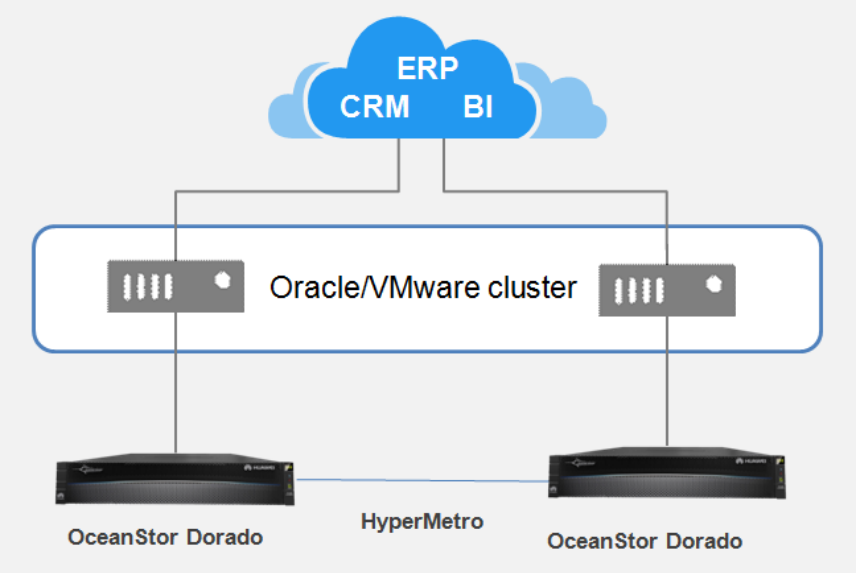
- Support for fan-in

HyperReplication of OceanStor Dorado V6 supports data replication from 64 storage devices to one storage device for central backup (64:1 replication ratio, which is four to eight times that provided by other vendors). This implements disaster recovery resource sharing and greatly reduces the disaster recovery cost.

5.6 HyperMetro (Active-Active Layout)

HyperMetro, an array-level active-active technology provided by OceanStor Dorado V6, enables two storage systems to work in active-active mode in two locations within 100 km from each other, such as in the same equipment room or in the same city. HyperMetro supports both Fibre Channel and IP networking (10GE). It allows two LUNs from separate storage arrays to maintain real-time data consistency and to be accessible to hosts. If one storage array fails, hosts automatically choose the path to the other storage array for service access. If the links between storage arrays fail and only one storage array can be accessed by hosts, the arbitration mechanism uses a quorum server deployed at a third location to determine which storage array continues providing services.

Figure 5-16 Active-active arrays



Technical Features of HyperMetro

- Gateway-free active-active solution

  Simple networking makes deployment easy. The gateway-free design improves reliability and performance because there is one less possible failure point and the 0.5 ms latency caused by a gateway is avoided.
- Active-active mode

  Hosts in different data centers can read or write data in the same LUN simultaneously, implementing load balancing across data centers.
- Site access optimization

UltraPath is optimized specifically for active-active scenarios. It can identify region information to reduce cross-site access, reducing latency. UltraPath can read data from the local or remote storage array. However, when the local storage array is working properly, UltraPath preferentially reads data from and writes data to the local storage array, preventing data read and write across data centers.

- FastWrite

  In a common SCSI write process, a write request goes back and forth between two data centers twice to complete two interactions, namely Write Alloc and Write Data. FastWrite optimizes the storage transmission protocol and reserves cache space on the destination array for receiving write requests. Write Alloc is omitted and only one interaction is required. FastWrite halves the time required for data synchronization between two arrays, improving the overall performance of the HyperMetro solution.

- Service granularity-based arbitration

  If links between two sites fail, HyperMetro can enable some services to run preferentially in data center A and others in data center B based on service configurations. Compared with traditional arbitration where only one data center provides services, HyperMetro improves resource usage of hosts and storage systems and balances service loads. Service granularity-based arbitration is implemented based on LUNs or consistency groups. Generally, a service belongs to only one LUN or consistency group.

- Automatic link quality adaptation

  If multiple links exist between two data centers, HyperMetro automatically balances loads among links based on the quality of each link. The system dynamically monitors link quality and adjusts the load ratio of the links to reduce the retransmission ratio and improve network performance.

- Compatibility with other features

  HyperMetro can work with existing features such as HyperSnap, SmartThin, SmartDedupe, and SmartCompression.

- Active and standby quorum servers

  The quorum servers can be either physical or virtual machines. HyperMetro can have two quorum servers working in active/standby mode to eliminate single point of failure and guarantee service continuity.

- Static Priority Mode

  Static priority mode allows active-standby between two data centers. The preferred site wins the arbitration and provides services.

- Expansion to 3DC

  HyperMetro can work with HyperReplication/A to form a geo-redundant architecture.

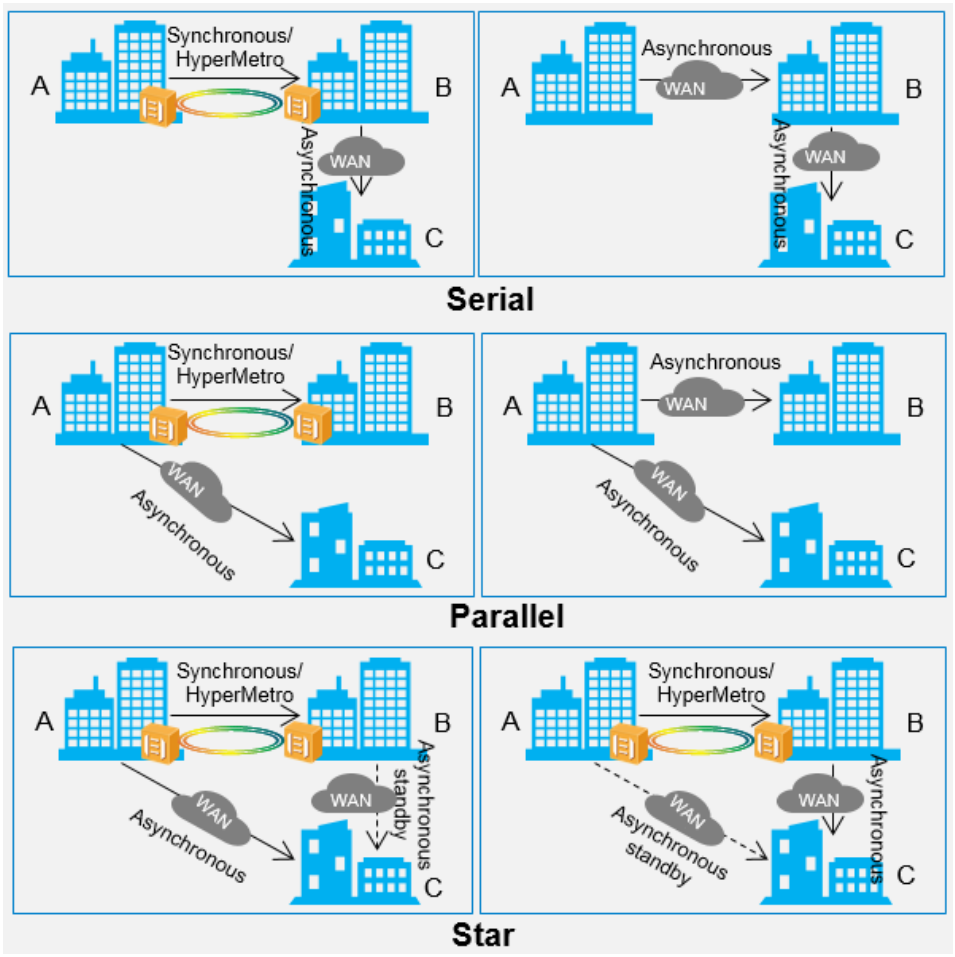
5.7 3DC for Block (Geo-Redundancy)

3DC supports flexible networking using HyperMetro, synchronous remote replication, and asynchronous remote replication, including:

- Cascading network in synchronous + asynchronous mode
- Parallel network in synchronous + asynchronous mode
- Cascading network in asynchronous + asynchronous mode
- Parallel network in asynchronous + asynchronous mode
- Star network in synchronous + asynchronous mode

- Star network in HyperMetro + asynchronous mode

Figure 5-17 3DC networking



Technical Highlights:

- Two HyperMetro or synchronous remote replication sites can be flexibly expanded to 3DC without requiring external gateways.
- In the star topology, only incremental synchronization is required in the event of any site failure.
- The star topology supports centralized configuration and management at a single site.

6 System Security and Data Encryption

6.1 Data Encryption

OceanStor Dorado V6 can work with self-encrypting drives (SEDs) and Internal Key Manager to implement static data encryption and ensure data security.

Internal Key Manager

Internal Key Manager is OceanStor Dorado V6's built-in key management system. It generates, updates, backs up, restores, and destroys keys, and provides hierarchical key protection. Internal Key Manager is easy to deploy, configure, and manage. Internal Key Manager is recommended if certification is not required and the key management system is only used by the storage systems in a data center.

SEDs

SEDs provide two-layer security protection by using an authentication key (AK) and a data encryption key (DEK).

- An AK authenticates the identity in disk initialization.
- A DEK encrypts and decrypts data in the event of writing data to or reading data from SEDs.

AK mechanism: After data encryption has been enabled, the storage system activates the AutoLock function of SEDs and uses AKs assigned by a key manager. SED access is protected by AutoLock and only the storage system itself can access its SEDs. When the storage system accesses an SED, it acquires an AK from the key manager. If the AK is consistent with the SED's, the SED decrypts the DEK for data encryption/decryption. If the AKs do not match, all read and write operations will fail.

DEK mechanism: After the AutoLock authentication is successful, the SED uses its hardware circuits and internal DEK to encrypt or decrypt the data that is written or read. When you write data, the data is encrypted by the DEK of the AES encryption engine into ciphertext, and then written to the system. When you read data, the system decrypts the requested data into plaintext using the DEK. The DEK cannot be acquired separately, which means that the original information on an SED cannot be read directly after it is removed from the storage system.

6.2 Role-based Access Control

OceanStor Dorado V6 supports role-based access control to authenticate users. Roles can be classified into default and user-defined ones.

- Default roles

Table 6-1 Default roles and permission

Default Role	Permission
Super administrator	Has all permissions of the system.
Administrator	Has all permissions except user management and security configuration permissions.
Security administrator	Has the security configuration permission, including security rule management, audit management, and KMC management.
Network administrator	Has the network management permission, including management on physical ports, logical ports, VLANs, and failover groups.
SAN resource administrator	Has the SAN resource management permission, including management on storage pools, LUNs, mapping views, hosts, and ports.
Data protection administrator	Has the data protection management permission, including management on local data protection, remote data protection, and HyperMetro schemes.
Backup administrator	Has the data backup management permission, including management on local data and mapping views.

- User-defined roles: The system allows you to define permissions as required. You can specify the role when creating a user account.

7 System Management and Compatibility

7.1 System Management

OceanStor Dorado V6 provides device management interfaces and integrated northbound management interfaces. Device management interfaces include a graphic management interface DeviceManager and a command-line interface (CLI). Northbound interfaces are RESTful interfaces, supporting SMI-S, SNMP, evaluation tools, and third-party network management plug-ins. For details, see http://support-open.huawei.com/ready/pages/user/compatibility/support-matrix.jsf.

7.1.1 DeviceManager

DeviceManager is a common HTML5-based management system for Huawei OceanStor systems. It provides wizard-based GUI for efficient management.

7.1.2 CLI

The CLI allows administrators and other system users to perform management and maintenance operations on storage systems. The CLI supports key-based SSH v2 user access permission, allowing users to execute scripts on a remote host. You are not required to save the passwords in the scripts.

7.1.3 Call Home Service

In traditional service support mode, technical support personnel provide local services manually. Faults may not be detected quickly and information may not be communicated correctly. Call Home enables a storage system to upload its alarms and logs to eService. eService uses artificial intelligence (AI) technologies to implement intelligent fault reporting, real-time health analysis, intelligent fault prevention, and intelligent optimization, minimizing device running risks and reducing operational costs. Call Home is a remote maintenance expert system. Using the secure and controllable network connections between devices and Huawei technical support centers, Call Home enables Huawei to monitor the health status of customers' devices, 24/7. If a fault occurs, the fault information is automatically and immediately sent to Huawei technical support, shortening fault discovery and handling time.

After the built-in Call Home service is enabled on the DeviceManager, the pre-installed eService Agent on devices periodically collects information and sends the information to Huawei technical support. Customers must ensure that devices can be connected to Huawei technical support over a network. HTTP proxy is supported.

The following information is collected:

- Device performance statistics
- Device running data
- Device log and alarm data

All data is sent to Huawei technical support in text mode over HTTPS. Records of sent information can be sent to the Syslog server for security audit. If data cannot be uploaded due to network interruption, devices can save the last day's data files (up to 5 MB per controller) and send them when the network recovers. The files that are not uploaded can be exported for troubleshooting by using the command line.

The information sent to Huawei technical support can be used to provide the following functions.

- Alarm monitoring: Device alarms are monitored 24/7. If a fault occurs on a device, Huawei technical support is notified within 1 minute and a troubleshooting ticket is dispatched to engineers. This helps customers locate and resolve problems quickly.
- In conjunction with big data analysis technologies and device fault libraries across the world, fault prevention and fast fault troubleshooting are supported.
- Based on industry application workload models, optimal device configurations and performance optimization suggestions are provided.

7.1.4 RESTful API

RESTful APIs of OceanStor Dorado V6 allow system automation, development, query, and allocation based on HTTPS interfaces. With RESTful APIs, you can use third-party applications to control and manage arrays and develop flexible management solutions for Dorado V6.

7.1.5 SNMP

SNMP interfaces can be used to report alarms and connect to northbound management interfaces.

7.1.6 SMI-S

SMI-S interfaces support hardware and service configuration and connect to northbound management interfaces.

7.1.7 Tools

OceanStor Dorado V6 provides diversified tools for pre-sales assessment and post-sales delivery. These tools can be accessed through eDesigner, SmartKit, DeviceManager, SystemReporter, and eService and effectively help deploy, monitor, analyze, and maintain OceanStor Dorado V6.

7.2 Ecosystem and Compatibility

7.2.1 Virtual Volume (VVol)

OceanStor Dorado V6 supports VVol 1.0, which includes new objects such as Protocol Endpoint (PE) LUN, VVol, and VVol SNAP. The VVol object supports cascading snapshot, differential bitmap, and LUN data copy. To quickly deploy VMs, you can create a VVol snapshot for the VM template and then create snapshots for the VVol snapshot to generate multiple VMs using the same data image.

When a VM that has snapshots is cloned, data can be copied by the host or storage system.

- When the host copies the VM data, it can query the area where the VVol object stores data and perform full copy. Then the host can query the differences between the snapshots and the VM and copy the differential data.
- When the storage system copies the VM data, it uses its own full copy and differential copy capabilities to copy the data to the new VM directly. Data can be copied between different controllers, controller enclosures, and storage pools.

VMware uses the VASA Provider plug-in to detect and use storage capabilities to deploy, migrate, and clone VMs quickly.

Each VM is stored in multiple VVols. VMware can clone, migrate, or configure traffic control policies for individual VMs. The storage system completes the data copy operations directly without occupying host bandwidth, greatly improving VM management efficiency.

7.2.2 OpenStack Integration

OceanStor Dorado V6 launches the latest OpenStack Cinder Driver in the OpenStack community. Vendors of commercial OpenStack versions can obtain and integrate OpenStack Cinder Driver, allowing their products to support OceanStor Dorado V6.

OceanStor Dorado V6 provides four versions of OpenStack Cinder Driver: OpenStack Juno, Kilo, Liberty, and Mitaka. In addition, OceanStor Dorado V6 supports commercial versions of OpenStack such as Huawei FusionSphere OpenStack, Red Hat OpenStack Platform, and Mirantis OpenStack.

For details, see
http://support-open.huawei.com/ready/pages/user/compatibility/support-matrix.jsf.

7.2.3 Virtual Machine Plug-ins

OceanStor Dorado V6 supports various VM plug-ins. For details, see
http://support-open.huawei.com/ready/pages/user/compatibility/support-matrix.jsf.

7.2.4 Host Compatibility

OceanStor Dorado V6 supports mainstream host components, including operating systems, virtualization software, HBAs, volume management, and cluster software. OceanStor Dorado V6 supports a wider range of operating systems and VM platforms for mainstream database software. For details, see
http://support-open.huawei.com/ready/pages/user/compatibility/support-matrix.jsf.

8 Best Practices

Huawei is continuously collecting requirements of important customers in major industries and summarizes the typical high-performance storage applications and challenges facing these customers. This helps Huawei provide best practices which are tested and verified together with application suppliers.

For best practices, visit http://storage.huawei.com/en/index.html.

9 Appendix

9.1 More Information

You can visit our official website to get more information about Huawei storage:

http://e.huawei.com/en/products/cloud-computing-dc/storage

For after-sales support, visit our technical support website:

http://support.huawei.com/enterprise/en

For pre-sales support, visit the following website:

http://e.huawei.com/en/how-to-buy/contact-us

You can also contact your local Huawei office:

http://e.huawei.com/en/branch-office

9.2 Feedback

Huawei welcomes your suggestions for improving our documentation. If you have comments, send your feedback to storagedoc@huawei.com.

Your suggestions will be seriously considered and we will make necessary changes to the document in the next release.

