OceanStor Dorado 3000, 5000, 6000 6.1.7

Technical White Paper

Issue 03

Date 2024-03-01

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: https://www.huawei.com

Email: support@huawei.com

Contents

1 Executive Summary

Huawei OceanStor Dorado all-flash storage systems (OceanStor Dorado for short) are designed to carry mission-critical services of enterprises, financial institutions, and data centers. The storage systems have the following highlights:

- Always-on applications with unique SmartMatrix reliable layout
  - Zero service interruption: The industry's only storage product that ensures service continuity in the event of a single point or dual points of failure.
  - Failover transparent to applications: In the event of a controller failure, services are switched over in seconds without interruption.
  - Global resource balancing: The end-to-end active-active architecture balances resources globally.
- No.1 performance with chip-powered architectures
  - Five chips for transmission, compute, storage, management, and intelligence lay the solid foundation for the industry's fastest storage.
  - The intelligent chip and cache algorithm allow the storage system to deeply learn service I/O patterns and intelligently accelerate data processing.
  - The industry's unique smart disk enclosure shares the computing load of the storage system, achieving linear expansion of performance and capacity.
- Efficient O&M with edge-cloud synergy
  - Intelligence throughout the service lifecycle: Intelligence participates in the end-to-end service process from resource provisioning to fault locating, allowing the system to predict the performance and capacity trend for the next 60 days, detect potential faulty disks 14 days in advance, and immediately provide solutions to 93% of problems upon detection.
  - On- and off-premises synergy with general-purpose cloud intelligence and customized edge intelligence: The built-in intelligent chip performs incremental training and deep learning of service patterns to enhance personalized experience.
  - FlashEver: The intelligent and elastic architecture allows module-based upgrade. Customers can use the latest-generation software and hardware without data migration, maximizing return on investment (ROI).

OceanStor Dorado meets the requirements of enterprise applications such as databases, virtual desktop infrastructure (VDI), virtual server infrastructure (VSI), and file sharing, helping finance, manufacturing, carrier, and other industries smoothly transition to all-flash storage. It also helps build virtualized, cloud-ready, and intelligent IT systems to effectively support the evolution to Industry 4.0.

- Online transaction processing (OLTP) applications are widely used in finance, carrier, manufacturing, and government sectors. As the amount of data increases, the response speed of the IT system decreases, especially at the beginning and end of a month, resulting in customer complains. OceanStor Dorado ensures high throughput and low latency for the IT system to bear more service load and respond quickly without degrading the SLA, improving customer satisfaction.
- For the online analytical processing (OLAP) services of the finance, carrier, manufacturing, and government sectors, the time window for batch processing becomes insufficient as the service volume grows. OceanStor Dorado provides faster processing to shorten the batch processing time, allowing a massive amount of data to be processed within the time window.

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

2 Overview

This chapter describes the product portfolio of OceanStor Dorado and its unique benefits to customers. The product models include OceanStor Dorado 3000, OceanStor Dorado 5000, and OceanStor Dorado 6000.

2.1 Customer Benefits

In the past few years, the explosive growth of data and mining of data values have led to the innovation of IT systems, especially storage devices. The main storage media changes from HDDs to SSDs, and the main storage protocol changes from SAS to NVMe. The end-to-end access latency of storage systems is shortened from 10 ms to 1 ms or even lower, and storage systems are becoming more intelligent with high performance and reliability. Business developments and technical innovations pose higher requirements on the design and construction of customers' IT infrastructure, and choosing a proper storage system is a crucial part in building a modern IT infrastructure. Stable storage performance and high reliability are the basis for building an intelligent and scalable IT system; efficient storage is a key factor in reducing IT system costs; efficient data flow, intelligent O&M, non-disruptive upgrade, and long-term supply assurance are crucial for long-term IT system evolution and development.

Guided by the concept of "Data + Intelligence", OceanStor Dorado redefines storage architecture. With advanced software and hardware, OceanStor Dorado implements intelligent, native all-flash storage to provide stable, high-performance, and highly reliable services and meets the storage requirements of intelligent IT systems.

- **Using core software and hardware technologies to achieve efficient and high-performance storage**
  - OceanStor Dorado provides low latency and large throughput from end to end.
  - The storage systems support 32 Gbit/s Fibre Channel, 100 Gbit/s RDMA, and NVMe for front-end interconnection, cluster switching, and back-end interconnection, and can offload specific tasks to hardware to free CPU resources, ensuring low latency and high bandwidth.
  - The core processing unit of the storage systems incorporates multiple cores and processors. With the key software designs, such as balanced distribution, lock-free design, and high scalability, a storage system can have up to 1000 cores to ensure efficient service processing.

  - FlashLink® enables close collaboration between controllers and SSDs. The use of multistreaming, full-stripe write, garbage collection, and read/write priority control effectively reduces write amplification on SSDs, making the most of the SSD performance throughout the lifecycle.
  - The intelligent system learns and analyzes the workload to identify long-interval sequential flows and data associations, which greatly improves the data prefetch capability of the cache.

- **Innovative SmartMatrix architecture providing highly reliable and stable storage services**
  - OceanStor Dorado employs a new-generation hardware platform and an ultra-stable SmartMatrix architecture to enhance data reliability and service continuity, ensuring always-on storage services.
  - In terms of data reliability, the storage systems provide end-to-end data redundancy protection, validity check, and recovery mechanisms. Data protection measures at various levels can be used on demand, such as multiple cache copies, data protection across controller enclosures, RAID, data integrity protection, snapshots, remote replication, and active-active data center solution. The system checks data integrity and rectifies any error in the end-to-end data read and write processes to prevent unexpected recoverable faults (such as bit reversal) in hardware. This effectively prevents data error spreading when devices are in an uncontrollable and intermittent sub-health state, ensuring data security.
  - In terms of service continuity, the system accurately monitors the device health status to quickly identify faults. If a fault is detected, the system isolates and attempts to rectify the fault through redundancy takeover. If the fault is rectified, the involved component continues providing services. If the fault fails to be rectified, an alarm is reported to prompt users to replace the faulty component.

- **Intelligent architecture realizing intelligent storage services**

  As the business of enterprises grows, a single storage system will carry multiple service systems, which have varied requirements on performance, capacity, data protection, and cost. This poses a challenge on storage reliability, performance, and capacity, as well as the customers' overall IT planning and management. OceanStor Dorado has balanced the capacity and performance for hybrid application models in which various services and workloads share storage resources. The system provides a default intelligent configuration management mode and supports user-defined configuration management modes to configure multiple devices on a single page based on the service logic. Customers can choose from the configuration modes as required. The system also supports report management to help customers grasp the health status, performance, and capacity of devices in real time and predict future trends. The reports can be used as references for service planning and adjustment. In adaptive mode, the system uses the optimal configuration. In the future, the system will continuously learn the workloads and make adjustments accordingly to optimize performance and handle periodic or burst service peaks. For sub-healthy or worn parts, the system adjusts their parameters to guarantee stability and prolong the service life of the devices after long-term operation.

3 Hardware Architecture

OceanStor Dorado employs a full-mesh architecture. All FRUs, such as front-end interface modules, controllers, back-end interface modules, power modules, BBUs, fan modules, and SSDs, are redundant and protected against single points of failure. All field replaceable units (FRUs) are hot-swappable and can be replaced online.

The RDMA high-speed network implements shared access to the global cache at a low latency. The smart disk enclosures (equipped with CPUs and memory to provide computing power) at the back end offload disk reconstruction tasks from controllers to save controller resources.

3.1 Hardware Description

3.1.1 Controller Enclosure of Mid-Range Devices

OceanStor Dorado mid-range devices include OceanStor Dorado 3000, OceanStor Dorado 5000, and OceanStor Dorado 6000. They use a 2 U controller enclosure that has two controllers. The controller enclosure can house 25 SAS disks, 25 NVMe disks, and 36 NVMe disks. The NVMe disk adopts a customized physical form, which allows a 36-slot enclosure to house 40% more NVMe disks than 2.5-inch disks. All FRUs are redundant and can be replaced online.

OceanStor Dorado mid-range devices provide SAN and NAS unified storage services. The devices use end-to-end low-latency SAN (FC-NVMe, FC, iSCSI, and NVMe over RoCE) and NAS (NFS, CIFS, S3 and NDMP) front-end protocols on various service interface modules.

- Front-end connection

  Front-end interface modules include 4-port 8 Gbit/s, 16 Gbit/s, and 32 Gbit/s FC/FC-NVMe interface modules, 4-port 25 Gbit/s NVMe over RoCE interface modules, 4-port 10GE and 25GE interface modules, 2-port 40GE and 100GE interface modules, 2-port 100 Gbit/s NVMe over RoCE interface modules, and 4-port GE electrical interface modules.
- Scale-out

The 4-port 25 Gbit/s RDMA interface modules are used for scale-out through direct connections among four controllers. The 2-port 100 Gbit/s RDMA interface modules are used for scale-out through switched networking among multiple controllers.

- Back-end connection

  Back-end interface modules include 4-port 12 Gbit/s SAS interface modules (for connecting to SAS disk enclosures) and 2-port 100 Gbit/s RDMA interface modules (for connecting to smart SAS and NVMe disk enclosures).

- Intelligence

  The intelligent accelerator card is supported to implement the intelligent read cache.

OceanStor Dorado mid-range devices support the following types of disk enclosures:

- SAS disk enclosure
- Smart SAS disk enclosure
- Smart NVMe disk enclosure

The two controllers in a controller enclosure of OceanStor Dorado mid-range devices are interconnected through RDMA mirror channels, and multiple controller enclosures can be directly connected through the scale-out interface modules. Each controller has two GE management and maintenance ports and one serial port. The following figures show the front and rear views of OceanStor Dorado mid-range devices.

Figure 3-1 Front view of a 36-disk NVMe controller enclosure

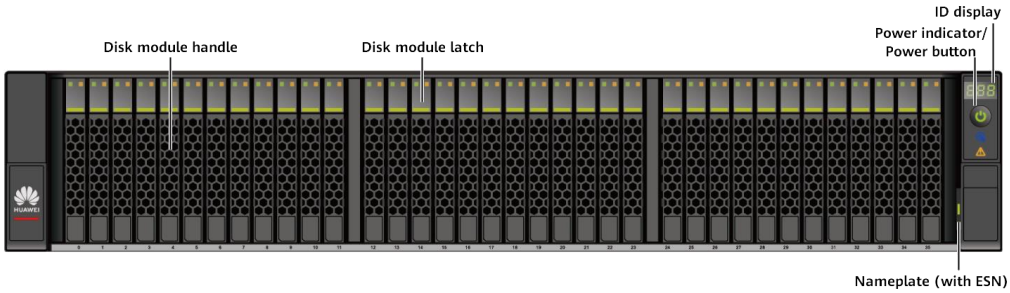


Figure 3-2 Front view of a 25-disk SAS controller enclosure

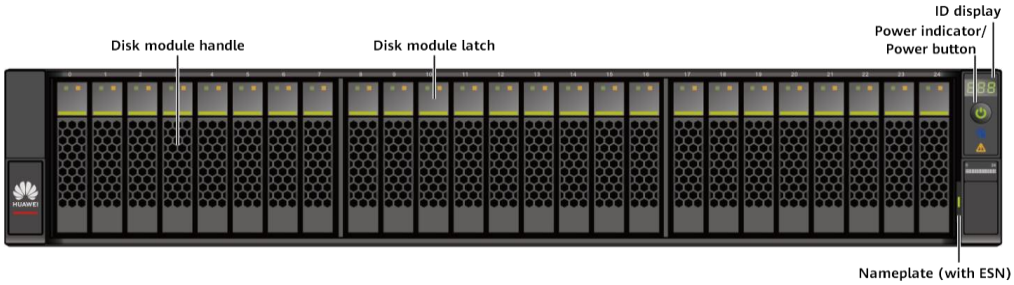


Figure 3-3 Front view of a 25-disk NVMe controller enclosure

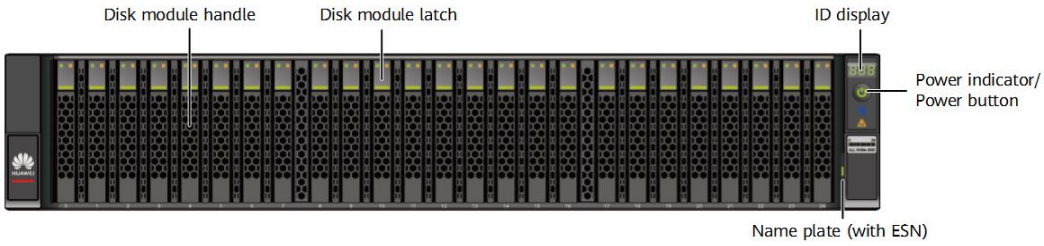


Figure 3-4 Rear view of a 36-disk NVMe controller enclosure

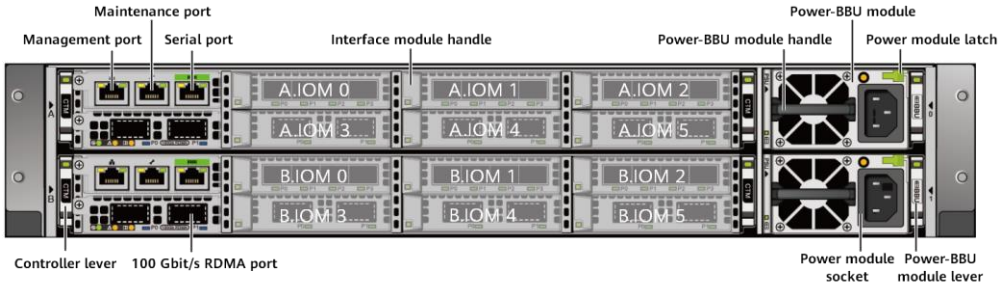


Figure 3-5 Rear view of a 25-disk SAS controller enclosure

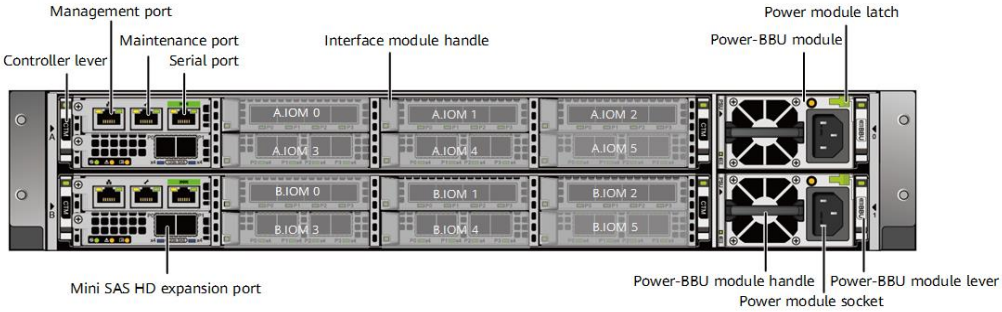
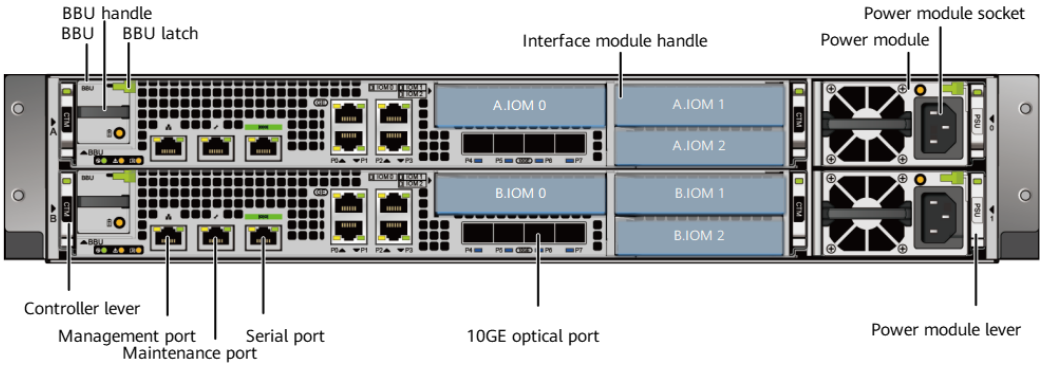


Figure 3-6 Rear view of a 25-disk NVMe controller enclosure



OceanStor Dorado mid-range devices use a 2 U NVMe or SAS controller enclosure that has two controllers and adopt a symmetric active-active architecture. Two controllers work in load balancing mode in normal situations and take over services in fault scenarios. The two controllers are interconnected by RDMA mirror channels. The following figures show the logical architectures. OceanStor Dorado mid-range devices use 2P or 1P boards, which use different CPU models but have the same service capability. The actual board depends on the delivery.

Figure 3-7 Logical architecture (1) of a mid-range storage system (NVMe)

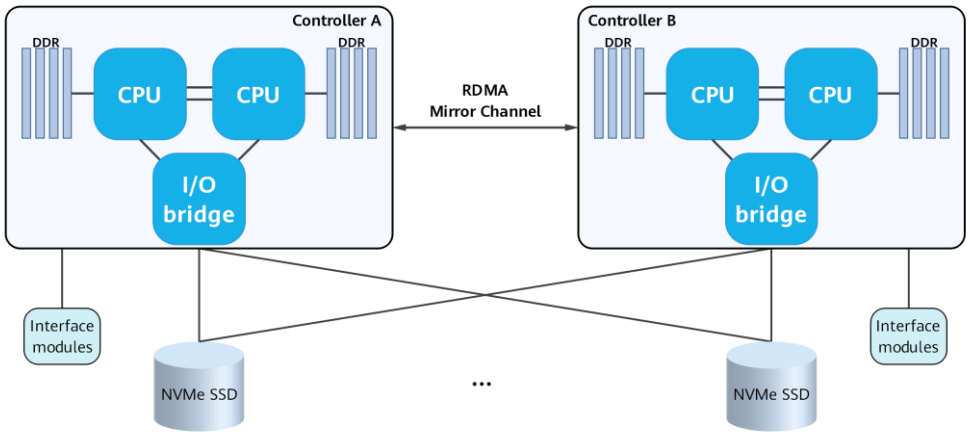


Figure 3-8 Logical architecture (2) of a mid-range storage system (NVMe)

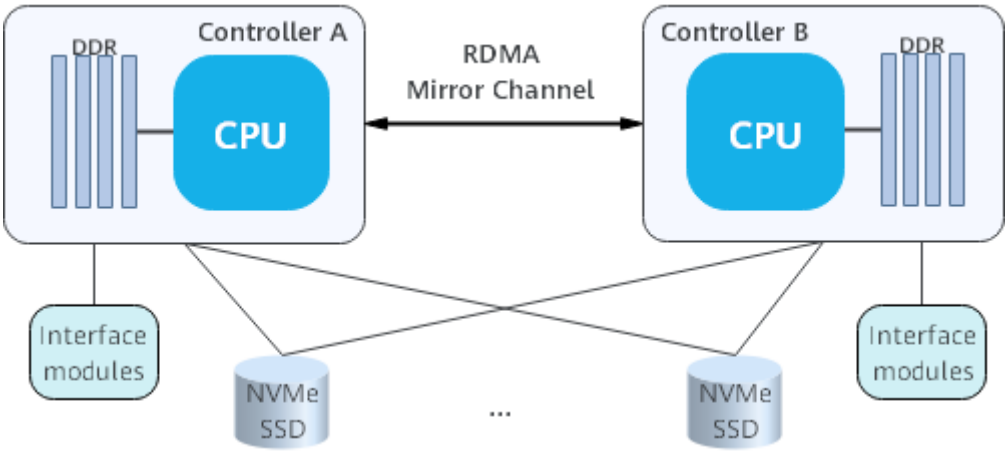


Figure 3-9 Logical architecture (1) of a mid-range storage system (SAS)

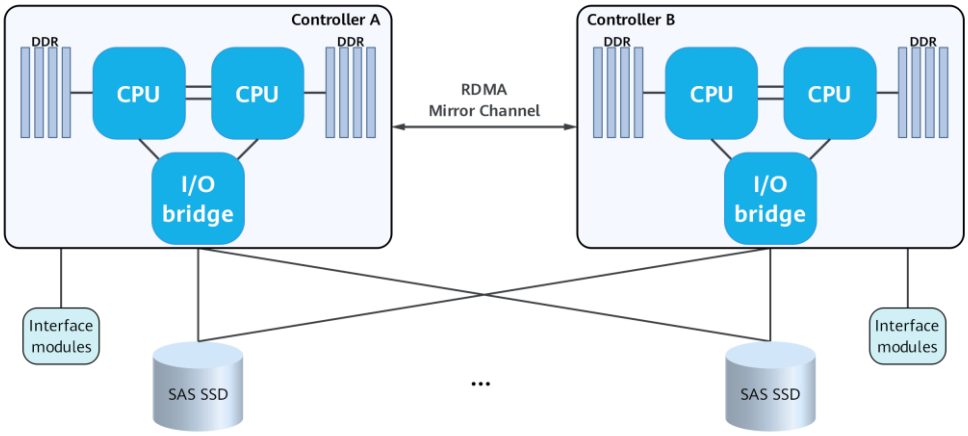
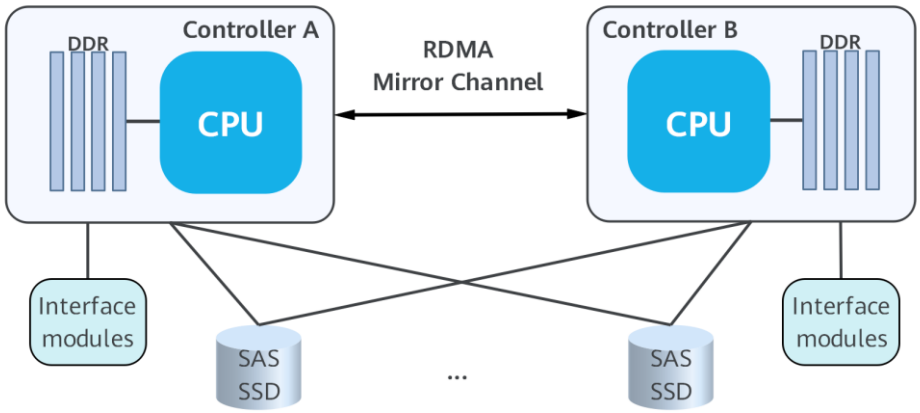


Figure 3-10 Logical architecture (2) of a mid-range storage system (SAS)



Improved Interface Module Density

To increase the number of interface modules per system, the interface modules use compact connectors and a golden finger design for PCIe ports to reduce the module thickness from 25 mm to 18.5 mm, allowing a 2 U controller enclosure to house up to 12 interface modules.

Table 3-1 Interface module density comparison

	Connector	Interface Module	Interface Module Density

	Connector	Interface Module	Interface Module Density
OceanStor Dorado	Connector width: 6 mm	Thickness: 18.5 mm	12 interface modules in 2 U space (6 interface modules per U)
OceanStor Dorado V3	Connector width: 14.5 mm	Thickness: 25 mm	6 interface modules in 2 U space (3 interface modules per U)
Specifications	Width reduced by 59%	Thickness reduced by 26%	Density increased by 100%

3.1.2 Disk Enclosure

OceanStor Dorado supports three types of disk enclosures.

Table 3-2 Disk enclosure types

Disk Enclosure	Disk Type	Port	Number of Disks
SAS disk enclosure	2-port SAS disk	4 x 12 Gbit/s SAS port	25
Smart SAS disk enclosure	2-port SAS disk	4 x 100 Gbit/s RDMA	25
Smart NVMe disk enclosure	2-port NVMe disk	4 x 100 Gbit/s RDMA	36

NOTE

OceanStor Dorado 3000 supports only SAS and smart NVMe disk enclosures. For details, see the product specifications list.

3.1.2.1 SAS Disk Enclosure

The SAS disk enclosure uses the SAS 3.0 protocol and supports 25 SAS SSDs. A controller enclosure connects to a SAS disk enclosure through the onboard SAS ports or SAS interface modules.

Figure 3-11 Front view of a 2 U 25-slot SAS disk enclosure

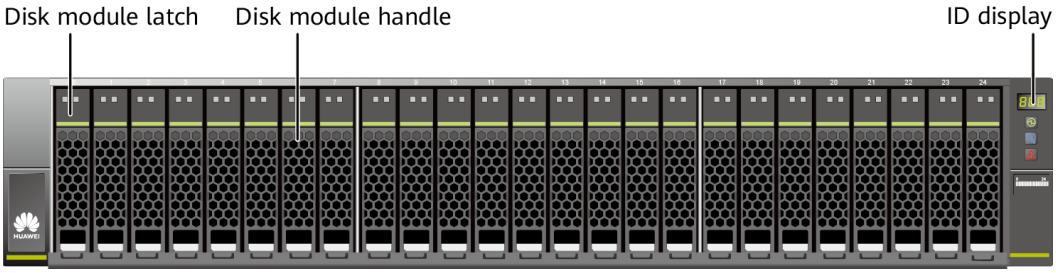
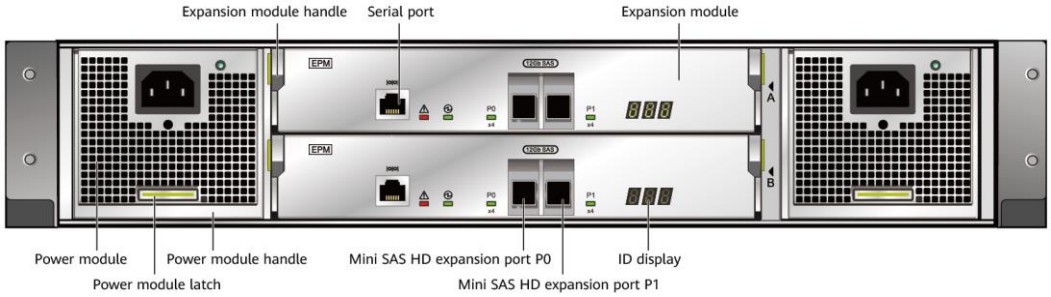


Figure 3-12 Rear view of a 2 U 25-slot SAS disk enclosure



3.1.2.2 Smart SAS and Smart NVMe Disk Enclosures

A smart SAS or smart NVMe disk enclosure has the CPU and DDR memory on its expansion modules, which provide computing capability for the smart disk enclosure to offload computing tasks from controllers. For example, the disk reconstruction task in the event of a disk failure can be offloaded to the smart SAS or smart NVMe disk enclosure, minimizing the impact of the reconstruction task on controller performance.

General PCIe RC systems have only 256 PCIe bus addresses, which are occupied by internal devices, interface modules, and NVMe devices. As a result, a single system generally supports no more than 128 NVMe SSDs. Huawei smart NVMe disk enclosures use onboard processors and independent PCIe address domains, effectively isolating the PCIe address domains of the controller enclosure and disk enclosures. The NVMe SSDs on a smart NVMe disk enclosure occupy the number of PCIe devices in this independent computer system, instead of that on the controller enclosure. In addition, smart NVMe disk enclosures connect to a controller enclosure through RDMA ports, allowing a system to support a large number of NVMe SSDs.

A controller enclosure connects to smart SAS or smart NVMe disk enclosures through onboard 100 Gbit/s RDMA ports or back-end 100 Gbit/s RDMA interface modules, which provide large-bandwidth and low-latency transmission channels. The smart SAS disk enclosure is 2 U high and has up to 25 SSDs, while the smart NVMe disk enclosure is 2 U high and has up to 36 SSDs. Figure 3-13 and Figure 3-14 show the front and rear views of a smart SAS disk enclosure, while Figure 3-15 and Figure 3-16 show the front and rear views of a smart NVMe disk enclosure.

Figure 3-13 Front view of a 2 U 25-slot smart SAS disk enclosure

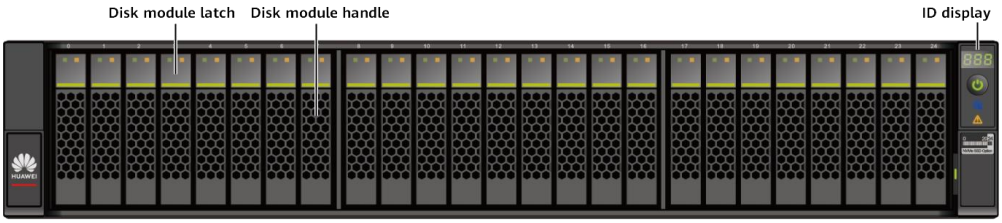


Figure 3-14 Rear view of a 2 U 25-slot smart SAS disk enclosure

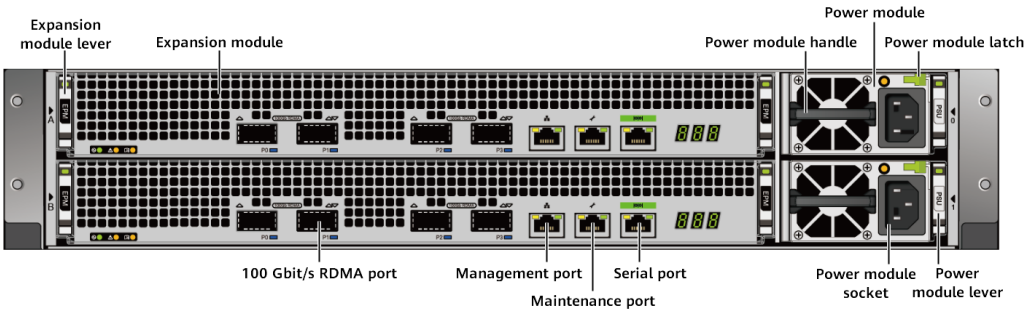


Figure 3-15 Front view of a 2 U 36-slot smart NVMe disk enclosure

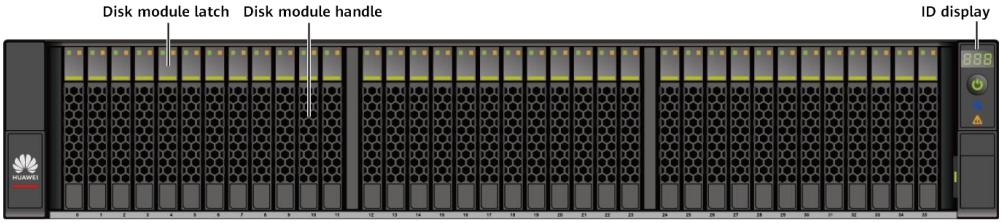
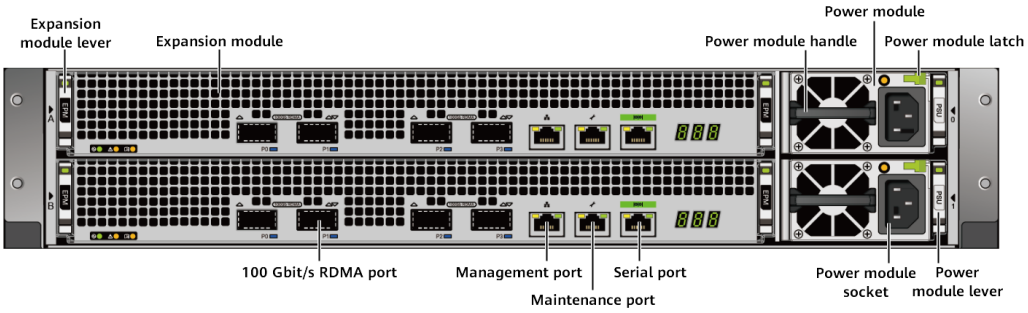


Figure 3-16 Rear view of a 2 U 36-slot smart NVMe disk enclosure



Improved Disk Density

OceanStor Dorado supports the highest density design in the industry thanks to the industry-leading heat dissipation. Traditional vertical backplanes provide small heat dissipation windows, and connectors on both sides interfere with each other. In addition, the thickness of a traditional 2.5-inch SSD reaches 14.8 mm. These factors limit the number of SSDs supported by a 2 U enclosure. Figure 3-17, Figure 3-18, and Figure 3-19 show the design of traditional vertical backplanes.

Figure 3-17 Traditional vertical backplane

Figure 3-18 Front view of the window areas on the vertical backplane

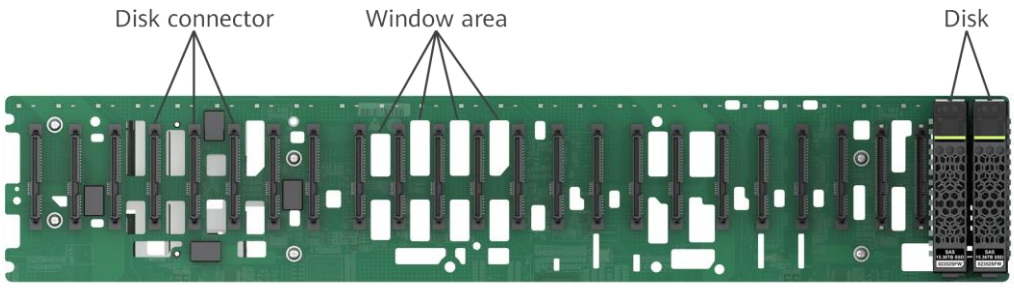
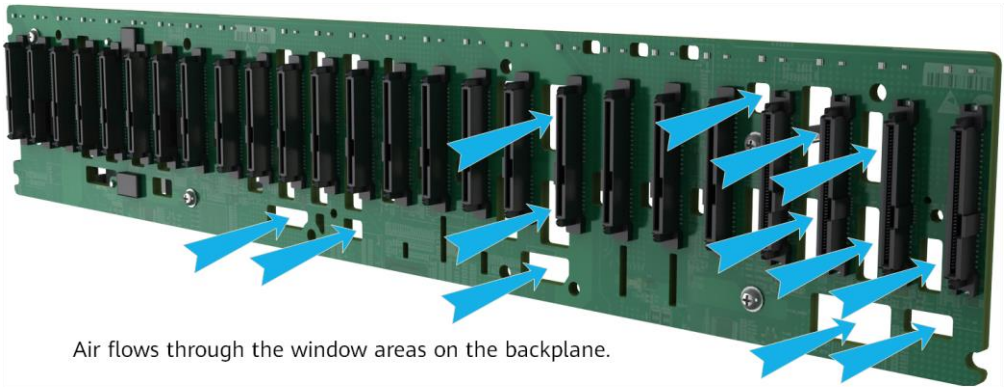


Figure 3-19 Air flow on the vertical backplane



Huawei introduces a horizontal backplane with the orthogonal connection structure, as shown in Figure 3-20 and Figure 3-21. Disk connectors and controller connectors are connected orthogonally, preventing mutual interference and improving the disk connector density. In addition, the thickness of palm-sized NVMe SSDs is reduced from 14.8 mm to 9.5 mm, allowing a 2 U enclosure to accommodate 36 SSDs, improving the disk density by 44%. Moreover, the horizontal backplane increases the window area by 50% and the heat dissipation capability by 25% to ensure heat dissipation of the SSDs, as shown in Figure 3-22 and Figure 3-23.

Figure 3-20 Horizontal backplane

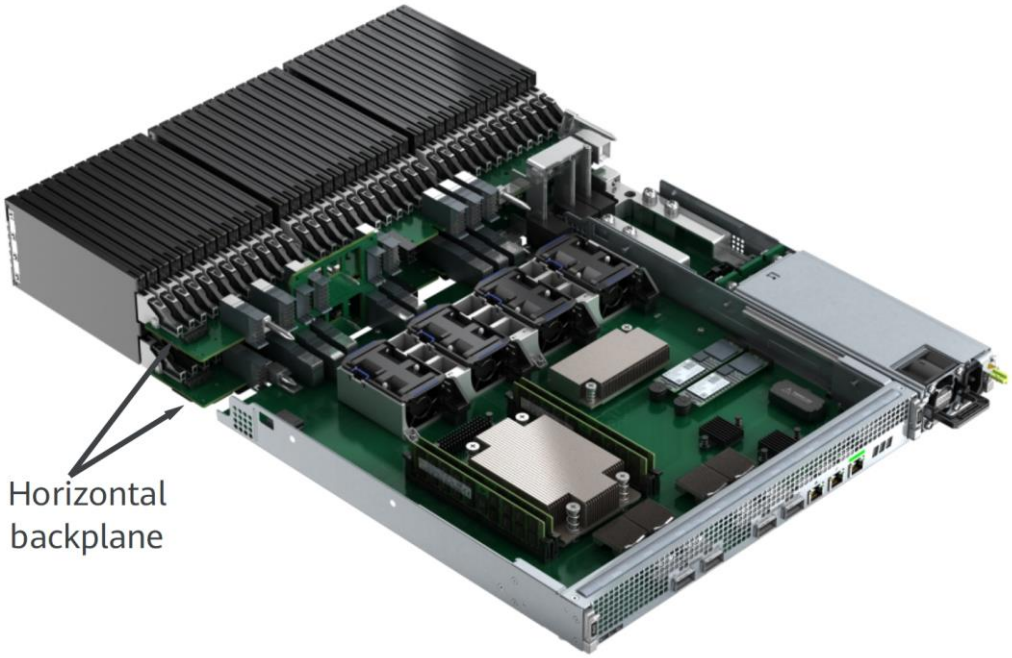


Figure 3-21 Orthogonal connection structure of the horizontal backplane (side view)

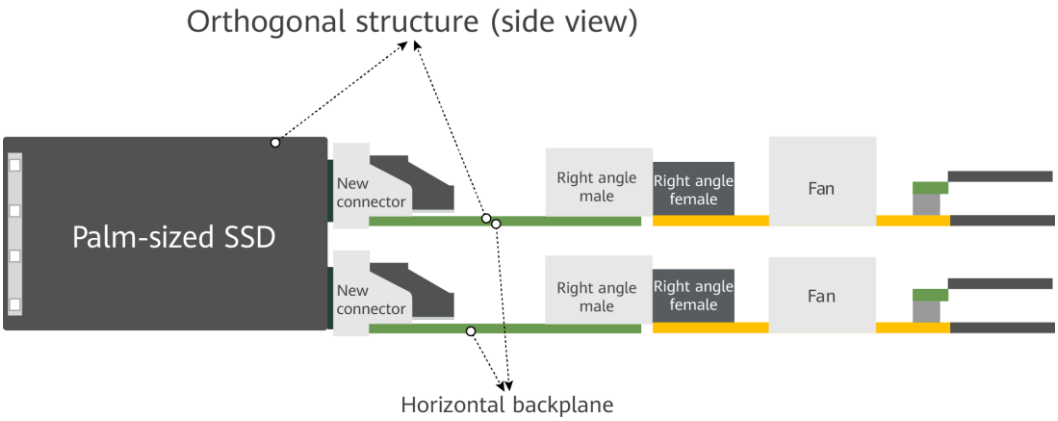


Figure 3-22 Front view of the horizontal backplane

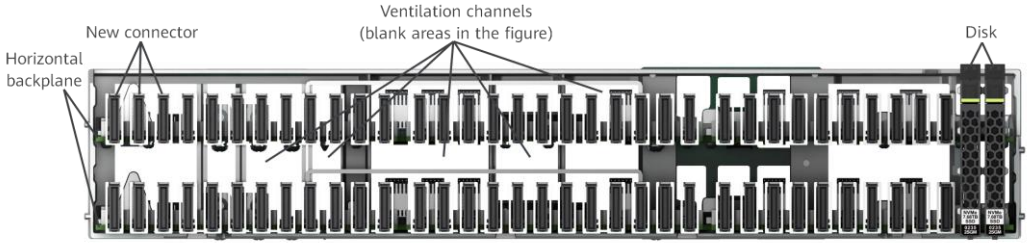
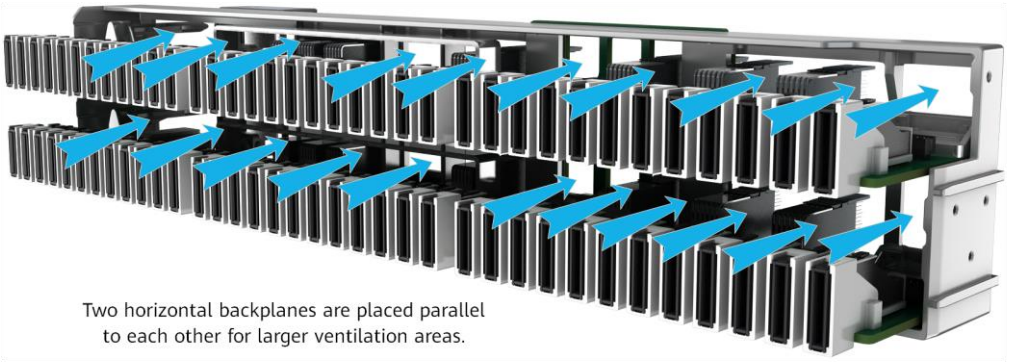


Figure 3-23 Air flow on the horizontal backplane



3.1.3 HSSDs

OceanStor Dorado uses Huawei SSDs (HSSDs) 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. The control unit contains an SSD controller, host interface, and dynamic random access memory (DRAM) module. The storage unit mainly contains 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 16 KB.

Operations on NAND flash include erase, program, and read. The program and read operations are implemented on pages, while the erase operations are implemented on blocks. 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.

Wear Leveling

The SSD controller uses software algorithms to monitor and balance the P/E cycles on blocks in the NAND flash. This prevents block failure caused by over-erasure 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.

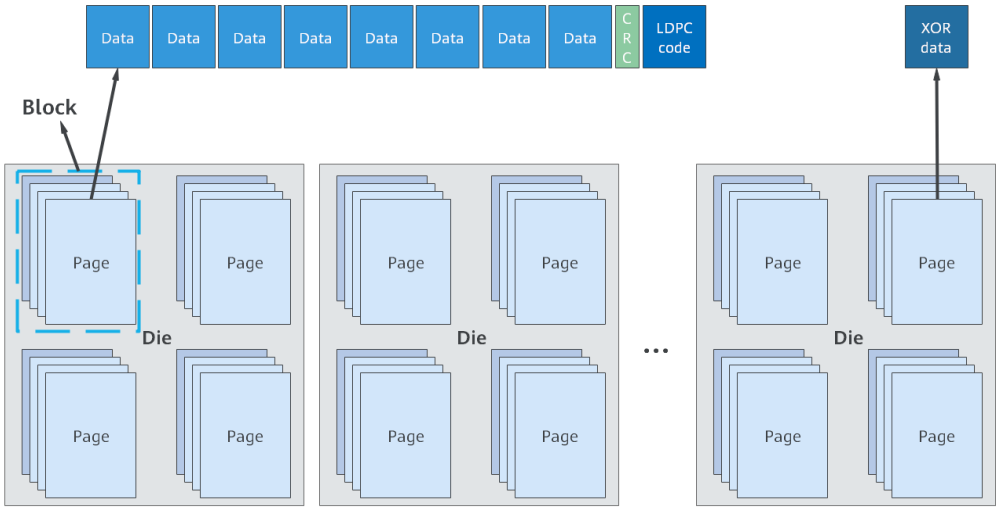
Bad Block Management

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) parity data between the NAND flash memories, and saves it to a new block. HSSDs have reserved space to replace bad blocks, ensuring sufficient available capacity and user data security.

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 protect data on pages; XOR redundancy is used between NAND flash memories to prevent data loss caused by flash failures.

Figure 3-24 Data redundancy protection



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

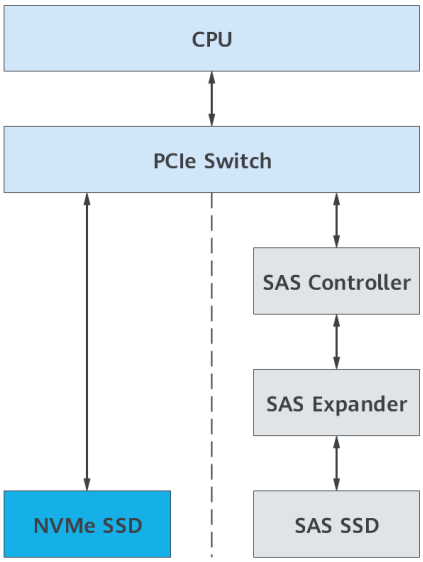
Background Inspection

After data has been 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.

Support for SAS and NVMe

HSSDs support SAS or 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, as shown in Figure 3-25. In addition, NVMe supports a larger concurrency and queue depth (64,000 queues, and up to 64,000 concurrent commands in each queue), fully exploiting SSD performance. The NVMe HSSDs provide dual ports and are hot swappable, improving system performance, reliability, and maintainability.

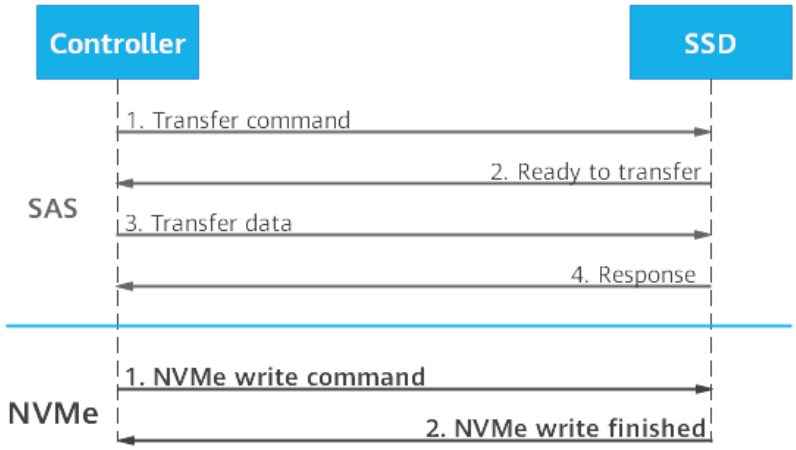
Figure 3-25 Transmission paths of NVMe and SAS SSDs



NVMe SSDs reduce the number of interactions in a write request from 4 (in a SAS protocol) to 2. In Figure 3-26:

- SAS requires four SCSI interactions to complete a write request.
- NVMe requires only two interactions to complete a write request.

Figure 3-26 SAS and NVMe protocol interactions



SAS SSD Module

A SAS SSD module consists of a handle and a 2.5-inch SSD, as shown in Figure 3-27.

Figure 3-27 SAS SSD appearance

NVMe SSD Module

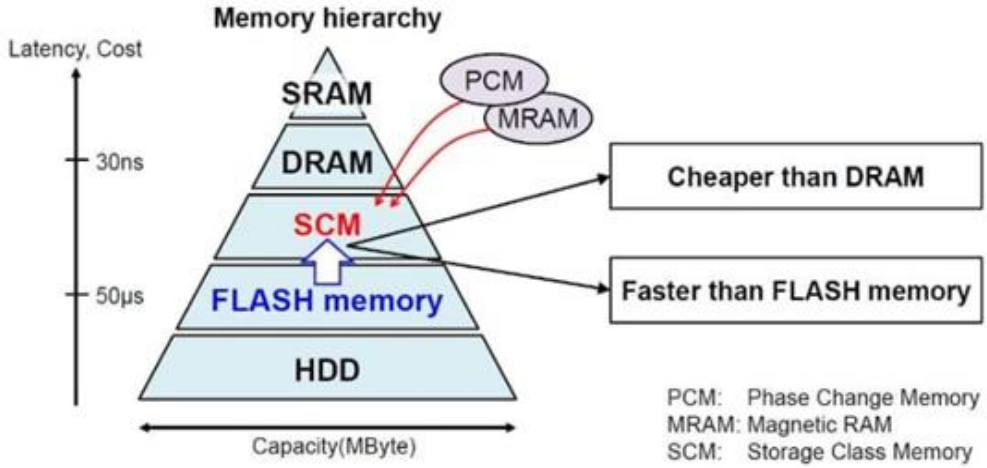
The dimensions of an NVMe SSD are 79.8 mm x 9.5 mm x 160.2 mm, as shown in Figure 3-28.

Figure 3-28 NVMe SSD appearance

3.1.4 SCM Drives

Introduction to Storage Class Memory (SCM)

Mainstream storage media include SRAM, DRAM, SCM, flash memory, and HDD (in ascending order of latency and descending order of cost). SCM provides a faster speed and lower latency than NAND flash, as well as a larger capacity and lower cost than DRAM. The use of SCM media will provide revolutionary storage solutions for big data applications and complex workloads. Currently, mainstream SCM media include XL-FLASH, Intel PCM 3D Xpoint, and Z-NAND.



XL-FLASH uses the 3D SLC stack with modifications to the SLC structure. XL-FLASH supports a plane size much smaller than common 3D NAND flash to deliver a read latency down to 4 µs and a write latency down to 75 µs. XL-FLASH media provides lower read and write latencies than 3D NAND flash media (the read latency is 10-fold lower).

SCM Models and Specifications of Huawei Storage

OceanStor Dorado supports both SCM drives and cards.

SCM drives have the same appearance as common SSDs and are installed in disk slots. They are further classified into 2.5-inch and palm-sized SCM drives. For more details, see the product specifications list.

3.1.5 Power Consumption and Heat Dissipation

OceanStor Dorado uses the following designs to meet the requirements for energy conservation and environment protection:

- Processor with high energy efficiency ratio
- Power module with the industry's highest conversion efficiency
- Fan speed adjustment algorithm for high heat dissipation efficiency

- Staggered power-on design, avoiding peak load on the power supply

The energy-efficient design reduces power supply and heat dissipation costs. The industry-leading heat dissipation technology and customized NVMe SSDs increase the SSD density in an enclosure by 44% (a 2 U enclosure can house up to 36 NVMe SSDs).

Processor with High Energy Efficiency Ratio

The differences between the Arm and x86 platforms are mainly in the internal design of the chip, including the instruction set, pipeline, core distribution, cache, memory, and I/O control. x86 uses the complex instruction set computer (CISC) to gain higher performance by increasing the processor complexity. The x86 instruction set has developed from MMX to SSE and AVX. Arm uses the reduced instruction set computer (RISC), which greatly simplifies the architecture and retains only the necessary instructions. This simplifies the processor and achieves a higher energy efficiency in a smaller size. Arm supports 64-bit operations, 32-bit instructions, 64-bit registers, and 64-bit addressing capability. The in-depth collaboration between hardware and software improves performance and the multi-processor architecture supports performance scalability, providing great advantages in energy efficiency.

Huawei uses the highly integrated SOC chip. The cache coherence bus allows symmetrical multiprocessing (SMP) of up to four processors. Each processor provides various types of I/O ports, storage controllers, and storage acceleration engines.

- I/O ports: Up to 16 x SAS 3.0 ports, 40 x PCIe 4.0 ports, and 2 x 100GE ports
- Storage acceleration engines: RAID engine, SEC engine, and ZIP compression engine

The processor integrates these ports and engines to reduce the number of peripheral chips, simplifying the system design and reducing power consumption.

Efficient Power Module

OceanStor Dorado uses 80 Plus Platinum and Titanium power modules, which provide up to 94% power conversion efficiency and 98% power factor when the power load is 50%, reducing power loss. The Titanium power module can even reach 96% conversion efficiency when the load is 50% and over 90% conversion efficiency when the load is light, minimizing power loss. The power modules have passed the 80 Plus certification (the certificate can be provided).

Table 3-3 80 Plus efficiency requirements

80 Plus Power Module Type	Power Conversion Efficiency (230 V Input)			
Load (%)	10%	20%	50%	100%
80 Plus Bronze	---	81%	85%	81%
80 Plus Silver	---	85%	89%	85%
80 Plus Gold	---	88%	92%	88%
80 Plus Platinum	---	90%	94%	91%
80 Plus Titanium	90%	94%	96%	91%

High-Voltage DC Power Input

OceanStor Dorado supports high-voltage direct current (HVDC) or AC/DC hybrid power input for better power supply reliability. This also reduces the UPS footprint and the equipment room construction and maintenance costs. The HVDC power supply cuts down the intermediate processes of power conversion, improving the power efficiency by 15% to 25%. This significantly saves electricity fees for large data centers every year. In comparison, when low-voltage DC (12 V/48 V) is supplied to high-power devices, thick cables must be used to increase the current, which causes trouble in cable layout. This problem is solved when HVDC is used.

PID Fan Speed Adjustment Algorithm

OceanStor Dorado uses the proportional integral derivative (PID) algorithm to adjust the fan speed, which solves the problems such as slow response of fan speed adjustment, high fan power consumption, great fan speed fluctuation, and loud noise. The PID algorithm allows the system to adjust the fan speed quickly, save energy, and reduce noise.

- The PID algorithm increases energy efficiency by 4% to 9% and prevents fan speed fluctuation.
- The PID algorithm increases the fan response speed by 22% to 53% and significantly reduces the noise.

Staggered Power-On

Staggered power-on prevents the electrical surge that would occur when multiple devices are powered on simultaneously, eliminating the risk on the power supply of the equipment room.

Deduplication and Compression

Deduplication and compression are commonly used data reduction techniques. The system compares the data blocks it receives and deletes duplicate data blocks to save storage space. Because less storage space is required, the power consumption of the system is reduced.

Energy Conservation Certification

The product has passed the RoHS energy efficiency certification.

3.2 SmartMatrix Balanced Architecture

The OceanStor Dorado entry-level and mid-range devices use the dual-controller architecture. Interface modules are single-homed.

Figure 3-29 Interconnection of four controllers

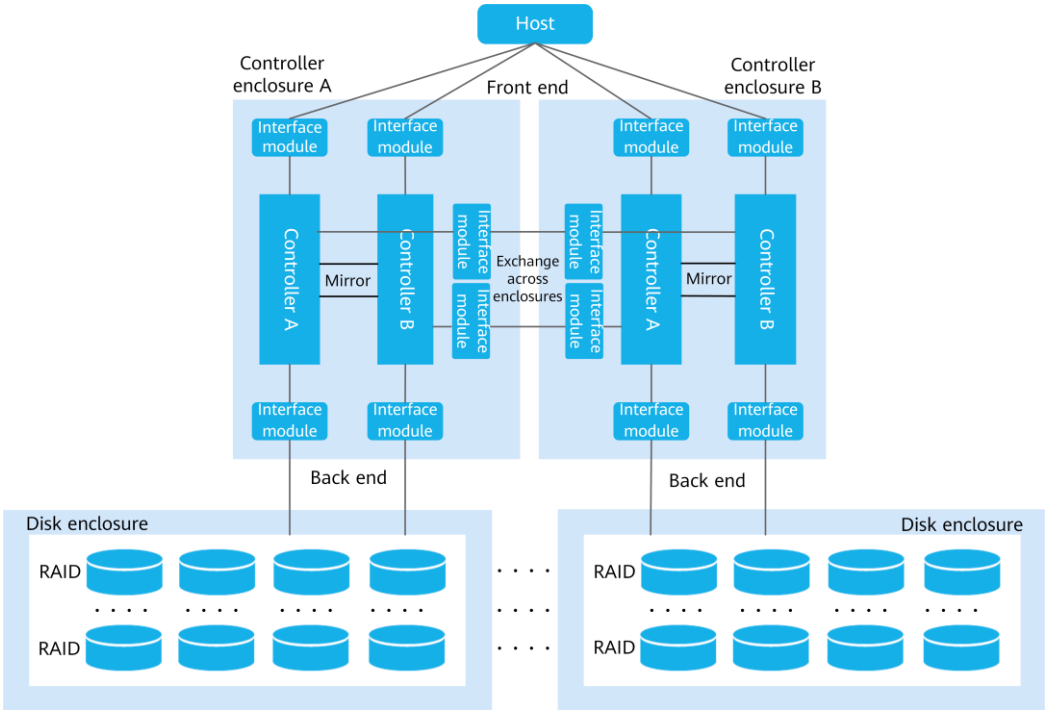
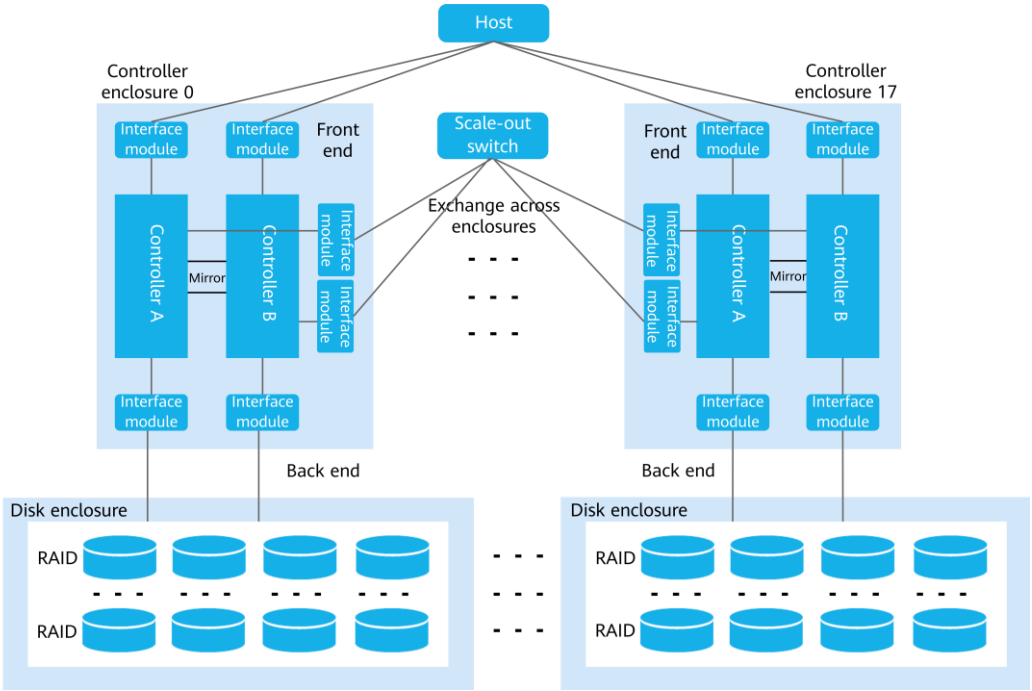


Figure 3-30 Interconnection of 32 controllers



3.2.1 Fully Interconnected Controllers

Fully Interconnected Controllers Within a Controller Enclosure

On OceanStor Dorado mid-range devices, a controller enclosure uses highly reliable hardware that integrates disks and controllers. Each controller enclosure has two built-in controllers that are interconnected through a passive backplane and communicate with each other through the RDMA protocol. Front-end and back-end interface modules are interconnected with the control board using PCIe 3.0 to provide front-end and back-end storage access.

Thanks to the full interconnection of controllers, data flows between controllers do not need to be forwarded by a third component, achieving balanced, fast, and efficient access. No external cables or switches are required for the four controllers within a controller enclosure. This simplifies deployment and eliminates the risk of human errors. In addition, the passive backplane uses only passive components, further improving reliability.

Fully Interconnected Controllers Across Controller Enclosures

Huawei OceanStor Dorado mid-range devices support scale-out to a maximum of 32 controllers by using non-shared interface modules. The devices support scale-out to four controllers through direction connections or switched networks. If more than four controllers are required, 100 Gbit/s DCB switches must be used. For direct connections, each controller enclosure uses 4-port 25 Gbit/s RDMA interface modules to connect to the other controller enclosure. On a switched network, each controller enclosure uses 2-port 100 Gbit/s RDMA interface modules to connect to the switches.

Huawei OceanStor Dorado entry-level devices support scale-out to a maximum of 16 controllers by using non-shared interface modules. The devices support scale-out to four controllers through direction connections or switched networks. If more than four controllers are required, 100 Gbit/s DCB switches must be used. For direct connections, each controller enclosure uses 4-port 25 Gbit/s RDMA interface modules to connect to the other controller enclosure. On a switched network, each controller enclosure uses 2-port 100 Gbit/s RDMA interface modules to connect to the switches.

Direct-connection networking cannot be changed to switched networking online.

3.2.2 RDMA for Low Latency

OceanStor Dorado uses RDMA for networking between controllers and between smart disk enclosures and controller enclosures. Data is directly transmitted between controllers over RDMA links without forwarding.

Data is remotely transferred over RDMA links by interface modules without intervention by the CPUs on either side. This greatly improves data transfer efficiency and reduces the access latency.

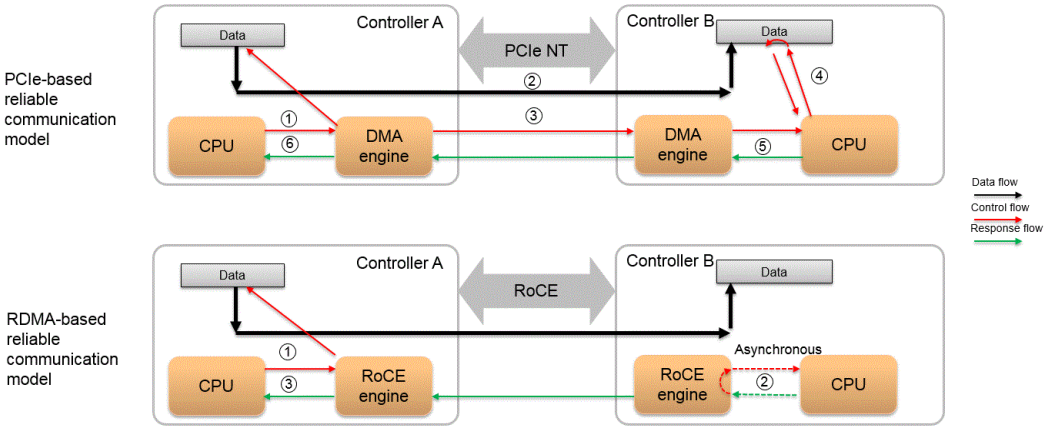
Generally, the data transmission process has two steps:

1. The transmit end sends a control message to the receive end. The receive end prepares memory resources and tells the transmit end to send data.
2. The transmit end sends data to the receive end.

With RDMA, the receive end prepares memory resources in advance and the transmit end sends the control message and data together. This reduces one interaction between the transmit and receive ends for a lower latency.

The RDMA technology provides higher reliability and lower communication latency than PCIe and SAS links. Figure 3-31 compares the I/O interaction processes over PCIe and RDMA links. Data transfer involves I/O request delivery, data transfer to the peer end, data reception at the peer end, data verification, and acknowledgement. In the PCIe communication model, which is bidirectional communication, after data has been transferred from controller A to controller B, the CPU of controller A must notify controller B of data arrival through the control flow to trigger an interrupt on controller B. Then controller B invokes interrupt processing, checks the data, and returns a response. In the RDMA communication model, which is unidirectional communication, after data has been sent successfully, controller A does not need to notify controller B of data arrival. Controller B polls and processes the received data, and returns a response. Therefore, RDMA eliminates the notification of data arrival to reduce the interactions, providing lower latency and higher bandwidth than PCIe.

Figure 3-31 Comparison between I/O interaction processes over PCIe and RDMA links



Reliable Communication Model	Number of Interactions	Total Round Trip Latency (μs)
PCIe communication model	6	About 50
RDMA communication model	3	About 30

When RDMA is used for back-end connections, it is superior to PCIe and SAS interconnection in helping storage systems achieving better performance and higher scalability. The following table compares system performance and scalability based on RDMA, PCIe, and SAS models.

Table 3-4 Comparison of system performance and scalability based on RDMA, PCIe, and SAS models

Item		PCIe (PCIe 3.0 x 4)	SAS (12 Gbit/s SAS 3.0 x 4)	RDMA (100 Gbit/s RDMA port)
Performance	Bandwidth (GB/s)	3.2	4	10

Item		PCIe (PCIe 3.0 x 4)	SAS (12 Gbit/s SAS 3.0 x 4)	RDMA (100 Gbit/s RDMA port)
	Latency (round trip) (μs)	About 50	About 60	About 30
Scalability	Connection distance (copper cable)	About 0.5 m	3–5 m	5–7 m
	Maximum number of NVMe SSDs	Less than 100	Not supported	Unlimited
	Hot swappable	Press the button before removal and insertion. Forcible removal and insertion may cause system breakdown.	Supported	Supported
	Number of controllers that can access a disk enclosure simultaneously	2	4	4

3.3 End-to-End NVMe

NVMe provides reliable NVMe commands and data transmission. NVMe over Fabrics (NVMe-oF) extends NVMe to diverse storage networks for reduced processing overhead for storage network protocol stacks, high concurrency, low latency, and adaptive flexibility for SSD architecture evolution. NVMe-oF maps NVMe commands and data to multiple fabric links, including Fibre Channel, InfiniBand, RoCE v2, iWARP, and TCP.

OceanStor Dorado supports E2E NVMe.

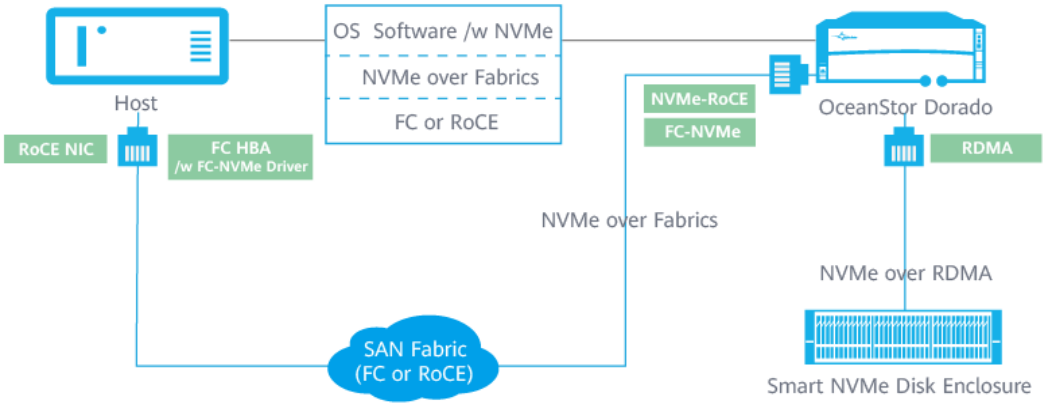
- NVMe over FC (FC-NVMe), NVMe over RoCE v2, and NVMe over TCP/IP (planned) are supported for the networks between hosts and storage systems.
- 32 Gbit/s FC-NVMe, 25 Gbit/s NVMe over RoCE, and 100 Gbit/s NVMe over RoCE ports are supported.
- iSCSI over TCP/IP reduces CPU consumption and latency from network protocol stacks.
- The NVMe multi-queue polling of multi-core CPUs provides lock-free processing of concurrent I/Os to fully realize the computing capacities of processors.
- Read requests to NVMe SSDs are prioritized, accelerating responses to read requests when NVMe SSDs are written.
- OceanStor Dorado reduces access latency to below 100 μs with an E2E NVMe design for a 50% latency reduction compared to the previous generation.

3.3.1 End-to-End NVMe-oF Deployment

Figure 3-32 illustrates how Huawei storage systems support E2E NVMe over the entire data path, including hosts, front-end networks, back-end disk enclosures, and SSDs.

- Host OS: mainstream OSs with the NVMe protocol, such as SUSE and Red Hat.
- Host NIC: FC HBA and RoCE card with an NVMe driver.
- SAN fabric: FC switch and DCB Ethernet switch, which are transparent to the NVMe protocol.
- Storage system: OceanStor Dorado series with FC-NVMe and NVMe over RoCE interface modules.
- Disk enclosure: smart NVMe disk enclosure with NVMe SSDs.

Figure 3-32 NVMe over Fabrics



FC-NVMe applications can be carried by the existing FC SANs and switches of data centers. Customers only need to install the NVMe driver on the FC HBAs of the hosts to support NVMe, protecting the original investment.

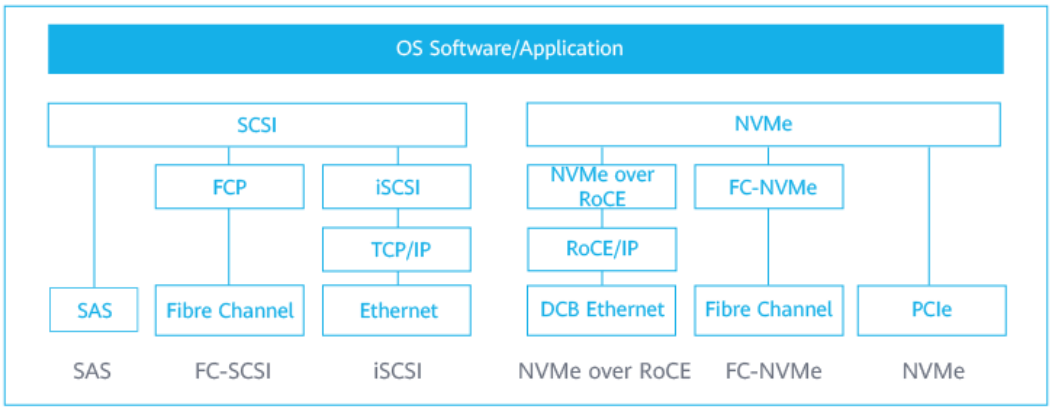
NVMe over RoCE applications require RoCE NICs and DCB Ethernet switches. They converge and unify the server clusters, front-end LANs, and storage SANs of new data centers for reduced TCO.

RoCE NICs and DCB switches form the basis for RDMA and NVMe applications by controlling congestion and backpressure with service flows to create an enhanced Ethernet without packet loss.

3.3.2 Comparison of Fibre Channel, iSCSI, and NVMe-oF

NVMe is a new protocol command set for block storage. Figure 3-33 illustrates its evolution from back-end direct-connected storage to network-connected storage throughout the data path.

Figure 3-33 Comparison of SCSI and NVMe protocol stacks



1. NVMe replaces SCSI by defining a new protocol command set of block storage. It uses PCIe transmission channels to greatly reduce latency and improve bandwidth.
2. NVMe-oF solves the problems of NVMe scalability and extension:
   a. PCIe NVMe is limited by the number of PCIe bus addresses for a maximum of 255 nodes and approximately 100 SSDs.
   b. PCIe NVMe only supports direct connections to controllers within limited distances.
3. NVMe over RoCE establishes E2E NVMe over DCB lossless Ethernet with the low latency and low CPU usage of RDMA. It converges the LANs and SANs of data centers. DCB controls congestion and backpressure based on service flows for lossless transmission of Ethernet packets.
4. The existing FC infrastructure supports NVMe over FC for quick deployment of E2E NVMe.

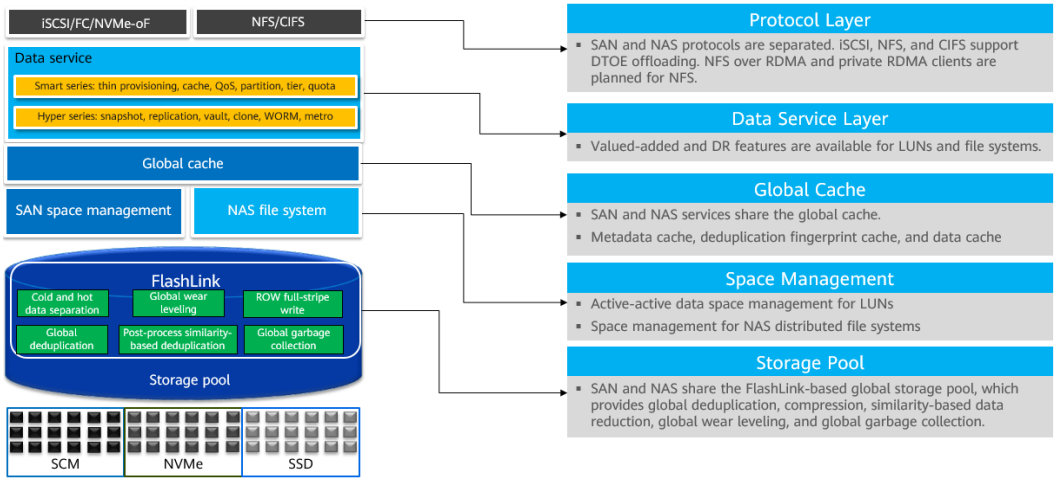
4 Software Architecture

OceanStor Dorado uses the symmetric active-active software architecture to implement load balancing and FlashLink® to optimize the system based on SSD characteristics, fully utilizing all-flash storage system performance. All models of OceanStor Dorado use the unified software architecture and support interconnection with each other.

4.1 Unified Storage Architecture for SAN and NAS

OceanStor Dorado integrates both SAN and NAS on one set of hardware and software without using independent NAS gateways. The systems support file access protocols such as NFS and CIFS, and block access protocols such as Fibre Channel, iSCSI, and NVMe-oF. Both SAN and NAS can scale out to multiple controllers. Hosts can access any LUN or file system from a front-end port on any controller.

Figure 4-1 OceanStor Dorado software architecture



The OceanStor Dorado system architecture consists of the following subsystems:

- Storage pool: globally unifies storage pool services. It uses redirect-on-write (ROW) to allocate space for LUN and file system data, globally deduplicates and compresses data, and identifies and distributes metadata and user data to SSDs. It also provides fast reconstruction with RAID 2.0+ and global garbage collection in the background.
- Space management: allocates and reclaims space for LUNs and file systems with thin provisioning.
- Global cache: provides the read/write and metadata caches for LUNs and file systems.
- Data service layer: provides disaster recovery capabilities such as remote replication and active-active configuration for LUNs and file systems. It provides unified data replication and manages the replication configuration and networks.
- Protocol layer: provides protocol parsing, I/O receiving and sending, and error processing for LUNs and file systems.

The storage pools of OceanStor Dorado directly allocate space for the file systems and LUNs, which directly interact with the underlying storage pools for parallel SAN and NAS architecture.

Parallel architecture streamlines storage with the shortest I/O paths for LUNs and file systems. In addition, space management of LUNs and file systems is independent of each other, enhancing reliability.

4.1.1 Active-Active Logical Architecture for SAN

In an asymmetrical logical unit access (ALUA) architecture, each LUN is owned by a specific controller. Customers need to plan the owning controllers of LUNs for load balancing. However, it is difficult for an ALUA architecture to implement load balancing on live networks because service pressures vary with LUNs and vary in different periods for a same LUN.

OceanStor Dorado uses the symmetric active-active software architecture, which uses the following technologies:

- Load balancing algorithm

  Balances the read and write requests received by each controller.
- Global cache

  Allows that LUNs and filesystems have no ownership. Each controller processes received read and write requests, achieving load balancing among controllers.
- RAID 2.0+

  Evenly distributes data to all disks in a storage pool, balancing disk loads.

4.1.1.1 Global Load Balancing

OceanStor Dorado hashes the logical block addressing (LBA) of each host read/write request to determine the controller that processes the request. Huawei multipathing software UltraPath, FIMs, and controllers negotiate the hash method and parameters to implement intelligent distribution of read and write requests. UltraPath and FIMs work together to directly distribute a read/write request to the optimal processing controller, avoiding forwarding between controllers.

If no FIMs are available (on OceanStor Dorado 8000 and OceanStor Dorado 18000 that use iSCSI front-end modules and OceanStor Dorado 3000 V6, OceanStor Dorado 5000, and OceanStor Dorado 6000), UltraPath sends I/O requests to an interface module on the optimal controller. Then, the interface module directly sends the requests to the corresponding

controller for processing. In this way, I/O requests do not need to be forwarded between controllers.

If neither FIMs nor UltraPath is available, a controller forwards received I/O requests to the corresponding controller based on the hash result of the I/O LBA, achieving load balancing among controllers. You are advised to install UltraPath to prevent overhead caused by I/O forwarding between controllers.

4.1.2 Active-Active Logical Architecture for NAS

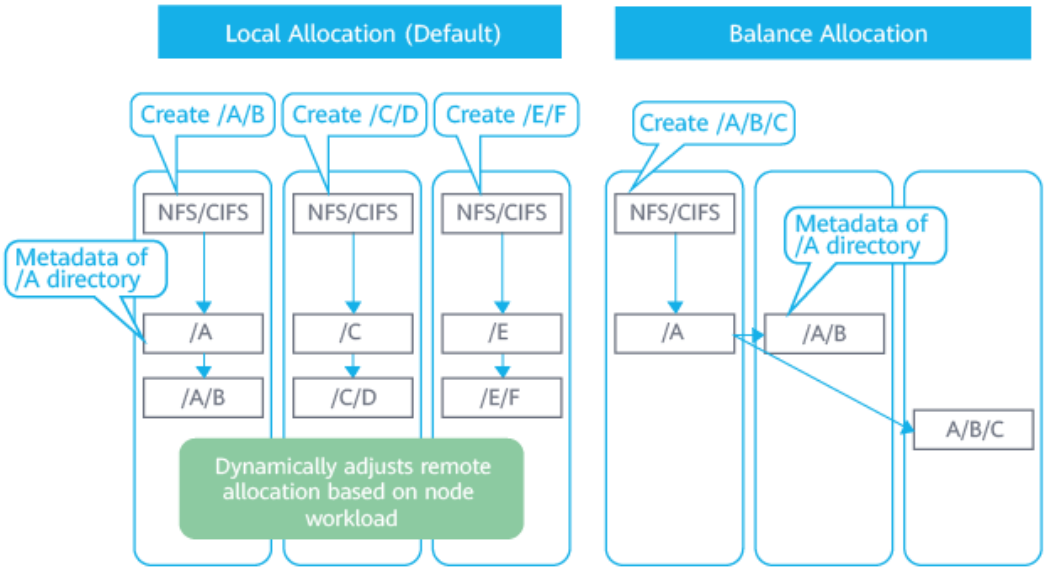
NAS file systems traditionally use active-passive architecture. Each file system is owned by a specific controller. When file systems are created, the administrator must plan their owning controllers so they run on different controllers to balance the load. As a result, a file system can only utilize the resources of its owning controller. Systems with only one file system waste resources on other idle controllers, resulting in limited system performance. This architecture therefore does not support a single namespace. When there are multiple file systems, they usually carry different service loads, making it difficult to balance the global loads.

OceanStor Dorado NAS uses distributed file systems to eliminate controller ownership. Directories and files in a file system are evenly distributed to all controllers by a balancing algorithm. Read and write requests are equally distributed on each controller so that one file system can fully utilize the resources of the entire storage cluster. Customers can use the file system in one namespace or multiple file systems based on their service plans.

4.1.2.1 Active-Active File System

The distributed file systems of OceanStor Dorado apply to file sharing scenarios with coexisting mass volumes of small and large files. Data in each directory is evenly distributed to each controller for load balancing. The same controller processes the I/Os of a directory and its files to eliminate forwarding across controllers and improve performance for directory traversal, attribute traversal, and batch attribute configuration. When large files are written to a storage pool, RAID 2.0+ globally distributes their data blocks to all SSDs in the storage pool for improved write bandwidth.

Figure 4-2 Directory balancing policies



OceanStor Dorado NAS supports two directory balancing policies:

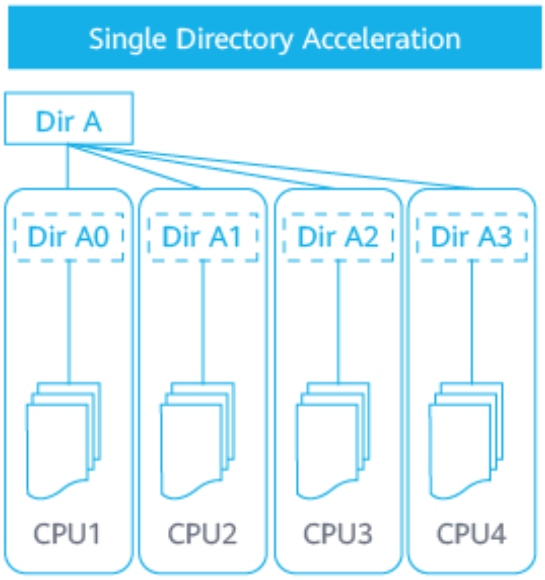
- Local Allocation

  The controller whose IP address is mounted to a host preferentially processes the directory access request of that host to optimize performance and latency. The system dynamically adjusts the distribution ratios of remote controllers in the background based on the data capacity, file quantity, and load of each controller. When one controller exceeds the others in one of these parameters and the difference reaches a threshold, the system automatically increases the distribution ratios of remote controllers to universally balance controller loads. This is the default distribution mode of OceanStor Dorado, and it can achieve the optimal performance when each IP address of each controller is evenly mounted to the hosts. Local allocation is best for scenarios with requirements for high performance and low latency, such as EDA simulation, HPC, software compilation, and parallel computing. It is best used with the built-in DNS of OceanStor Dorado, which automatically balances the IP address mounting of hosts.

- Balance Allocation

  The directory access requests from hosts are evenly distributed to each controller for best load balancing. The system also performs load balancing in the background based on the capacity of each controller. Balance allocation is best for hosts with uneven mounting of storage IP addresses, unbalanced service loads, or scenarios with low requirements for latency.

Figure 4-3 Parallel processing

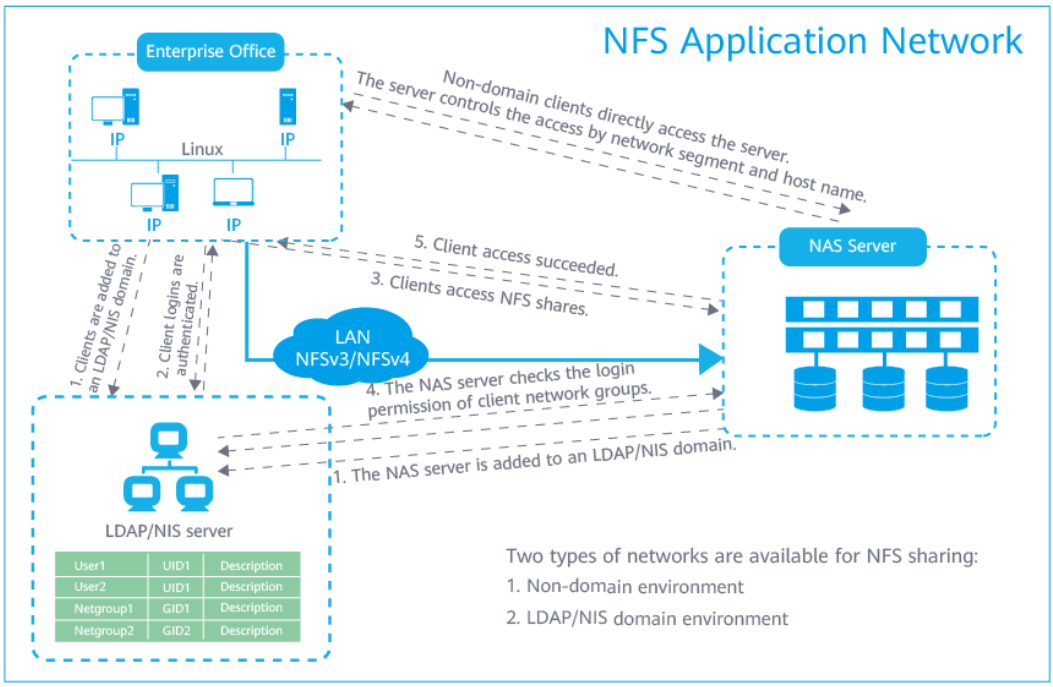


When a directory and its files are frequently accessed, OceanStor Dorado distributes the load of this directory to multiple cores of multiple CPUs for parallel processing and improved efficiency.

4.1.2.2 NAS Protocols

4.1.2.2.1 NFS Protocol

Figure 4-4 NFS application network



Network File System (NFS) is a common network file sharing protocol used in Linux and Unix environments.

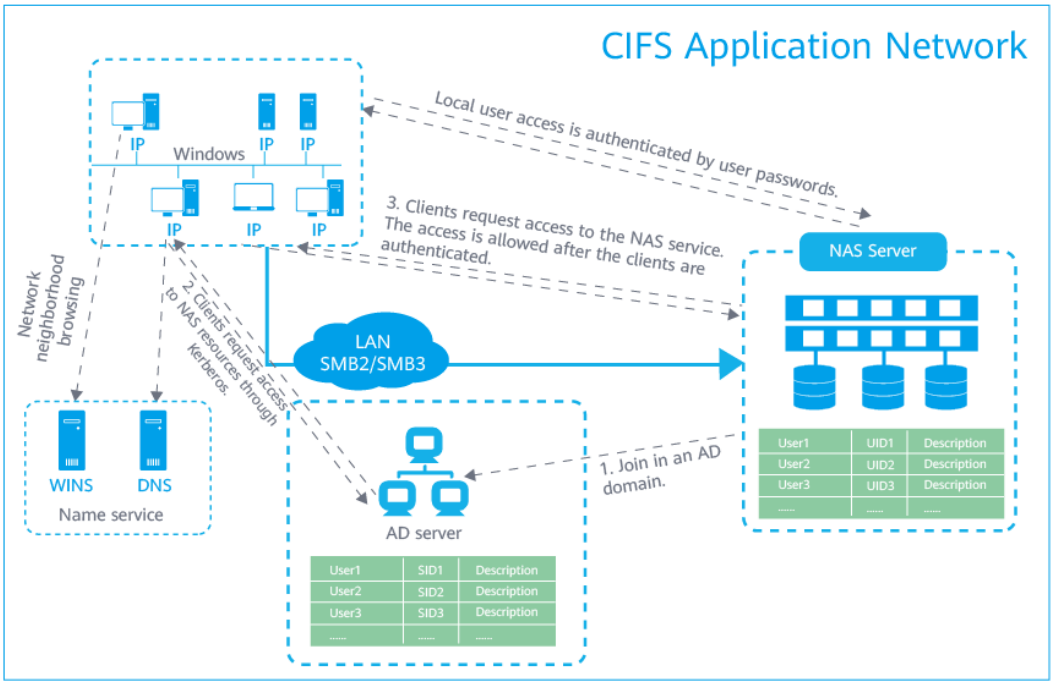
NFS is mainly used for:

- Network file sharing in Unix and Unix-like OSs such as Linux, AIX, and Solaris
- VMware and Xen VMs
- SAP HANA and Oracle databases

OceanStor Dorado supports the NFSv3 and NFSv4.1 protocols (see the product specifications list for details) and usage in local user environments (non-domain environments) and LDAP/NIS domain environments. Users can import LDAP certificates for secure domain transmission with LDAPS. In a multi-tenant environment, the LDAP/NIS service can be configured separately for each tenant.

4.1.2.2.2 CIFS Protocol

Figure 4-5 CIFS application network



Server Message Block (SMB), also known as Common Internet File System (CIFS), is a network file sharing protocol widely used in Windows.

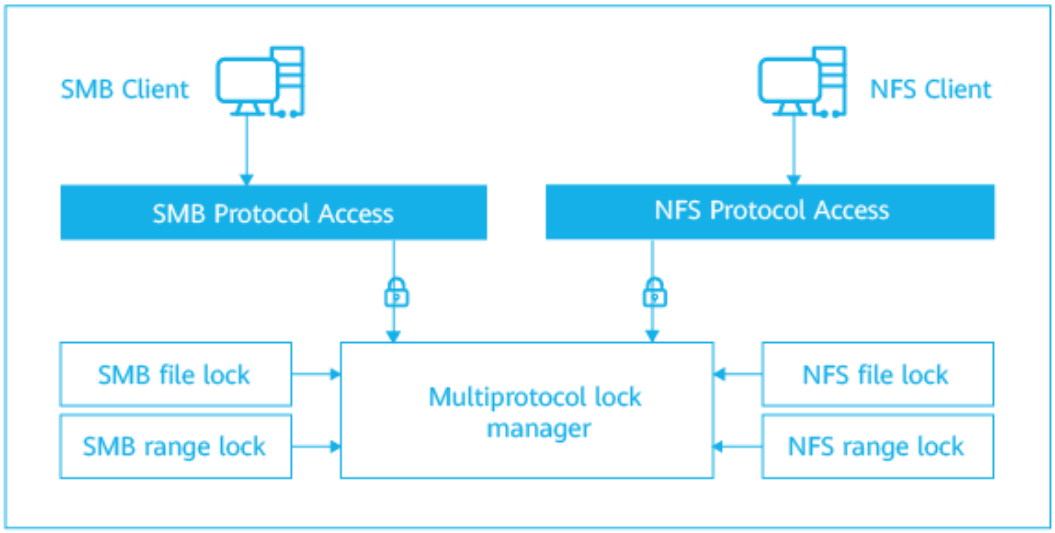
CIFS is mainly used for:

- Network file sharing in Windows
- Hyper-V scenarios

OceanStor Dorado supports SMB 2 and SMB 3, and can be used in local user environments (non-domain environments) and AD domain environments. Kerberos and NTLM authentication by AD domain is also supported. The AD domain environments can be individual, parent-child, or trusted domains. In a multi-tenant environment, an independent AD domain can be configured for each tenant.

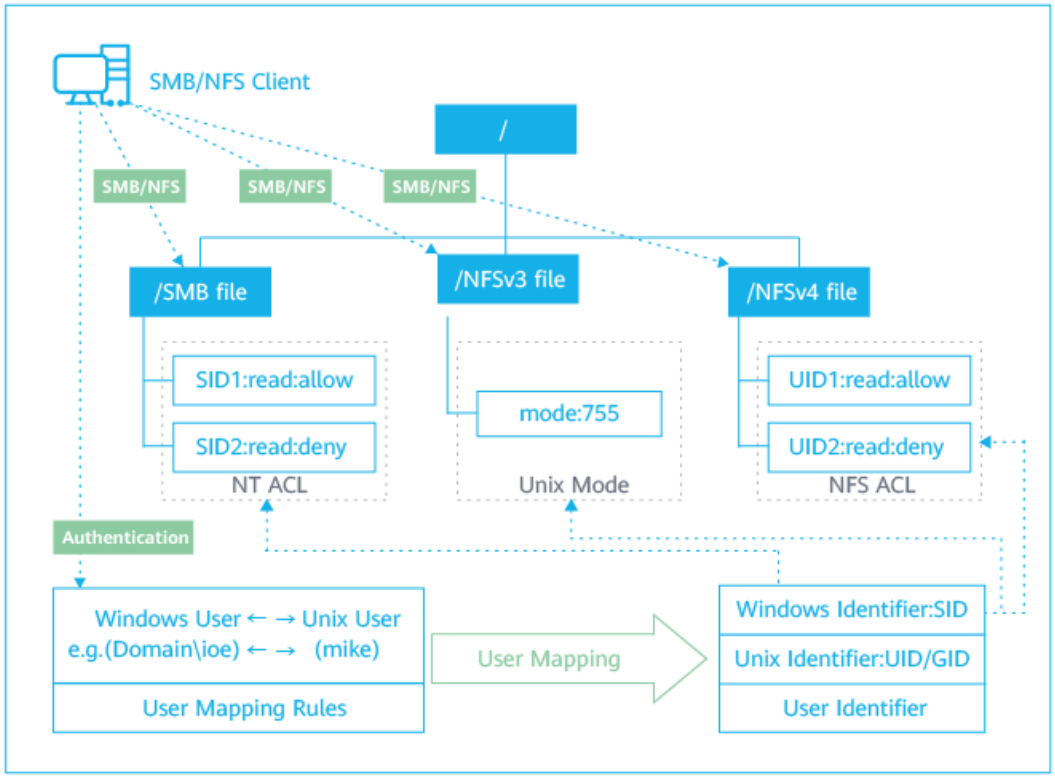
4.1.2.2.3 Multi-Protocol Access

OceanStor Dorado supports access across the NFS and SMB protocols. Both NFS and SMB shares are configurable for a file system. The system uses a multi-protocol lock manager with NFS and SMB for exclusive file access without data corruption or inconsistency.



OceanStor Dorado supports the following security styles for multi-protocol access:

- NT mode: File attributes and ACL permissions can only be set on Windows clients with SMB. The system automatically maps the file permissions of SMB users to the NFS users on Linux clients based on the user mapping relationship for successful authentication of NFS users during file access. NFS users are prohibited from setting the mode or ACL permissions for files.
- Unix mode: File permissions can only be set on Linux/Unix clients with NFS. The system automatically maps the file permissions of NFS users to the SMB users based on the user mapping relationship for successful authentication of SMB users during file access. SMB users are prohibited from setting the ACL permissions of files.
- Mixed mode: Users can set permissions on both Unix and Windows clients, which will overwrite each other. The last configured permissions prevail.
- Native mode: Users can set permissions on both Unix and Windows clients. NFS and SMB permissions are saved and authenticated separately.



4.1.2.2.4 NDMP Protocol

Network Data Management Protocol (NDMP) is an open protocol used to manage enterprise data. It was initiated by Network Appliance and Legato Systems. Currently, the Storage Network Industry Association (SNIA) has set up a work group to take responsibility for the development of the protocol standard.

NDMP defines a network-based control mechanism to control data backup and recovery, as well as data transmission between storage systems and tape libraries. Also, NDMP allows storage systems to directly transmit data to tape libraries or backup servers over a network, without the need of backup clients or consumption of a large amount of network and server resources.

The NDMP network architecture consists of the NDMP client, NDMP server, and tape library. The architecture separates the data and control planes to implement direct data transmission between storage systems and efficient server-free data backup. Both the backup data source and target devices must support NDMP. OceanStor Dorado supports NDMP-based NAS backup and 2-way (as shown in Figure 4-6) and 3-way (as shown in Figure 4-7) backup network topologies. An NDMP client is a backup server on which backup software is installed. It allows you to configure and manage backup jobs. An NDMP server is an OceanStor Dorado device. It provides the NDMP capability to read/write the source file system, the capability of obtaining snapshots and the snapshot difference ratio, and the capability to access tape libraries. A tape library is a third-party tape library or VTL device that provides Fibre Channel or iSCSI driver access interfaces to store, read, and write backup data.

Figure 4-6 Two-way network topology

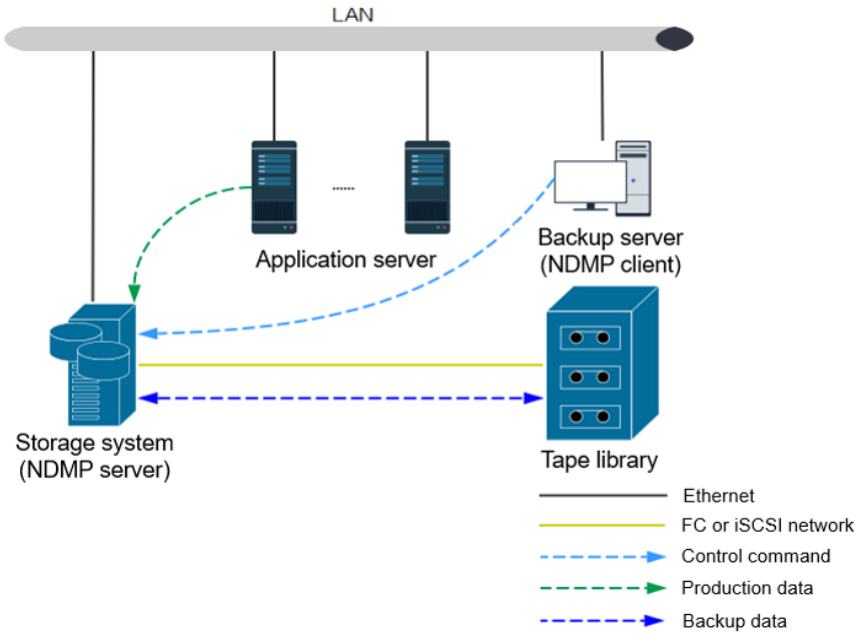
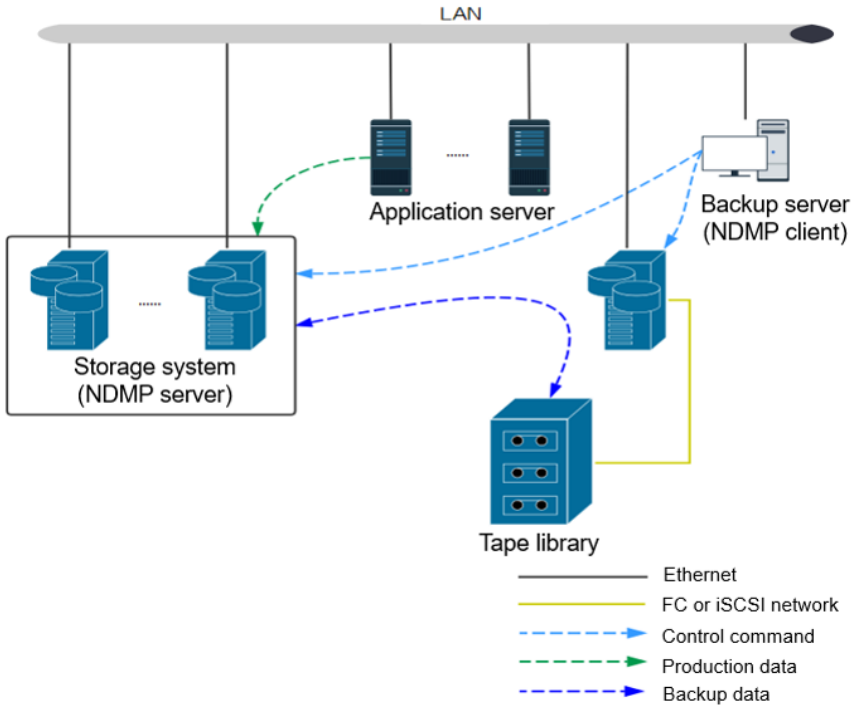


Figure 4-7 Three-way network topology



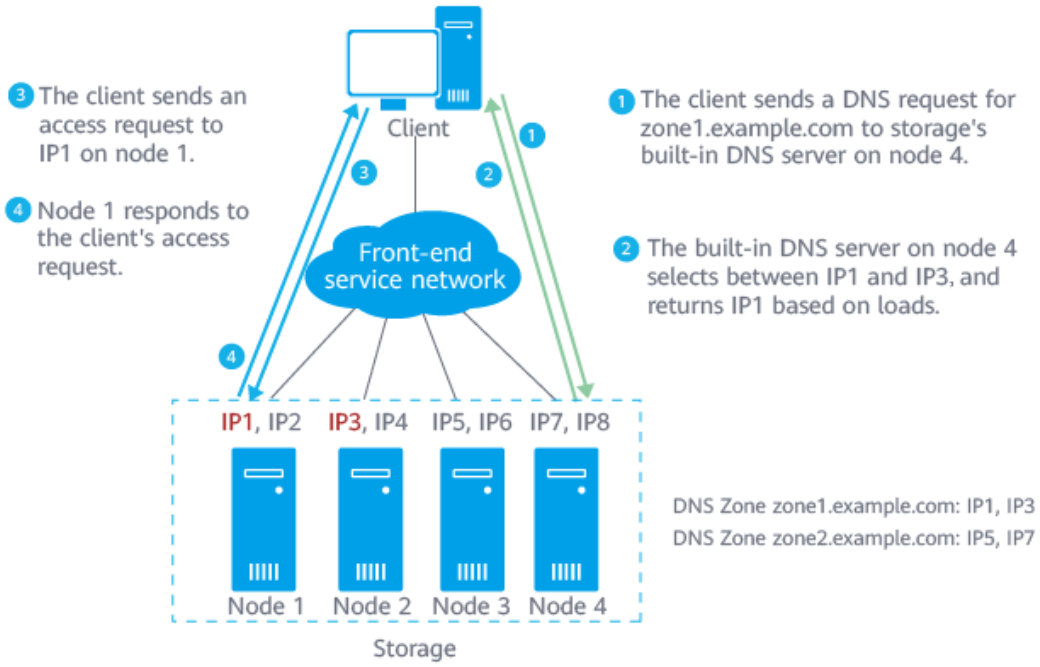
4.1.2.2.5 S3 Protocol

OceanStor Dorado provides the object storage service (OBS), for which the NAS file system provides underlying storage capabilities. The OBS is compatible with the Amazon Simple Storage Service (S3) protocol.

This simplified storage protocol uses a three-layer model consisting of accounts, buckets, and objects where a bucket and object can be regarded as a directory and file respectively. Users can use the uniform resource identifier (URI) to locate and use their own data. The protocol discards the directory tree structure, and simplifies the read and write semantics. It is suitable for storing mass unstructured data, or the data that is frequently read than written.

The OceanStor Dorado OBS features high scalability, security, reliability, efficiency, and integration. It applies to mass data storage and centralized backup, and provides users with the following benefits: high reliability, easy maintenance, and easy expansion.

4.1.2.3 Built-in DNS Load Balancing



A host can use a domain name to access NAS services on a storage system. A domain name system (DNS) can balance the loads of multiple IP addresses for a domain name. External DNS servers cannot detect the CPU usage of the node or the bandwidth usage of the port where each IP address resides. They generally use the round-robin policy, which cannot provide the optimal load balancing effect.

Traditional external DNS servers have the following problems:

- Their general use of round-robin policies results in suboptimal load balancing. They cannot detect the CPU usage of nodes or the bandwidth usage of ports for each IP address.
- DNS services for the storage system are unreliable due to their dependence on external DNS servers.

OceanStor Dorado provides built-in DNS load balancing, which can detect the load of each IP address in the storage system for better load distribution, performance, and system reliability.

DNS load balancing of OceanStor Dorado supports the round-robin policy or load balancing by node CPU usage, number of connections, node bandwidth usage, or overall load.

Table 4-1 DNS load balancing policies

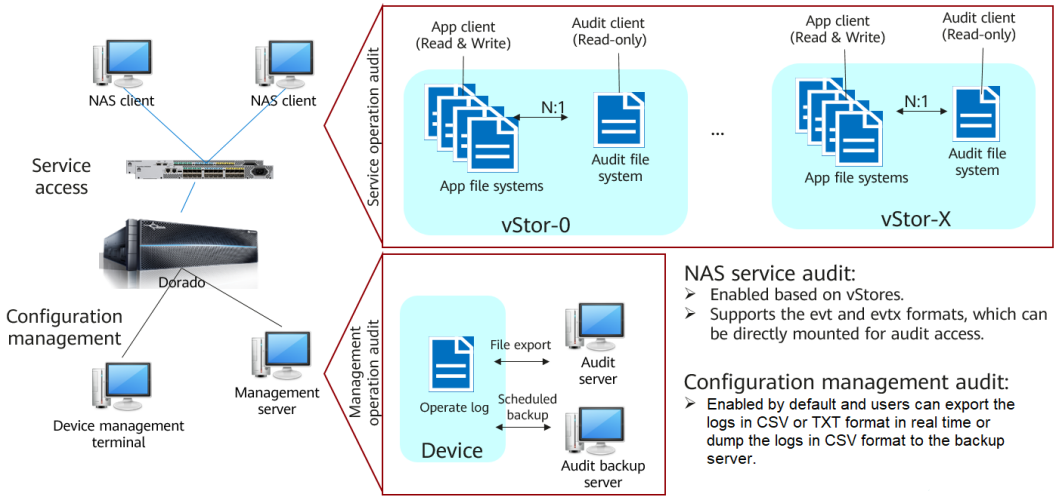
Policy	Description	Advantage	Disadvantage
Weighted round robin	Performance data determines the weight. IP addresses which process loads and are under the same domain name have an equal chance of being selected for processing.	This policy optimizes load balancing when the clients deliver similar NAS services.	• The load is balanced only among IP addresses. The real-time load for each node is undetectable. • Load balancing is disrupted by DNS requests (**ping**, **nslookup**, **showmount**, and other client commands), service timeouts, and authentication failures.
CPU usage	The CPU usage of each node determines the weight. The storage system uses the weight to select a node to process client services.	The node with the lowest CPU usage is selected to process client services. This policy can be used to manage concurrent service requests.	• CPU usage only significantly changes when nodes carry services. Clients must first deliver services before load balancing can occur.
Bandwidth usage	The total bandwidth usage of each node determines the weight. The storage system uses the weight to select a node to process client services.	The node with the lowest total port bandwidth usage is selected to process client services. This policy can be used to manage concurrent service requests.	• Load balancing is implemented among nodes based on the total bandwidth usage, not among the physical ports. • Bandwidth usage only significantly changes when nodes carry services. Clients must first deliver services before load balancing can occur.
NAS connections	The NAS connections of each node determine the weight. The storage system uses the weight to select a node to process client services.	The node with the fewest NAS connections is selected to process client services. This policy can be used to manage concurrent service requests.	• Load balancing can be disrupted because the connection quantity does not reflect real-time performance data. • A node clears the connection but retains the mount point if NFS fails to send packets in a specified period of time. Load balancing may be disrupted if a new client is mounted to this node based on the number of connections.
Overall load	The overall load of CPU usage, bandwidth usage, and number of NAS connections determines node selection. Less loaded nodes are more	• This policy considers the CPU usage and throughput of each node and selects the node with the lightest load to carry	• The overall load only significantly changes when nodes carry services. Clients must first deliver services before

Policy	Description	Advantage	Disadvantage
	likely to be selected.	client services. • This policy does not always select the node with the lightest load. Instead, it maximizes the service load of the lightest node to balance the entire system.	load balancing can occur.

4.1.2.4 Audit Log

The audit log feature records device operations in logs and provides the log audit function to manage device access security and analyze service patterns and trends. OceanStor Dorado provides audit logs for management operations and NAS service operations, as shown in Figure 4-8.

Figure 4-8 Audit log



Service access audit is to log the NAS service access operations on NAS clients. An independent audit log file system is used to record the logs. This function can be separately configured for each vStore for isolation and supports the guarantee and non-guarantee modes. Specific audit objects can be configured. For details about the audit objects, see the feature guide. After the audit log function is enabled for a vStore, operation logs must be recorded for operations to be audited before the operations are performed. In guarantee mode, if an operation fails to be recorded in the operation log, the operation will not be performed, which may cause service interruption. In non-guarantee mode, the operation can still be performed but the operation log is incomplete. The system uses the guarantee mode by default, and users can modify the mode online. The audit log file system stores logs in evt format and can be shared and accessed by the audit log server in real time.

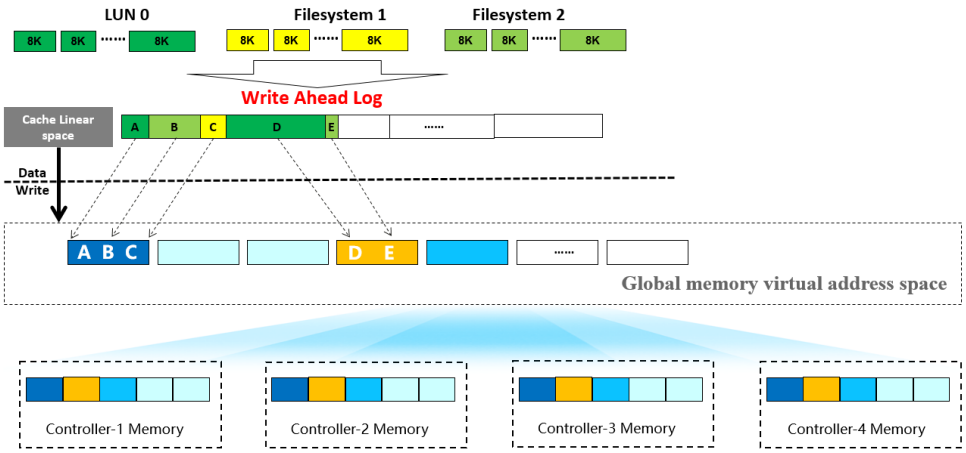
Management operation audit is to log device configuration and management operations performed by device administrators. Only configuration modifications are recorded (query operations are not recorded). The log file is stored in the system and does not support online

access. This function is enabled by default and cannot be disabled. Users can export the logs in CSV or TXT format in real time or dump the logs in CSV format to the backup server.

4.1.3 Global Cache

On OceanStor Dorado, caches of all controllers constitute a global cache. Data on each LUN or Filesystem is evenly distributed to the caches of all controllers and processed by them. In this way, LUNs and Filesystems do not have any ownership. Figure 4-9 shows the principle of the global cache of OceanStor Dorado with four controllers.

Figure 4-9 principle of the global cache



OceanStor Dorado employs global unowned LUNs and file systems. Blocks of LUNs and file systems are distributed to the memory of each controller using distributed algorithms. In this way, a LUN or file system can use cache resources of the entire storage cluster.

For a write request, data is first written into a controller , and the data is divided into many data blocks by using a virtual address layer of a global cache, and the data blocks are dispersed into memories of all controllers.In addition, the storage system mirrors data in the write cache once to ensure reliability. For entry-level and mid-range OceanStor Dorado storage, each cache block additionally mirrors one piece of data to the other controller. For high-end storage, each cache block can additionally mirror two pieces of data to different controllers.

For a read request, a controller searches a global cache virtual address layer after receiving a read request,so that valid data is found in the controller or another controllers, and then data is returned to the host. If the data cannot be found in the cache, the system obtains the data from the storage pool.

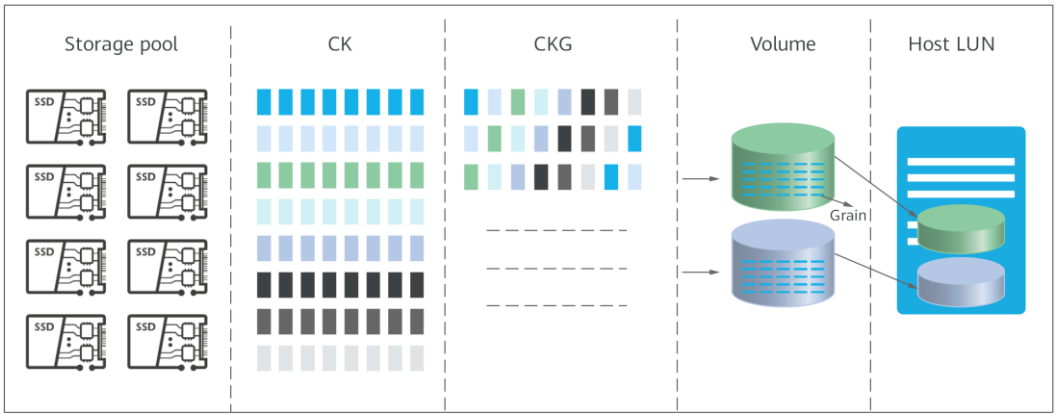
4.1.4 RAID 2.0+

If data is not evenly stored on SSDs, some heavily loaded SSDs may become a system bottleneck. OceanStor Dorado uses RAID 2.0+ to implement fine-grained division of SSDs and evenly distributes data to all LUNs on each SSD, balancing load among disks. The storage systems implement RAID 2.0+ as follows:

- Multiple SSDs form a storage pool.
- 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 based on the customer-configured RAID policy.
- A chunk group is further divided into grains (typically 8 KB per grain), which are the smallest unit for volumes.

Figure 4-10 RAID 2.0+ mapping



4.2 FlashLink®

FlashLink® associates storage controllers with SSDs by using a series of technologies for flash media, ensuring performance of flash storage. The key technologies of FlashLink® include the intelligent multi-core technology, ROW full-stripe write, multistreaming, end-to-end I/O priority, smart disk enclosures, intelligence, and end-to-end NVMe. These techniques ensure a consistent low latency and high IOPS of OceanStor Dorado.

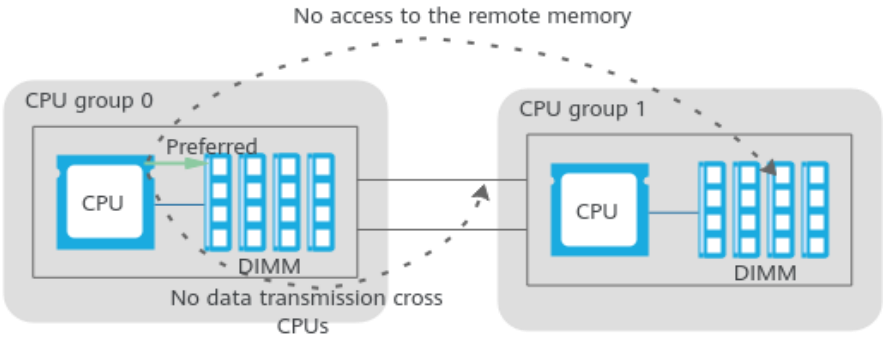
4.2.1 Intelligent Multi-Core Technology

A storage system needs to provide powerful compute capabilities to deliver high throughput and low latency as well as support more value-added features, such as data deduplication and compression. On OceanStor Dorado, each controller has four high-performance CPUs, providing up to 192 cores. OceanStor Dorado provides the industry's most CPUs and cores in a controller, offering powerful compute capabilities. The intelligent multi-core technology brings CPUs' compute capabilities into full play, allowing performance to increase linearly with the number of CPUs. The intelligent multi-core technology consists of CPU grouping, service grouping, and lock-free design between cores.

CPU Grouping

In a multi-CPU architecture, that is, a non-uniform memory access (NUMA) architecture, each CPU can access either a local or remote memory. Accessing the local memory involves a lower delay and less overhead than accessing a remote one. OceanStor Dorado considers each CPU in a controller and its local memory as a CPU group. A CPU group processes received host read/write requests from end to end if possible. This eliminates the overhead of communication across CPUs and accessing the remote memory. CPU grouping enables each CPU group to process different host read/write requests, allowing performance to increase linearly with the number of CPUs.

Figure 4-11 CPU grouping



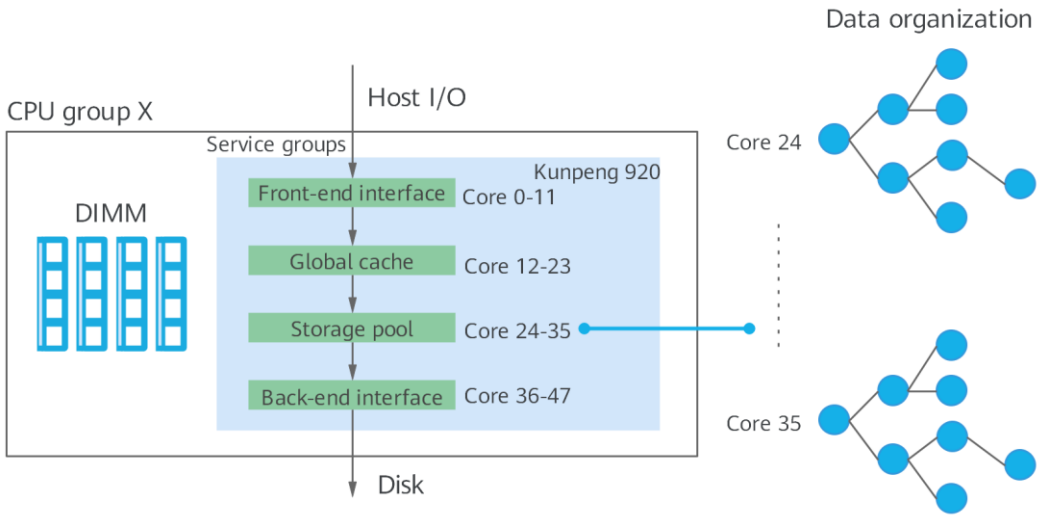
Service Grouping

CPU grouping enables different CPUs to process different read/write requests. However, various service processes running in each CPU compete for CPU cores, causing conflicts and hindering the linear increase of performance. OceanStor Dorado classifies services into front-end interface service, global cache service, back-end interface service, and storage pool service groups, and allocates dedicated CPU cores to each service group. In this way, different service groups run on different CPU cores. For example, if a CPU has 48 cores, each of the front-end interface service, global cache service, back-end interface service, and storage pool service uses 12 cores. The system dynamically adjusts the number of cores allocated to each service group based on the service load. Service grouping prevents CPU resource contention and conflicts between service groups.

Lock-free Design Between Cores

In a service group, each core uses an independent data organization structure to process service logic. This prevents the CPU cores in a service group from accessing the same data structure, and implements lock-free design between CPU cores. The following figure shows lock-free design between CPU cores. In the example, CPU cores 24 to 35 are allocated to the storage pool service group and run only the storage pool service logic. In the storage pool service group, services are allocated to different cores, which use independent data organizations to prevent lock conflicts between the cores.

Figure 4-12 Lock-free design between CPU cores

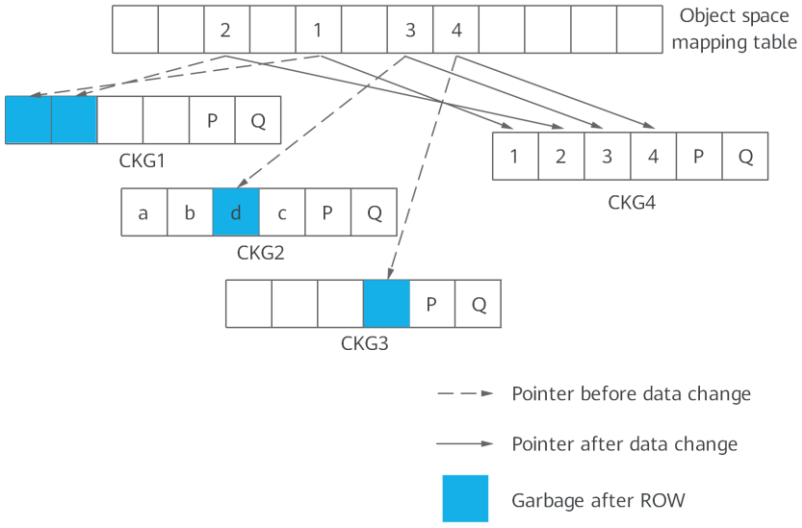


The CPU grouping, service grouping, and lock-free technologies enable system performance to increase linearly with the number of controllers, CPUs, and CPU cores.

4.2.2 ROW Full-Stripe Write

Flash chips on SSDs can be erased for a limited number of times. In traditional RAID overwrite (write in place) mode, hot data on an SSD is continuously rewritten, and its mapping flash chips wear out quickly. OceanStor Dorado uses redirect on write (ROW) full-stripe write for both new data writes and old data rewrites. It allocates a new flash chip for each write, balancing the number of erase times of all flash chips. 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 4-13 ROW full-stripe write



In Figure 4-13, 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, a storage 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. During this process there is no need to read any existing data.

For traditional RAID, for example, RAID 6, when D0 is changed, the system must first read D0, P, and Q, and then write new nD0, nP, and nQ. Therefore, both the read and write amplifications are 3. Generally, the read and write amplification of small random I/Os in traditional RAID (xD+yP) is y+1.

Figure 4-14 Write amplification of traditional RAID 6

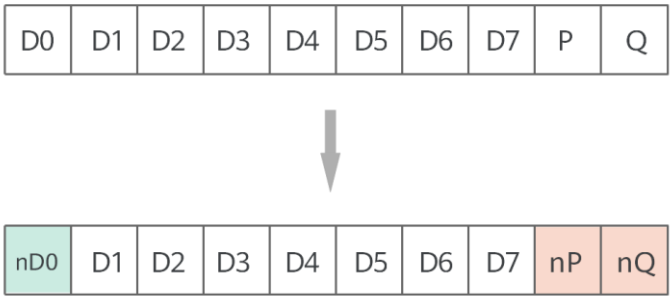


Table 4-2 lists the write amplification statistics of various traditional RAID levels.

Table 4-2 Write amplification of traditional RAID levels

RAID Level	Write Amplification of Random Small I/Os	Read Amplification of Random Small I/Os	Write Amplification of Sequential I/Os
RAID 5 (7D+1P)	2	2	1.14 (8/7)
RAID 6 (14D+2P)	3	3	1.14 (16/14)
RAID-TP (not available in traditional RAID)	-	-	-

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

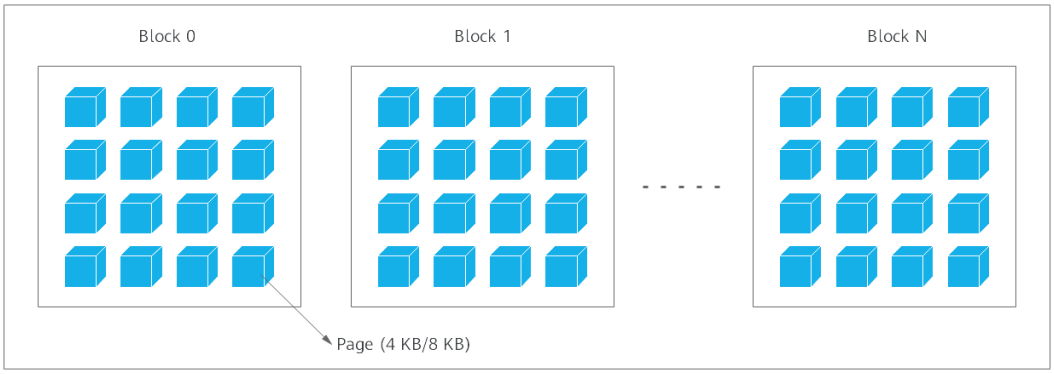
Table 4-3 Write amplification in ROW full-stripe write

RAID Level	Write Amplification of Random Small I/Os	Read Amplification of Random Small I/Os	Write Amplification of Sequential I/Os
RAID 6 (22D+2P)	1.09 (24/22)	0	1.09
RAID-TP (21D+3P)	1.14 (24/21)	0	1.14

4.2.3 Multistreaming

SSDs use NAND flash. Figure 4-15 shows the logical structure of an SSD. Each SSD consists of multiple NAND flash chips, each of which contains multiple blocks. Each block further contains multiple pages (4 KB or 8 KB). The blocks in NAND flash chips must be erased before being written. Before erasing data in a block, the system must migrate valid data in the block, which causes write amplification in the SSD.

Figure 4-15 Logical structure of an SSD



The multistreaming technology classifies data and stores different types of data in different blocks. There is a high probability that data of the same type is valid or garbage data at the same time. Therefore, this technology reduces the amount of data to be migrated during block erasure and minimizes write amplification on SSDs, improving the performance and service life of SSDs.

During garbage collection, an SSD must migrate any 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, according to change frequency. For example, metadata (hot) is updated more frequently and is more likely to cause garbage than user data (cold). The multistreaming technology enables SSD drives and controllers to work together to store hot and cold data in different blocks. This increases the possibility that all data in a block is invalid, reducing valid data to be migrated during garbage collection and improving SSD performance and reliability.

Figure 4-16 shows data migration for garbage collection before separation of hot and cold data, in which a large amount of data needs to be migrated. Figure 4-17 shows data migration

for garbage collection after separation of hot and cold data, in which less data needs to be migrated.

Figure 4-16 Data migration for garbage collection before separation of hot and cold data

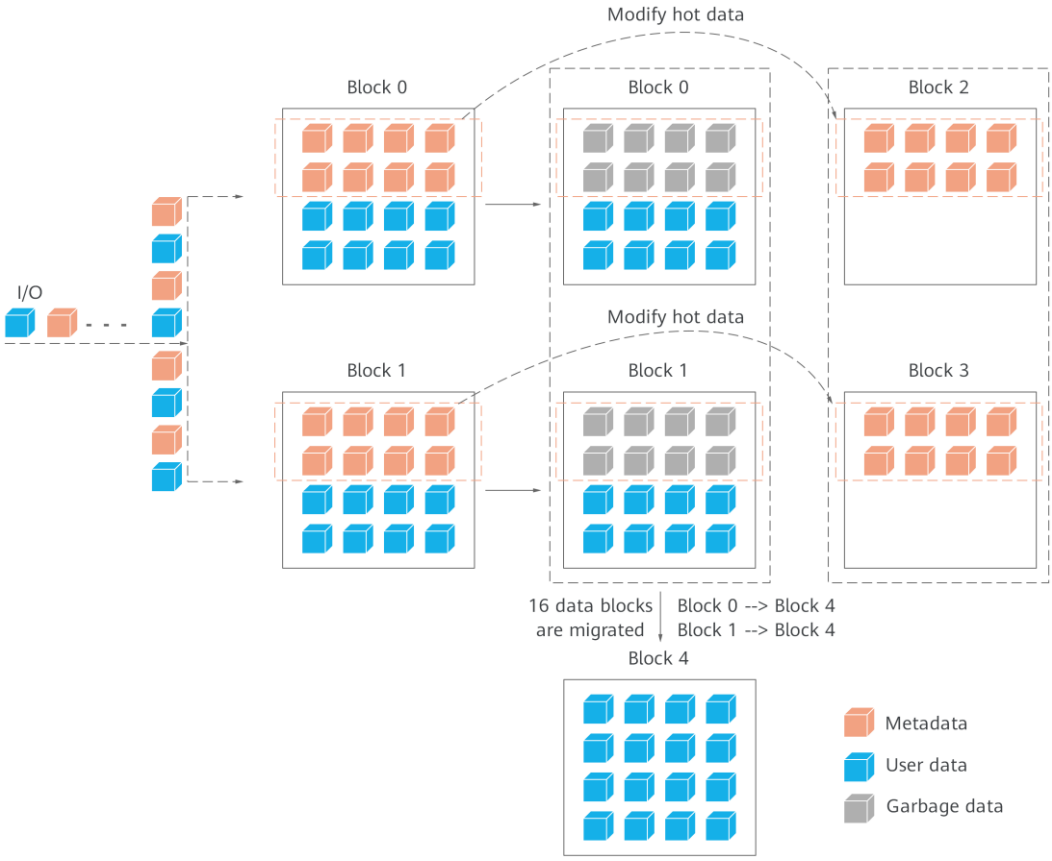
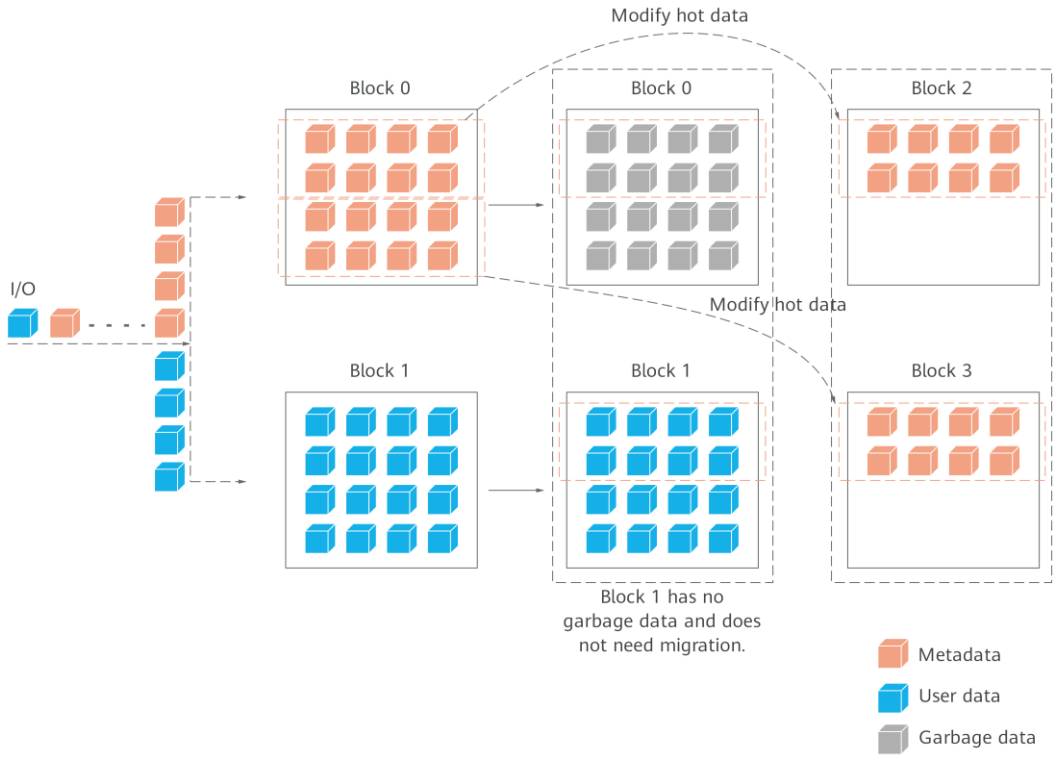


Figure 4-17 Data migration for garbage collection after separation of hot and cold data



To identify hot and cold data, OceanStor Dorado distinguishes between metadata, newly written data, and valid data migrated by garbage collection. Metadata is the hottest. For newly written data, if it is not modified in a long time, it will be migrated by garbage collection. The migrated data has the lowest probability of changes and is the coldest. Separating these types of data reduces write amplification of SSDs, greatly improves garbage collection efficiency, and reduces the number of erasures on SSDs, thereby prolonging the service life of SSDs.

4.2.4 End-to-End I/O Priority

On OceanStor Dorado, controllers label each I/O with a priority according to its type to ensure stable latency for specific types of I/Os. 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 classifies I/Os into the following five types and assigns their priorities in descending order, achieving optimal internal and external I/O response.

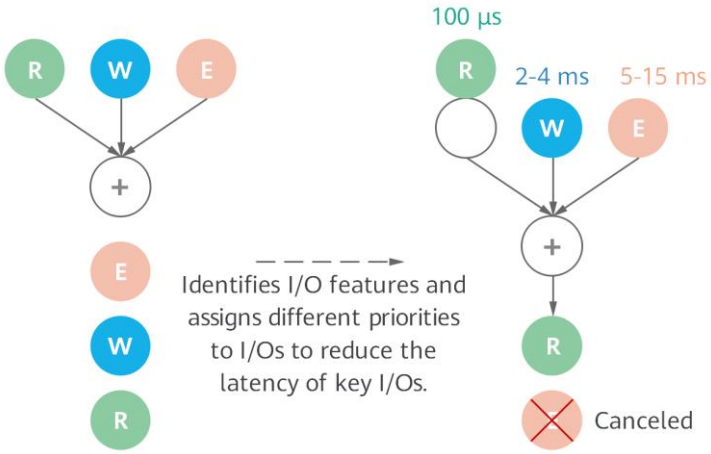
- Read and write I/Os
- Advanced feature I/Os
- Reconstruction I/Os
- Cache flush I/Os
- Garbage collection I/Os

Figure 4-18 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.

On each disk, in addition to assigning priorities to I/Os, OceanStor Dorado also allows high-priority read requests to interrupt ongoing write and erase operations. When a host writes data to a storage system, a write success is returned to the host after the data is written to the global cache. When a host reads data from a storage system, data must be read from SSDs if the cache is not hit. In this case, the disk read latency directly affects the read latency of the host. OceanStor Dorado is equipped with the latest generation of SSDs and uses the read first technology to ensure a stable latency. Generally, there are three operations on the flash media of an SSD: read, write, and erase. The erase latency is 5 ms to 15 ms, the write latency is 2 ms to 4 ms, and the read latency ranges from dozens of µs to 100 µs. When a flash chip is performing a write or an erase operation, a read operation must wait until the current operation is finished, which causes a great jitter in read latency.

Figure 4-19 Read first on SSDs



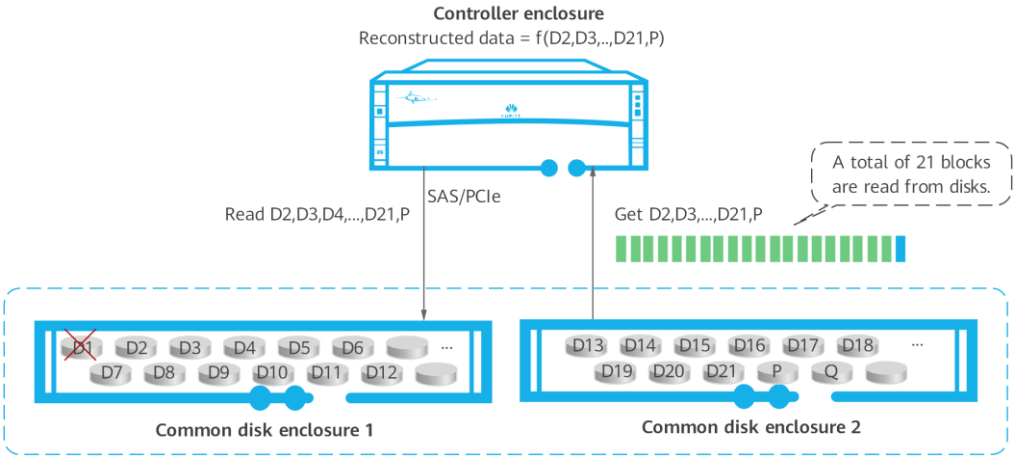
As shown in Figure 4-19, if a read request with a higher priority is detected during an erase operation, the system cancels the current operation and preferentially processes the read request. This greatly reduces the read latency on SSDs.

4.2.5 Smart Disk Enclosure

The smart disk enclosure of OceanStor Dorado is equipped with CPU and memory resources. It can offload tasks, such as disk reconstruction upon a disk failure, from controllers to reduce the load on the controllers. This greatly relieves pressure on controllers in the event of data

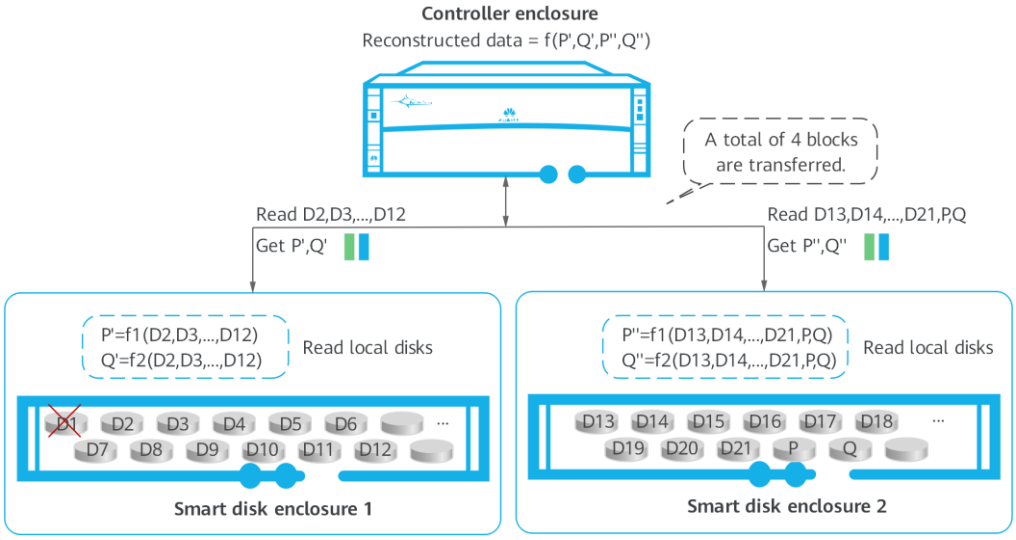
reconstruction due to disk failures. The following figure shows the reconstruction process of a common disk enclosure, using RAID 6 (21+2) as an example. If disk D1 is faulty, the controller must read D2 to D21 and P, and then recalculates D1. A total of 21 data blocks must be read from disks. The read operations and data reconstruction consume great CPU resources.

Figure 4-20 Data reconstruction process of a common disk enclosure



When the smart disk enclosure is used, it receives the reconstruction request and reads data locally to calculate the parity data. Then, it only needs to transmit the parity data to the controller. In Figure 4-21, only four blocks of parity data need to be transmitted between the controller and the smart disk enclosure, saving network bandwidth by 5.25 folds.

Figure 4-21 Data reconstruction process of a smart disk enclosure



Reconstruction tasks are offloaded to smart disk enclosures, reducing the drop in system performance to less than 10% in the event of data reconstruction due to disk failures (system performance drops by 20% if traditional disk enclosures are used).

4.2.6 Intelligence Technology

OceanStor Dorado builds an intelligent learning framework with the powerful computing capability of the intelligent chip to continuously learn the service load characteristics and device health status. This provides the following enhancements to the system:

- Intelligent read cache

  Accurately predicts service models to improve the hit ratio of the read cache and ensure high system performance in complex service models.
- Intelligent quality of service (QoS)

  Identifies and classifies different system loads, and suppresses the traffic of non-critical services to guarantee stable running of critical services.
- Enhanced data reduction

  Performs inline or post-process deduplication according to data models, and chooses proper reduction algorithms for specific data models to achieve the optimal reduction ratio and performance.

The intelligent read cache uses deep learning to identify and prefetch service flows based on space, time, and semantics, improving the read cache hit ratio.

- From the perspective of space, there are sequential flows (accessing adjacent LBAs in sequence), simple interval flows (with regular interval between two access operations), and complex interval flows (with no obvious regular interval between two access operations).
- From the perspective of time, hotspots exist (centralized access to one area in a short period of time).
- From the perspective of semantics, there is a semantic pattern (the accessed data is logically related).

Complex interval flows and semantic pattern flows require intelligent chips.

Figure 4-22 Service flow patterns

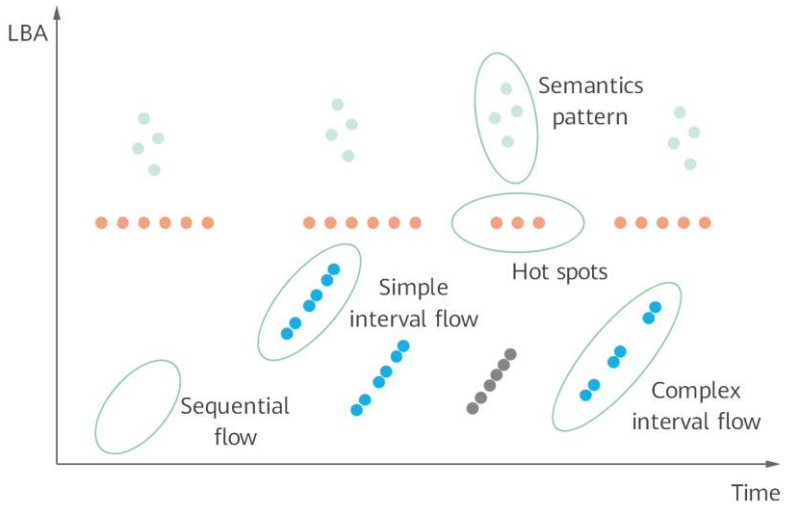


Table 4-4 Identification methods for various service flow patterns

Pattern	Computing Resource	Identification Method	Flow Identification Accuracy
Sequential flow	CPU	Statistical learning	100%
Hotspots	CPU	Statistical learning (MQ)	100%
Simple interval flow	CPU	Statistical learning	100%
Complex interval flow	Intelligent chip	Machine learning (Bayesian network, EM, LZ77, and others)	98%
Semantic pattern	Intelligent chip	Deep learning (including CNN and RNN)	95%

4.3 Rich Software Features

OceanStor Dorado provides the Smart series software for efficiency improvement, Hyper series for data protection, implementing data management throughout the lifecycle.

- The Smart series software includes SmartThin, SmartDedupe, SmartCompression, SmartQoS, SmartVirtualization, SmartMigration, SmartErase, SmartQuota, SmartCache, SmartTier, and SmartMulti-Tenant, which improve storage efficiency and reduce the TCO.
- The Hyper software series includes HyperSnap, HyperCDP, HyperClone, HyperReplication, HyperVault, HyperMetro, 3DC, HyperEncryption, and HyperDetect, which provide disaster recovery, data backup, and security functions.

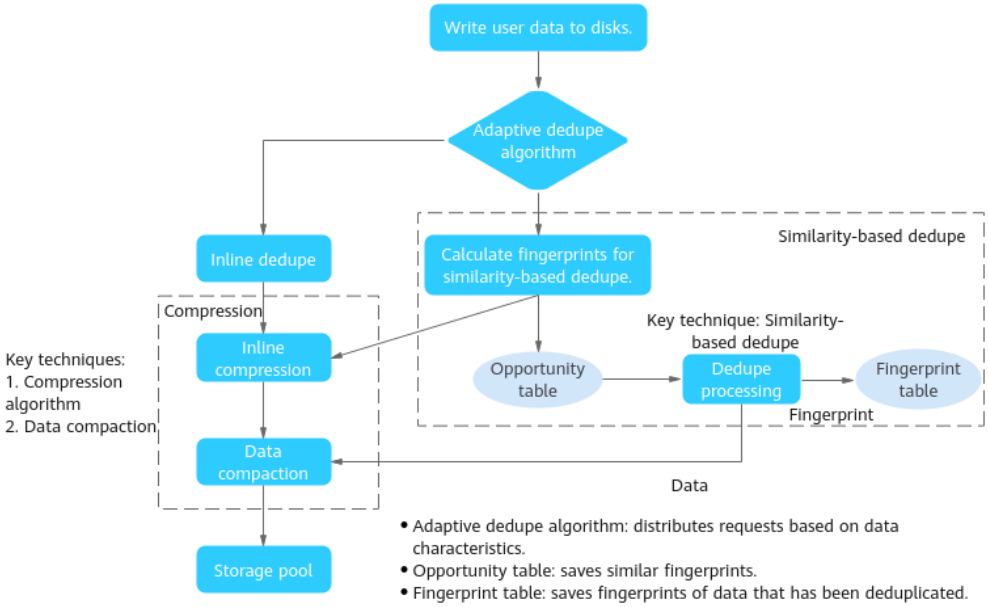
5 Smart Series Features

OceanStor Dorado provides the Smart series software, including SmartThin, SmartDedupe, and SmartCompression for better space utilization, SmartQoS, SmartCache, SmartTier, SmartMulti-Tenant (for SAN and NAS), and SmartQuota (for NAS only) for better performance and improved service quality, SmartVirtualization (for SAN only), SmartMigration (for SAN only), and SmartErase for system lifecycle and security management, as well as SmartContainer for a comprehensive storage solution capability.

5.1 SmartDedupe and SmartCompression (Data Reduction)

OceanStor Dorado automatically performs adaptive deduplication and compression based on user data characteristics, maximizing the reduction ratio. Figure 5-1 shows the adaptive deduplication and compression process.

Figure 5-1 Adaptive deduplication and compression process



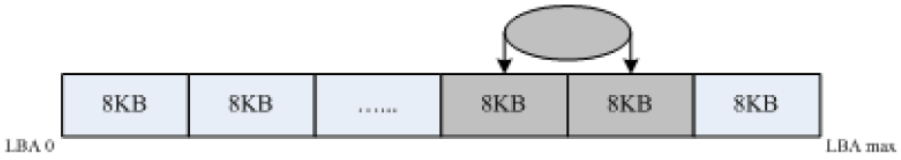
1. When a user writes data, the adaptive deduplication algorithm identifies data suitable for inline deduplication based on data characteristics and directly performs inline deduplication.
2. The adaptive deduplication algorithm identifies data suitable for similarity-based deduplication based on data characteristics, calculates similar fingerprints (SFPs), and adds the SFPs to the similarity-based deduplication opportunity table. Then the system compresses the user data, writes the compressed data to the storage pool, and returns a success message.
3. The background deduplication task finds similar data in the opportunity table and reads the data from disks for similarity-based deduplication. After the deduplication is complete, the fingerprint table is updated.

Note:

In 6.1.5 and later versions, you can run the change disk_domain general disk_domain_id=? dedup_method=? command on the CLI in developer mode to change the deduplication mode, which can be inline (inline deduplication) or offline (similarity-based deduplication), in the inline deduplication mode, only inline deduplication and inline compression will perform.

5.1.1 Deduplication

The storage system divides the written data into blocks based on the block granularity and calculates fingerprints for the blocks. The default block granularity is 8 KB. You can also set it to 16 KB or 32 KB. The following figure shows how data is divided into blocks.



As shown in the figure, blocks start from logical block address (LBA) of the LUN or file system and are divided into the configured size.

If the start address and the length taken for writing user data cover two blocks, the user data must be divided into two blocks. If the length of a block after the division does not reach the configured length, such as 8 KB, the system checks whether the start LBA of the current block was written previously. If it has been written, the system reads the existing data and combines it with the data to be written to form a new block. If the block has not been written, the system pads the block with 0s and performs subsequent operations.

5.1.1.1 Inline Deduplication Procedure

In inline deduplication mode, the process is as follows:

After I/O data blocks are written to the system, the system checks whether these blocks are duplicate based on the existing fingerprints in the system. If the same fingerprint is found, the system reads the data corresponding to the fingerprint and compares that existing 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 disks. If the fingerprint is not found or byte-by-byte comparison is not passed, the system writes the new data block to disks and records the mapping between the fingerprint and storage location.

5.1.1.2 Similarity-based Deduplication Procedure

Similarity-based deduplication identifies similarities between data based on the SFPs and then performs deduplication coding on similar data. For two sets of data that are similar but not completely the same, the same SFP can be obtained by using dedicated algorithms.

For example, the same SFP can be calculated from the following sentences:

Nick likes red apple.

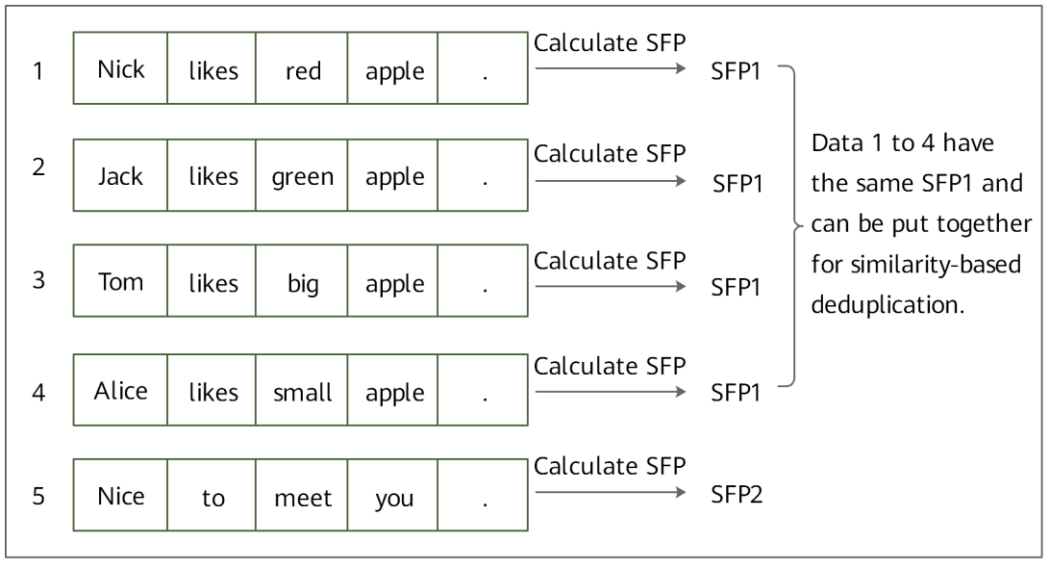
Jack likes green apple.

Tom likes big apple.

When calculating SFPs, similarity-based deduplication configures different parameters to obtain multiple SFPs for the same data block (granularity: 8 KB, 16 KB, or 32 KB). This is like drawing different pictures of a person from the front and the side. Multiple similar fingerprints are identified by similarity and sorted by gradient. The system performs gradient iterative deduplication on similar data based on the similarity, and then compresses the data after deduplication. This provides a higher reduction ratio than fixed-length deduplication.

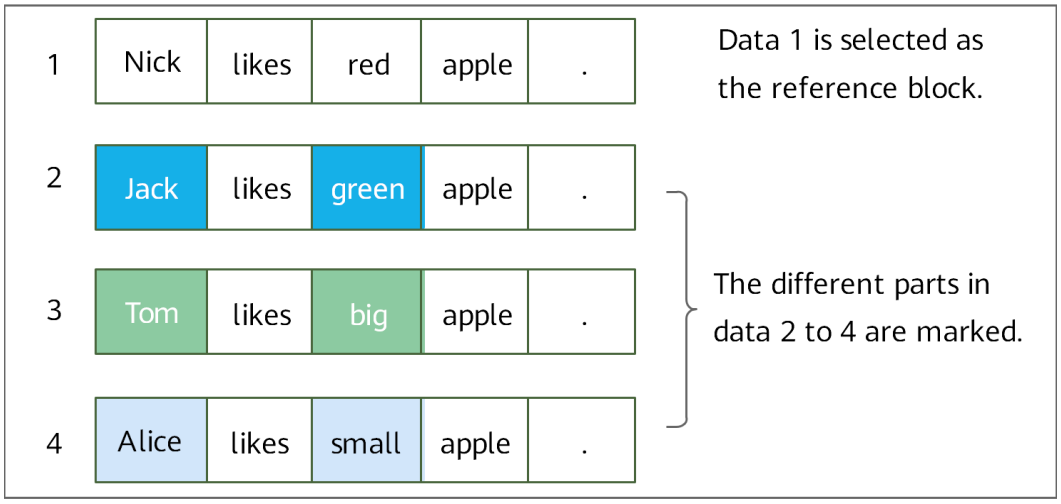
Step 1 The system calculates SFPs of data and identifies data with the same SFP.

Figure 5-2 Identifying data with the same SFP



Step 2 The system gathers data with the same SFP and selects a reference block. For the other data, the system identifies the parts different from the reference block. The data on the reference block and the data different from the reference block are both compressed.

Figure 5-3 Selecting a reference block and marking different parts



Step 3 The system retains all content of the reference block. For the other data, only the different parts and their locations are recorded. Figure 5-4 shows the deduplication result.

Figure 5-4 After deduplication

Step 4 The system saves the deduplicated data to a storage pool.

----End

5.1.1.3 Global Deduplication

To achieve an optimal data reduction ratio, global deduplication is performed by storage pool to ensure that the same or similar data of all LUNs and file systems in a storage pool can be deduplicated.

Theoretically, a larger range of scanning for duplicate data indicates a higher probability of locating the same or similar data and a higher deduplication ratio.

OceanStor Dorado introduces the global deduplication function, which deletes data that is the same as or similar to existing data saved on other LUNs/file systems in a storage pool, eliminating redundancy.

Generally, the system requires a large amount of memory to cache fingerprint information, so that a long time window can be supported to identify fingerprints whose corresponding data can be deduplicated. OceanStor Dorado adopts the deduplication opportunity table to store deduplication fingerprint information and persistently saves the table in the storage pool. Only a small amount of memory is required to analyze and process massive deduplication fingerprints, and the memory overhead does not increase linearly with the capacity. Therefore, OceanStor Dorado empowers global deduplication of massive amount of data. The upper limit of the deduplication capacity is the maximum capacity of the storage pool.

5.1.1.4 Secure Deduplication

SmartDedupe uses the weak hash algorithm and byte-by-byte comparison to implement data deduplication.

During inline deduplication, the system uses the weak hash algorithm to calculate the fingerprint of each data block and then compares the fingerprint with all the fingerprints in the opportunity table. If the same fingerprint exists, the system reads the corresponding data block and perform byte-by-byte comparison between the two data blocks. Deduplication is performed only when the two data blocks are the same. This ensures that data blocks with the same weak hash values but different data contents are not deduplicated, guaranteeing absolutely secure data consistency in inline deduplication.

During post-process deduplication, identical data is deduplicated in the same way as inline deduplication, and delta compression is performed on similar data. In delta compression, redundant data is removed only after comparison with complete data to ensure data consistency.

5.1.2 Compression

SmartCompression involves two processes. Input data blocks are first compressed to smaller sizes using a compression algorithm and then compacted before being written to disks.

5.1.2.1 Data Compression

OceanStor Dorado performs inline compression by using Huawei's dedicated algorithm that combines LZ matching and Huffman entropy encoding. This provides a compression ratio 20% higher than that of the LZ4 algorithm while providing the same performance.

Preprocessing

Before data compression, OceanStor Dorado uses Huawei's preprocessing algorithm to identify the data blocks that are difficult to compress by the general compression algorithm based on the data format. Then the system rearranges the data to be compressed for a higher compression ratio.

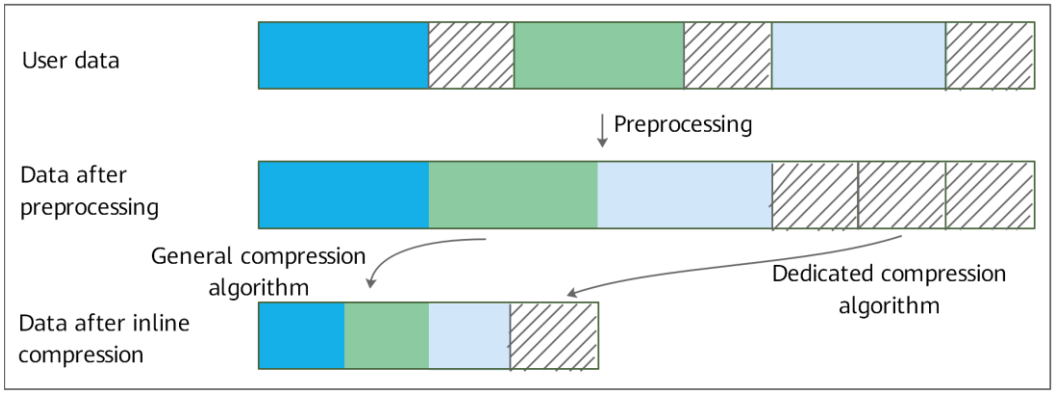
After preprocessing, the data to be compressed is divided into two parts:

- For the part that is difficult to compress, the system compresses it using the dedicated compression algorithm.
- For the other part, the system uses the general compression algorithm to compress it.

Dedicated Compression

Based on the preprocessing result, OceanStor Dorado uses a proprietary compression algorithm to compress the data that is difficult to compress. This dedicated compression algorithm uses special coding rules to compress the data without adding the metadata. It features high performance and does not affect read and write operations. Figure 5-5 explains preprocessing and dedicated compression.

Figure 5-5 Preprocessing and dedicated compression

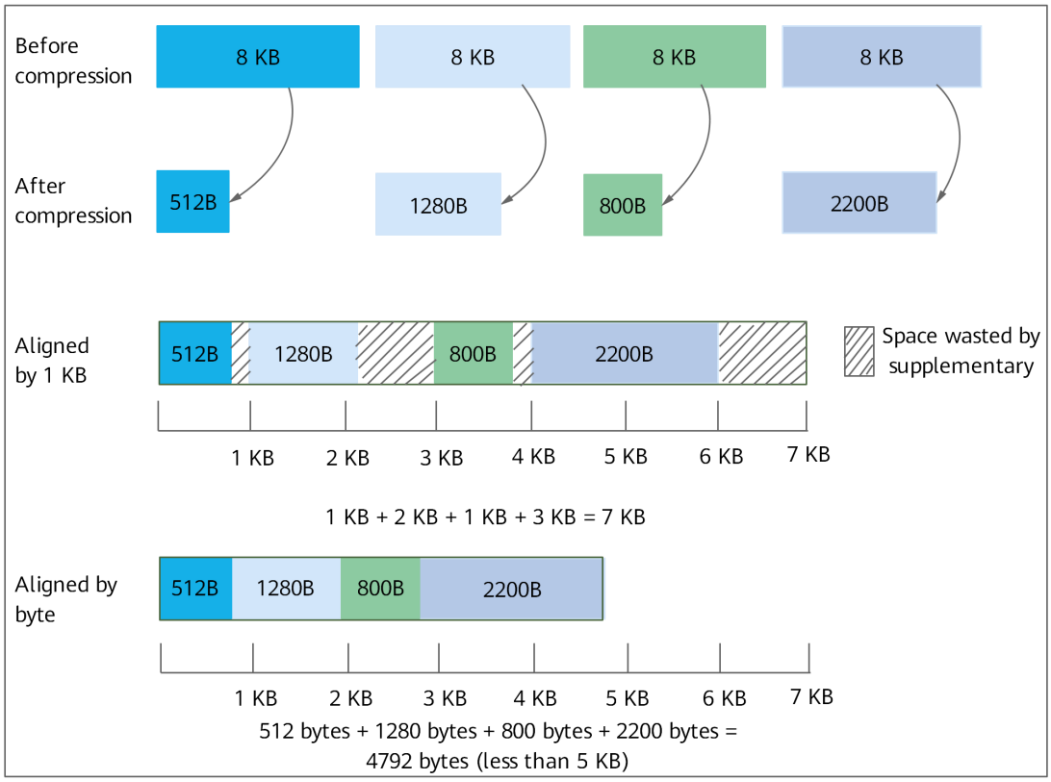


5.1.2.2 Data Compaction

Compressed user data is aligned by byte and then compacted to reduce the waste of physical space. This provides a higher reduction ratio than the 1 KB alignment granularity generally used in the industry.

As the example shown in Figure 5-6, byte-level alignment saves 2 KB physical space compared with 1 KB granularity alignment, improving the compression ratio.

Figure 5-6 Alignment of compressed data by byte



5.1.3 Flexible Configurations of Deduplication and Compression Granularities

OceanStor Dorado is applicable to scenarios where services with different I/O sizes are involved. If the block granularity configured for the system is different from the I/O size of services on the storage device, previously written data needs to be read or unwritten space needs to be written with zeros to fill the data blocks, causing a negative impact on the storage performance. For example, in Oracle systems, the size of most I/Os is 8 KB.

SmartDedupe and SmartCompression allow the configuration of 8 KB, 16 KB, and 32 KB data block granularities. This allows the granularities to match the granularity of storage device blocks with the I/O size of services on the device, and to reduce performance loss caused by reading previously written data or filling zeros into unwritten space.

5.2 SmartQoS (Intelligent Quality of Service Control)

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

SmartQoS uses object-specific I/O priority scheduling and I/O traffic control (including upper limit control and minimum performance guarantee) to guarantee the service quality.

Table 5-1 SmartQoS policy and controlled objects

Function	Controlled Object	Configuration Item
Upper limit control (including burst traffic control)	SAN: LUN, snapshot, LUN group, host, and vStore NAS: file system and vStore	IOPS and bandwidth
Lower limit guarantee	SAN: LUN, snapshot, and LUN group NAS: file system	IOPS, bandwidth, and maximum latency

5.2.1 Functions

SmartQoS supports upper limit control and minimum performance guarantee. Upper limit control prevents traffic of some services from affecting the normal running of other services. Minimum performance guarantee is mainly used to guarantee resources allocated to critical services, especially latency-sensitive services, thereby ensuring their performance.

5.2.1.1 Upper Limit Control

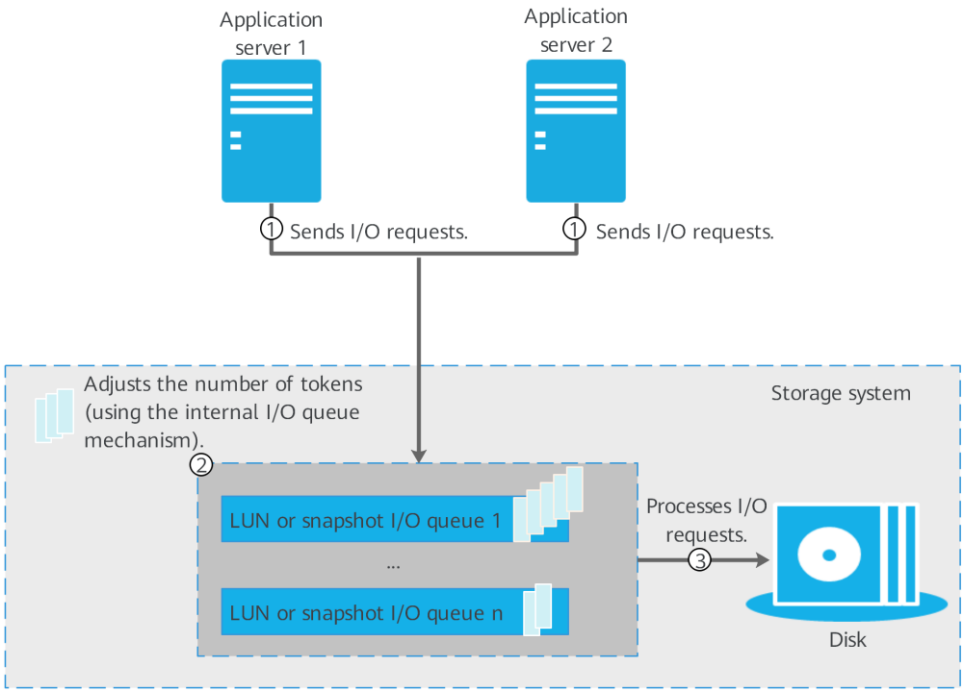
I/O traffic control is implemented based on a user-configured SmartQoS policy that contains SAN resource objects (LUNs/snapshots, LUN groups, hosts/host groups, and vStores) and NAS resource objects (file systems and vStores) to limit their IOPS and bandwidth. This prevents some specific applications from affecting the performance of other services due to heavy burst traffic.

I/O traffic control is implemented by I/O queue management, token allocation, and dequeue control.

After an upper limit objective is set for a SmartQoS policy, the system allocates tokens based on the objective to control traffic. If the objective is to limit IOPS, an I/O is converted to a number of normalized 8 KB I/Os and a token is allocated to each of the 8 KB I/Os. If bandwidth is limited, a token is allocated to each byte.

I/O queue management allocates storage resources by tokens. The more tokens owned by an I/O queue, the more resources will be allocated to that queue. Figure 5-7 explains the implementation process.

Figure 5-7 Upper limit control



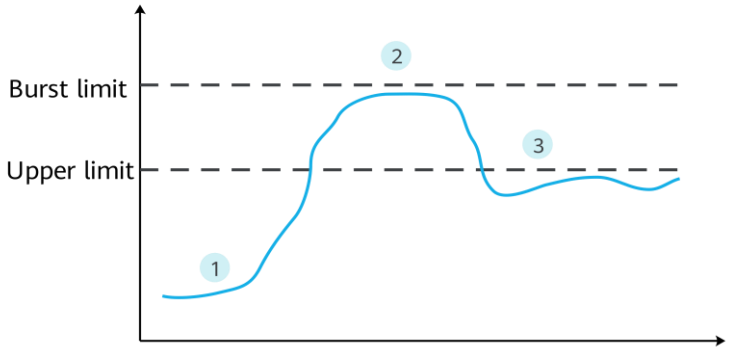
1. The application servers send I/O requests to the target I/O queues.
2. The storage system converts the traffic control objective into a number of tokens. You can set upper limit objectives for low-priority objects to guarantee sufficient resources for high-priority objects.
3. The storage system processes the I/Os in the queues by tokens.

Burst Traffic Control Management

For latency-sensitive services, you can allow them to exceed the upper limit for a specific period of time. SmartQoS supports burst traffic control management to specify the burst IOPS, bandwidth, and duration for the controlled objects.

The system accumulates the unused resources during off-peak hours and consumes them during traffic bursts to break the upper limit for a short period of time. To achieve this, the long-term average traffic of the service should be below the upper limit.

Figure 5-8 Traffic burst



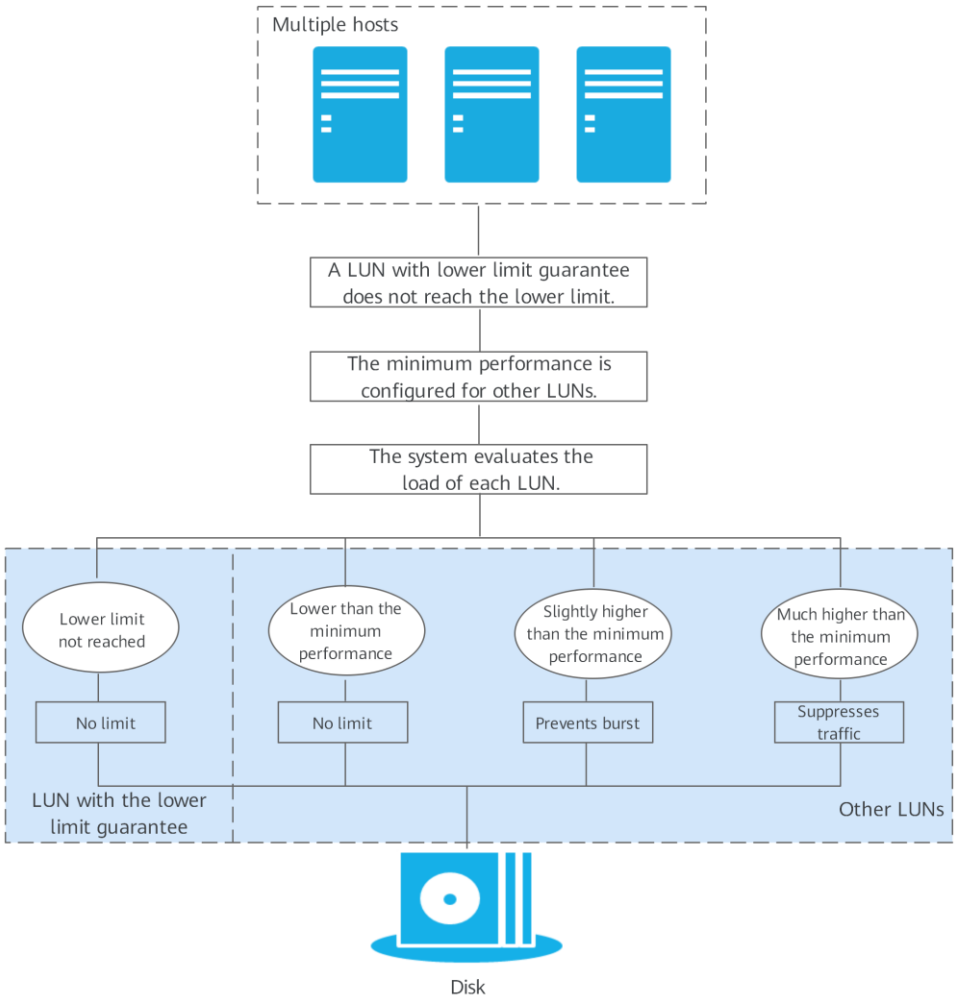
1. If the traffic of an object does not reach the upper limit in a past period of time, the service traffic of the object can temporarily exceed the upper limit set by SmartQoS in a future period of time as long as the system is not overloaded. This meets the requirements of the burst service traffic during peak hours. The maximum duration and ratio of a burst are configurable.
2. Burst traffic control is implemented by accumulating burst durations. If the traffic of a LUN, snapshot, LUN group, or host is below the upper limit in a second, the system accumulates this second for the burst duration. When the service load surges, performance can break the upper limit to reach the specified burst limit for a duration accumulated earlier (this duration will not exceed the maximum value specified).
3. When the accumulated duration or the specified maximum duration is reached, the performance drops below the upper limit.

5.2.1.2 Lower Limit Guarantee

The lower limit guarantee function is dedicated to ensuring critical services. It takes effect for all service objects (such as LUNs, snapshots, LUN groups, and file systems) in a system to release resources for objects whose lower limit objectives are not fulfilled. The working principle of the lower limit guarantee function is to ensure the quality of services as much as possible. Therefore, this function is applicable to only a few critical services. If this function is configured for all services, the quality of all services cannot be guaranteed.

If a controlled object does not reach the lower limit, the system evaluates the load on all controlled objects in the system. For heavily loaded objects, the system suppresses their traffic based on the lower limit until sufficient resources are released for all objects in the system to reach the lower limit. For objects that have reached the lower limit but are not heavily loaded, the system prevents burst traffic on these objects. For LUNs that do not reach the lower limit, the system does not limit their traffic and allows them to preempt resources from heavily loaded LUNs.

Figure 5-9 Lower limit guarantee for LUNs



Latency assurance is to prioritize requests of objects that have latency objectives. If latency assurance is not achieved, the latency objectives are converted into traffic control objectives and guaranteed in the same way as lower limits.

5.2.2 Policy Management

5.2.2.1 Hierarchical Management

SmartQoS supports both common and hierarchical policies.

- A common policy contains only controlled objects. It controls the traffic from a single application to the controlled objects. For example, VDI application startup causes a temporary boot storm. You can configure a common policy during the startup to prevent the boot storm from affecting other services.
- A hierarchical policy can contain common policies. It controls the traffic when there are multiple applications running in the system. For example, in a VMware environment, the customer wants to control the upper limit of a specific VMDK on a VM. To do this, the customer can set a hierarchical policy for the VM and a common policy for the VMDK.

NOTE

- The types of objects in a SmartQoS policy are classified by LUNs/snapshots, LUN groups, hosts, file systems, and vStores. A common policy can have only one type of object.
- A hierarchical policy can have multiple common policies, each of which can contain a different type of object.
- Each LUN or snapshot can be added to only one SmartQoS policy.
- Each LUN group can be added to only one SmartQoS policy.
- Each host can be added to only one SmartQoS policy.
- Each file system can be added to only one SmartQoS policy.
- Each vStore can be added to only one SmartQoS policy whose owner is the system.
- Each vStore can be added to one SmartQoS policy whose owner is the system and one SmartQoS policy whose owner is the vStore at the same time.

The following figures show the relationship between the common and hierarchical policies (using LUNs as an example).

Figure 5-10 Adding 1 to 512 LUNs or snapshots to each policy

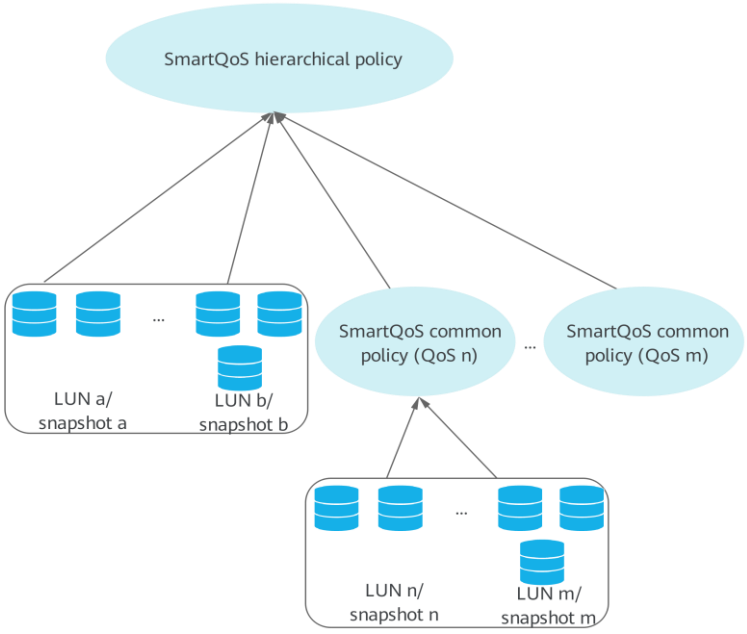
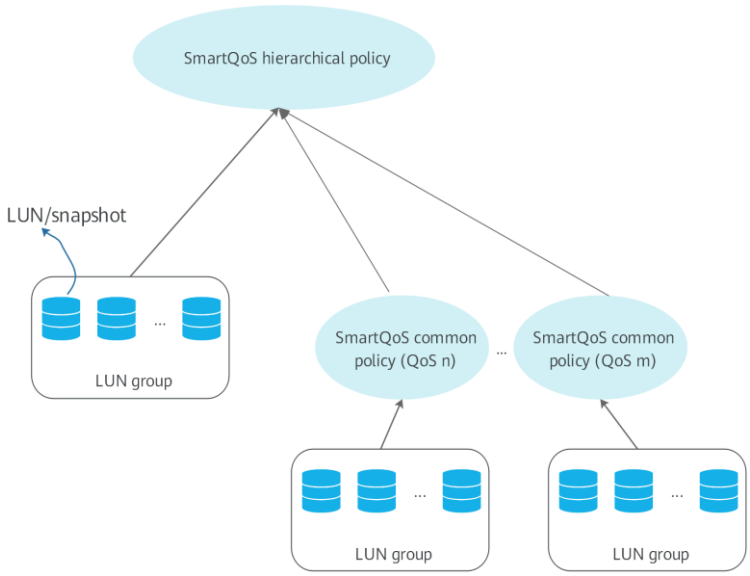


Figure 5-11 Adding a LUN group to each policy



For details about the policy specifications, see the product specifications list.

5.2.2.2 Objective Distribution

All objects in a SmartQoS policy share the upper limit objectives. The SmartQoS module periodically collects the performance and requirement statistics of all objects in a traffic control policy, and distributes the traffic control objective to each object.

Currently, the optimized weighted max-min fairness algorithm is used for objective distribution. It determines the traffic control objective for each object based on the policy's overall objective (including the upper and lower limits) and each object's resource requirement. This algorithm preferentially meets the requirement of objects. The requirement refers to the number of requests received by an object, including the number of successful requests and the number of rejected requests. Then, the remaining resources are distributed to each object based on the object's weight. In addition, it uses a filtering mechanism to ensure a relative stable objective for each object.

Note:

You can add each LUN/file system or snapshot to a traffic control policy separately, or add LUNs/file systems or snapshots to a LUN group and then add the LUN group to a traffic control policy. When a LUN/file system or snapshot is added to multiple traffic control policies, the smallest value among the upper limits takes effect for the LUN/file system or snapshot.

5.2.2.3 Recommended Configuration

Upper Limit Configuration

- In a multi-tenant scenario, you can configure different hierarchical policies for different tenants to ensure that the resources occupied by a single tenant do not exceed the limit. Common policies can be configured for different services of a single tenant to specify the upper and lower performance limits for these services.

- When hybrid service loads are carried, you can set upper limits for non-critical services, especially services with great fluctuations in loads, and set lower limits for critical and latency-sensitive services.
- You are advised to configure a burst policy for service loads that are not evenly distributed and are sensitive to latency.

Lower Limit Configuration

For a few mission-critical services in the system, you can configure a lower limit policy to set the minimum IOPS, minimum bandwidth, and maximum response latency. When system resources are insufficient, the lower limit policy works to ensure the quality of the critical services by limiting the performance of non-critical services. When the lower limit policy fails to work due to insufficient system resources, the system generates an alarm. You can adjust the lower limit policy based on the alarm information. Therefore, ensure that the overall performance of the services configured with the lower limit policy does not exceed 50% of system performance. If the lower limit policy works for all services, it works for nothing.

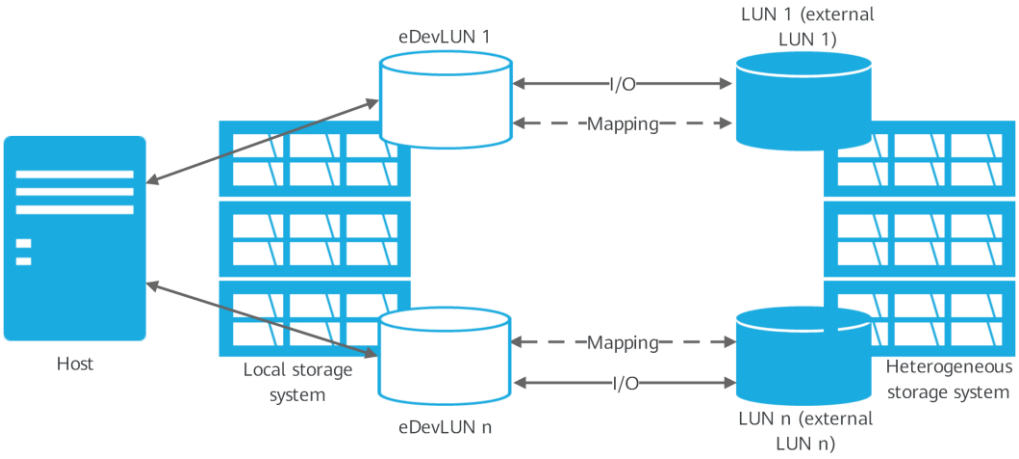
5.3 SmartVirtualization (Heterogeneous Virtualization)

OceanStor Dorado uses SmartVirtualization to take over LUNs from heterogeneous storage systems. In addition, SmartVirtualization can work with SmartMigration to migrate data from heterogeneous storage systems online, facilitating device replacement.

Working Principles

As shown in the following figure, 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 only a small amount of 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.

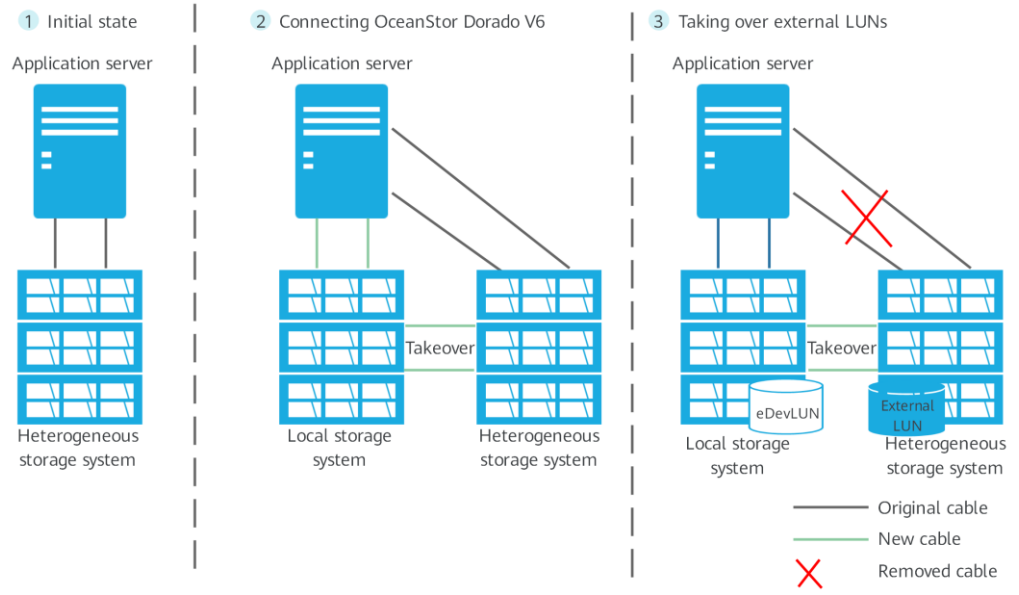
Figure 5-12 Heterogeneous storage virtualization



Online Takeover of Heterogeneous Storage Systems

SmartVirtualization uses LUN masquerading to set the WWN of the eDevLUN on OceanStor Dorado to be the same as that of the LUN on the heterogeneous storage system. After a host accesses an eDevLUN, the multipathing software considers the eDevLUN the same LUN as that on the heterogeneous storage system, but adds the access path. After the links between the heterogeneous storage system and the host are removed, the host's multipathing software switches the links to the eDevLUN to achieve online takeover. Figure 5-13 shows the online takeover process.

Figure 5-13 Online takeover process



Huawei storage systems can be taken over online. To take over third-party storage systems online, you must query the online takeover compatibility list on Huawei Storage Interoperability Navigator.

Online Data Migration from Heterogeneous Storage Systems

If some devices in the data center are out of warranty or the performance and capacity cannot meet service requirements, customers need to upgrade storage devices. SmartVirtualization and SmartMigration can migrate customer data to OceanStor Dorado online without interrupting host services.

5.4 SmartMigration (Intelligent Data Migration)

OceanStor Dorado 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: data synchronization and LUN information exchange.

Data Synchronization

1. Before the migration, you must configure the source and target LUNs.
2. When migration starts, the source LUN replicates data to the target LUN in the background.
3. During migration, the host can still access the source LUN. When the host writes data to the source LUN, the system records the request in a log, which only records the address information but no data content.
4. The system writes the incoming data to both the source and target LUNs.
   - The system waits for the write response from the source and target LUNs. If writing to both LUNs is successful, the system deletes the log. If writing to either LUN fails, the system retains the log and replicates the data again in the next synchronization.
   - The system returns the write result of the source LUN to the host.
5. The system performs the preceding operations until all data is replicated to the target LUN.

LUN Information Exchange

After data replication is complete, the source LUN and target LUN exchange information. In the information exchange, source and target LUN IDs and WWNs remain unchanged but the data volume IDs of the source LUN and target LUN are exchanged. This creates a new mapping relationship between the source LUN ID and target volume ID. After the exchange, the host can still identify the source LUN using the source LUN ID but accesses physical space of the data volume corresponding to the target LUN.

SmartMigration can meet the following requirements:

- 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. For example, a LUN can be migrated from one storage pool to another.

Value-Added Configurations for Target LUNs

The duration of data migration across devices varies from hours to weeks or even months, depending on the amount of data to be migrated. Cross-array remote replication and HyperMetro features can be configured for target LUNs to replicate data across data centers during migration for disaster recovery. This ensures immediate data protection upon completion of the migration, improving solution reliability.

NOTE

During data migration, data on the target LUNs is incomplete. Therefore, the DR site cannot take over services before the migration and DR data synchronization are complete.

5.5 Intelligent File Migration (SmartMigration for NAS)

Intelligent file migration mainly applies to heterogeneous migration of peer vendors' data. Currently, NAS data migration at the host layer is the most common migration solution. However, the following problems may occur:

1) A Windows or Linux host must be used as the migration server.

2) The cutover time is long.

To solve the preceding problems, two intelligent file migration modes are provided:

Copy-first

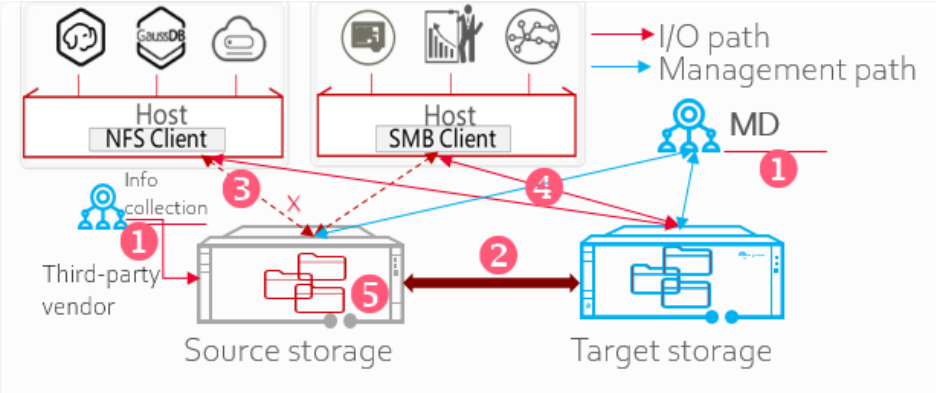
This mode is similar to host migration. Specifically, the built-in migration module of a storage system reads remote data and synchronizes the data to the local storage system. However, this mode does not require the customer to provide a Windows or Linux host as the migration server. In this mode, the built-in migration module and a NAS protocol client directly access the remote storage data. Similar to host migration, the cutover time in this mode depends on the time of the last incremental synchronization.

Takeover-first

Different from traditional host migration, this mode shortens the cutover time to several minutes. In this mode, a customer stops services on the remote storage, creates a synchronization task between the local and remote storage, and switches I/Os to the local storage. The customer's services can continue, and the I/O data is written to the local storage and synchronized to the remote storage. In this way, data consistency between the two storage systems is ensured. After a background data synchronization task migrates all data from the remote storage to the local storage, the synchronization task can be deleted. In this way, services can be seamlessly switched to the local storage. In this mode, customer services run on the local storage without waiting for the completion of data migration.

5.5.1 Copy-First Mode

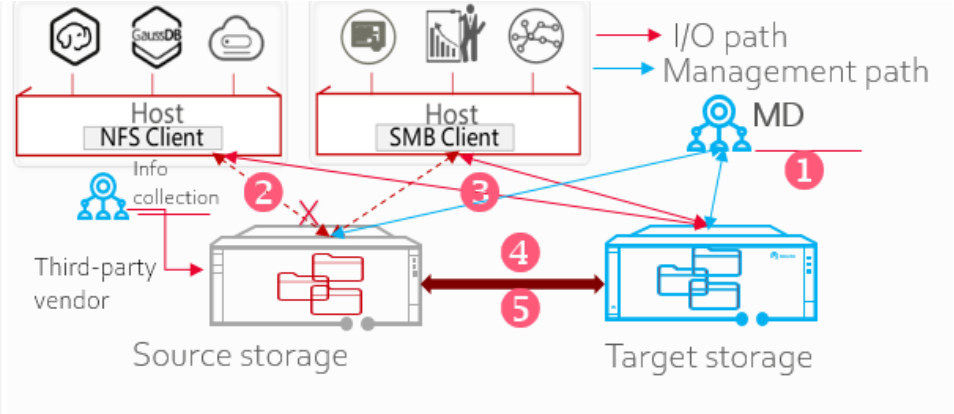
The following figure illustrates the copy-first mode.



1. Preparation	Collect configuration information about the source storage and use MD to configure the target storage. Set up a connection between the source storage and the target storage and create a file system to be migrated.
2. Initial migration	Start initial migration using MD. The source storage continues to process client services.
3. Shutdown for cutover	Shut down the service client and start incremental migration based on difference comparison. After the migration, disconnect the two storage systems.
4. Service recovery	Complete IP address configuration and DNS modification, mount the shared file system to the service client again, and use the target storage for read and write services.
5. Migration completion	Remove the source storage from the network.

5.5.2 Takeover-First Mode

The following figure illustrates the takeover-first mode.



1. Preparation	Collect configuration information about the source storage and use MD to configure the target storage. Set up a connection between the source storage and the target storage and create a file system to be migrated.
2. Shutdown for cutover	Shut down the service client.

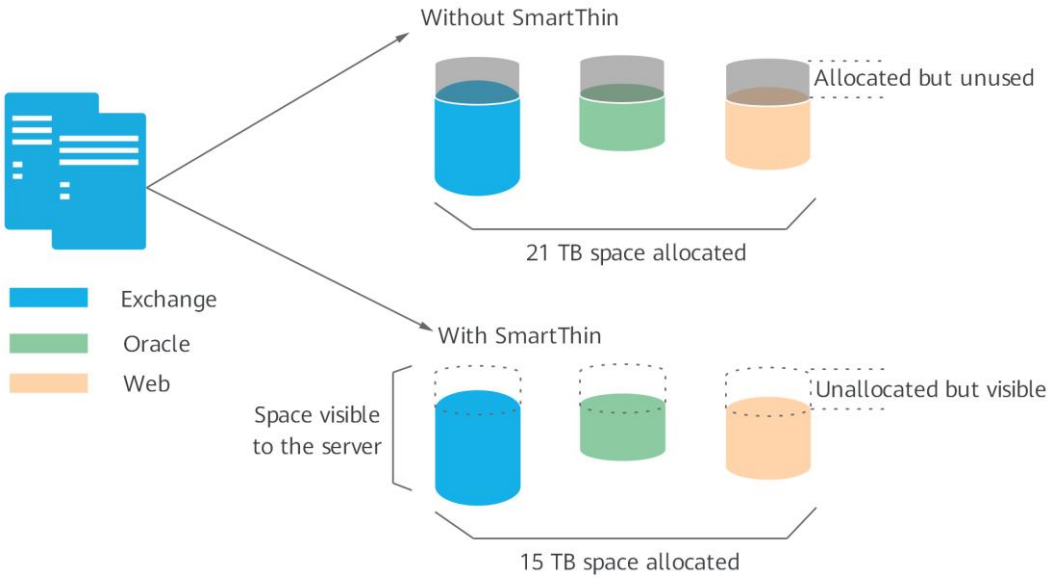
3. Service recovery	Complete IP address configuration and DNS modification, mount the shared file system to the service client again, and write new services in dual-write mode (to both the target and source storage).
4. Service migration	Use MD to migrate data on the file system of the source storage in the background.
5. Service completion	Disconnect the two storage systems and remove the source storage from the network.

5.6 SmartThin (Intelligent Thin Provisioning)

OceanStor Dorado supports thin provisioning, which enables the storage systems 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 be used up, SmartThin triggers storage resource pool expansion to add more space. The expansion process is transparent to users and causes no system downtime.

The following figure shows the benefits of SmartThin (using LUNs as an example).

Figure 5-14 Comparison between the thin LUN and traditional LUN



5.7 SmartErase (Data Destruction)

If a disk is no longer used in its original scenario, data on the disk is not needed. If the data is not properly processed, unauthorized users may use the residual data to reconstruct the original data, causing information leakage risks. Therefore, data on disks must be thoroughly erased to ensure data security.

Data destruction of OceanStor Dorado is implemented based on disks and is not specific to service types. It is applicable to both SAN and NAS applications. Data destruction complies with DoD standards.

During data erasure, the storage system sends one or more SANITIZE commands (defined by the SCSI or NVMe protocol) to the disk according to the data erasure security standard. After receiving the command, the disk returns a success message. At the same time, the data erasure task is executed in the background. The storage system periodically queries the task progress until it ends. All data, including data in the over provisioning (OP) space, will be erased, and the data cannot be restored.

Security standards for data erasure include:

- DoD 5220.22-M (E)

  This standard suggests a software method to erase data from writable storage media, including three overwrites:

  a. Using a 1-byte character to overwrite all addresses
  b. Using one's complement of the character to overwrite all addresses
  c. Using a random number to overwrite all addresses

- DoD 5220.22-M (ECE)

  This standard is an extended version of the DoD 5220.22-M (E). It runs the DoD 5220.22-M (E) twice and uses a random number once to overwrite all addresses.

  a. Using the DoD 5220.22-M (E) standard to overwrite all addresses for three times
  b. Using a random number to overwrite all addresses once
  c. Using the DoD 5220.22-M (E) standard to overwrite all addresses for three times

- VSITR

  The VSITR data sanitization method was originally defined by the German Federal Office for Information Security and is implemented in the following way:

  a. Write zero bytes (0x00).
  b. Write high bytes (0xFF).
  c. Write zero bytes (0x00).
  d. Write high bytes (0xFF).
  e. Write zero bytes (0x00).
  f. Write high bytes (0xFF).
  g. Write pseudo random bytes.

- User-Defined

  In the user-defined overwrite mode, the system uses the data mode specified by the user. The value is a 1-byte hexadecimal number starting with r or 0x. A maximum of three parameters separated by commas (,) can be entered. The system uses the data to overwrite all the addresses of the disk for specified times. You can set the times of overwriting the disk to a value ranging from 1 to 15. The default value is 1.

5.8 SmartQuota (Quota)

SmartQuota is a file system quota technology. It controls the storage resources for directories, users, and user groups to prevent overuse of storage resources by specific users.

Directory trees (dtrees) are critical for SmartQuota. A dtree is a special directory that can be created at any level of a file system. A dtree manages the quotas of all of its sub-directories and files (including recursive common sub-directories and files). Dtrees cannot be created in another dtree. Dtrees can be created, deleted, or modified only on the GUI or CLI of the maintenance terminal rather than on the client hosts. Directory, user, and group quotas can be configured only on dtrees. The root directory of a file system is also a dtree, so it supports quotas.

A dtree differs from a common directory in the following ways:

- Dtrees can be created, deleted, and modified only by administrators on the CLI or GUI. Dtrees can be created at any level of a file system.
- Dtrees can be shared using protocols and cannot be renamed or deleted during sharing.
- Files cannot be moved (NFS) or cut (SMB) between two dtrees. This is because a file or directory can belong to only one dtree.
- Hard links cannot be established between two dtrees.

SmartQuota supports the following quota types:

- Space soft quota: sets the limit for the space usage alarm. When the used space exceeds this quota, the system reports an alarm and prompts users to delete unnecessary files or increase the quota. Users can continue to write data after the alarm.
- Space hard quota: restricts the maximum available space of a quota. When the used space reaches this quota, the system generates an error stating insufficient space.
- File quantity soft quota: sets a limit for the file quantity alarm. When the number of files exceeds this quota, the system reports an alarm and prompts users to delete unnecessary files or increase the quota. Users can continue to create files or directories after the alarm.
- File quantity hard quota: restricts the maximum available file quantity of a quota. When the number of files reaches this quota, the system generates an error stating insufficient space.

SmartQuota supports the following quota objects:

- Dtree quota: limits the total quota of all files and directories in a dtree, including the capacity and number of files.
- User quota: limits the total quota of files and directories created by a user, including the capacity and number of files.
- User group quota: limits the total quota of files and directories created by a user group, including the capacity and number of files.

5.9 SmartCache (Intelligent Cache)

SmartCache uses SCM as the read cache. Storage systems identify and store hot data to the SmartCache pool to accelerate read requests and improve overall system performance.

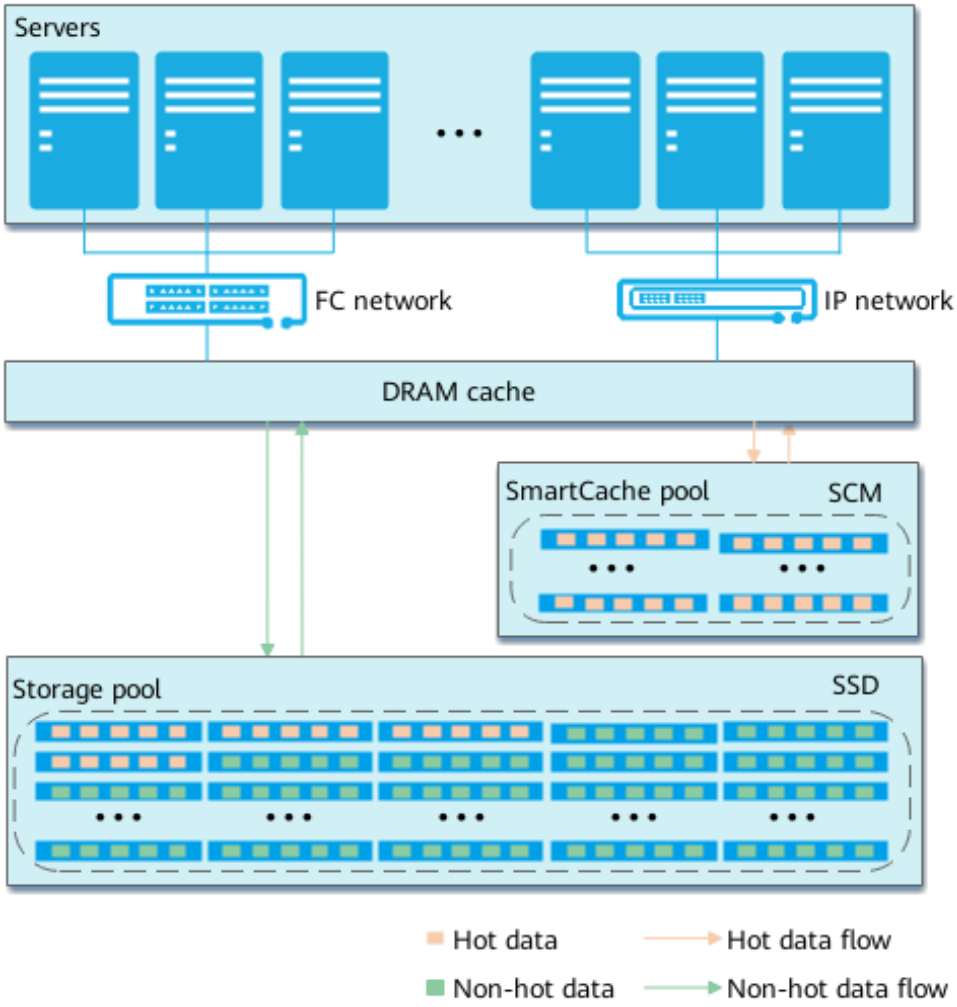
5.9.1 Working Principles

SmartCache can be used by each LUN or file system separately to improve hot data performance. If SmartCache is enabled for a LUN or file system, the DRAM cache delivers hot data to the SmartCache pool. SmartCache saves the data to SCM media and establishes the mapping between the data and SCM media in the memory.

Upon reception of a read I/O request, the storage system first attempts to read data from the DRAM cache or the SCM of SmartCache. If the requested data is found, the storage system returns the data to the host. If the requested data is not found, the storage system reads the data from the storage pool and returns to the host.

When the cached data in the SmartCache pool reaches the eviction threshold, the system evicts non-hot data based on the Least Recent Used (LRU) algorithm to release space for new hot data. By repeating this operation, the system ensures the hit rate of hot data in the SmartCache pool.

Figure 5-15 SmartCache working principles



When the data scale is large and there is no outstanding hotspots in the service model (the SmartCache capacity is less than the data volume in the service hotspots), you can configure the metadata cache policy to ensure the metadata access performance (requiring a lower ratio of SCM drives than SmartTier). This ensures a steady 500 μs latency for latency-sensitive services.

5.9.2 Application Scenarios

SmartCache applies to random read-intensive services in hotspot areas, such as OLTP and VDI.

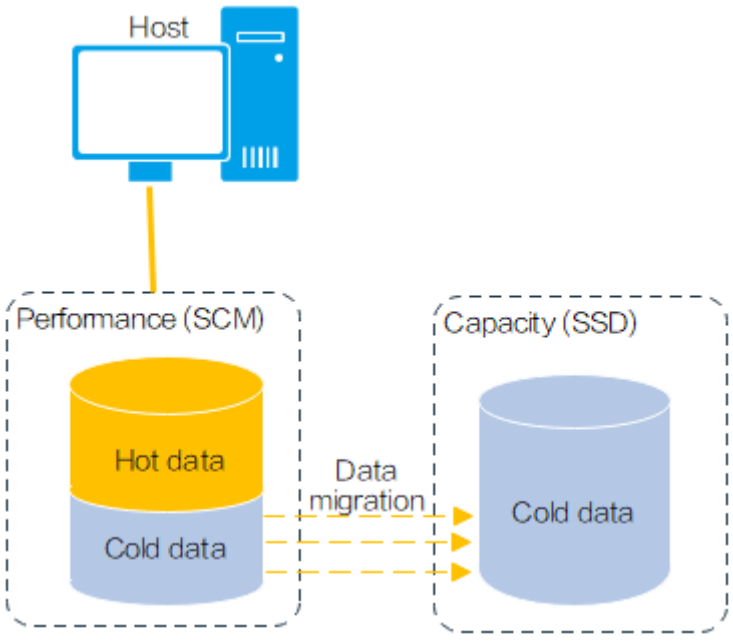
SmartCache is not recommended in the following scenarios:

- SmartCache is not recommended for services with sequential or large I/Os. In large I/O read scenarios, data may be stored in different locations on SCM media. Therefore, the SCM media will be accessed for multiple times, increasing the access concurrency.
- SmartCache is not recommended for services with a large proportion of write requests.

5.10 SmartTier (Intelligent Tiered Storage)

SmartTier simplifies data lifecycle management, improves media usage and storage performance, and reduces the cost. It dynamically stores hot and cold data to specific storage media to effectively balance performance.

SCM provides better performance than SSDs. When a storage pool has both SCM and SSDs, the SCM space is used as the performance tier and SSD space is used as the capacity tier. When SmartTier is used, new data from hosts is preferentially written to the performance tier to improve the real-time access performance. Cold data that is less frequently accessed is migrated to the capacity tier through intelligent background scheduling.



SmartTier features:

- Access acceleration

  Metadata acceleration: Metadata is stored at the performance tier (SCM) to ensure high-performance access in large-capacity scenarios.

  User data acceleration: By default, user data is stored at the performance tier (SCM). When the used capacity of the performance tier reaches a certain threshold, cold data is migrated from the performance tier to the capacity tier in the background, so the performance tier can store more hot data to improve the data access speed.

- Cost-effectiveness

  SmartTier enables tiered storage. The storage system saves data on SCM drives and SSDs, ensuring service performance at a lower cost in comparison with a storage system that only uses SCM drives.

5.11 SmartMobility (Intelligent File Tiering)

With the rapid growth of data, more and more enterprises want to put infrequently accessed data to cheaper devices. However, enterprises also want to prevent existing applications from being affected and require that applications be unaware of the data tiering of file systems. Thus, SmartMobility is developed to put infrequently accessed data to object storage devices (including cloud) and NAS devices with upper-layer applications being unaware of data tiering. The biggest advantage is that the local system space is saved to reduce costs.

SmartMobility is to automatically migrate cold files in a file system to a remote device (object storage or NAS device) based on a specified policy.

The remote device of SmartMobility can be an object storage or a NAS device.

SmartMobility supports recall. When a user reads or modifies an object that has been migrated to the remote device, the system automatically copies the object from the remote device to the local device in the background and deletes the object from the remote device.

SmartMobility is mainly used to migrate file data that is not frequently accessed to a remote device, so the usable capacity of the local device is increased, saving storage costs for customers.

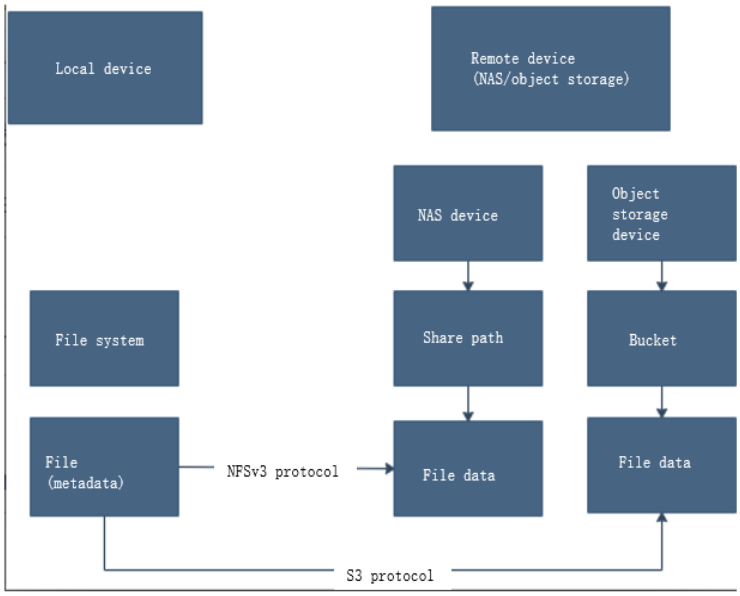
SmartMobility supports automatic recall. Files that have been migrated to a remote device can be read and written in real time and recalled in the background. After files are recalled, user's file read and write performance is improved.

5.11.1 Migration Principles

SmartMobility implements file-level migration instead of block-level migration. File data is migrated to the remote device by file. The local device stores only file metadata.

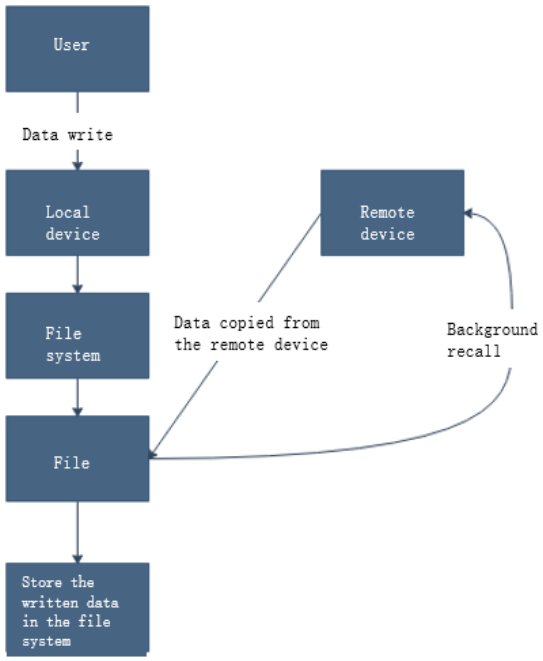
When data is migrated to a NAS device, the remote device must support the NFSv3 protocol.

When data is migrated to an object device, the remote device must support the S3 protocol.



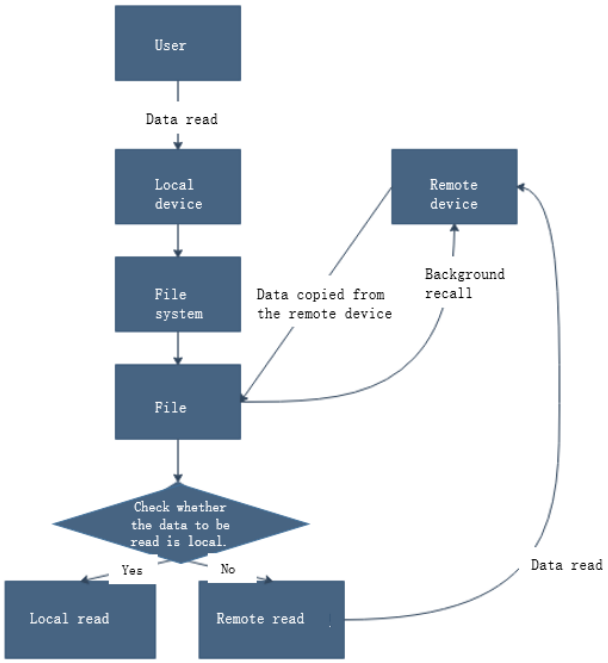
5.11.2 Recall upon Write

When a user writes data to a file that has been migrated to a remote device, the data is first stored locally. Then the file is copied from the remote device to the local device in the background. In this way, the user does not need to wait for the file to be recalled before writing it, providing a fast write speed.



5.11.3 Recall upon Read

Generally, the remote device has lower performance than the local device. When a user reads a file that has been migrated to the remote device, the data is first read from the local device. If the data is not found, it is read from the remote device and returned to the user. Finally, the file is recalled in the background. After the recall, the user no longer needs to read data from the remote device, and the read performance is improved.



5.12 SmartMulti-Tenant (Multi-Tenancy)

With improved service capability, a single storage system is carrying more and more customer service systems. In this case, customers want to isolate these service applications. OceanStor Dorado uses SmartMulti-Tenant to isolate SAN and NAS services.

SmartMulti-Tenant isolates logical resources among vStores, including services and networks. Users cannot access data across vStores, ensuring security isolation.

- **Service isolation**: Each vStore has its own storage services and user access authentication. Users can access the services through the logical interfaces (LIFs) of the vStore.
- **Network isolation**: VLANs and LIFs separate the networks among vStores, preventing unauthorized access of storage resources. For SAN services, the Fibre Channel protocol implements point-to-point communication. Users can specify Fibre Channel ports available to vStores to implement network isolation.

Service Isolation

SmartMulti-Tenant isolates SAN and NAS services by vStore. File systems and NAS users are independently configured and managed in each vStore, achieving mutual isolation. NAS vStores' resource objects include file systems, dtrees, NAS share protocols, NAS user authentication (local user and domain user), and vStore-level features (such as audit logs and quota). For SAN services, a vStore cannot access the LUNs of other vStores, ensuring data isolation by vStore.

NOTE

vStores only isolate SAN and NAS service resources and do not isolate storage pools. SAN isolation will be available in later versions. For details, see the product roadmap. Storage pool isolation is implemented using the multi-pool technology. That is, users can plan and configure multiple pools on a device to isolate storage space.

Network Isolation

vStores' network resources are managed using LIFs, implementing port virtualization, management, and isolation as well as flexible and secure use of resources.

For SAN services, users can specify Fibre Channel ports available to vStores to implement network isolation.

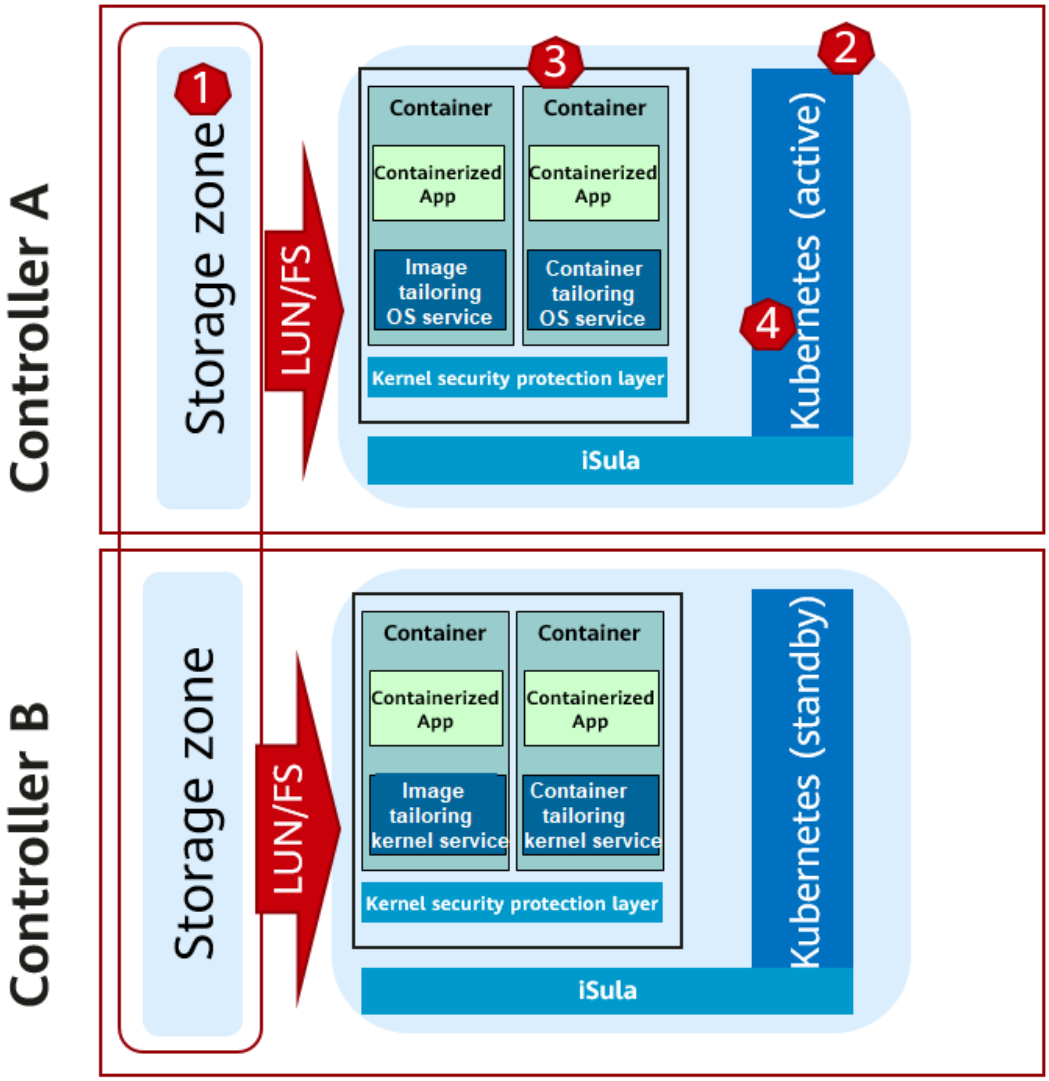
5.13 SmartContainer (Container)

SmartContainer allocates idle resources from the existing hardware resources of the device to run container services, implementing an integrated service solution. SmartContainer uses the Huawei iSula container framework and integrates the container management service of Kubernetes. It provides secure container functions to load and run containerized applications, and supports resource configurations such as container CPU, memory, network, and port. Users can enable the container service as required, import image files, and configure container running policies.

Basic Principles

SmartContainer of OceanStor Dorado uses the iSula container architecture of Huawei EulerOS and integrates the Kubernetes container management service. This feature requires license authorization. Figure 5-16 shows the logical diagram of the system after the feature is activated.

Figure 5-16 Basic principles of SmartContainer



As shown in Figure 5-16, after SmartContainer is activated, storage resources are divided into a storage zone (1 in the figure) and a container zone (2 in the figure). The storage zone provides SAN and NAS services for external applications or the containers. The container system can access the storage space through SAN and NAS protocols.

SmartContainer uses the secure container deployment. The secure container has a lightweight kernel and provides kernel security isolation to ensure isolation between storage and container services. OS service files of the container service can be packaged and imported into containerized applications together with the container image file suite to implement on-demand customization.

The container management system uses Kubernetes to manage the lifecycle of containerized applications, including container deployment, migration, health status monitoring, cross-node HA switchover, and cross-node deployment balancing. By default, the system has the built-in Kubernetes servers (one active node and multiple standby nodes) and clients of each container, forming a complete container O&M ecosystem. If container service deployment needs to be associated with the customer's production system, the Kubernetes server of the device can be disabled. The Kubernetes clients of the container can be directly connected to the external

Kubernetes server of the customer for unified scheduling and management, implementing optimal deployment and running of containerized applications on demand.

Using SmartContainer

Using SmartContainer will re-allocate system resources. To maximize resource usage, the process of using SmartContainer is different from that of using other features.

1. Ensure that the software version supports this feature, and import the feature license.
2. Activate SmartContainer. If the storage system has been providing services for external systems when the feature is activated, the system evaluates whether there is sufficient space for the container based on the current service load. If the space is sufficient, the system performs a rolling restart to release the resources required by the container while ensuring storage service continuity.
3. Import the container image that has passed the storage signature authentication, complete the service configuration of the container, and start the container service.

NOTE

SmartContainer licenses are classified into different levels based on the available resources. For details, see the product specifications list.

Application Scenarios

A storage system provides limited hardware resources, and the resources allocated to the container will affect the original service specifications and performance of the storage system. Therefore, the container images must be authenticated and signed by the storage system before they can be used. SmartContainer is recommended for data access or near-data computing container services, for example, storage access protocol extension, FTP, HTTP, S3, CloudBackup, and HyperDetect.

5.14 SmartMove (Intelligent File System Migration)

The SmartMove migration process consists of two phases:

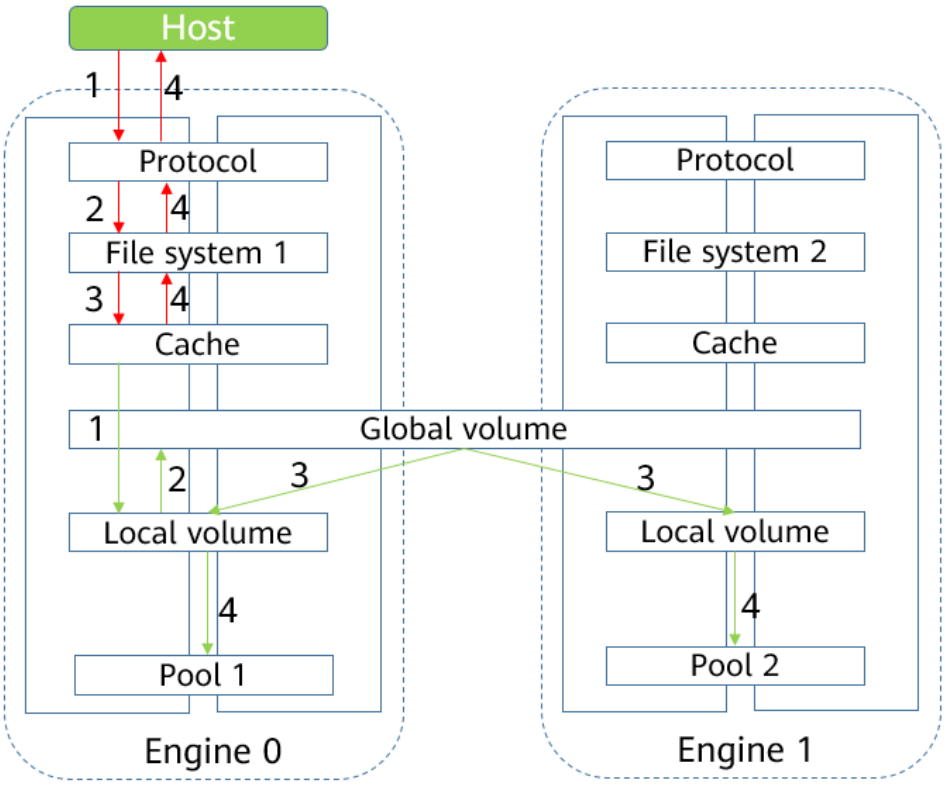
1. Data synchronization between the source and target file systems. This phase may take a long time, depending on the file system data volume and synchronization rate.

2. Services switchover to the target storage pool after data synchronization is complete. This phase takes a short time, which is about 1 minute typically.

5.14.1 SmartMove I/O Process

- Host I/O process of migrating a file system

Figure 5-17 Host I/O process



After SmartMove is configured, a migration cluster is generated. In addition, a global volume is generated and works in the migration cluster. As shown in the figure, the foreground I/O process during data migration is the same as that of local services. Cache flushing I/Os are written to the local volume of the source file system and forwarded to the global volume. The global volume writes the data to the local volumes of the source and target file systems, and then to the pool space of the source and target file systems.

- Working principles of background copy

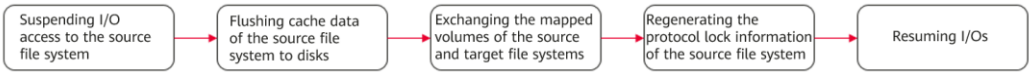
The migration task copies data in the background at the volume layer. Incremental copy is supported. Data is read from the pool of the source file system and written to the target file system based on the differential data record according to the point in time. This ensures that data at each point in time on the source and target file systems is consistent. In this way, value-added features such as HyperSnap, HyperReplication, and HyperMetro are supported. The copy process is divided into two phases. In the initial phase, cache data is flushed only to the storage pool of the source file system. In the final phase, cache data is flushed to the storage pools of both the source and target file systems, reducing the impact of the copy on host performance.

The logic of data distribution in background copy tasks of SmartMove is as follows: Copy tasks are created for each dtree in the file system and each copy task is distributed to all nodes for copy. In this way, resources of all nodes can be fully utilized to achieve more efficient copy performance. The distribution logic for copy tasks is the same as that for the file system, so that data forwarding is not required when file system read/write occurs during the copy.

5.14.2 SmartMove Service Cutover Process

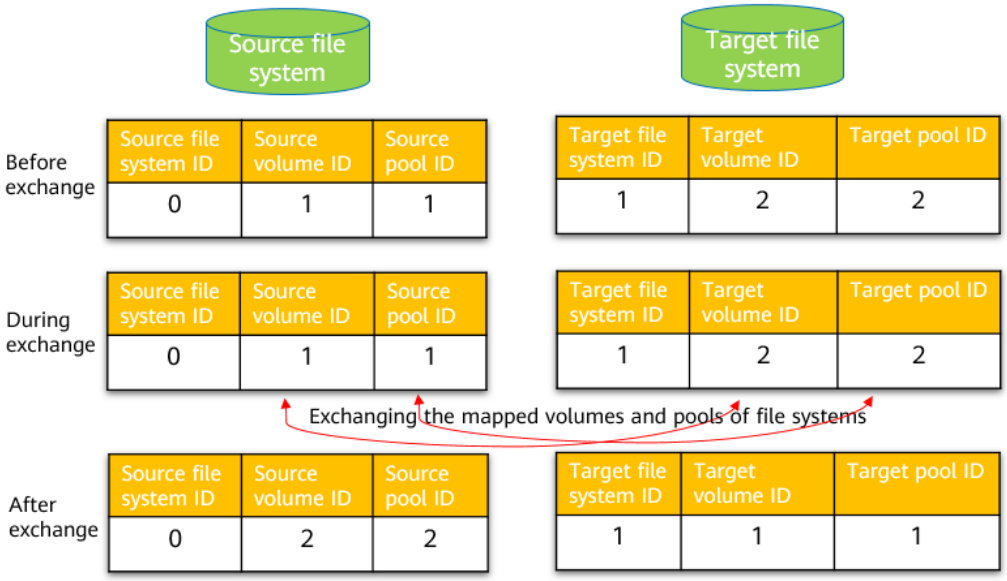
After the pool-layer data synchronization between the source and target file systems is complete, the mapped volumes and pools of the source and target file systems can be exchanged. As shown in the following figure, the service cutover process includes: suspending I/Os, flushing cache data of the source file system to disks, exchanging the mapped volumes of the source and target file systems, generating host access information, and resuming I/Os.

Figure 5-18 SmartMove service cutover process



The following figure shows the process of exchanging mapped objects between the source and target file systems. Specifically, the mapped volumes and pools of the source and target file systems are exchanged in the process. After the exchange, the source file system uses the volume and storage pool of the target file system, and the target file system uses the volume and storage pool of the source file system.

Figure 5-19 Mapped object exchange of file systems



5.14.3 Applicable Scenarios of SmartMove

SmartMove applies to scenarios where space utilization optimization, controller load balancing, or service performance optimization is required.

5.14.3.1 Space Utilization Optimization

If the usage of storage pools in a service cluster is inconsistent, you can migrate file systems between storage pools to optimize space utilization. If the space of the storage pool where a file system resides is insufficient, you can migrate the file system to another storage pool with sufficient space to meet future space growth requirements.

5.14.3.2 Controller Load Balancing

If the controllers of a controller enclosure in a storage system are heavily loaded, you can migrate some file systems that are frequently read and written to the storage pools of other unbusy controller enclosures to balance the load among controllers. This improves the overall performance of the storage system and reduces the failure rate.

5.14.3.3 Service Performance Optimization

If the performance of the storage pool where a file system resides is low, you can migrate the file system to an all-flash storage pool with higher performance to improve service performance. Similarly, when services are adjusted and high-performance access to the file system is not required, you can migrate the file system to a hybrid-flash storage pool with a lower performance to release the high-performance disk space occupied by the file system.

6 Hyper Series Features

To meet customers' requirements for local protection and remote disaster recovery, OceanStor Dorado provides the Hyper series software. HyperSnap and HyperCDP help you recover from local logic errors. HyperClone creates a complete data copy. Data integrity of the parent object has no impact on the data integrity of the clone object, isolating fault domains. HyperReplication and 3DC are used to implement remote disaster recovery. HyperMetro not only ensures service continuity but also provides disaster recovery capabilities.

6.1 HyperSnap (Snapshot)

HyperSnap is a snapshot feature of OceanStor Dorado. It provides different snapshot functions for SAN and NAS services. Snapshots for SAN services are readable and writable. They are considered as independent LUN objects by the hosts and must be mapped separately. Snapshots for NAS are read-only and are deployed in the snapshot directories of the file systems. They can be accessed after the file systems are mounted.

6.1.1 HyperSnap for SAN (Snapshot for SAN)

This section describes the technical principles and key functions of HyperSnap for SAN.

6.1.1.1 Basic Principles

OceanStor Dorado provides readable and writable snapshots for SAN services. A snapshot must be separately mapped to a host. This section details the time point (TP) technology, which is crucial to HyperSnap.

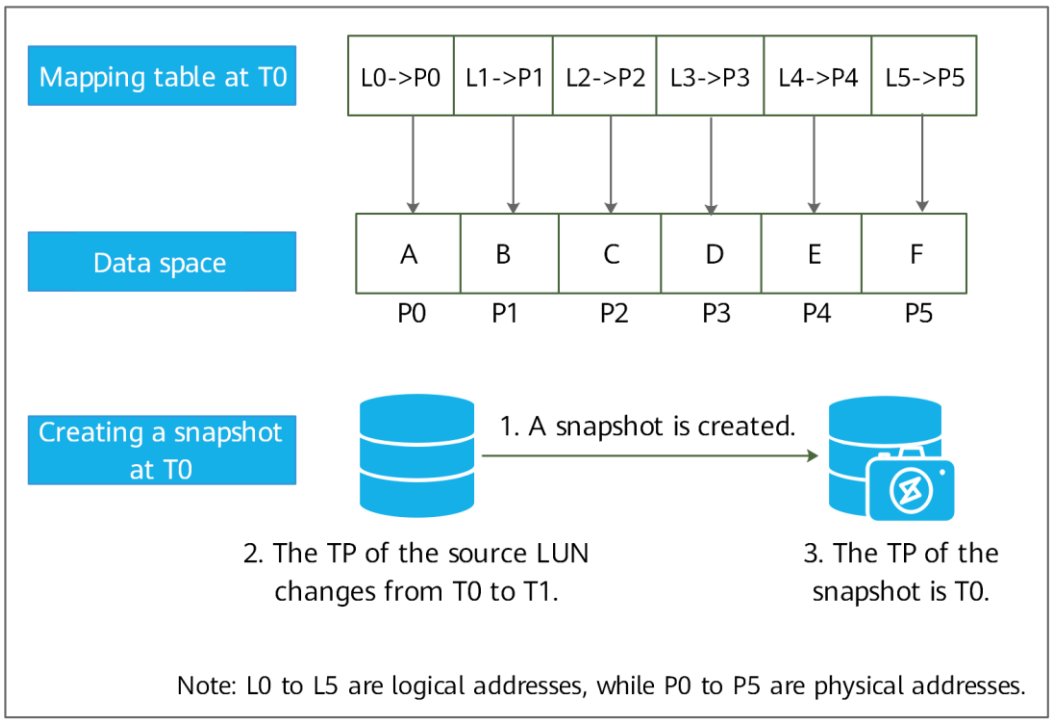
TP

OceanStor Dorado uses the multi-TP technology to implement basic data protection features. All local and remote data protection features use this technology to obtain data copies and ensure consistency.

This technology adds a TP attribute to LUNs. When a snapshot is created for a LUN, the value of TP attribute of the source LUN is incremented, while the TP attribute of the snapshot is the original TP at the snapshot creation.

In the example of Figure 6-1, the current TP of the source LUN is T0 and the user creates a snapshot for the source LUN.

Figure 6-1 Snapshot principles

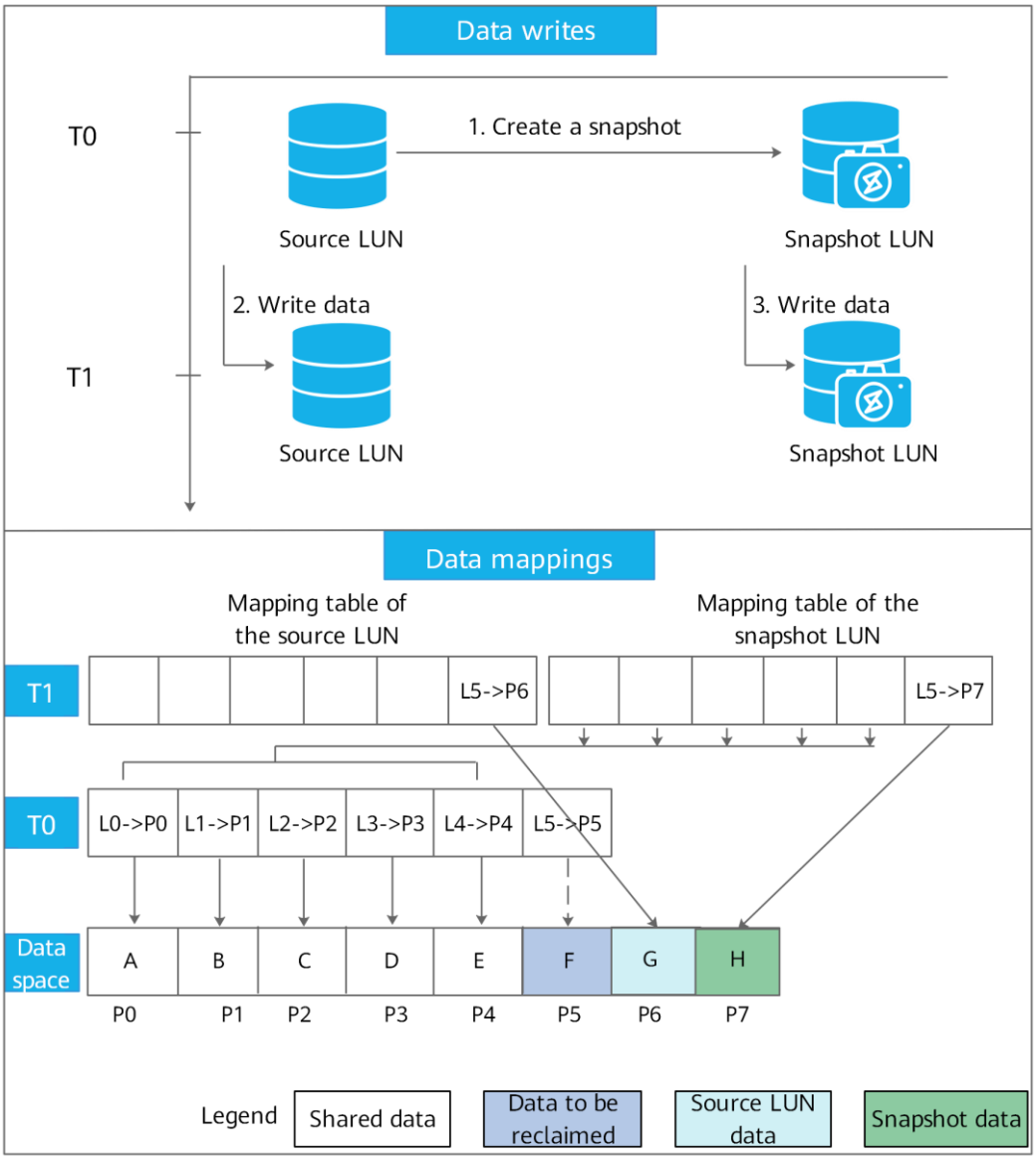


Because a snapshot is created at T0, the TP attribute of the source LUN changes from T0 to T1. The TP attribute of the snapshot is T0. When the source LUN is read, data at T1 is read; when the snapshot is read, data at T0 is read (ABCDEF in this example).

Reading and Writing a Snapshot

The host reads and writes data on the source LUN at the latest TP, or reads and writes data on the snapshot at the original TP, as shown in Figure 6-2.

Figure 6-2 Reading and writing a snapshot



- Reading the source LUN

  After a snapshot is created for the source LUN, the TP of the source LUN is updated from T0 to T1. The read requests on the source LUN will read the data within the [T0, T1] time range on the source LUN (from the latest data to the old data according to mapping entries in the mapping table). This will not cause any new performance overhead.

- Reading the snapshot

  The latest TP of the snapshot is T0. When the snapshot is read, the data at T0 is returned if the mapping table of the snapshot is not empty. If the mapping table of the snapshot is empty, a TP redirection is triggered and the data at T0 of the source LUN is read.

- Writing the source LUN

When new data is written to the source LUN, the write requests carry the latest T1. The system uses the logical address of the new data and T1 as the key, and uses the address where the new data is stored in the SSD storage pool as the key value.

- Writing the snapshot

  When new data is written to the snapshot, the write requests carry the latest T0 of the snapshot. The system uses the logical address of the new data and T0 as the key, and uses the address where new data is stored in the SSD storage pool as the value.

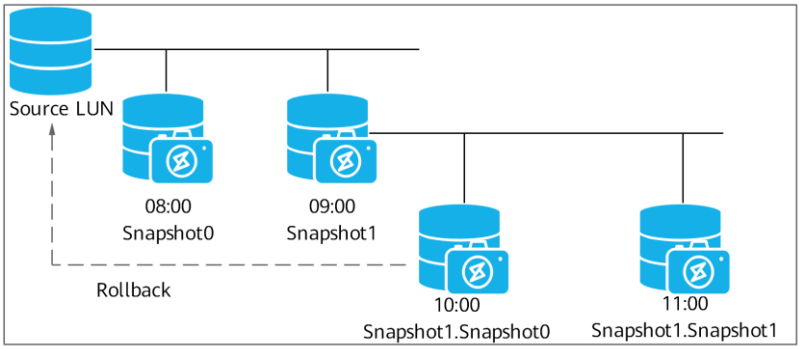
Because the read and write requests on the source LUN or snapshot carry corresponding TPs, the metadata can be quickly located, minimizing the impact on performance.

6.1.1.2 Cascading Snapshot

To protect writable snapshots, you can use cascading snapshots. Cascading snapshot is to create child snapshots for a parent snapshot. On OceanStor Dorado, HyperSnap supports up to eight levels of cascading snapshots.

Cascading snapshots support cross-level rollback. For multi-level cascading snapshots that share a source LUN, they can roll back to each other regardless of their cascading levels. In Figure 6-3, **Snapshot1** is created for the source LUN at 9: 00, and **Snapshot1.Snapshot0** is a cascading snapshot of **Snapshot1** at 10:00. The system can roll back the source LUN using **Snapshot1.Snapshot0** or **Snapshot1**, or roll back **Snapshot1** using **Snapshot1.Snapshot0**.

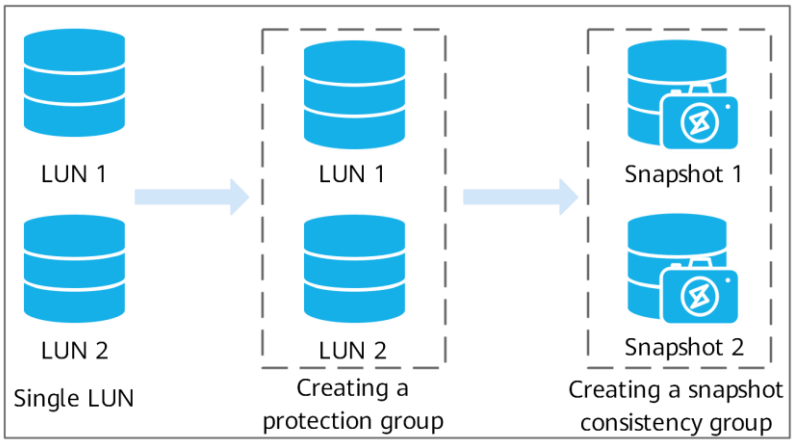
Figure 6-3 Cascading snapshot and cross-level rollback



6.1.1.3 Snapshot Consistency Group

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 6-4 Working principles of the snapshot consistency group



1. Create a LUN protection group and add LUNs to it.

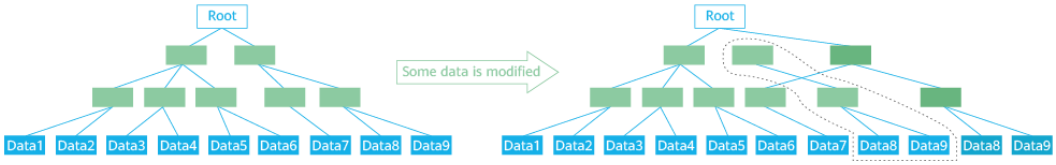
NOTE

A maximum of 4096 LUNs can be added to a LUN protection group.

2. Create a snapshot consistency group for the protection group. The snapshots in the snapshot consistency group have the same TP.

6.1.2 HyperSnap for NAS (Snapshot for NAS)

HyperSnap for NAS is also based on the TP technology. It works with the NAS service object management model to implement thin space and lossless snapshot capabilities. NAS snapshots are read-only. The writable snapshot capability is provided by the clone feature.



OceanStor Dorado uses the redirect-on-write (ROW) mechanism to create snapshots for file systems. ROW allows the file system to write newly created or modified data to a new space instead of overwriting the original data, ensuring high reliability and scalability of the file system. With ROW, file system snapshots can be created in seconds and do not occupy additional space unless the source files are deleted or modified.

A file system snapshot only copies and stores the root node of the file system. No user data is copied, so creation can be completed in 1 to 2 seconds. The snapshot shares space with its source file system before data modification to eliminate the need for additional space.

A new snapshot does not contain any user data but only a group of pointers for locating user data of the source file system. As a result, users accessing snapshot data are actually accessing the data of the source file system. When data in the source file system is modified, the snapshot retains the space occupied by the original data for protection. The protected space will not be reclaimed unless the snapshot is deleted.

Updates to the source file system cause the snapshot to retain more of the original data blocks and increase the snapshot space. However, the snapshot only retains the data at the point in time when it was created. New data change after that time point is not protected by the

snapshot and does not occupy snapshot space. Snapshot rollback restores the data to its original state at the time of creation to prevent data loss from misoperations or viruses.

Exercise caution when performing snapshot rollback because its changes are irreversible. A rollback can restore data to a specified point in time, but data written after that time will be lost with the snapshot after a rollback. This loss of data may result in services being interrupted if data is accessed during rollback. Manually copy individual files from the snapshot to the source file system for small-scale restoration to avoid rolling back the entire file system.

NAS Secure Snapshot

You can set security attributes to convert read-only snapshots of a file system into secure snapshots to prevent accidental or intentional deletion of the snapshots.

The security attributes of a snapshot include:

- Whether it is a secure snapshot.
- Protection period of the secure snapshot. The value ranges from 1 day to 20 years.
- Whether the secure snapshot is automatically deleted after expiration.

You can set the security attributes of a snapshot in the following ways:

- Set the security attributes when creating a snapshot.
- Modify the security attributes after a snapshot has been created.

After security attributes are set for a file system snapshot, the snapshot cannot be deleted during the protection period.

SnapDiff (NAS Snapshot Comparison)

SnapDiff is a snapshot comparison engine provided by OceanStor Dorado in the form of RESTful APIs. By comparing the differential data of the read-only snapshots of two file systems, you can identify changes such as file creation, deletion, and modification and form a differential file list, which can be used for incremental data backup.

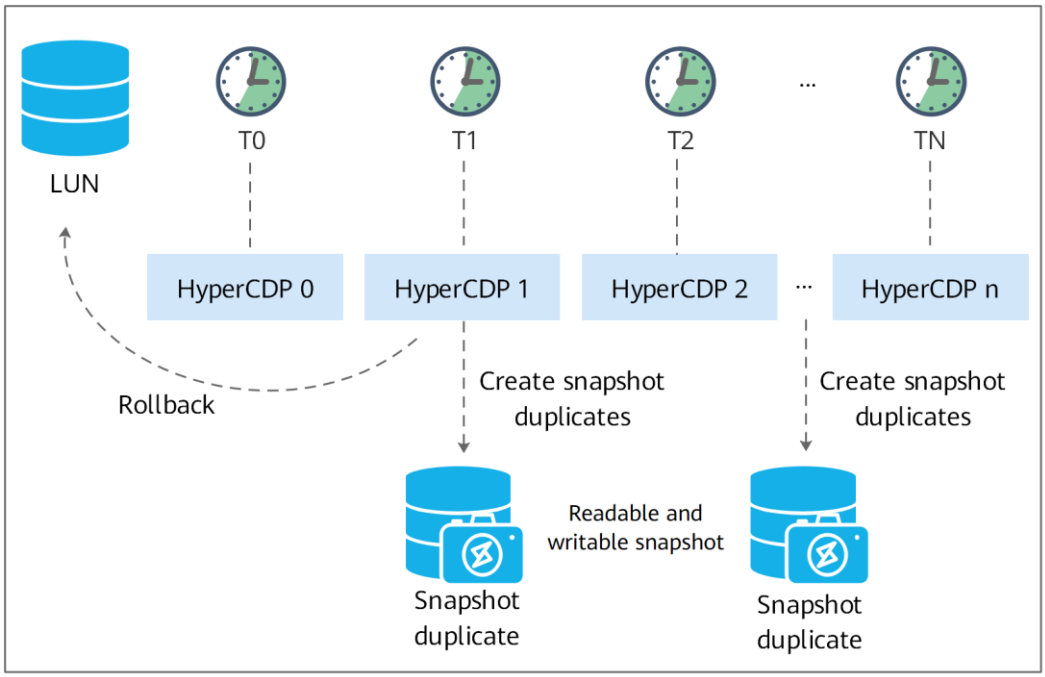
6.2 HyperCDP (Continuous Data Protection)

HyperCDP creates high-density snapshots on a storage system to provide continuous data protection (CDP). This section provides an overview of this feature.

Misoperations and virus attacks may cause data corruption. Continuous data protection is to create snapshots at a short interval to help customers restore data.

HyperCDP allows OceanStor Dorado to continuously protect LUNs. HyperCDP is based on the lossless snapshot technology (TP and ROW). Each HyperCDP object matches a time point of the source LUN. Figure 6-5 illustrates how HyperCDP is implemented.

Figure 6-5 HyperCDP principles



HyperCDP Schedule

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

Table 6-1 HyperCDP schedule

Schedule Type	Description
Fixed period	• If you set the interval by second, the HyperCDP schedule is executed every 10 seconds by default. • If you set the interval by minute, the HyperCDP schedule is executed every 1 minute by default. • If you set the interval by hour, the HyperCDP schedule is executed every 1 hour by default. **NOTE** SAN: A single LUN supports a maximum of 60,000 HyperCDP objects. The system supports a maximum of 2,000,000 HyperCDP objects. NAS: A single file system supports a maximum of 4096 HyperCDP objects. The system supports a maximum of 128,000 HyperCDP objects.
Execute daily	Set the point in time every day for the storage system to create HyperCDP snapshots. The value ranges from 00:00 to 23:59 at the local time zone. **NOTE** The number of retained HyperCDP objects ranges from 1 to 256.
Execute weekly	Set the point in time and day in a week for the storage system to create HyperCDP snapshots. The value ranges from 00:00 Sunday to 23:59

Schedule Type	Description
	Saturday at the local time zone. **NOTE** The number of retained HyperCDP objects ranges from 1 to 256.
Execute monthly	Set the point in time and day in a month for the storage system to create HyperCDP snapshots. The value ranges from 00:00 Day 1 to 23:59 Day 31 (or last day) at the local time zone. **NOTE** The number of retained HyperCDP objects ranges from 1 to 256.

Intensive and Persistent Data Protection

A single LUN supports 60,000 HyperCDP objects. The minimum interval is 3 seconds. A single file system supports a maximum of 4096 HyperCDP objects. The minimum interval is 15 seconds. You can configure the retention policy of HyperCDP in the scheduling policy.

HyperCDP Consistency Group (for SAN Only)

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

Secure Snapshot

In financial, securities, or bank applications, HyperCDP objects are configured to back up critical data for long term retention. To prevent HyperCDP objects from deletion, you can set a retention period. The HyperCDP objects cannot be deleted within the retention period. After the period expires, they can be deleted manually or automatically.

- You can create secure snapshots for individual LUNs or file systems and secure snapshot CGs for LUN CGs. The retention period ranges from 1 day to 20 years. You can configure whether to automatically delete the snapshot upon expiration.
- You can change HyperCDP objects to secure snapshots and HyperCDP CGs to secure snapshot CGs. The retention period ranges from 1 day to 20 years. You can configure whether to automatically delete the snapshot upon expiration.
- You can adjust the retention period and the automatic deletion policy for secure snapshots and secure snapshot CGs. The retention period can be extended but cannot be shortened.
- You can create schedule policies for secure snapshots.

NOTE

- Secure snapshots cannot be deleted within the retention period. Given that the retention period cannot be shortened, exercise caution when configuring it.
- Even if **Protection Data Auto Deletion** is enabled, secure snapshots will not be deleted if the used capacity of the storage pool reaches the **Capacity Used Up Alarm Threshold** or the protection capacity of the storage pool reaches the upper threshold.
- Secure snapshots have an independent clock free from the system time change. The clock is updated once every minute. After the system is shut down, the clock stops.

6.3 HyperClone (Clone)

On OceanStor Dorado, HyperClone allows the system to create a complete physical copy of the source LUN's or file system's data on the target LUN or file system. The target LUN or file system can be an existing one or automatically created when the clone pair is created. The source and target LUNs or file systems that form a clone pair must have the same capacity. The target LUN or file system can either be empty or have existing data. If the target LUN or file system has data, the data will be overwritten by the source LUN or file system of HyperClone. Data access of the clone LUN or file system is independent from that of the source LUN or file system. That is, changes to one LUN or file system do not affect the data of the other LUN or file system.

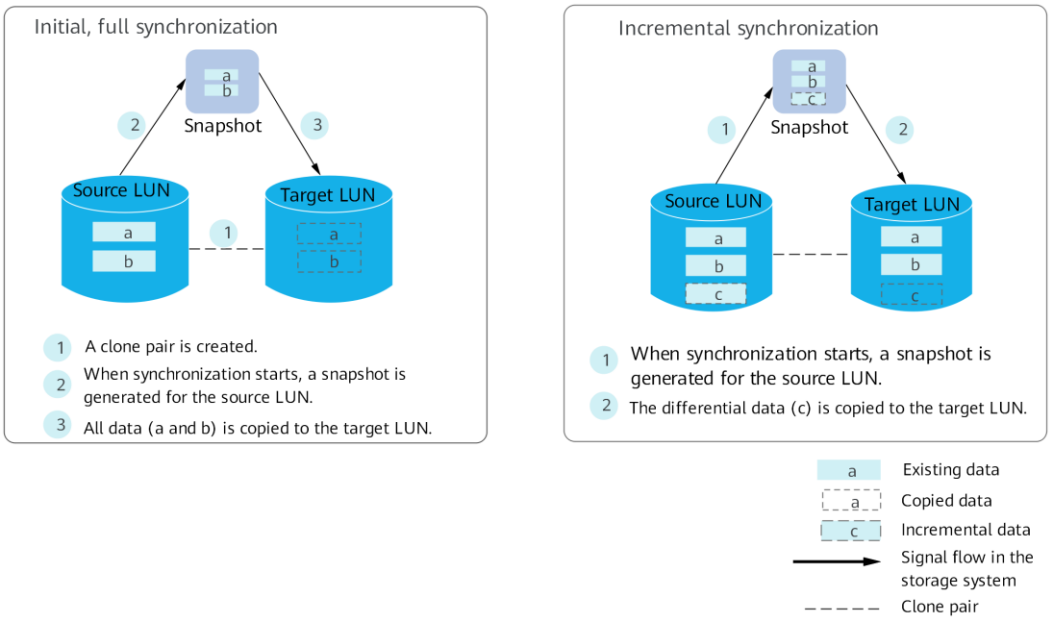
6.3.1 HyperClone for SAN (Clone for SAN)

After a clone LUN is created, it shares the same data with the source LUN. The data read/write model is the same as that of the snapshot LUN and source LUN. Users can split a clone LUN to start background data replication. The data synchronization status does not affect the read/write status of the target LUN. The target LUN can be read and written immediately without waiting for the background replication to complete. HyperClone for SAN supports incremental synchronization and reverse synchronization. You can create a HyperClone consistency group using a LUN protection group to protect data consistency on a group of source LUNs.

6.3.1.1 Data Synchronization

When a clone pair starts synchronization, 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 the DCL. 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.

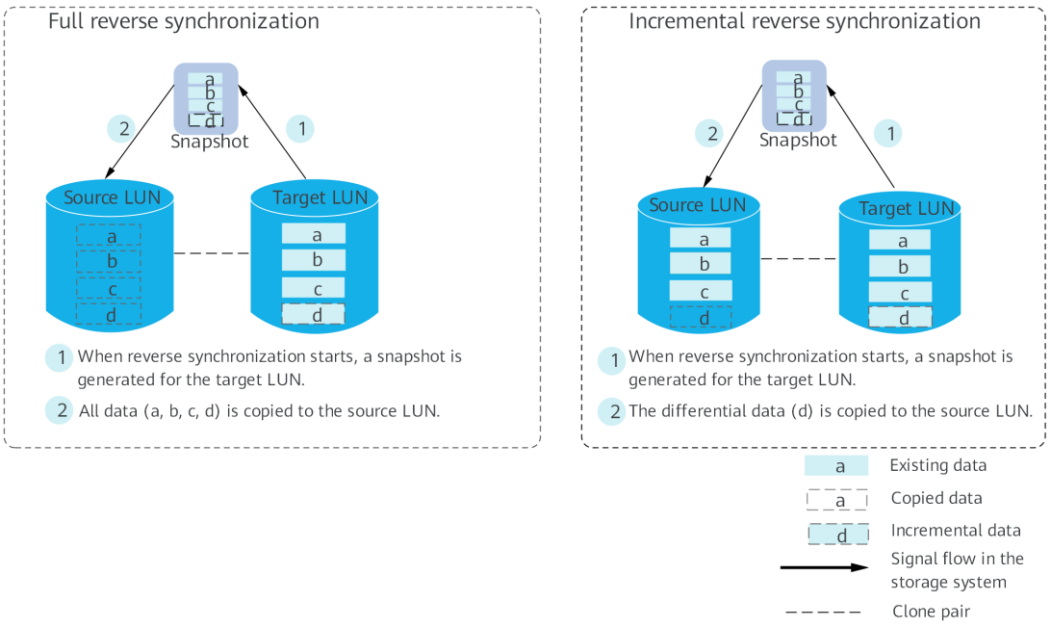
Figure 6-6 Data synchronization from the source LUN to the target LUN



6.3.1.2 Reverse Synchronization

If the source LUN is damaged, data on the target LUN can be reversely synchronized to the source LUN. Both full and incremental reverse synchronizations are supported. When reverse synchronization starts, the system generates a snapshot for the target LUN and synchronizes the snapshot data to the source LUN. For incremental reverse synchronization, the system compares the data of the source and target LUNs, and only synchronizes the differential data.

Figure 6-7 Reverse synchronization from the target LUN to the source LUN



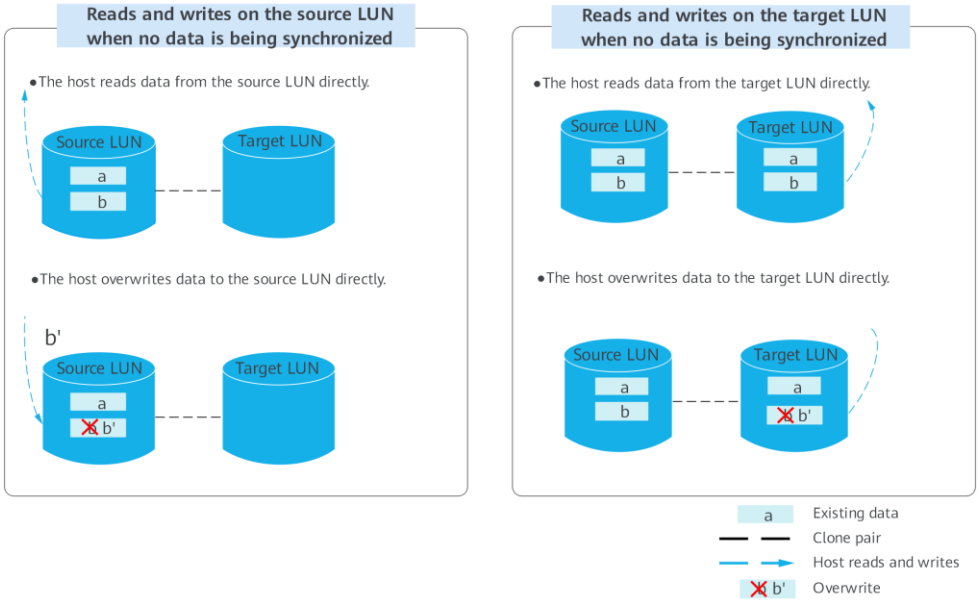
6.3.1.3 Immediately Available Clone LUNs

The HyperClone data synchronization status includes **Synchronizing**, **Sync paused**, **Unsynchronized**, or **Normal**. In different status, the read and write I/Os of the source LUN and target LUN are processed in different ways.

1. When HyperClone is in the normal or unsynchronized state:

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

   Figure 6-8 Reads and writes when HyperClone is in the normal or unsynchronized state



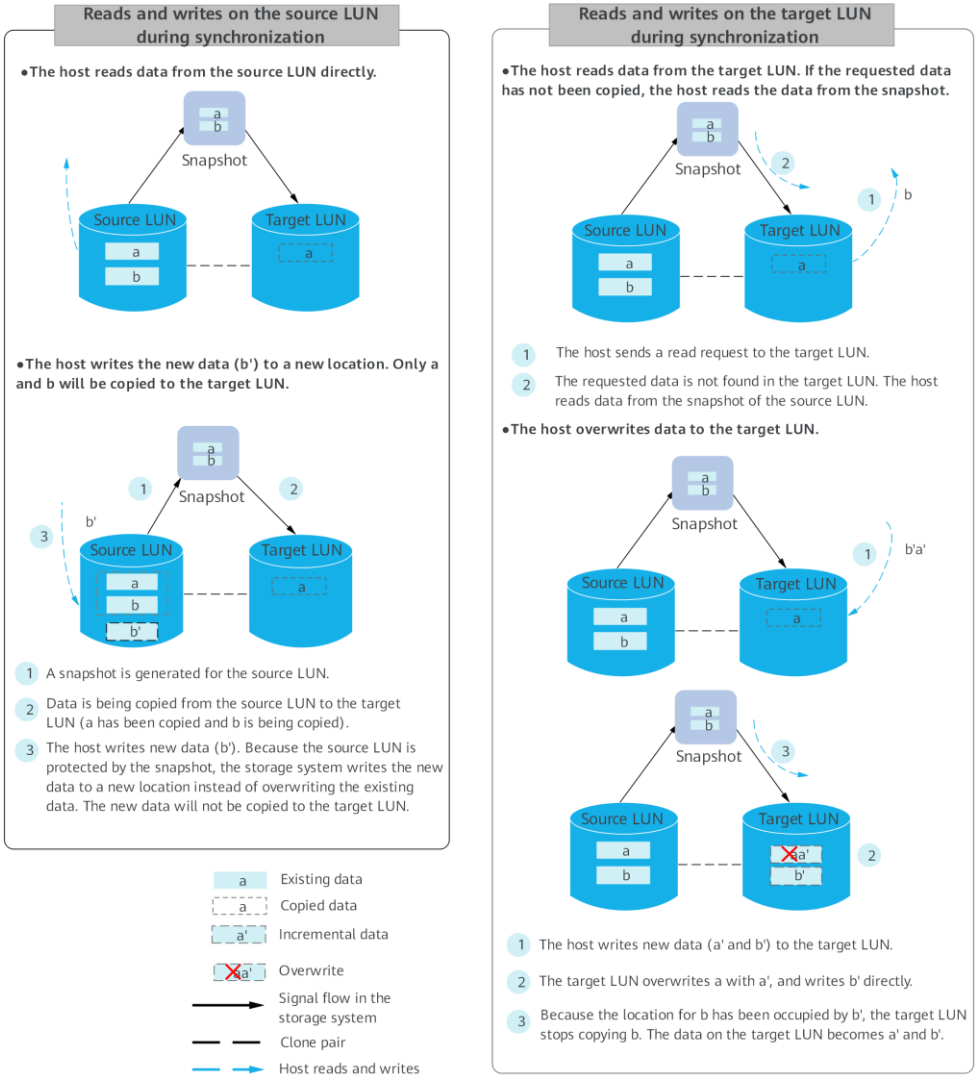
2. When HyperClone is in the synchronizing or paused state:

   The host reads and writes the source LUN directly.

   For read operations on the target LUN, if the requested data is found on the target LUN (the data has been synchronized), the host reads the data from the target LUN. If the requested data is not found 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 6-9 Reads and writes when HyperClone is in the synchronizing or paused state



6.3.1.4 HyperClone Consistency Group

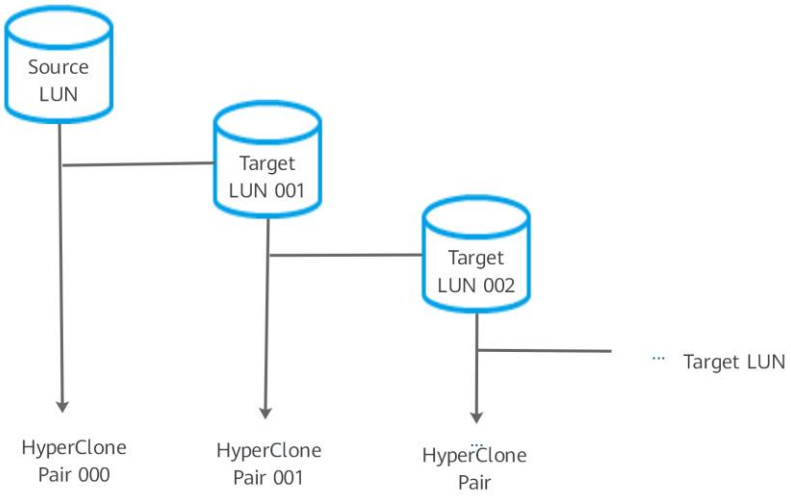
HyperClone allows you to create a consistency group for a LUN protection group. A HyperClone consistency group contains multiple clone pairs. When you synchronize or reversely synchronize a consistency group, data on all of its member LUNs is always at a consistent point in time, ensuring data integrity and availability.

A HyperClone consistency group supports a maximum of 4096 members.

6.3.1.5 Cascading Clone Pairs

After data has been synchronized to the target LUN of a clone pair, you can create another clone pair for the target LUN by using it as a new source LUN, as shown in Figure 6-10.

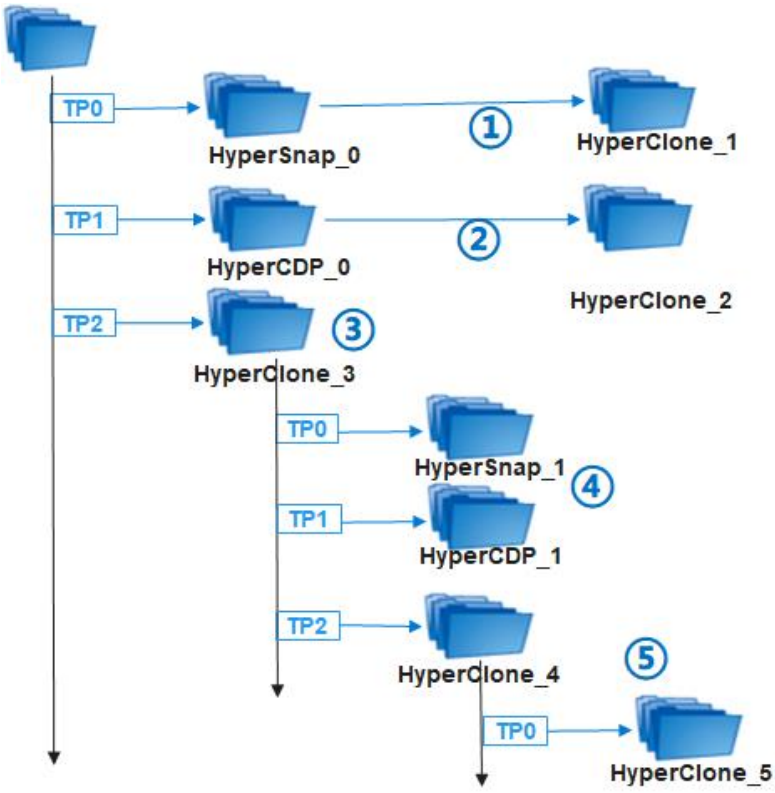
Figure 6-10 Cascading clone pairs



HyperClone has no restriction on the cascading levels.

6.3.2 HyperClone for NAS (Clone for NAS)

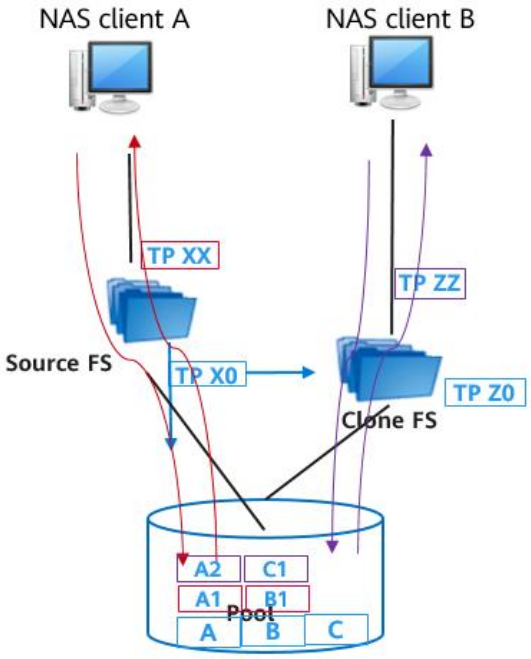
NAS clone includes file system-level clone and file-level clone. Currently, OceanStor Dorado supports only file system-level clone. File system-level clone allows users to create a clone file system using a specified source file system or file system snapshot. The data in the clone file system is the same as that in the parent file system at the creation time. The clone file system can be read and written immediately. HyperClone can be used together with HyperSnap and HyperCDP to flexibly protect and manage data locally.



1. Creating a clone based on a user snapshot
2. Creating a clone based on a HyperCDP object
3. Creating a clone based on a source file system
4. Creating a snapshot or HyperCDP object for a clone file system
5. Cascading clone (not available in the current version)

Read and Write Principles of a Clone File System

A clone file system is an independent file system that shares the existing data and pool with the parent file system. The clone file system inherits the attributes of the sub-objects from the parent file system, but does not inherit the user management configurations, such as the value-added and protocol sharing configurations. Therefore, before reading and writing a clone file system, you must configure a share and mount the clone file system to the client.



The object dependency of the clone file system is as follows:

A clone file system is created for the source file system at TP X0. Data of the file system at TP X0 matches that of the clone file system at TP Z0. In addition, the time point of the source file system is changed to TP XX, and the time point of the clone file system is changed to TP ZZ. All data is stored in the pool (ABC).

Accessing the source file system:

- Write: New and modified data (A1/B1) is written to the pool at TP XX.
- Read: The requested data is searched in the pool at TP XX (A1/B1). If the data is not found at TP XX, the data at the nearest time point to TP XX is read. That is, the read data is A1/B1/C.

Accessing the clone file system:

- Write: New and modified data (A2/C1) is written to the pool at TP ZZ.
- Read: The requested data is searched in the pool at TP ZZ. If the data is not found at TP ZZ, the data (A2/C1) at a time point between TP ZZ and TP Z0 and nearest to TP ZZ is read. If the data is still not found, it means the data has not been modified in the clone

file system, and the system obtains the data (B) of TP X0 and earlier time points from the source file system corresponding to TP Z0. That is, the read data is A2/B/C.

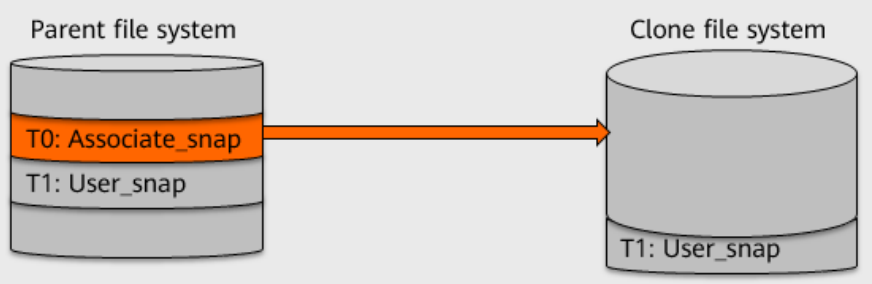
Splitting a Clone File System

After a clone file system has been created based on the snapshot of the parent file system, the clone file system inherits all data of the snapshot (which is called an associated snapshot of the clone file system). The clone file system can access the data inherited from the parent file system without occupying extra space. If you split the clone file system, it becomes a common file system independent of its parent file system and does not share any data in the parent file system.

Working principle of splitting:

- When a clone is split, all data at the point in time preserved by the associated snapshot is copied to the clone file system at the corresponding point in time. Data that has been modified by the clone file system will not be copied.

Figure 6-11 Splitting a clone file system



6.4 HyperReplication (Remote Replication)

6.4.1 HyperReplication for SAN (Remote Replication for SAN)

HyperReplication is a remote replication feature provided by OceanStor Dorado to implement synchronous or asynchronous data replication, supporting intra-city and remote disaster recovery solutions.

HyperReplication supports the following two modes:

1. **Synchronous remote replication**. Data on the primary LUN is synchronized to the secondary LUN in real time. No data is lost if a disaster occurs. However, production service performance is affected by the data transfer latency.
2. **Asynchronous remote replication**. Data on the primary LUN is periodically synchronized to the secondary LUN. Production service performance is not affected by the data transfer latency. However, some data may lose if a disaster occurs.

HyperReplication provides the storage system-based consistency group function for synchronous and asynchronous remote replication to ensure the crash consistency of cross-LUN applications in disaster recovery replication. The consistency group function protects the dependency of host write I/Os across multiple LUNs, ensuring data consistency between secondary LUNs.

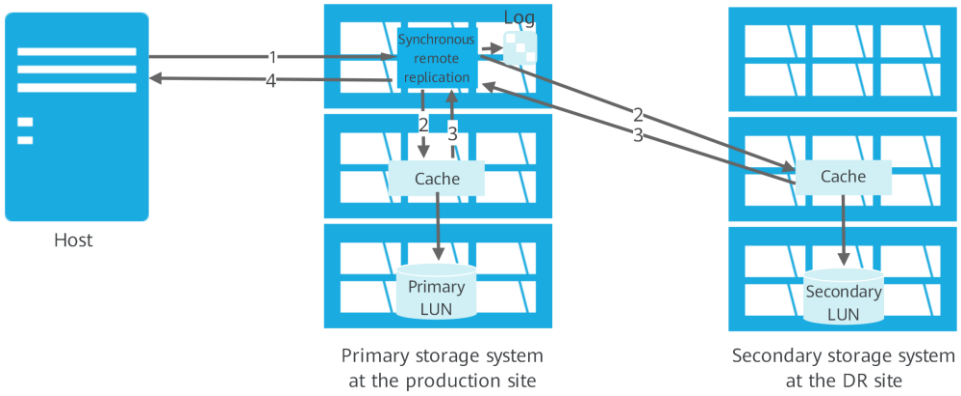
HyperReplication enables data to be replicated using Fibre Channel and IP networks. Data can be transferred between the primary and secondary storage systems using Fibre Channel or IP links.

6.4.1.1 HyperReplication/S (Synchronous Remote Replication)

HyperReplication/S supports the short-distance data disaster recovery of LUNs. It applies to same-city disaster recovery that requires zero RPO.

It concurrently writes each host write I/O to both the primary and secondary LUNs of the remote replication pair and returns a write success acknowledgement to the host after the data is successfully written to the primary and secondary LUNs. Therefore, the RPO is zero.

Figure 6-12 Working principles



Description:

1. The production storage system receives a write request from the host. HyperReplication logs the address information instead of data content.
2. The data of the write request is written to both the primary and secondary LUNs. If LUNs are in the write-back state, a write result will be returned after the data is written to the cache.
3. HyperReplication waits for the data write results from 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 enters the interrupted state. The data will be replicated in the next synchronization.
4. HyperReplication returns the data write result. The data write result of the primary LUN prevails.

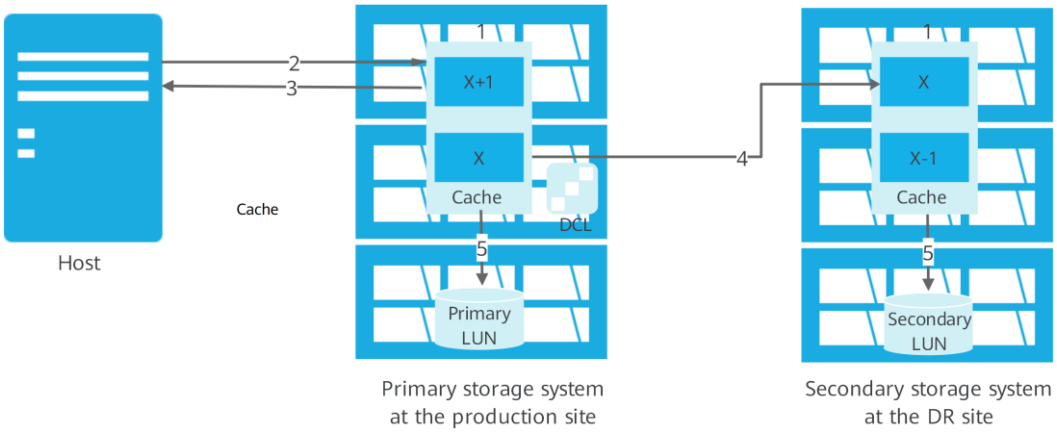
6.4.1.2 HyperReplication/A (Asynchronous Remote Replication)

HyperReplication/A of OceanStor Dorado adopts the multi-time-segment caching technology (patent number: PCT/CN2013/080203). The working principle of the technology is as follows:

1. After an asynchronous remote replication relationship is established between a primary LUN at the production site and a secondary LUN at the DR site, an initial synchronization is implemented by default to copy all data from the primary LUN to the secondary LUN.

2. When the initial synchronization is complete, the data status of the secondary LUN becomes **Consistent** (data on the secondary LUN is a copy of data on the primary LUN at a certain point in time in the past). Then I/Os are processed as follows:

Figure 6-13 Working principles of HyperReplication/A



Description:

1. When an asynchronous remote replication task is started, snapshots are generated for the primary and secondary LUNs and the snapshots' TPs are updated. (The primary snapshot TP is X, and the TP is updated to X+1. The secondary snapshot TP is Y, and the TP is updated to Y+1.)
2. New data from the host is stored in the primary LUN cache using TP X+1.
3. The host receives a write success.
4. Data at X is directly replicated to the secondary LUN at Y+1 based on the DCL.
5. The primary and secondary LUNs write the received data to disks. After the synchronization is complete, the data at the latest TP Y+1 on the secondary LUN is the data at the TP X on the primary LUN.

6.4.1.3 Technical Highlights

- Load balancing

  When a remote replication task is started for a LUN, the task is executed by all controllers concurrently. The workload of the task is distributed to all controllers based on the data layout, improving the replication bandwidth and reducing the impact on front-end services.

- Data compression

  In asynchronous remote replication, both Fibre Channel and IP links support data compression by using the fast 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 OLTP application with 100 Mbit/s bandwidth, data compression saves half of the bandwidth.

- Quick response to host requests

  In asynchronous remote replication, 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 supports splitting, synchronization, primary/secondary switchover, and recovery after disconnection. Disaster recovery tests and service switchover are supported in various fault scenarios.

- 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.

- Support for fan-in/fan-out

  HyperReplication supports data replication from 64 storage devices to one storage device for central backup (64:1 fan-in/fan-out), greatly reducing the disaster recovery cost.

- Support for various types of replication links

  HyperReplication supports both Fibre Channel and IP replication links. In synchronous replication scenarios, Fibre Channel (FastWrite) or RDMA is recommended for replication links.

- Entry-level, mid-range, and high-end storage interworking

  OceanStor Dorado allows the HyperReplication to be deployed across high-end, mid-range, and entry-level storage systems.

- Cross-array snapshot data synchronization

  User snapshots of the primary LUN can be synchronized to the secondary LUN in sequence. This function is recommended when host data consistency is required.

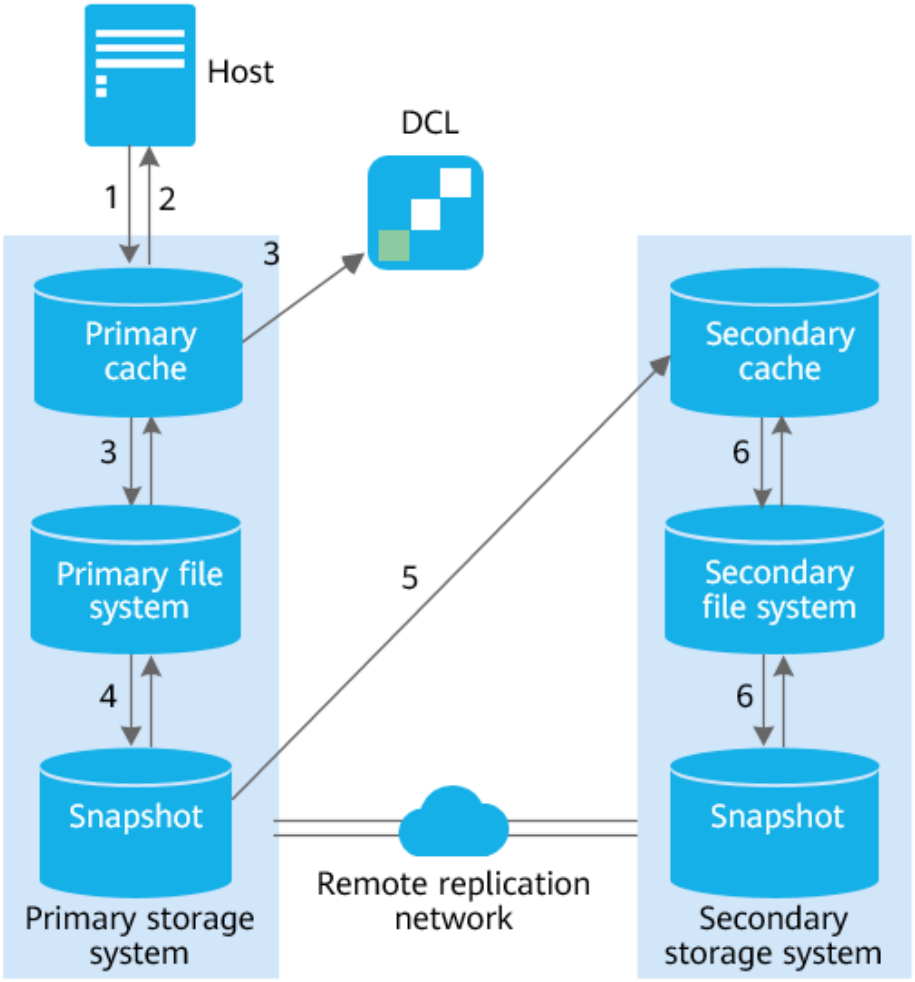
- Interoperability across product generations

  OceanStor Dorado 6.1.2 and later 6.x.x versions can communicate with OceanStor Dorado V3 series, OceanStor V3 series, and OceanStor V5 series.

6.4.2 HyperReplication for NAS (Remote Replication for NAS)

HyperReplication for NAS implements asynchronous remote replication, which applies to scenarios where zero RPO is not required. If zero RPO is required, use the NAS HyperMetro feature and select the synchronous mode or active-active mode based on the customer's switchover policy for service continuity.

Similar to asynchronous remote replication for SAN, asynchronous remote replication for NAS periodically replicates data based on ROW snapshots. The difference lies in the logic of service objects. The following figure shows the I/O logic.



Description:

0: After the asynchronous replication object is successfully configured, the system starts the synchronization task in the background. The snapshot time point of the file system is incremented by 1.

1: Host data is written to the primary file system.

2: The primary file system returns the client access result.

3-4: The primary file system records the new data in a data change log (DCL) and uses a private snapshot to retain the original data.

4-6: The asynchronous replication task replicates the data in the task period to the secondary file system based on the DCL.

The data is periodically replicated to the secondary file system in the background. Replication periods are defined by users. The system records the addresses of incremental data in each period but does not record the data content. At the beginning of each period, a snapshot is created for the primary file system. The system reads the incremental data from the end of the last period to the present and replicates it to the secondary file system. After the incremental replication is complete, the data in the secondary file system is consistent with that in the primary file system. Then a snapshot is created for the secondary file system for data protection. If the next replication is interrupted by a production site fault or link failure, the system uses the snapshot to roll back the secondary file system to the state at the last successful replication for guaranteed data consistency.

HyperReplication/A for file systems has the following highlights.

Splitting and Incremental Resynchronization

You can split a remote replication pair to stop data replication from the primary file system to the secondary file system. Splitting stops the current replication process and all planned replications in the future. After splitting, new data written by the host is recorded as incremental data. If you synchronize the remote replication pair again, the system replicates only the differential data to the secondary file system. Data that already exists in the secondary file system will not be replicated again.

Splitting applies to planned device maintenance scenarios, such as storage array upgrades and replication link changes. In such scenarios, various system tasks will be paused for better reliability and resumed or restarted after the maintenance.

Replication Interruption and Automatic Recovery

Remote replication enters the interrupted state if data replication from the primary file system to the secondary file system is interrupted by a fault such as a link failure. In this state, new data written by the host is recorded as incremental data. After the fault is rectified, remote replication automatically recovers and performs incremental resynchronization without manual intervention.

Automatic recovery is configurable. You can change it to manual recovery if needed.

Readable and Writable Secondary File System and Incremental Failback

In the normal state, you cannot read data from or write data to the secondary file system, and data on any snapshot of the secondary file system is read-only.

The secondary file system can be readable and writable when the following conditions are met:

1. The remote replication pair is split or interrupted.
2. Data on the secondary file system is complete. HyperReplication/A provides complete data on the secondary file system after initial synchronization.

If replication to the secondary file system is incomplete (inconsistent data) when you set the secondary file system to be readable and writable, the system rolls back the secondary file system to the point of the snapshot generated at the last complete replication.

Data written to a readable and writable secondary file system will be recorded as incremental data for subsequent incremental resynchronization. When replication recovers, you can choose to replicate incremental data from the primary to the secondary file system or from the secondary to the primary file system (a primary/secondary switchover is required before synchronization). Before replication starts, the system first uses a snapshot to roll back the target side data to make it consistent with the source file system data at the snapshot creation time. Then the system replicates the incremental data between the snapshot creation time and the current time.

Readable and writable secondary file systems are commonly used in disaster recovery (DR) and backup. After a failover in the event of a primary site fault, you can set the secondary file system to be readable and writable so that the secondary host can access it to take over services. After fault recovery and failback, incremental data is replicated from the secondary file system to the primary file system.

Primary/Secondary Switchover

Primary and secondary file systems support role switching when the asynchronous remote replication pair is split or interrupted. The original primary file system becomes secondary, and the original secondary file system becomes primary. Data is replicated from the primary to the secondary file system, so these roles determine the direction of replication.

Readable Secondary File System

Principles:

The secondary file system of asynchronous replication is readable even if write protection is not disabled. Data from the private snapshot of the secondary file system will be read.

Steps:

1. Configure a logical interface (LIF) for each vStore on the secondary file system.

2. Create shares and share permissions for the secondary file system.

3. Mount the secondary file system to the host and read data from the secondary file system.

Constraints:

1. The read function is supported in asynchronous replication at least after initial synchronization is complete (private consistency snapshots are generated on the secondary file system).

2. IP addresses and share authentication cannot be synchronized during configuration synchronization of asynchronous replication.

6.5 HyperVault (All-in-One Backup)

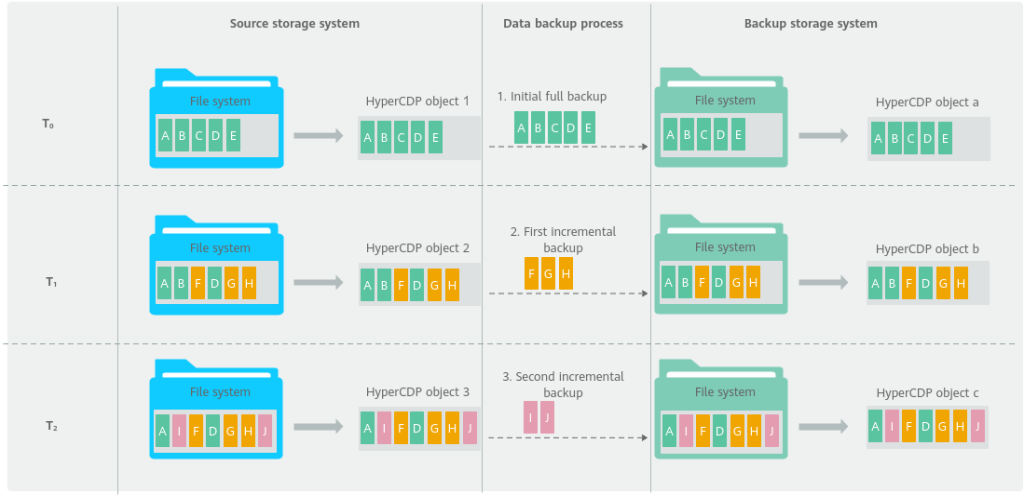
The file system HyperCDP and remote replication snapshot synchronization functions are used to implement file system data backup and recovery between OceanStor Dorado and the new-gen OceanStor hybrid flash storage. Users do not need to purchase additional commercial backup software, saving the backup solution cost. In addition, HyperVault provides basic local and remote backup functions without any need for deploying extra backup servers and media servers, saving investments and simplifying management.

Basic principles:

HyperVault backs up data within a storage system or between storage systems by file system. The backup copies are implemented based on HyperSnap. Hosts and applications are unaware of the backup process and copy generation process. On the backup storage system, each copy contains full service data of the source file system at a backup point-in-time and the copies are independent from one another. The deletion of any copy will not affect the usability of other copies. HyperCDP of the primary site provides various backup policies. You can configure a short interval for local backup to generate backup copies frequently on the local storage system, and configure a longer backup interval for remote backup to gradually migrate old data to the remote storage system, saving storage space and reducing workloads of the local storage system.

HyperSnap generates copies of file systems on the primary and backup storage systems. After a snapshot is generated on the primary storage system during a backup process, the differential data between this snapshot and the last synchronized snapshot is transmitted to the backup storage system. After the backup is complete, a snapshot is generated on the backup

storage system, which is called a copy. The snapshots generated in the backup storage system contain full data of the primary storage system at backup points in time and therefore are remote snapshots of the primary storage system.



HyperVault only backs up the changed data between two backup points in time. For example, during initial backup, HyperCDP object 1 on the file system of the primary storage is fully backed up to the backup storage. After the backup is complete, HyperCDP object a (copy) is generated on the file system of the backup storage.

In the next backup point in time, HyperCDP object 2 is generated for the file system of the primary storage. Between HyperCDP objects 2 and 1, blocks F, G, and H are changed. When HyperCDP object 2 is backed up, only changed data blocks (F, G, and H) are transferred. After the backup is complete, HyperCDP object b (copy) is generated on the file system of the backup storage. When HyperCDP object 1 is deleted from the primary storage, user hosts or applications can still access the full set of data at the backup point in time of HyperCDP objects a and b. That is, HyperCDP objects a and b on the backup storage system are the full service data.

In the next backup point in time, HyperCDP object 3 is generated for the file system of the primary storage. Between HyperCDP objects 3 and 2, blocks I and J are changed. When HyperCDP object 3 is backed up, only changed data blocks I and J are transferred. After the backup is complete, HyperCDP object c (copy) is generated on the backup storage. When HyperCDP object 2 is deleted from the primary storage, user hosts or applications can still access the full set of data at the backup point in time of HyperCDP objects a, b, and c. That is, HyperCDP objects a, b, and c on the backup storage system are the full service data.

HyperVault does not move or back up unchanged data on the source file system. A full backup is performed only for the first backup and incremental backups are performed subsequently, saving physical bandwidth and improving backup efficiency. In this way, backups can be performed more frequently with less data transmitted each time.

HyperVault provides the following functions:

1. Local backup: backs up data within the primary storage system.
2. Remote backup: backs up data from the primary storage system to the backup storage system.
3. Local recovery: recovers data within the primary storage system.

4. Remote recovery: recovers data from the backup storage system to the primary storage system.

6.6 HyperMetro (Active-Active Deployment)

HyperMetro, an array-level active-active technology provided by OceanStor Dorado, allows two LUNs from separate storage systems to maintain real-time data consistency and to be accessible to hosts.

HyperMetro supports both Fibre Channel and IP networking. The two storage systems in a HyperMetro deployment can be at two locations within 300 km from each other, such as in the same equipment room or in the same city. It is recommended that the quorum server should be deployed at a third site.

If one storage system fails, hosts automatically choose the paths to the other storage system for service access. If the replication links between the storage systems fail, only one storage system can be accessed by hosts, which is determined by the arbitration mechanism of HyperMetro.

6.6.1 HyperMetro for SAN

6.6.1.1 Read and Write Processes

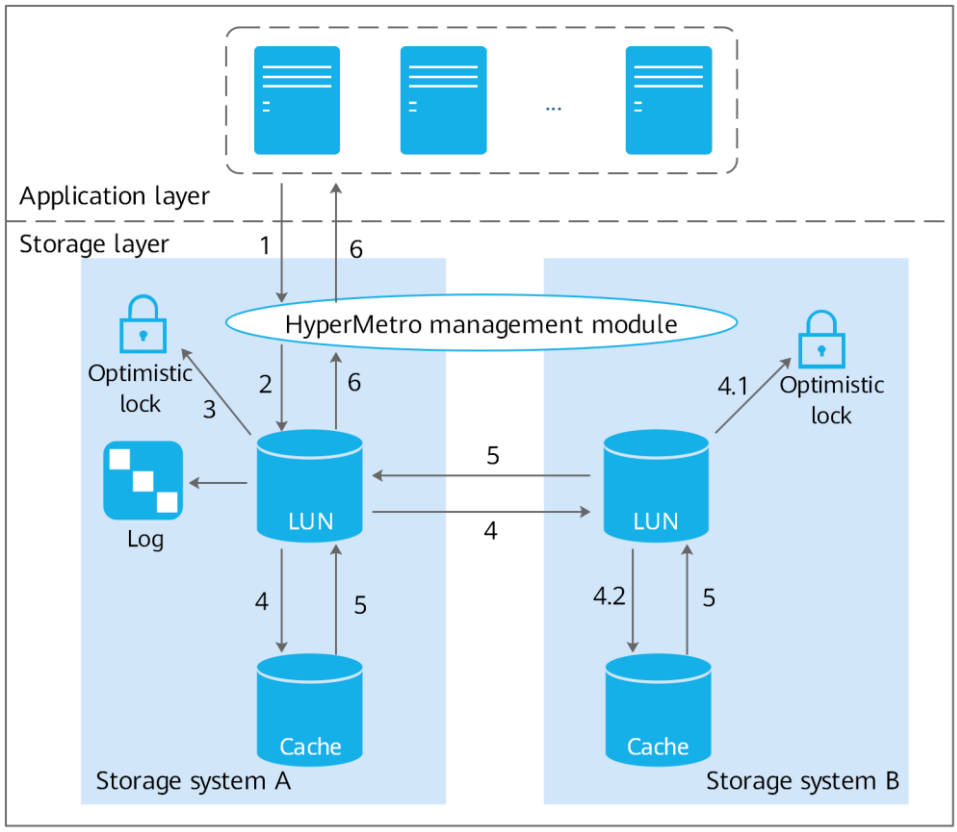
Read Process

When a LUN receives a read request, the storage system reads its local cache and returns the requested data to the application.

Write Process

When a LUN receives a write request, mutual exclusion is performed for parallel access paths. After the write permission is obtained, the requested data is written to the caches of both the local and remote LUNs of the HyperMetro pair. After the write operation is complete at both ends, a write success is returned to the application. Figure 6-14 illustrates the write process of HyperMetro.

Figure 6-14 SAN HyperMetro write process



1. The host delivers a write request.
2. Storage system A receives the request.
3. Storage system A uses the local optimistic lock to apply for write permission in the I/O address range corresponding to the HyperMetro pair. After the write permission is obtained, the HyperMetro pair records the request in a log, which records only the address information but no data content into memory space with power failure protection to achieve better reliability.
4. The system writes the request to the caches of both the local and remote LUNs separately. When receiving a write request, the remote LUN also needs to apply for the write permission in the I/O address range corresponding to the HyperMetro pair. Data can be written to the cache only after the remote LUN obtains the write permission.
5. The LUNs at both ends report the write results.
6. The HyperMetro pair returns a write success acknowledgement to the host.

Optimistic Lock

In the active-active storage systems, read and write operations are concurrently performed at both sites. If hosts deliver read and write requests to the same LBA of a LUN simultaneously, data at both sites must be consistent at the storage layer.

In the conventional solution, a cross-site distributed lock service is required. When a site receives a data write request, the site applies for a lock from the lock server. If the lock service is at the peer site, data can be written to the two sites only after a cross-site lock is obtained.

Because hosts in the upper-layer application cluster seldom send write requests to the same LBA concurrently, HyperMetro uses the optimized optimistic lock. If no lock conflict exists, HyperMetro directly initiates a write request to the peer storage system. After the data is written, a lock is added to the local storage system. When hosts concurrently access data at the same LBA, the storage system can also convert concurrent access requests into serial queuing requests to ensure data consistency between the two sites. In this solution, the cross-site lock server is not required, which reduces the architecture complexity. This solution also allows direct dual write operations when there is no conflict. It eliminates the interaction with the lock server (or even the remote lock server) and improves performance.

Cross-Site Bad Block Repair

Disks may have bad blocks due to abnormalities such as power failure. If repairable bad blocks fail to be repaired by the local end, HyperMetro automatically obtains data from the remote end to repair them, which further enhances the system reliability.

6.6.1.2 HyperMetro Consistency Group

HyperMetro provides and manages services by pair or consistency group.

A consistency group contains multiple HyperMetro pairs. It ensures data consistency between multiple associated LUNs on the storage systems.

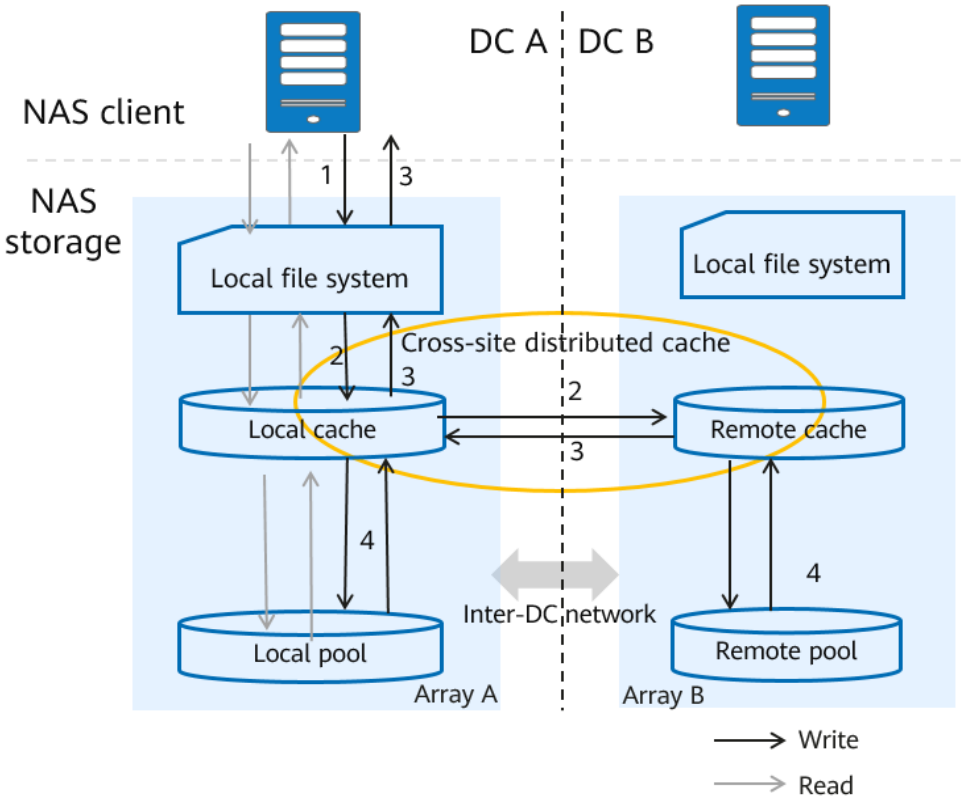
When you split or synchronize a consistency group, all HyperMetro pairs in the group are split or synchronized at the same time. If a link fault occurs, all member pairs are interrupted simultaneously. After the fault is rectified, data synchronization is implemented for all pairs to ensure data availability.

6.6.2 HyperMetro for NAS

Working Principles of HyperMetro I/Os

The OceanStor Dorado NAS HyperMetro solution uses cross-site distributed cache to implement active-active service access of a single file system. Figure 6-15 shows the I/O access process of NAS HyperMetro.

Figure 6-15 NAS HyperMetro I/O process



HyperMetro write:

Front-end host write process:

1. The client writes data to the file system.
2. The data is written to the local cache and mirrored to the local cache and remote cache.
3. All cache copies are successfully written, and a write success is returned to the host.

Background disk flushing process:

4. The local cache initiates cache flushing. The local cache data is written into the local pool, and the remote cache data is written into the remote pool.

HyperMetro read:

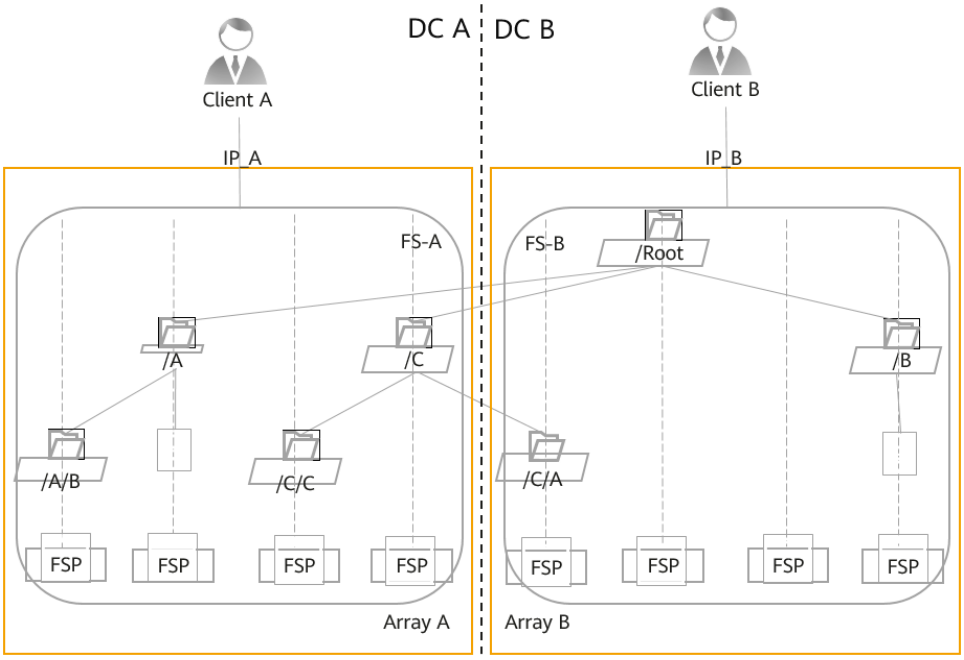
1. The client reads data from the file system.
2. The requested data is searched in the local cache. If the data is found in the local cache, it is returned to the host. If the data is not found in the local cache, the data is read from disks and returned to the host.

The front-end network of NAS HyperMetro supports Layer 2 or Layer 3 networks. Common LIFs are used for Layer 2 networks, and VIP LIFs for Layer 3 networks.

For NAS HyperMetro in active-active mode, a single file system can be mounted to two data centers using different IP addresses for shared access. In synchronous mode, a single file system can be mounted and accessed only in the primary DC. The multi-copy and dual-side disk flushing policies of the distributed cache ensure real-time data synchronization and

consistency across sites. Figure 6-16 shows the load balancing policy when two data centers access the same file system.

Figure 6-16 Load balancing of NAS HyperMetro file systems



As shown in Figure 6-16, NAS HyperMetro uses the local cache policy and consistency control policy to implement cross-site load balancing.

HyperMetro configuration process:

1. Create file system FS-A on array A and FS-B on array B.
2. Create a HyperMetro pair for FS-A and FS-B.
3. Mount FS-A to client A and FS-B to client B.

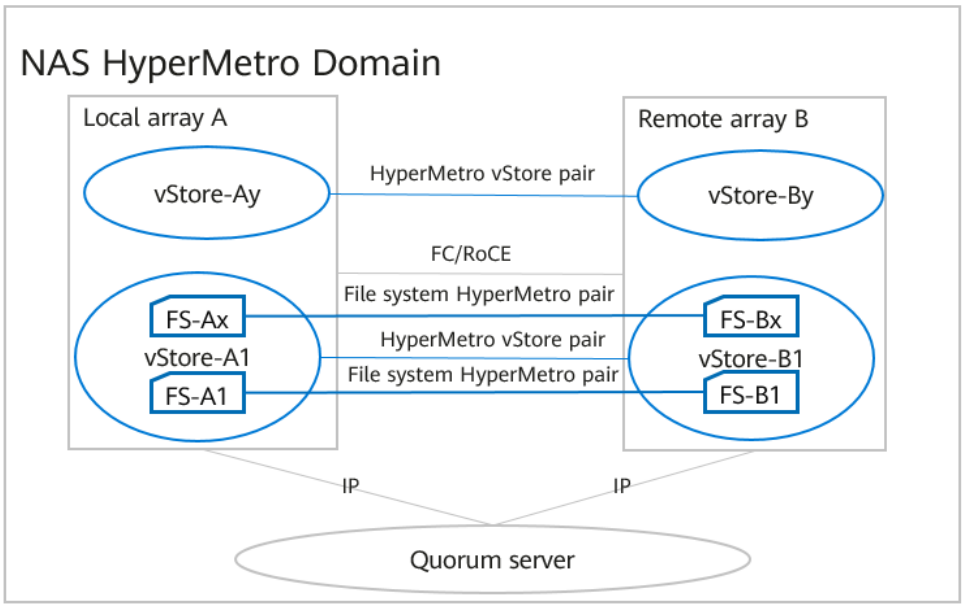
Cache balancing: site preference and intra-site balancing

1. Client A accesses array A through IP_A. The customer creates directories **/A**, **/C**, **/A/B**, and **/C/C**, which are evenly distributed in array A.
2. Client B accesses array B through IP_B. The customer creates directories **/B** and **/C/A**, which are evenly distributed in array B.
3. If client A wants to access a directory created by client B, the directory is hit on array B first based on the consistency requirement. If the directory is not hit, data is read from and written to the pool.

HyperMetro configuration model:

The NAS HyperMetro configuration objects include file system HyperMetro domains, HyperMetro vStore pairs, and HyperMetro file systems. The HyperMetro file systems are shared to provide active-active access for customers. Figure 6-17 shows the relationship between NAS HyperMetro objects.

Figure 6-17 NAS HyperMetro objects



File system HyperMetro domain: local storage array, remote storage array, network, and quorum server. Storage arrays can be interconnected through FC or RoCE networks (common TCP/IP networks are not recommended due to large network latency). Storage arrays communicate with the quorum server over common IP networks. The quorum server is required only in active-active mode and is not required in synchronous replication mode.

HyperMetro vStore pair: After a HyperMetro pair is created between two vStores, all configurations of one vStore are synchronized across the storage arrays in real time and the configuration status is controlled and managed based on the HyperMetro service logic. The configurations in a vStore include local user configuration, domain user configuration, shared service network configuration (LIF configuration), and protocol policy configuration. In active-active mode, these configurations are automatically synchronized. In synchronous replication mode, you can select whether to synchronize the configurations.

File system HyperMetro pair: You are advised to configure HyperMetro for all file systems in a HyperMetro vStore pair to implement real-time dual-write synchronization of file system access permissions and data across storage arrays. A single file system can be accessed in active-active mode on the local and remote storage arrays simultaneously.

Failover: If a storage array is faulty or the link between storage arrays is down, NAS HyperMetro initiates arbitration in the HyperMetro domain. The storage array that wins the arbitration continues providing services and takes over the services from the other storage array. After the fault is rectified, the system starts data synchronization and service recovery.

6.6.3 HyperMetro Technical Features

6.6.3.1 Gateway-free Active-Active Solution

The HyperMetro deployment groups two storage systems into a cross-site cluster without any additional virtual gateway.

This solution has a simplified architecture and is well compatible with other value-added features. It delivers the following values to customers:

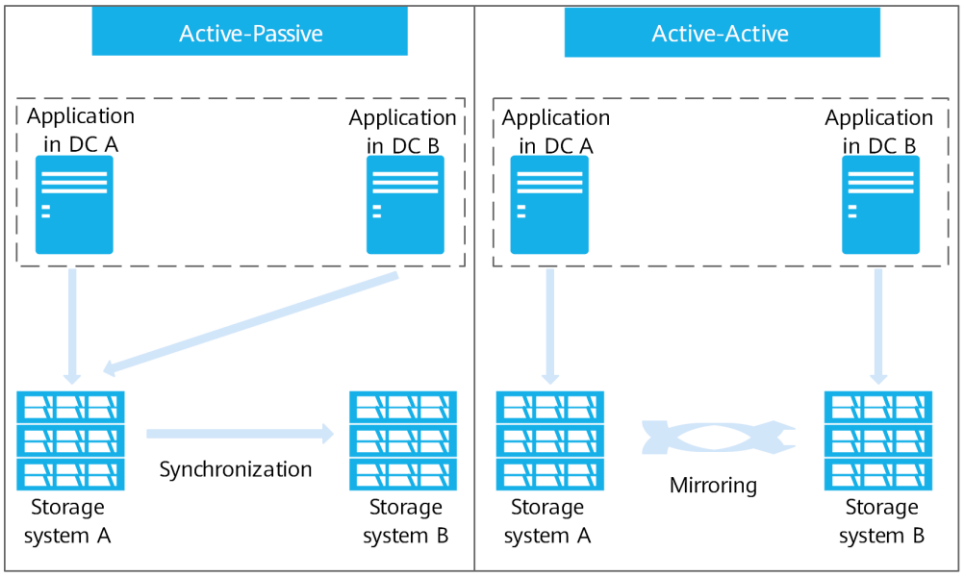
- Reduced gateway-related fault points and enhanced solution reliability
- Quicker I/O response (Latency caused by gateway forwarding is eliminated because I/Os are not forwarded by gateways.)
- Superb compatibility with existing storage features. HyperMetro can work with other Smart- and Hyper-series features on OceanStor Dorado to deliver a wide range of data protection and DR solutions.
- Simplified network and easier maintenance

6.6.3.2 Parallel Access

HyperMetro delivers active-active service capabilities on two storage systems. Data is synchronized in real time between the HyperMetro service objects on both storage systems, and both storage systems process read and write I/Os from application servers to provide non-differentiated parallel active-active access. If either storage system fails, services are seamlessly switched to the other system without interrupting service access.

In comparison with the active-passive mode, the active-active solution fully utilizes computing resources, reduces communication between storage systems, and shortens I/O paths, thereby ensuring better access performance and faster failover. Figure 6-18 compares the active-passive and active-active solutions.

Figure 6-18 Active-passive and active-active storage architectures



6.6.3.3 Reliable Arbitration

If links between two HyperMetro storage systems are disconnected, real-time mirroring will be unavailable to the storage systems and only one system can continue providing services. To ensure data consistency, HyperMetro uses the arbitration mechanism to determine which storage system will continue providing services.

HyperMetro supports arbitration by pair or consistency group. If services running on multiple pairs are mutually dependent, you can add the pairs into a consistency group. After arbitration, only one storage system provides services.

Arbitration Modes

HyperMetro provides the following arbitration modes:

- Static priority mode

  The static priority mode is used when no quorum server is deployed. You can set either side of a HyperMetro pair or consistency group as the preferred site and the other side the non-preferred site. If heartbeats between the storage systems are interrupted, the preferred site wins the arbitration.

- Quorum server mode

  In quorum server mode, when heartbeats between the storage systems are lost, each of them sends an arbitration request to the quorum server, and only the winner continues providing services. You can set one site as the preferred site, which takes precedence in arbitration.

Automatic Switch of Arbitration Modes

If all quorum servers fail or their links to storage systems fail but the heartbeats between the storage systems are normal, the system automatically switches to the static priority mode.

6.6.3.4 Strong Scalability

HyperMetro can work with other Smart- and Hyper-series features on OceanStor Dorado to provide various data protection and DR solutions.

- Online expansion to active-active storage systems

  HyperMetro is paired with Huawei UltraPath multipathing software, which supports LUN aggregation and shields physical differences at the storage layer. When users want to expand a single storage system to active-active storage systems, UltraPath can seamlessly take over the new storage system and HyperMetro member LUNs for online expansion.

- Compatibility with existing features

  HyperMetro can be used together with existing features such as HyperReplication, HyperSnap, HyperClone, SmartThin, SmartDedupe, and SmartCompression.

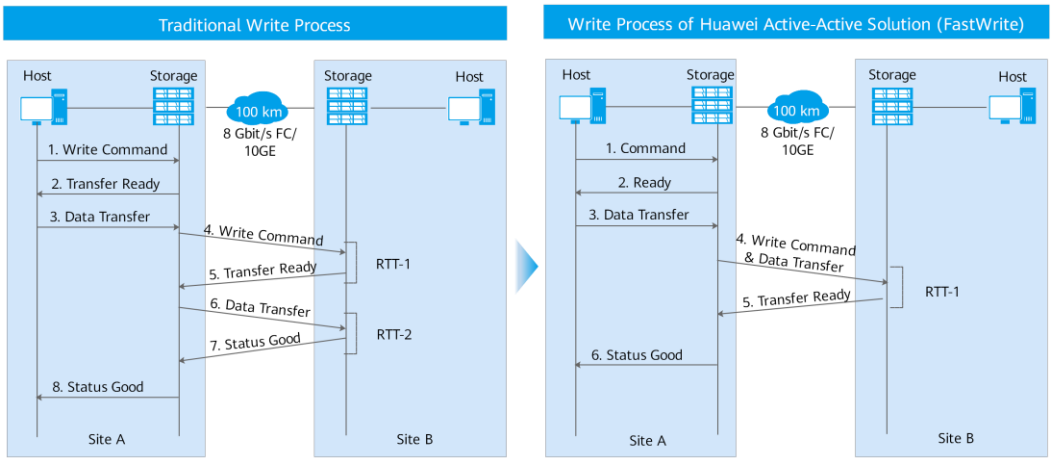
6.6.3.5 High Performance

A series of optimization designs are used to enhance HyperMetro performance.

FastWrite

HyperMetro uses FastWrite to optimize data transmission over the replication links between storage systems. With SCSI's First Burst Enabled function, the data transmission interactions involved in a data write process are reduced by half. In a standard SCSI process, transmission of a write I/O undergoes multiple interactions between two ends, such as write command delivery, write allocation completion, data write, and write execution status return. FastWrite optimizes the write I/O interaction process by combining command delivery and data transfer and canceling write completion acknowledgement. This reduces the interactions involved in writing data across sites by half.

Figure 6-19 Transmission protocol optimization



Optimized Cross-Site Access

For active-active services, the distance between two sites is essential to I/O access performance. SAN HyperMetro works with OceanStor UltraPath and NAS HyperMetro works with DNS to provide two I/O access policies based on the distance between the active-active sites.

- Load balancing mode

  This mode is mainly used when HyperMetro storage systems are deployed in the same data center. In these scenarios, both storage systems deliver almost the same access performance to a host. To maximize resource usage, host I/Os are delivered in slices to both storage systems.

- Preferred storage system mode

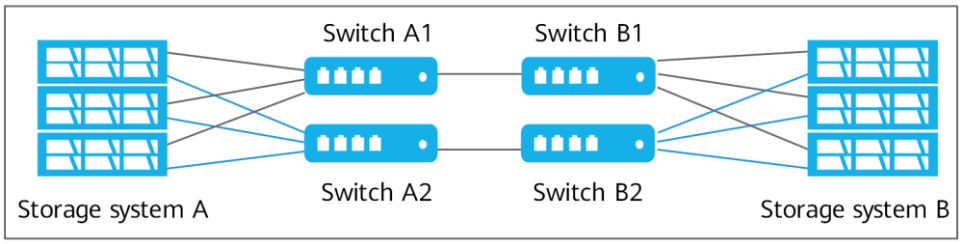
  This mode is mainly used when the HyperMetro storage systems are deployed in different data centers over distance. In these scenarios, cross-DC access will increase the latency. If the link distance between the DCs is 100 km, the round-trip time (RTT) is approximately 1 ms. Reducing cross-DC communication improves I/O performance.

Optimal Path

You are advised to use a fully interconnected network for Fibre Channel and IP replication links between OceanStor Dorado storage systems. That is, each controller pair of the two storage systems has a direct logical link with another controller pair.

For read and write I/O requests sent between storage systems, a storage system searches for the optimal path to the peer storage system based on the balancing policy of the service object, and preferentially sends I/O requests over a link that directly connects to the owning node on the peer storage system. This reduces the cross-controller forwarding latency within storage nodes and improves end-to-end performance.

Figure 6-20 Fully interconnected network between storage systems



6.7 Geo-Redundancy (Multi-DC)

This section describes the multi-DC networks supported by the OceanStor Dorado series and their features.

6.7.1 3DC (Geo-Redundancy)

The 3DC solution involves a production center, an intra-city DR center, and a remote DR center. Data of the production center is synchronously replicated to the intra-city DR center, and is asynchronously replicated to the remote DR center. The intra-city DR center has the same service processing capability as the production center. Applications can be switched to the intra-city DR center without any data loss, achieving zero RPO and second-level RTO. If a low-probability large-scale disaster, such as an earthquake, occurs and causes both the production center and intra-city DR center to be unavailable, applications can be switched to the remote DR center. Based on routine DR drills, applications can be recovered in the remote DR center within the tolerable time to ensure service continuity and RPO within seconds.

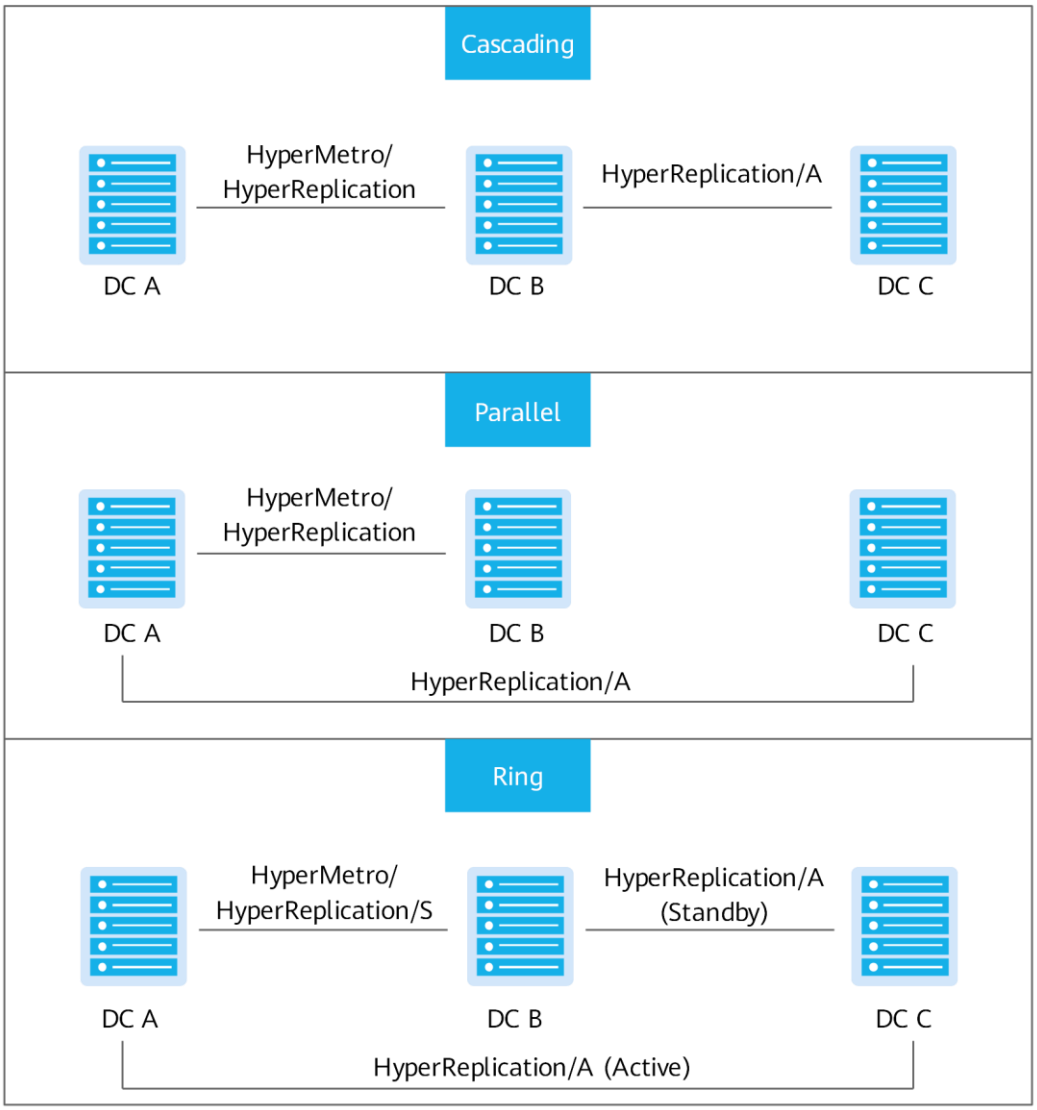
Compared with the solutions where either only an intra-city DR center or a remote DR center is deployed, the geo-redundant 3DC DR solution can cope with larger-scale disasters by combining their advantages. In this way, the DR system can efficiently respond to both small-scale regional disasters and large-scale natural disasters to prevent loss of service data as far as possible and provide superior RPO and RTO. Therefore, the geo-redundant 3DC solution has been widely used.

Flexible networks of 3DC are supported using HyperMetro, HyperReplication/S (synchronous remote replication), and HyperReplication/A (asynchronous remote replication), including: (For the roadmap of NAS 3DC networking topologies, see the product roadmap.)

- Cascading network topology with HyperMetro and HyperReplication/A
- Parallel network topology with HyperMetro and HyperReplication/A
- Ring network topology (DR star) with HyperMetro and HyperReplication/A
- Cascading network topology with HyperReplication/S and HyperReplication/A
- Parallel network topology with HyperReplication/S and HyperReplication/A
- Ring network topology (DR Star) with HyperReplication/S and HyperReplication/A
- Cascading network topology with only HyperReplication/A
- Parallel network topology with only HyperReplication/A

Figure 6-21 shows the cascading, parallel, and ring networking topologies.

Figure 6-21 3DC network topologies



The 3DC solution features cost effectiveness, elastic scalability, robust reliability, high efficiency, and enhanced security.

- Interoperability among entry-level, mid-range, and high-end storage systems

  Huawei storage systems of different performance levels can be used in one DR solution via remote replication. Customers can choose proper storage systems based on project conditions, reducing more than 50% of investment in DR construction. However, the active-active storage systems configuring HyperMetro must use devices of the same model.
- Elastic scalability

  The active-active DC solution can be online upgraded to a geo-redundant 3DC solution for higher reliability.
- Diverse networking modes

  Multiple networking modes, such as cascading and parallel topologies (with HyperReplication/A and HyperMetro, HyperReplication/A and HyperReplication/S, and

HyperReplication/A and HyperReplication/A) and ring topologies (with HyperReplication/A and HyperMetro, and HyperReplication/A and HyperReplication/S), are supported to build a DR system that is most suitable for the production environment and improve the cost-effectiveness of the DR system.

- Consistency protection

  The DR system provides application-based data consistency protection, strengthening application reliability.

- Visualization

  DR can be displayed in a topology and end-to-end real-time monitoring is supported, greatly simplifying maintenance.

- One-click operations

  The solution supports one-click testing, switchover, and recovery, and automated scripts of service clusters are used to replace manual operations, remarkably improving the work efficiency and DR success rate of the DR system.

- Security

  User authentication, encrypted data transfer, and service isolation between storage systems ensure security of the entire DR system.

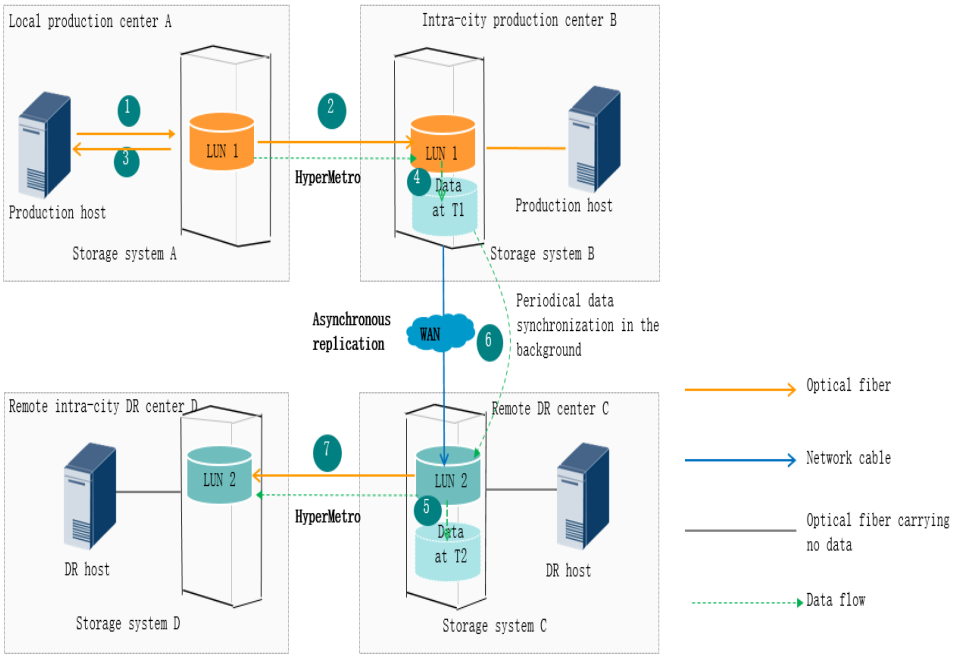
6.7.2 4DC (Geo-Redundancy)

This section describes Huawei SAN 4DC networking and features.

The Huawei geo-redundant 4DC DR solution includes one local production center, one intra-city production center, one remote DR center, and one remote intra-city DR center. The two active-active storage systems in the production centers provide the SAN active-active function. Data in the production centers is periodically and asynchronously replicated to the remote DR center. The remote DR center mirrors data to the remote intra-city DR center in real time. The intra-city production center has the same service processing capability as the local production center. Applications can be switched to the intra-city production center without any data loss to ensure service continuity. In case of low-probability large-scale disasters, such as earthquakes, which cause both the intra-city DR center and production center to be unavailable, applications can be failed over to the remote DR center and remote intra-city DR center to ensure service continuity and high protection level. Based on routine DR drills, applications can be recovered in the remote DR center and remote intra-city DR center within the tolerable time limit to ensure service continuity. However, a small amount of data may be lost during remote recovery. If a disaster occurs in a remote DC, the remaining remote DC can still provide services.

Compared with the solutions that include only one intra-city DR center or include one intra-city DR center and one remote DR center, the geo-redundant 4DC DR solution combines their advantages to continue providing services even if three DCs in two cities fail and to address disasters that affect broader areas. If a disaster that affects a small area or a natural disaster that affects a large area occurs, the solution responds quickly to prevent data loss to the maximum extent and achieve smaller recovery point objective (RPO) and recovery time objective (RTO).

OceanStor supports the 4DC solution consisting of HyperMetro, HyperReplication/A, and HyperMetro. The networking is as follows:



The process of handling write I/O requests for the HyperMetro + HyperReplication/A + HyperMetro solution is as follows:

1. The production host delivers a write request to the HyperMetro LUN.
2. The LUN writes data to the HyperMetro data LUNs in both local and intra-city production centers. A write success message is returned to the host.
3. When the synchronization period of the asynchronous replication from the intra-city production center to the remote DR center starts, storage system B generates a snapshot of LUN 1 at the point in time (such as T1) and notifies storage system C in the remote DR center of generating a snapshot of LUN 2 at the point in time (such as T2). In the background, the differential data between storage system B's LUN 1 at T1 and storage system C's LUN 2 at T2 is periodically synchronized. Data on LUN 2 of storage system C in the remote DR center is mirrored to LUN 2 of storage system D in the remote intra-city DR center in real time.
4. If data synchronization fails for the asynchronous replication and data on LUN2 in storage system C in the remote DR center must be used to run services, set LUN2 on storage system C to the writable state (this requires that the HyperMetro relationship between remote DCs C and D be suspended and storage system C take over services). After LUN2 on storage system C is set to the writable state, the system starts a task in the background to roll back data to T2 to ensure data availability on storage system C. After the rollback task is complete, start data synchronization for the HyperMetro pair between storage systems C and D. After the data synchronization is complete, the HyperMetro pair works in active-active mode. (If the storage system in a DC is faulty, the remaining DC can take over services. Before the standby HyperMetro pair between storage systems C and D works in active-active mode, storage system D cannot provide services independently.)

The 4DC DR solution features cost effectiveness, elastic scalability, robust reliability, and enhanced security.

- Multi-DC and multi-copy

  Multiple DCs are built and four copies are created to provide multi-region multi-copy DR protection, further improving the data protection level.

- Elastic scalability

  An active-active or geo-redundant 3DC DR architecture can be smoothly upgraded to the geo-redundant 4DC DR architecture without interrupting production services.

- Multi-level protection

  The geo-redundant 4DC DR architecture ensures that remote applications can still run in HA mode for a long time if both DCs in a city are faulty. The local production center and remote DR center can back up each other, achieving two service centers in two cities, service continuity, and high reliability.

- Consistency protection

  The DR system provides application-based data consistency protection, strengthening application reliability.

- Security

  User authentication, encrypted data transfer, and service isolation between storage systems ensure security of the entire DR system.

6.8 Storage-Optical Connection Coordination (Hyperlink)

Due to fiber quality deterioration caused by the environment and urban infrastructure construction, Data Center Interconnect (DCI) links are unstable with frequent jitters, intermittent disconnections, and bit errors. As a result, core system service failure and performance deterioration occur in the network sub-health scenarios. Transaction failures and frame freezes caused by DCI network sub-health have become a global common issue to be resolved in the financial industry.

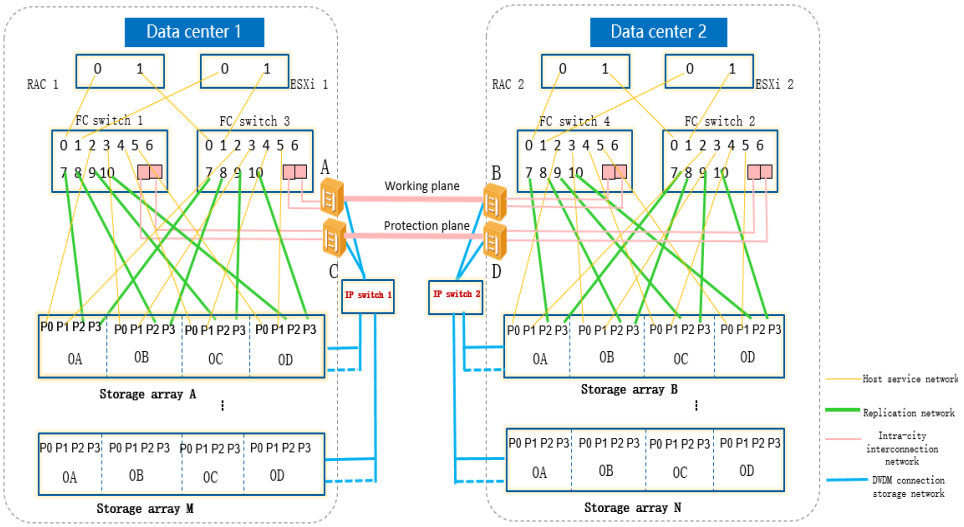
To meet the customer's requirement for high availability for core services in DCI network sub-health scenarios, Huawei launched the Storage-Optical Connection Coordination (SOCC) solution, which is unique in the industry. This solution leverages the advantages of the product portfolio of storage and optical transmission to prevent transaction failures caused by network sub-health and ensure stable and reliable transactions in the financial industry.

To reduce the DCI network sub-health impact on financial transactions, the Huawei SOCC solution enables direct handshakes through the SOCC channel between the industry-leading financial DCI optical transmission device (Huawei OptiXtrans DC908 or E96 series) and the storage system. In this solution, the optical network is used to detect link faults. When the link jitter exceeds the threshold, OptiXtrans notifies the storage system to switch I/O links within 5 seconds (down from the previous minute-level switchover time), greatly slashing the exception duration of financial transactions and ensuring zero financial transaction failure.

Parallel networking is supported.

Working principles of parallel networking:

Each switch of the storage replication links connects to only one DWDM device in the local data center.



Solution highlights:

- High service reliability

  When network jitter occurs, the coordination between the storage and DWDM devices implements a fast I/O switchover on sub-healthy links, improving efficiency by over 10 times compared with the industry average and ensuring high service reliability.

- Efficient O&M

  Integrated storage and DWDM management planes and visualized management enable O&M personnel to efficiently operate and quickly locate faults.

- Elastic deployment

  The storage and DWDM devices can be smoothly upgraded to the SOCC solution without interrupting production services.

6.9 HyperEncryption (Array Encryption)

OceanStor Dorado supports data encryption. You can purchase SEDs to encrypt data. If no SED is available, the storage system implements array encryption to ensure data security.

Internal Key Manager

The internal key manager is the storage system's built-in key management system. It generates, updates, backs up, restores, and destroys keys, and provides hierarchical key protection. The internal key manager is easy to deploy, configure, and manage. You are advised to use the internal key manager if security certification by the cryptographic module is not required and the key management system is only used by the storage systems in a data center.

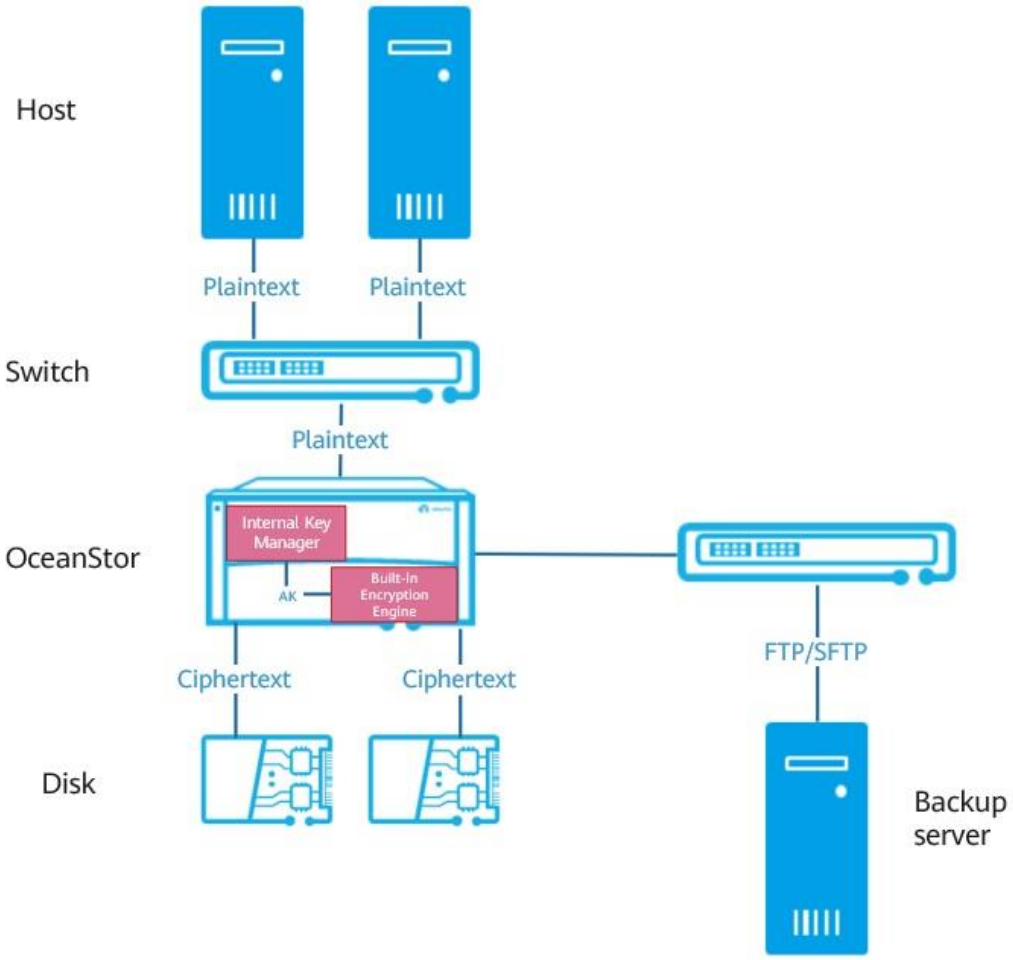
External Key Manager

The standard KMIP+TLS protocol is used to support interconnection with the external key manager. The external key manager supports key generation, update, destruction, backup, and restoration. Two external key managers can be deployed, which synchronize keys in real time for enhanced reliability.

Array Encryption Principles

Array encryption of the storage system uses the built-in encryption engine of the controller processor to implement encryption and decryption. The independent built-in encryption engine leverages the encryption algorithm of Arm hardware to offload encryption workloads. The encryption and decryption algorithms of the storage system are offloaded to the hardware for execution, without involving software. During data encryption on OceanStor Dorado storage systems, the block device management subsystem generates a data encryption key (DEK) on each disk, and the key manager provides an authentication key (AK). The AK is used to encrypt the DEK. After service I/Os are delivered, encryption and decryption are offloaded to the built-in encryption engine for execution. The encryption engine supports the AES-256-XTS and SM4-128-XTS (only for the Chinese mainland) algorithms. The algorithm used by the key manager must match that used by the encryption engine. The following figure shows the topology with an internal key manager (as an example):

Figure 6-22 Built-in encryption engine



- Data encryption: After the built-in encryption engine is enabled, the block device management subsystem uses the built-in encryption engine to encrypt and decrypt data with the DEK during data writes and reads.

  When the storage system receives a write request, the built-in encryption engine encrypts the plaintext data, and the block device management subsystem then writes the encrypted data into storage media.

When the storage system receives a read request, the block device management subsystem reads the encrypted data, which is then decrypted by the built-in encryption engine into plaintext.

- Data destruction: When the external key manager is used, data can be destroyed by destroying corresponding keys.
- AK update: AKs must be regularly updated to avoid cracking or leakage. The system encrypts and stores the DEKs again based on the new AK2 delivered by the key manager. The updates can be performed periodically (every year) or manually.

6.10 HyperDetect (Ransomware Detection)

There are two types of ransomware attacks:

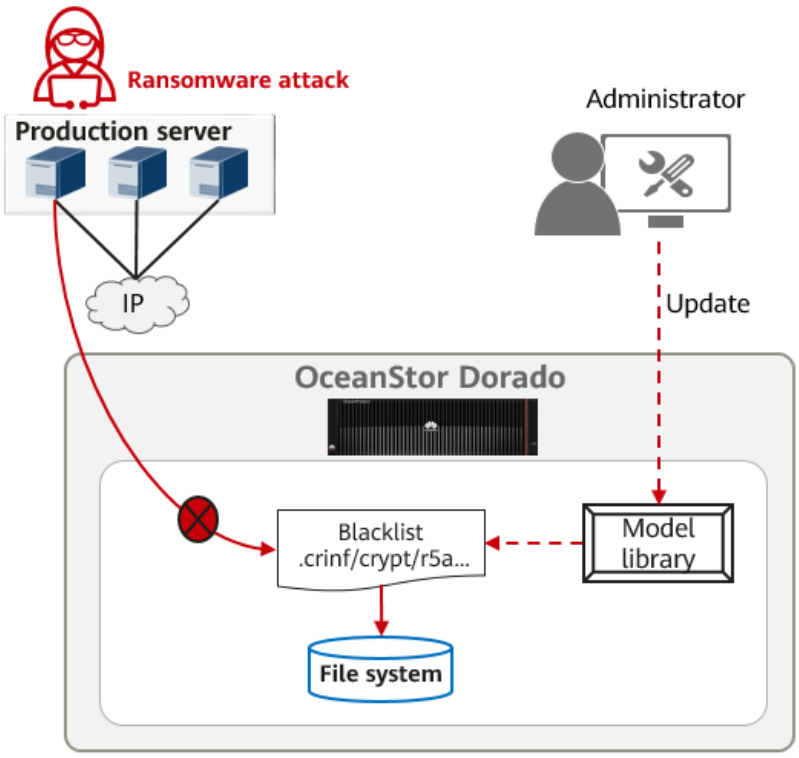
1. Strong encryption algorithms, such as AES and RSA, are used to encrypt user data. Users must pay the ransom to obtain the keys to restore and access data. If the ransom is not paid within a specified period, the file data will be lost permanently.
2. Sensitive and important user data is stolen. Attackers threaten to disclose the data to the public unless the user pays the ransom.

HyperDetect handles the first type of ransomware attack. It provides ransomware file interception, real-time ransomware detection, and intelligent ransomware detection to protect data from ransomware threats. The following table describes the basic concepts.

Concept	Description
File name extension filtering rule	Each file name extension filtering rule corresponds to a file name extension (that is, a file suffix) and is used to intercept operations of files with the same extension.
I/O behavior	The creation, read, write, deletion, and renaming of files by users.
Ransomware file interception	When launching attacks, ransomware usually generates encrypted files with special file name extensions. In light of this, the system intercepts the write of files with the specific file name extensions to block the extortion from known ransomware and protect file systems in the storage system.
Real-time ransomware detection	Ransomware attacks have similar I/O behavior characteristics. By analyzing file I/O behavior characteristics, the system quickly filters out abnormal files and performs deep content analysis on the abnormal files to detect files attacked by ransomware. Then, secure snapshots are created for file systems where files have been attacked, and alarms are reported to notify the data protection administrator, limiting the impact of ransomware and reducing losses.
Intelligent ransomware detection	The system detects known ransomware features to identify whether the file systems are attacked by ransomware. If no ransomware attack is identified, the system analyzes and compares the changes in file system snapshots, and uses machine learning algorithms to further check whether the file systems are infected by ransomware.

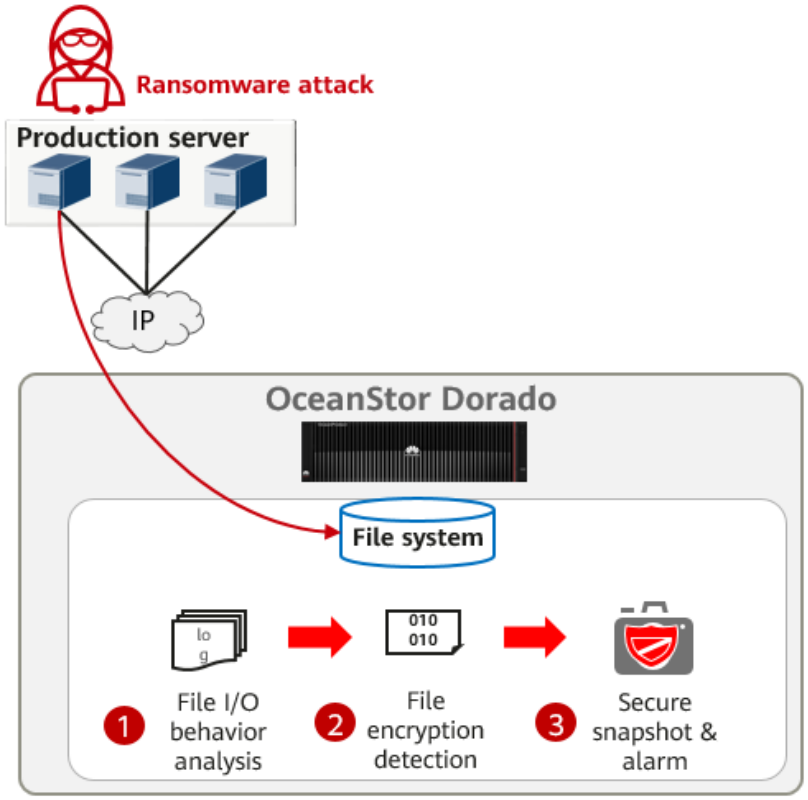
Ransomware File Interception

1. The system intercepts the writes of files infected by known ransomware based on the FileBlock function.
2. The model library is upgraded independently to update the interception rules.
3. Users or administrators can add or delete blacklist items and update interception rules on the management UI at any time.



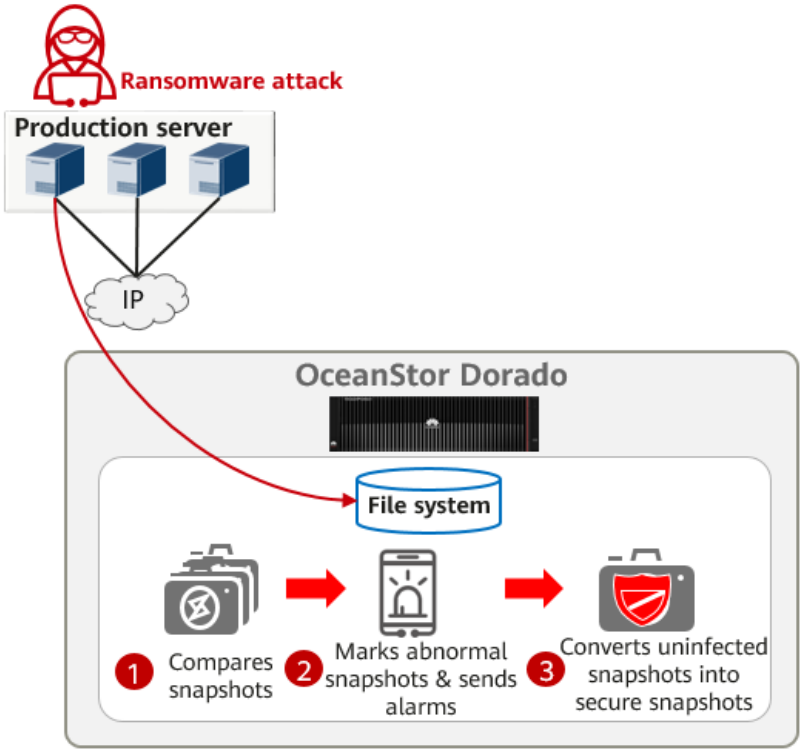
Real-time Ransomware Detection

1. Analyzes file I/O behaviors (such as read, write, renaming, and deletion) based on file operation logs to quickly identify abnormal files.
2. Performs encryption detection on the abnormal files to detect files infected by ransomware.
3. If a file encrypted by ransomware is detected, the system creates a secure snapshot for the file system immediately to minimize the impact of ransomware. In addition, an alarm is sent to the data protection administrator for confirmation and data recovery.



Intelligent Ransomware Detection

1. The system creates a common snapshot of a file system and compares it with last uninfected snapshot to analyze the changes to the file system snapshot. Then the system uses machine learning algorithms to determine whether the file system is infected by ransomware.
2. If the file system is infected by ransomware, the system marks the snapshot as abnormal and sends an alarm to the data protection administrator for data recovery.
3. If the file system is not infected, the system converts the common snapshot into a secure snapshot for the next comparison or data recovery.



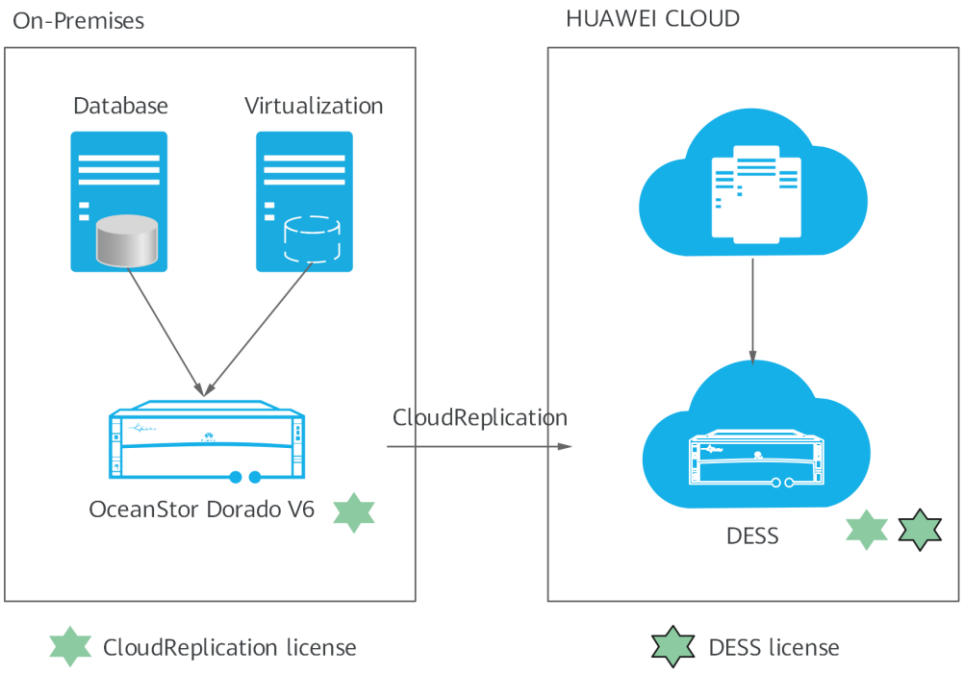
7 Cloud Series Features

In addition to on-premises DR, OceanStor Dorado provides cloud-based DR capabilities using the CloudReplication and CloudBackup features.

7.1 CloudReplication (Cloud Replication)

OceanStor Dorado supports CloudReplication, which works with Dedicated Enterprise Storage Service (DESS) on HUAWEI CLOUD to constitute cloud DR solutions. You can purchase HUAWEI CLOUD resources on demand to build your DR centers without the need for on-premises equipment rooms or O&M teams, reducing costs and improving efficiency.

Figure 7-1 CloudReplication architecture



Technical highlights:

- Data is replicated to the cloud in asynchronous mode. CloudReplication inherits all functions of HyperReplication/A.
- With Huawei DESS solution, no on-premises DR center or O&M team is required. Cloud DR resources can be purchased or expanded on demand.

Application scenarios:

- If you only have a production center, you can set up a remote DR center on HUAWEI CLOUD at a low cost, implementing remote protection for production data.
- If you have a production center and a DR center, you can upgrade the protection level to 3DC with a remote DR center on HUAWEI CLOUD.

NOTE

For the delivery plan of CloudReplication, see the product roadmap.

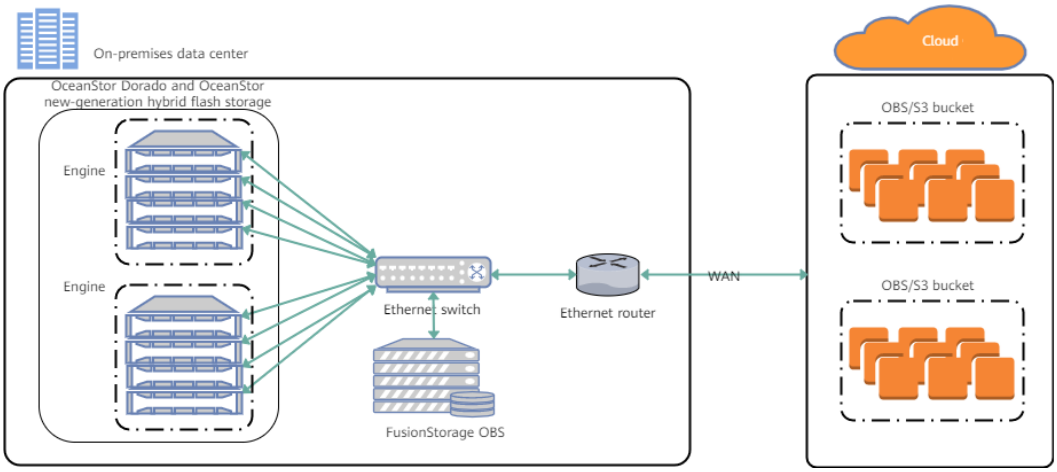
7.2 CloudBackup (Cloud Backup)

CloudBackup is a Huawei solution based on its storage systems for data backup to the cloud with no need for additional backup servers. CloudBackup can back up NAS data to the object storage over all-IP networks. There are two application scenarios based on the location of the backup storage:

1. Data backup to the on-premises DC. The backup storage is deployed in the same DC as OceanStor Dorado, providing better network performance for high-speed backup and recovery. The backup storage supported by CloudBackup in the on-premises DC includes FusionStorage OBS.

2. Data backup to the object storage on the remote public cloud. The object storage on the public cloud is used as the backup storage. Customers do not need to purchase backup storage devices, reducing the procurement and maintenance costs. The public cloud storage supported by CloudBackup.

Figure 7-2 Typical CloudBackup networking

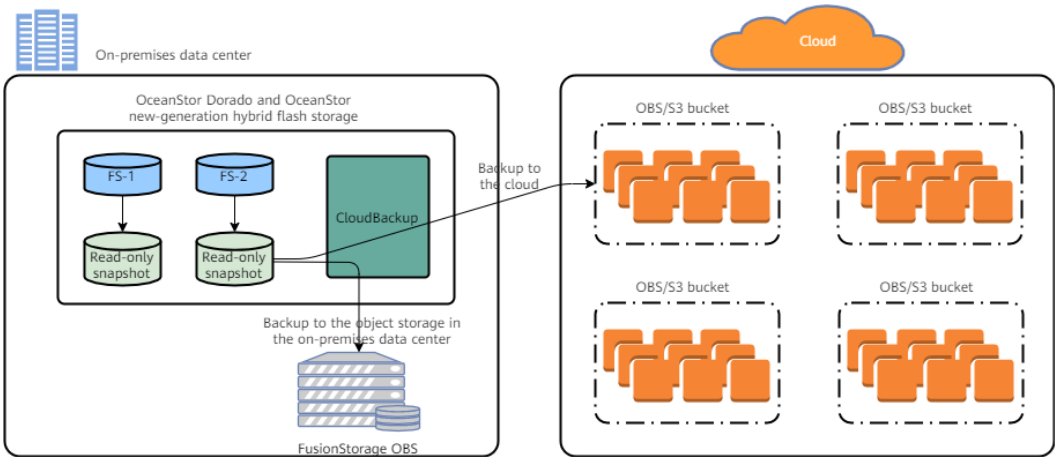


Key features and functions:

1. Simple, efficient, and secure: CloudBackup is deployed inside the storage system and backs up data to the cloud without external backup devices. It supports file-level incremental backup, which sends only changed files to the cloud. It also supports full recovery using a complete copy or fine-grained recovery of specific files. The transmission channels use HTTPS for encryption and security.
2. NAS backup of file systems: CloudBackup scans for storage file systems and displays them on the ProtectManager resource management page. You can add protection measures for a file system to back up its data. The system uses file system snapshots to back up data to the cloud through NAS sharing.
3. Periodic incremental backup: After the first full backup, incremental backup can be performed periodically, which is faster and requires less bandwidth and space than full backup.
4. Periodic full backup: You can set periodic full backup in the backup policy. For example, if the interval for periodic full backup is set to 10 backups, a full backup will be performed at the eleventh backup. This avoids long dependency paths for incremental backups.
5. Data restoration on the cloud: In scenarios such as disaster recovery, service migration, data development, testing, and utilization, you can restore data copies that have been backed up to the object storage to new storage devices, ensuring data availability and maximizing data value.
6. QoS control: CloudBackup is deployed in containers, so the impact of backup and recovery services on host performance is controllable. Network proxy services to the cloud are supported, meeting the requirements of various network scenarios. Network optimization and acceleration are supported to improve network transmission performance. Rate limiting policies can be configured to limit the network bandwidth occupied by backup services.

The data backup flow and principles are shown in the following figure:

Figure 7-3 Data backup flow



Step 1 A read-only snapshot is created for the file system to be backed up.

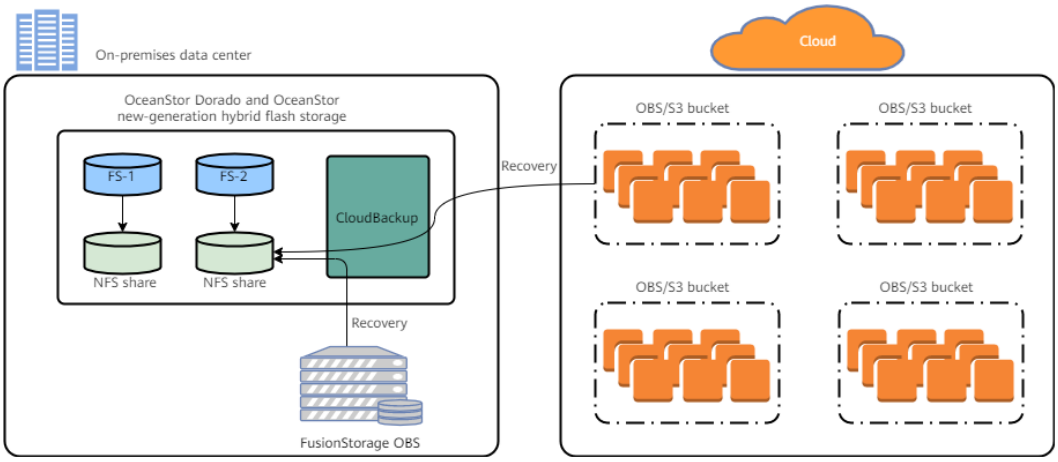
Step 2 The system scans the snapshot to obtain the differential data to be backed up. The first backup is a full backup of the entire file system.

Step 3 Based on the data obtained in step 2, the system creates and mounts a share for the snapshot, reads the data and metadata to be backed up, and transfers the data and metadata to the object storage in the DC or the public cloud.

----End

The data recovery flow and principles are shown in the following figure:

Figure 7-4 Data recovery flow



Step 1 Select the local file system you want to recover and query the backup copies.

Step 2 Select the backup copy you want. A backup copy is a set of data generated by backing up a file system. A file system may have multiple backup copies at different points in time. The backup copy can be in the remote public cloud or in the object storage of the on-premises DC.

Step 3 Determine whether to recover the entire file system using the selected copy or recover specific files.

Step 4 Determine whether to recover data to the original file system or a new file system.

Step 5 The system performs the recovery. During recovery, CloudBackup reads data from the specified backup copy in the object storage and writes the data to the share of the target file system.

CAUTION

During recovery, ensure that the target file system is not being read or written by any host service or other value-added features.

----End

7.3 CloudTier (SmartMobility)

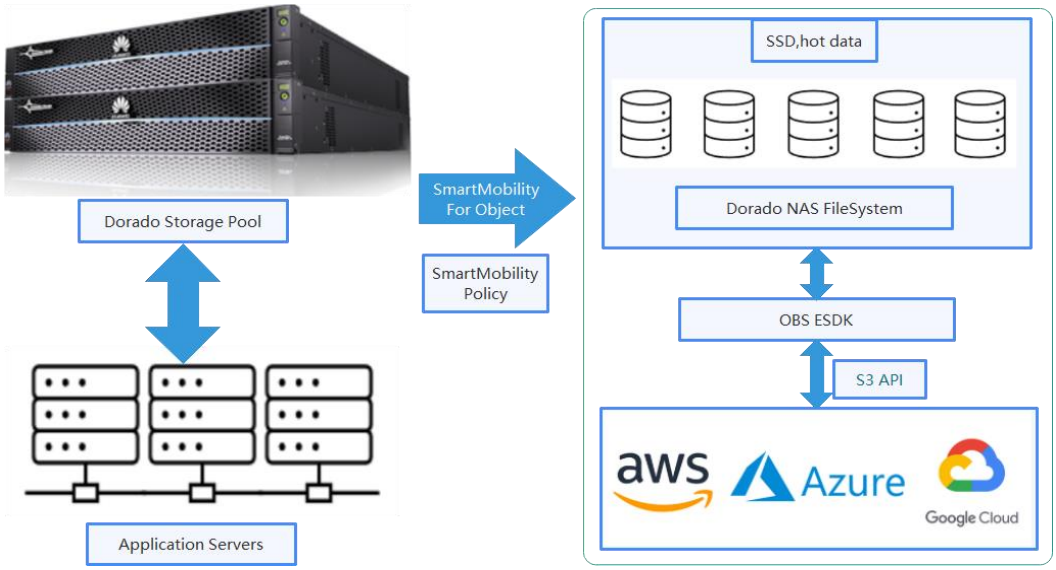
SmartMobility is to automatically migrate cold files in a file system to a remote device (object storage,cloud or NAS device) based on a specified policy.

SmartMobility is mainly used to migrate file data that is not frequently accessed to a remote device, so the available capacity of the local device is increased. Files that have been migrated to the remote device can be read and written in real time and recalled in the background, which achieves transparent access to users and also balances space and performance.

SmartMobility applies to two scenarios:

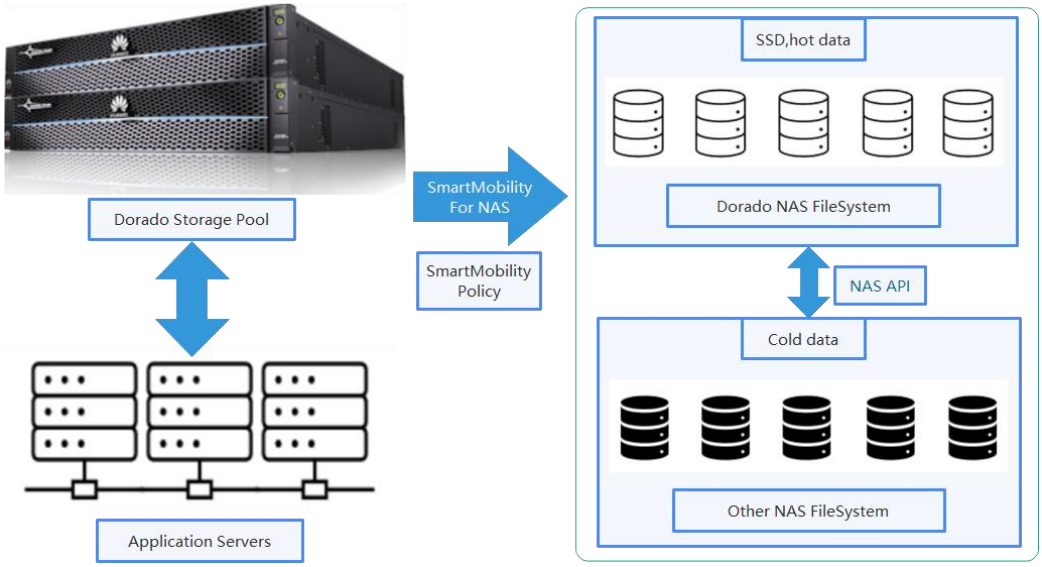
- Tiered file migration to the cloud: SmartMobility connects to cloud S3 object storage (including AWS S3, Google GCS, and Microsoft Azure) over the Internet and uses it as the cold data layer. Customers can configure policies to define what files are migrated to the cloud. Files can also be easily recalled from the cloud to local storage when the data is accessed again. This solution takes full advantage of the large and cost-effective object storage space and migrates seldom accessed data to release high-performance local storage space without adding too much overhead. Therefore, this solution helps enterprises reduce primary storage costs.

Figure 7-5 Tiered file migration to the cloud



- Tiered file migration to NAS devices: This solution is to reuse existing resources. Cold data is migrated from high-performance storage devices to less expensive NAS devices, reusing the legacy NAS device resources.

Figure 7-6 Tiered file migration to NAS devices



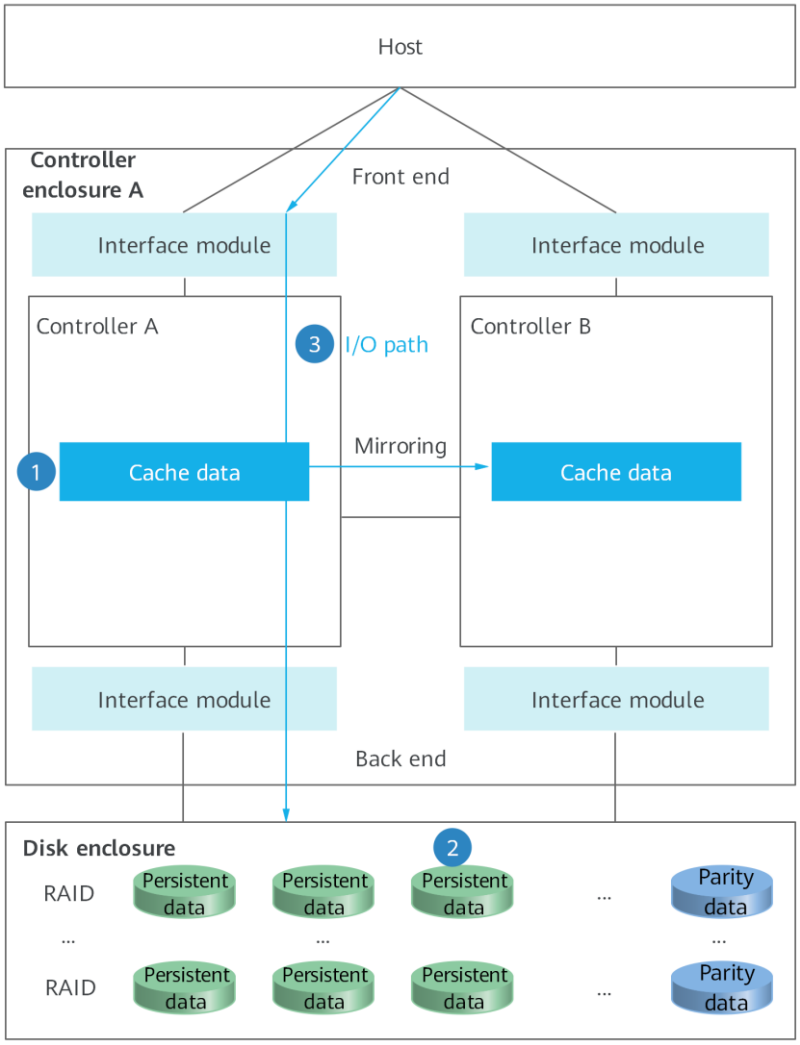
8 System-level Reliability Design

OceanStor Dorado 3000, 5000 and 6000 provides 99.999% availability via data reliability and service availability designs. Powered by the DR solution, the system availability can reach 99.9999%.

8.1 Data Reliability

For data written by hosts into storage systems, OceanStor Dorado undergoes three processes: data caching, data persisting on disks, and data path transmitting. The following describes the data reliability measures in the three processes.

Figure 8-1 Data reliability overview



8.1.1 Cache Data Reliability

To improve the speed of writing data, OceanStor Dorado provides the write cache mechanism. That is, after data is written to the memory cache of the controller and mirrored to the peer controller, a success message is returned to the host and then cache data is destaged to disks in the background.

User data stored in the controller memory may be lost if the system is powered off or the controller is faulty. To prevent the data loss, the system provides multiple cache copies across controllers and power failure protection to ensure data reliability.

8.1.1.1 Multiple Cache Copies

8.1.1.2 Power Failure Protection

OceanStor Dorado has built-in BBUs (backup power). When a power outage occurs in the storage systems, BBUs in controllers provide extra power for moving the cache data in the

memory to the coffer. After the power supply is recovered, the storage systems restore the cache data in the coffer to the memory during startup to prevent data loss.

8.1.2 Persistent Data Reliability

OceanStor Dorado uses the intra-disk RAID technology to ensure disk-level data reliability and prevent data loss. The RAID 2.0+ technology and dynamic reconstruction ensure system-level data reliability. As long as the number of faulty disks does not exceed that of the redundant disks, data will not be lost and the redundancy will not decrease.

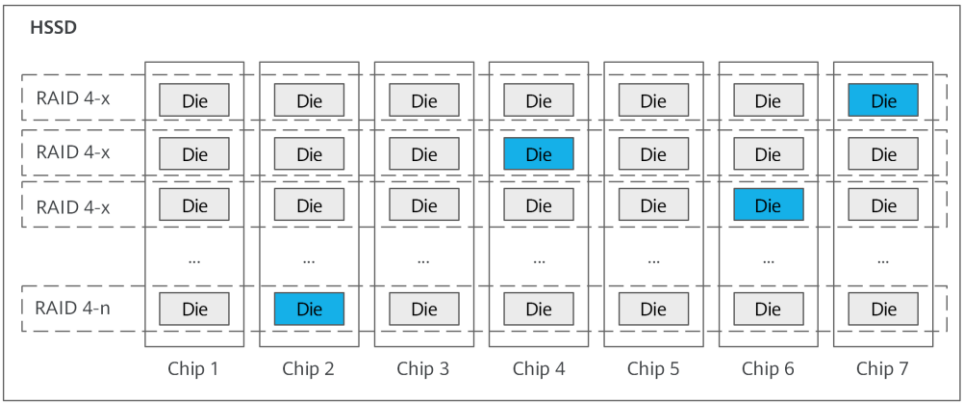
8.1.2.1 Intra-disk RAID

In addition to overall disk faults, regional damage may occur on the chips used for storing data. This is called the silent failure (bad block). These bad blocks do not cause the failure of the entire disk, but cause the data access failure on the disk.

Common bad block scanning can detect silent and invalid data in advance and repair the data. However, disk scanning occupies a large number of resources. To prevent impact on foreground services, the scanning speed must be controlled. Therefore, when the disk capacity and quantity are large, it takes weeks or even months to scan all disks. In addition, if both the bad block and disk failure occur in the interval between two scans, data may fail to be recovered.

Based on bad block scanning, OceanStor Dorado uses Huawei SSDs (HSSDs) to provide the intra-disk RAID feature to prevent silent failures in scanning isolation. Specifically, RAID 4 groups are created for data on SSDs in the unit of dies to implement redundancy, tolerating the failure of a single die without any data loss on SSDs.

Figure 8-2 Intra-disk RAID on an SSD



8.1.2.2 RAID 2.0+

In a conventional RAID storage system that uses fixed physical disks in RAID groups, LUNs or file systems used by users are divided from the RAID groups. Because the access frequency of each LUN or file system in the system is different, disks in some RAID groups are busy and become hot spots. Disks in other RAID groups cannot share the workloads even if they are idle. In addition, if a disk works for a longer time than others, its failure rate increases sharply and may be faulty in a shorter time than other disks. Therefore, hot disks in conventional RAID storage systems are at the risk of being overloaded.

With RAID 2.0+, OceanStor Dorado divides each SSD into fixed-size chunks (CKs, generally 4 MB). CKs from different SSDs are joined into a chunk group (CKG) based on the RAID groups. RAID 2.0+ has the following advantages over traditional RAID:

- Balanced service loads for zero hotspot. Data is evenly distributed to all disks in a storage resource pool, eliminating hotspot disks and lowering the disk failure rate.
- Quick reconstruction for a lowered data loss risk. If a disk is faulty, its data will be reconstructed to all the other disks in the resource pool. This many-to-many reconstruction is rapidly implemented, significantly shortening the non-redundancy period of data.
- All member disks in a storage resource pool participate in reconstruction, and each disk only needs to reconstruct a small amount of data. Therefore, the reconstruction process does not affect upper-layer applications.

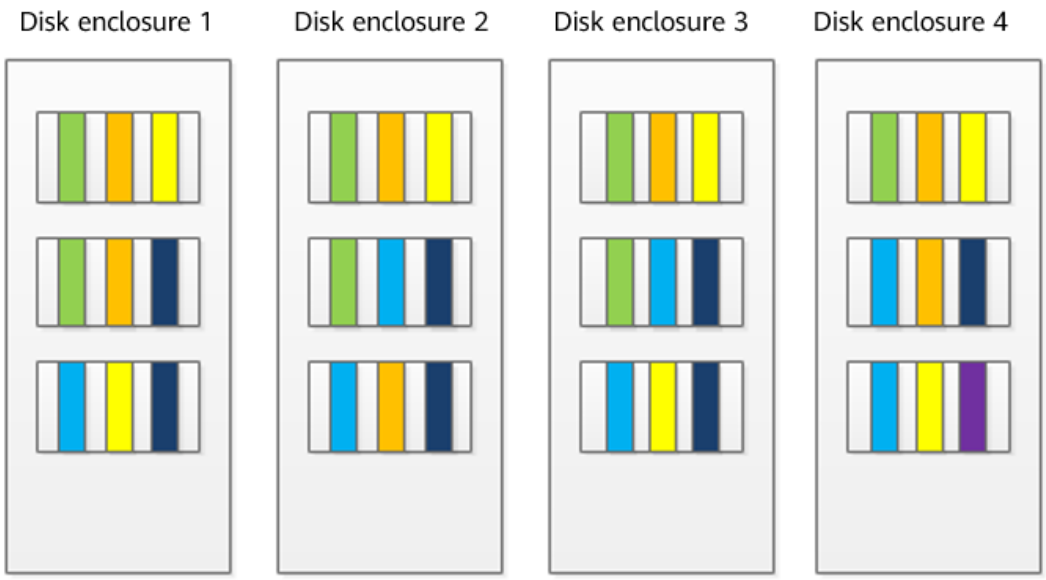
RAID for Disk Redundancy

RAID for disk redundancy leverages RAID 2.0+ to randomly select disks and evenly distribute data to selected disks. Each selected disk provides one chunk to form a chunk group. This ensures that no data is lost when a specific number of disks fails. For example, with RAID 6, a storage system that has four 25-slot disk enclosures can tolerate the failure of any two disks without service interruption. If you want to tolerate the failure of a disk enclosure, use RAID for enclosure redundancy.

RAID for Enclosure Redundancy

RAID for enclosure redundancy is implemented based on disk redundancy. The system selects a maximum of two chunks (on two disks) from each disk enclosure to form a chunk group. The number of chunks in a chunk group is determined by the number of disk enclosures. Enclosure redundancy applies to RAID levels with 2 disks (RAID 6). With enclosure redundancy, only two copies of data in each chunk group are lost in the event of a disk enclosure fault. These chunks can be recovered using the redundant disks for service continuity.

The following figure shows how chunks and chunk groups are distributed on a storage system with four disk enclosures (each with three disks). Chunks of the same color form a chunk group.



Enclosure redundancy uses the patented disk selection algorithm and has the following advantages:

- It inherits the advantages of RAID 2.0+ for load balancing, fast reconstruction, and global disk participation in reconstruction.
- It supports flexible configuration and can dynamically add disks to a RAID group.
- If a disk fails, the system can still tolerate the failure of any single disk enclosure after data reconstruction is completed, without the need of replacing the faulty disk.
- RAID-TP can tolerate the concurrent failure of one disk enclosure and any single disk in the remaining disk enclosures without interrupting services.
- Disks in a faulty disk enclosure can be removed and inserted into new or other functional disk enclosures to recover data without reconstruction.
- After the disk enclosure fault is rectified, enclosure redundancy is automatically restored. The patented disk selection algorithm minimizes the amount of data to be migrated and ensures fast recovery.
- Disks owned by different controller enclosures can form a chunk group to improve capacity usage.

8.1.2.3 Dynamic Reconstruction

If the number of available disks in a storage pool is fewer than *N*+*M* (due to consecutive failures of disks or disk replacement), the reconstruction cannot be performed and user data redundancy cannot be ensured. To cope with the problem, OceanStor Dorado uses dynamic reconstruction by reducing the number of data columns (*N*) and retaining the number of parity columns (*M*) during reconstruction. This method reduces the number of data columns, but retains the number of parity columns. After the reconstruction is complete, the number of member disks in the RAID group decreases, but the RAID redundancy level remains unchanged.

After the faulty disks are replaced, the system increases the number of data columns (*N*) based on the number of available disks in the storage pool, and new data will be written to the new *N*+*M* columns. Data that has been written during the fault will also be converted into the new *N*+*M* columns.

NOTE

Dynamic reconstruction reduces the total available capacity of the system. If multiple disks are faulty, handle the disk faults in time and pay attention to the storage pool usage.

8.1.3 Data Reliability on I/O Paths

During data transmission within a storage system, data passes through multiple components over various channels and undergoes complex software processing. Any problem during this process may cause data errors. If the errors cannot be detected immediately, error data can be written to persistent disks and the customer may obtain the error data, causing service exceptions.

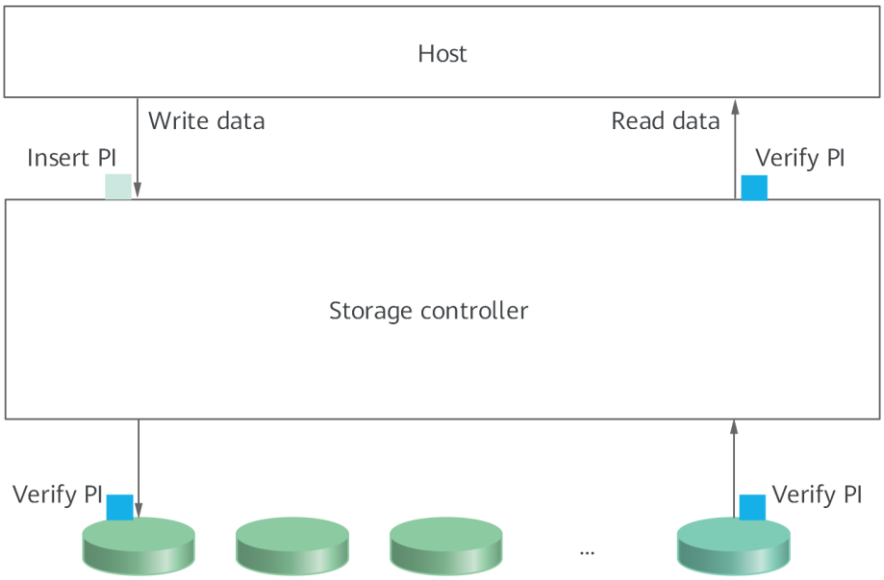
To resolve the preceding problems, OceanStor Dorado uses the end-to-end Protection Information (PI) function to detect and correct data errors (internal changes to data) on the transmission path. The matrix verification function ensures that changes to the whole data block (the whole data block is overwritten by old data or other data) can be detected. The preceding measures ensure data reliability on I/O paths.

8.1.3.1 End-to-end PI

OceanStor Dorado supports ANSI T10 PI. Upon reception of data from a host, the storage system inserts an 8-byte PI field to every 512 bytes of data before performing internal processing.

After the data is written to disks, the disks verify the PI fields of the data to detect any change to the data between reception and destaging to the disks. In the following figure, the green block indicates that a PI is inserted to the data. The blue blocks indicate that a PI is calculated for the 512-byte data and compared with the saved PI to verify data correctness.

Figure 8-3 End-to-end PI

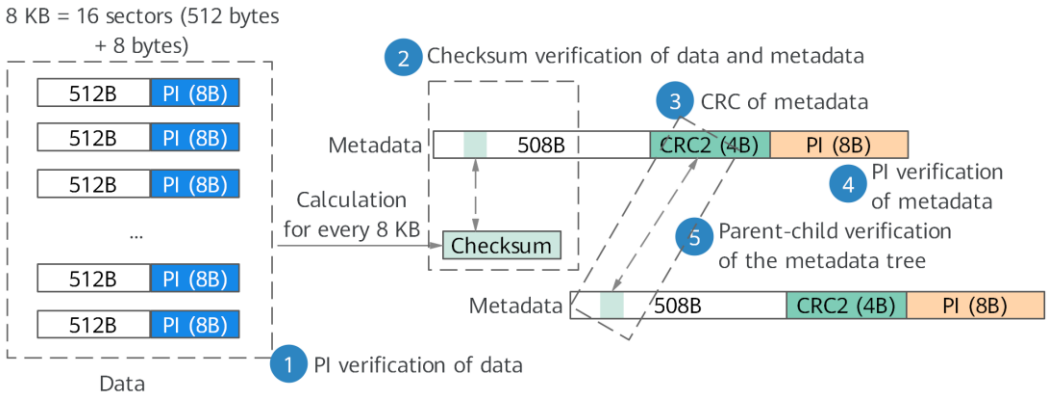


When the host reads data, the disks verify the data to prevent changes to the data. If any error occurs, the disks notify the upper-layer controller software, which then recovers the data by using RAID. To prevent errors on the path between the disks and the front end of the storage system, the storage system verifies the data again before returning it to the host. If any error occurs, the storage system recovers the data using RAID to ensure end-to-end data reliability from the front end to the back end.

8.1.3.2 Matrix Verification

Because the internal structure of disks is complex or the read path is long (involving multiple hardware components), various errors may occur due to software defects. For example, a write success is returned but the data fails to be written to disks; data B is returned when data A is read (read offset); or data that should be written to address A is actually written to address B (write offset). Once such errors occur, the PI check of the data is passed. If the data is still used, the incorrect data (such as old data) may be returned to the host.

Figure 8-4 Matrix verification



OceanStor Dorado provides matrix verification to cope with the write failure, read offset, and write offset that may occur on disks. In the preceding figure, each piece of data consists of 512-byte user data and 8-byte PI. Two bytes of the PI are used for cyclic redundancy check (CRC) to ensure reliability of the 512-byte data horizontally (protection point 1). The CRC bytes in 16 PI sectors are extracted to calculate the checksum, which is then saved in a metadata node. If offset occurs in a single or multiple pieces of data (512+8), the checksum of the 16 pieces of data is also changed and becomes inconsistent with that saved in the metadata. This ensures data reliability vertically. After detecting data damage, the storage system uses RAID redundancy to recover the data. This is matrix verification.

8.2 Service Availability

The storage system provides multiple redundancy protection mechanisms for the entire path from the host to the storage system. That is, when a single point of failure occurs on the interface module or link (1), controller (2), and storage media (3) that I/Os pass through, redundant components and fault tolerance measures can be used to ensure that services are not interrupted. In an active-active scenario, the peer storage system can take over services even if a single storage system fails (4) without interrupting host services.

Figure 8-5 Multi-layer redundancy and fault tolerance design (high-end)

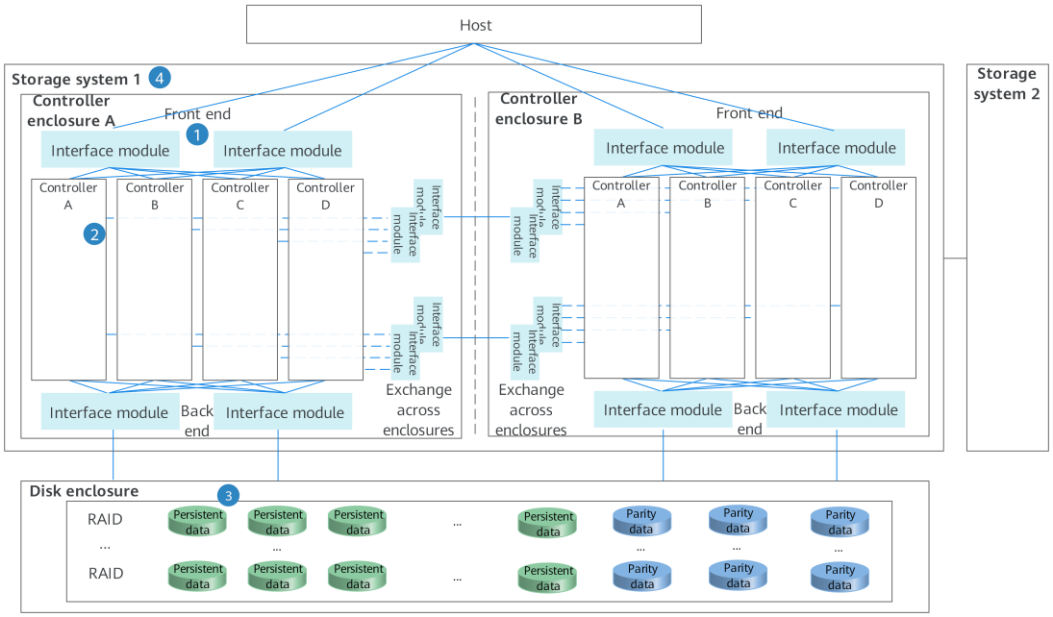
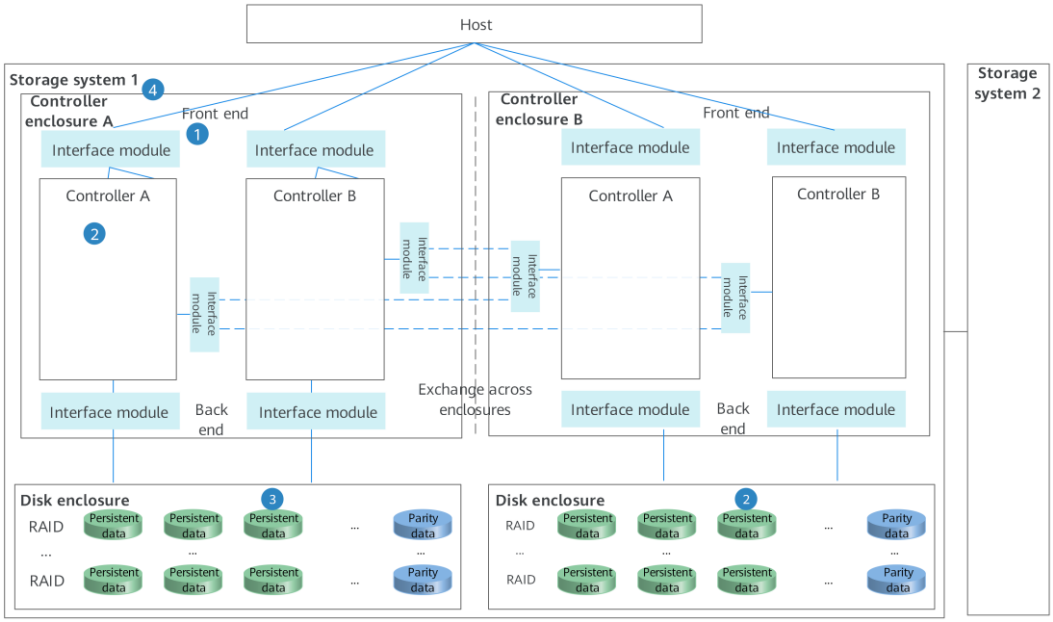


Figure 8-6 Multi-layer redundancy and fault tolerance design (entry-level and mid-range)



8.2.1 Interface Module and Link Redundancy Protection

OceanStor Dorado supports full redundancy. Link and interface module redundancy protection is provided for the front end for interconnection with the host, the back end for connecting disks, and the communication between controllers.

8.2.2 Controller Redundancy

OceanStor Dorado provides redundant controllers to ensure reliability. In typical scenarios, cache data is stored on the current controller and the copy of cache data is stored on another controller. If a controller fails, services can be switched to the controller to which the cache data copy belongs, ensuring service continuity.

8.2.3 Storage Media Redundancy

OceanStor Dorado not only ensures high reliability of a single disk, but also uses the multi-disk redundancy capability to ensure service availability if a single disk is faulty. That is, disk faults or sub-health is detected in a timely manner by using algorithms, and faulty disks are isolated in a timely manner to avoid long-term impact on services. Then, data of the faulty disk is recovered by using the redundancy technology. In this case, services can be continuously provided.

8.2.3.1 Fast Isolation of Disk Faults

When disks are running properly, OceanStor Dorado monitors the in-position and reset signals. If a disk is removed or faulty, the storage system isolates it. New I/Os are written to other disks and a read success is returned to the host after I/Os are read by using RAID.

In addition, continuous, long-time operation causes disks to wear and increases the chance of particle failures. As a result, disks respond more slowly to I/Os, which can affect services. Therefore, slow disks will be detected and isolated in a timely manner so that they cannot further affect services.

A model that compares the average I/O service time of disks is built for OceanStor Dorado based on common features of disks, including the disk type, interface type, and owning disk domain. With this model, slow disks can be detected and isolated within a short period of time, shortening the time when host services are affected by slow disks.

8.2.3.2 Disk Redundancy

OceanStor Dorado supports three RAID configuration modes, which ensure service continuity in the event of disk failures. The storage systems can tolerate the simultaneous failure of three disks in the storage pool at most, ensuring zero data loss without service interruption.

- RAID 5 uses the EC-1 algorithm and generates one copy of parity data for each stripe. The failure of one disk is allowed.
- RAID 6 uses the EC-2 algorithm and generates two copies of parity data for each stripe. The simultaneous failure of two disks is allowed.
- RAID-TP uses the EC-3 algorithm and generates three copies of parity data for each stripe. The simultaneous failure of three disks is allowed.

8.2.4 Array-level Redundancy

In addition to providing intra-array high availability protection for services requiring high reliability, OceanStor Dorado provides array-level (site) active-active protection to ensure service continuity in case of a power failure or disaster such as earthquake and fire.

HyperMetro, an array-level active-active technology provided by OceanStor Dorado, allows two LUNs or file systems from separate storage systems to maintain real-time data consistency and to be accessible to hosts.

HyperMetro supports both Fibre Channel and IP networking. The two storage systems in a HyperMetro deployment can be at two locations within 300 km from each other, such as in the same equipment room or in the same city. The quorum server is generally deployed at a third site.

If one storage system fails, hosts automatically choose the paths to the other storage system for service access. If the replication links between the storage systems fail, only one storage system can be accessed by hosts, which is determined by the arbitration mechanism of HyperMetro.

9 System Performance Design

OceanStor Dorado uses brand-new hardware design and optimizes I/O paths from multipathing software, networks, CPU computing, and SSDs to provide optimal high IOPS and low latency for customers. Table 9-1 describes the key performance design by I/O process for addressing the current problems and pain points.

Table 9-1 Key performance design

I/O Process	Challenge	Key Design Point	Performance Design Principles
Host path selection	SAN: After a read/write request is delivered to the controller, forwarding the request again increases the CPU overhead and latency.	Global load balancing	The path selection mode of UltraPath is changed from the load balancing mode to the mode in which UltraPath negotiates with controllers and delivers I/Os to controllers that eventually process the I/Os, reducing controller forwarding.
	NAS: Each client selects a service IP address to establish an access path. Multiple clients may select the same IP address, which causes the NIC of the IP address to become a performance bottleneck.	Dynamic load balancing by the DNS service	The device has the built-in DNS service, which supports different IP address balancing policies to meet the requirements of service load balancing in various service models.
Front end	The native Ethernet protocol has multiple layers, resulting in high latency.	Direct TCP/IP Offloading Engine (DTOE), a technology optimized by Huawei	• I/Os bypass the kernel mode to reduce cross-mode overheads. • Offload protocols using hardware, reducing CPU usage.
	Reducing the system scheduling latency is required.	Driver polling scheduling of Fibre Channel, iSCSI, NFS, SMB, and switching networks	• The working thread periodically checks the receiving queue of the interface module in polling mode to reduce the latency caused by waking up the

I/O Process	Challenge	Key Design Point	Performance Design Principles
			working thread upon a request. • Load balancing of front-end, mirroring, and back-end networks is supported, fully utilizing CPU capabilities.
Controller	How to make full use of the computing capability of multi-core CPUs	Intelligent multi-core technology	• I/Os are distributed among CPUs by CPU group to reduce the latency of cross-CPU scheduling. • A CPU is divided into different zones based on services to reduce service interference. • No lock is designed in the service partition to reduce lock conflicts.
	Load imbalance between CPU groups in different scenarios	Global load balancing	Tasks with high-density computing overheads are scheduled in load balancing mode among service groups in a CPU.
Back end	Write amplification causes short SSD life time and low performance.	Multistreaming	Hot and cold data is separated to reduce write penalty within disks.
		ROW full-stripe write	The ROW full-stripe write design reduces random write amplification.
	The background erasing and write operations of SSDs affect the foreground read latency.	Read first on SSDs.	Collaboration of SSD hardware and software improves the read I/O priority and reduces the read latency.
	System performance deteriorates in fault scenarios (for example, SSD faults).	Smart disk enclosure	Smart disk enclosures are used to offload SSD reconstruction overhead from controllers, ensuring minimal impact on system performance in fault scenarios.

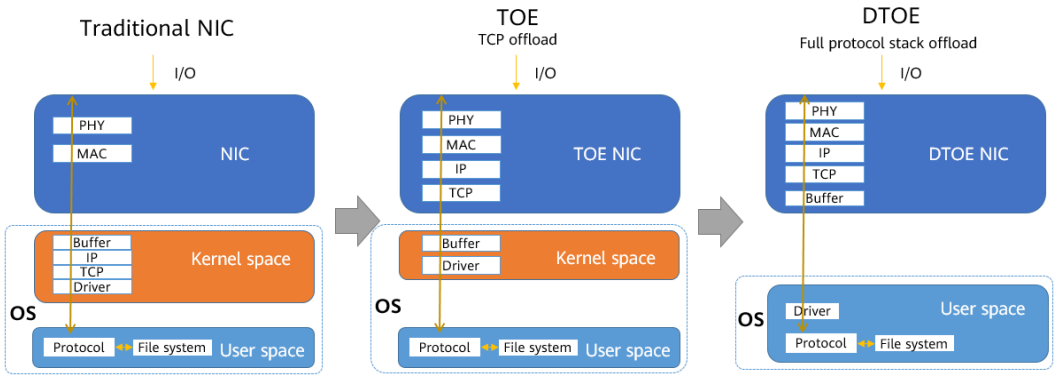
9.1 Front-end Network Optimization

Front-end network optimization mainly refers to the optimization of latency between applications and storage devices, including the optimization of the path selection algorithm of the multipathing software on the server side, protocol offloading optimization in iSCSI scenarios, and scheduling optimization in common scenarios.

Protocol Offloading

The performance bottleneck of network adapters (iSCSI) lies in the long I/O path. The overhead of TCP and IP protocols is extremely high. Huawei uses the highly-optimized user-mode iSCSI protocol stack and DTOE to offload TCP and IP protocols, as shown in Figure 9-2.

Figure 9-1 DTOE technology



If controllers use traditional NICs, network protocol stacks processed by the controllers have deep layers. As a result, every time a data packet is processed, an interruption is triggered, causing high CPU overhead.

By using the TOE technology, NICs offload the TCP and IP protocols. An interruption is triggered after an application implements a complete data processing, which significantly reduces the interruption overhead. However, in this case, some drivers are running in the kernel mode, resulting in the latency caused by the overhead of system calls and thread switchover between user mode and kernel mode.

The DTOE technology adopted by OceanStor Dorado offloads iSCSI/NAS TCP data paths to NICs. The transport layer processing, including the DIF check function, is offloaded to the network adapter microcode, eliminating the CPU overhead. In addition, the working thread in the system checks the receiving queue of the interface module periodically in polling mode. If there is a request, the thread processes the request immediately. This reduces the latency overhead caused by waking up the working thread after a request is received.

9.2 CPU Computing Optimization

Intelligent Multi-Core Technology

OceanStor Dorado uses high-performance processors. Each controller contains more CPUs and cores than any other controller in the industry.

For the symmetric multiprocessor (SMP), the biggest challenge is to keep the system performance growing linearly as the number of CPUs increases. The SMP system has the following two key problems:

1. The more CPUs, the more overhead of communication between CPUs, and the more memory access across CPUs.
2. The more the number of cores, the more likely that conflicts may be caused by program mutual exclusion, and the longer time for conflict handling.

OceanStor Dorado uses the intelligent multi-core technology to allow performance to increase linearly with the number of CPUs. The key optimization technologies for the key problems are as follows:

1. The CPU grouping and distribution technologies are used among CPUs. Each I/O is scheduled within one CPU during the process of being delivered to the controllers by the multipathing software and arriving at the back-end disk enclosures. In addition, memory is allocated on the current memory channel, minimizing the communication overhead between CPUs.
2. CPU cores within a CPU are grouped based on service attributes. The front-end, back-end, and node interconnection networks are scheduled in separate CPU core groups. The I/O stack for processing a task is scheduled only within one CPU core group, which effectively controls the conflict impact and processing overhead, improves the scheduling efficiency, and accelerates the processing of the I/O stack.

Dynamic Load Balancing

The traditional CPU grouping technology can solve the conflict domain problem of each service, but also brings the problem of unbalanced resource usage between CPU groups in different scenarios. The dynamic load balancing technology of OceanStor Dorado defines differentiated scheduling policies based on the computing overheads of tasks. In this way, tasks are balanced among the CPU core groups. High-density computing tasks are distinguished from common density computing tasks and are used as scheduling units for load balancing among groups. This prevents scheduled tasks from interfering with each other, improving task execution efficiency.

9.3 Back-end Network Optimization

Multistreaming

OceanStor Dorado uses the multistreaming technology. The SSD driver works with the controller software to effectively distinguish data with different change frequencies and store the data in different blocks. For example, metadata (hot data) and user data (warm data) are stored in different blocks. This increases the probability that data in the blocks becomes invalid at the same time, reduces the amount of valid data to be migrated during GC, and improves SSD access performance and the service life.

ROW Full-Stripe Write

OceanStor Dorado adopts the ROW full-stripe write design, which writes all new data to new blocks. This avoids the write amplification caused by data reads and parity check in a traditional RAID write process and eliminates the risk when multiple blocks are changed simultaneously. In addition, it effectively reduces the CPU overhead of the storage controller and the read/write workload on SSDs during the write process, and simplifies the processing

logic for a better error tolerance capability. Compared to the traditional Write In Place mode, the ROW full-stripe write mode delivers higher performance and fault tolerance efficiency in random write scenarios.

Read First on SSDs

OceanStor Dorado uses the latest generation of SSDs, reducing the average read latency by over 50 μs. Generally, there are three types of operations on the flash media of an SSD: read, write, and erase. The erase latency is 5 ms to 15 ms, the write latency is 2 ms to 4 ms, and the read latency ranges from dozens of μs to 100 μs. When a flash chip is performing a write or an erase operation, a read operation must wait until the current operation is finished, which causes a great jitter in read latency. By using the read first on SSDs technology, if a read request with a higher priority is detected during an erase or write operation, the system cancels the current operation and preferentially processes the read request. This greatly reduces the read latency on SSDs.

Smart Disk Enclosure

In the current storage system, reconstruction and disk enclosure event processing share the CPU and memory with I/Os. When reconstruction or abnormal events occur, the CPU and memory resources are consumed, affecting I/O performance.

The smart disk enclosure of OceanStor Dorado is equipped with CPU and memory resources. It can offload tasks, such as disk reconstruction upon a disk failure, from controllers to reduce the load on the controllers. When the smart disk enclosure is used, it receives the reconstruction request and reads data locally to calculate the parity data. Then, it only needs to transmit the parity data to the controller. This reduces the consumption of controller resources and the impact of reconstruction on application performance.

10 System Serviceability Design

This chapter describes how to manage OceanStor Dorado through various interfaces (including DeviceManager, CLI, RESTful API, SNMP, and SMIS) and the lifecycle of devices through FlashEver as well as describes transparent upgrade mode of hosts.

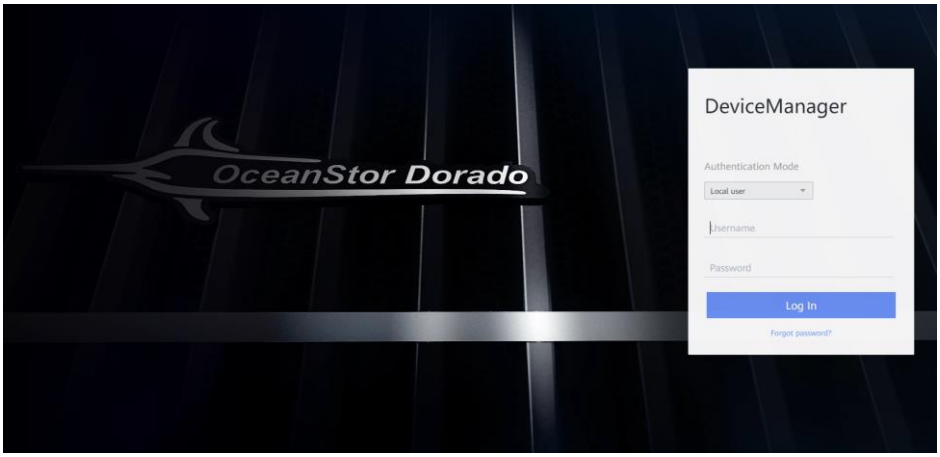
10.1 System Management

OceanStor Dorado 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 mainly RESTful interfaces, supporting 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.

10.1.1 DeviceManager

DeviceManager is a built-in HTML5-based management system for OceanStor Dorado. It provides wizard-based GUI for efficient management. Users can enter **https://***storage management IP address***:8088/** on the browser to use DeviceManager. On DeviceManager, you can perform almost all required configurations. The following figure shows the DeviceManager login page.

Figure 10-1 DeviceManager login page



You can use the following functions on DeviceManager:

- Storage space management: This includes storage pool management, LUN management, and mapping between LUNs and hosts.
- Data protection management: LUN data is protected using snapshot, clone, backup, replication, and active-active.
- Configuration task: Background tasks for complex configuration operations are provided to trace the procedure of the configuration process.
- Fault management: The status of storage devices and management units on storage devices are monitored. If faults occur, alarms will be generated and troubleshooting suggestions and guidance will be provided.
- Performance and capacity management: The performance and capacity of storage devices is monitored in real time. You can view the collected historical performance and capacity data and analyze associated performance data.
- Security management: DeviceManager supports the management of users, roles, permissions, certificates, and keys.

DeviceManager uses a new UI design to provide a simple interactive interface. Users can complete configuration tasks only in a few operations, which improves user experience.

10.1.1.1 Storage Space Management

Flexible Storage Pool Management

OceanStor Dorado manages storage space by using storage pools. A storage pool consists of multiple SSDs and can be divided into multiple LUNs for hosts to use. Users can use only one storage pool to manage all the space or divide multiple storage pools.

- One storage pool for the entire storage system

  This is the simplest division method. Users only need to create one storage pool using all disks during system initialization.

- Multiple storage pools to isolate applications

  If you want to use multiple storage pools to manage space and isolate fault domains of different applications, you can manually create storage pools at any time in either of the following ways:

- Specify the number of disks used to create a storage pool. The system automatically selects qualified disks to create a storage pool.
- Manually select specific disks to create a storage pool.

Mappings Between LUNs and Hosts

To facilitate LUN management and provide storage volumes for hosts, OceanStor Dorado defines the following types of management objects:

Object	Function
LUN	A storage volume that can be accessed by hosts
LUN group	A LUN group consists of multiple LUNs. If the data of an application comes from multiple LUNs, these LUNs can be added to a LUN group. Operations on the LUN group apply to all LUNs in the LUN group.
Host	A host that can access the storage system. The host can be a physical host or a VM.
Host group	A host group consists of multiple hosts. If an application is deployed on a cluster consisting of multiple hosts and these hosts access the data volumes of the application simultaneously, you can create a host group for these hosts.

DeviceManager provides multiple simple and flexible mechanisms. No matter the application is simple or complex, a proper mapping scheme is available for users.

- Mappings between LUN groups and host groups

  If an application has multiple LUNs and is deployed on a cluster consisting of multiple hosts, you are advised to manage the LUNs using a LUN group, manage the hosts using a host group, and create mappings between the LUN group and host group.
- Mappings between the LUN group and host

  If an application has multiple LUNs and is deployed on only one host, you are advised to manage these LUNs using a LUN group and create the mapping between the LUN group and the host.
- Mapping between the LUN and host

  If an application uses only one LUN and is deployed only on one host or you are not used to using LUN groups, create the mapping between the LUN and host.

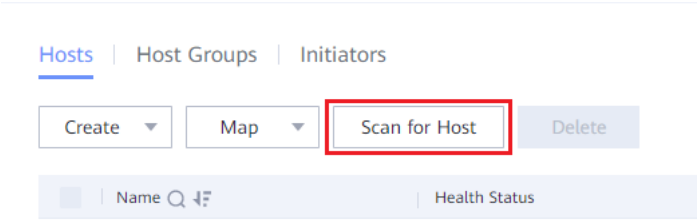
Automatic Host Detection

In addition to manually creating a host using the WWN or IQN, hosts can be automatically created and scanned if Huawei's multipathing software UltraPath is installed.

As shown in the following figure, after the physical network connection between the host and the storage device is set up, you can click **Scan for Host** on the host management page. The system scans for all hosts connected to the storage device, identifies their WWNs or IQNs, and automatically creates hosts. If a host has multiple WWNs or IQNs, the system can automatically identify them as on one host.

When a large number of hosts exist, managing their WWNs or IQNs is time-consuming. The automatic host detection function simplifies the management. For details about the operation requirements and environment requirements for automatic host detection, see the online help provided by OceanStor Dorado.

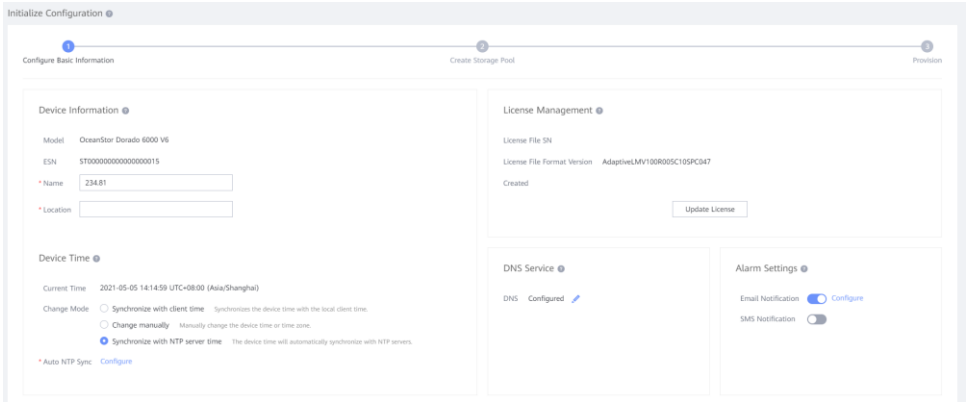
Figure 10-2 Automatic host detection



Quick Configuration Wizard

OceanStor Dorado provides an initial configuration wizard for comprehensive scenarios, simplifying the configuration process from unpacking to use. According to the configuration sequence, the configuration items are as follows:

- Basic device information, license, time, DNS, and alarm notification



- Storage pool creation

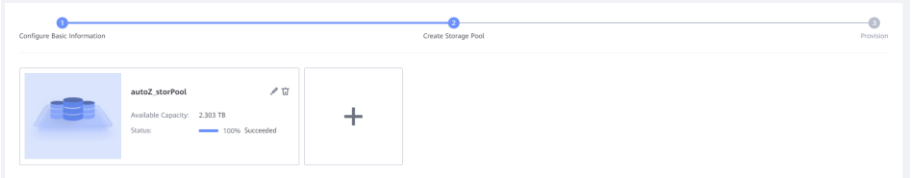
  OceanStor Dorado manages storage space by using storage pools. A storage pool consists of multiple SSDs and can be divided into multiple LUNs for hosts to use. You can use one storage pool to manage all of the space or create multiple storage pools.

  – Only one storage pool on the storage system

    This is the simplest method. You only need to create one storage pool with all disks during system initialization.

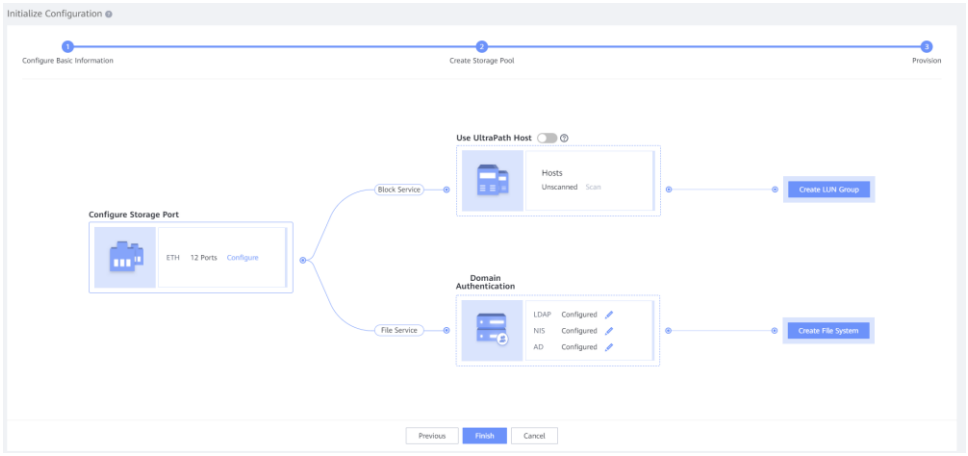
  – Multiple storage pools to isolate applications

    If you want to use multiple storage pools to manage space and isolate fault domains of different applications, you can manually create storage pools at any time in either of the following ways. You can specify the number of disks used to create a storage pool, and the system automatically selects the disks that meet the requirements. Alternatively, you can manually select specific disks to create a storage pool.



- Resource allocation

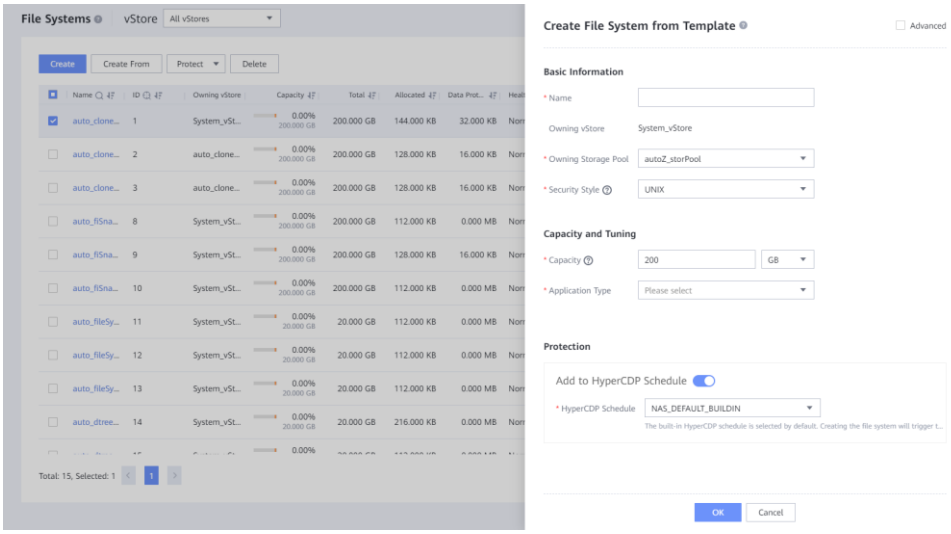
  The wizard allows for automatic discovery of hosts using Huawei UltraPath and NAS domain authentication configuration, and provides links for service provisioning. The entire process is streamlined, helping users get started quickly.



Rapid Provisioning of NAS Resources Based on Existing Objects

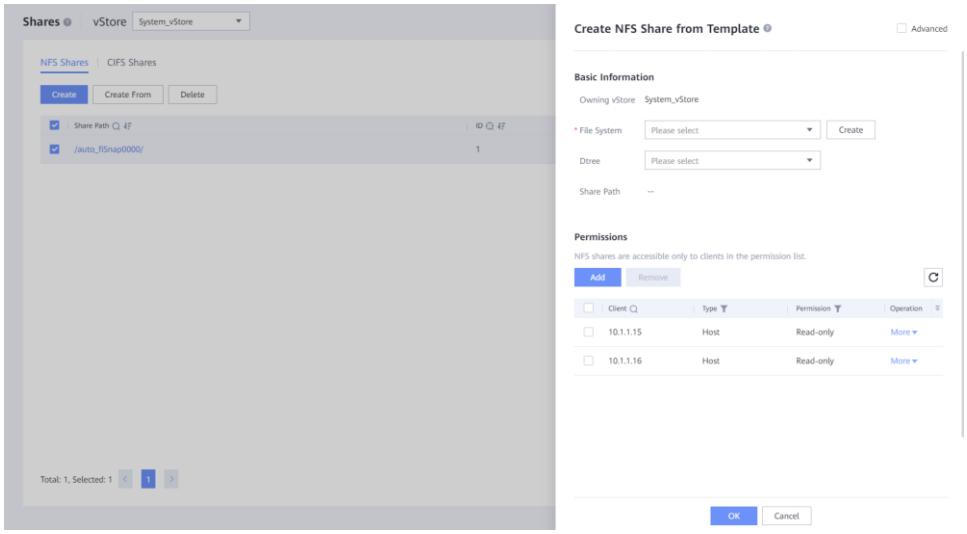
- Creating a file system using a template

  Use the selected file system to automatically fill in parameters except the name.



- Creating a share using a template

  Use the selected share to automatically fill in all parameters.



10.1.1.2 Data Protection Management

Data Protection Based on Protection Groups

If an application has multiple LUNs, data protection for the application is to protect its LUNs simultaneously and ensure their data consistency. OceanStor Dorado introduces protection groups to protect the LUNs of an application in groups.

A protection group consists of multiple LUNs. If you implement data protection on a protection group, such as creating snapshots, the operation applies to all LUNs in the protection group and ensures data consistency between LUNs.

The following features support batch LUN data protection based on the protection group:

- Protection group-based snapshot

  When you create a snapshot for a protection group, a snapshot is created for each LUN in the protection group. A snapshot consistency group is automatically created and the snapshots are added to the snapshot consistency group.

- Protection group-based clone

  When you create a clone for a protection group, a clone is created for each LUN in the protection group. A clone consistency group is automatically created and the clone LUNs are added to the clone consistency group.

- Protection group-based remote replication

  When you create remote replication for a protection group, a remote replication pair relationship is established for each LUN in the protection group. A remote replication consistency group is automatically created and the remote replication pairs are added to the remote replication consistency group.

- Protection group-based HyperMetro

  When you configure HyperMetro for a protection group, a HyperMetro pair relationship is established for each LUN in the protection group. A HyperMetro consistency group is automatically created and the HyperMetro pairs are added to the HyperMetro consistency group.

Like protecting an independent LUN, the protection group allows users to configure protection for multiple LUNs of an application. Users do not need to configure LUNs

separately. Managing a batch of LUNs is as simple as managing one LUN and data consistency is ensured at the same time.

Flexible Use of LUN Groups and Protection Groups

As previously mentioned, OceanStor Dorado uses LUN groups and protection groups to manage volumes of applications in batches. The following table describes the application scenarios.

Group	Application Scenario	Relationship with Applications
LUN group	Mapping	1. Manages all volumes of an application. 2. Manages all volumes of a host (involving multiple applications).
Protection group	Snapshot, clone, replication, and active-active	Manages all volumes of an application.

- Using a LUN group only to manage LUNs of an application

  Most recommended method: Use a LUN group to manage LUNs of an application. In this way, LUNs are mapped and protected on the basis of LUN groups.

  This management model is simple. The system will automatically create a unique protection group and implement various types of data protection based on the protection group.

  This method makes the management of multiple LUNs as simple as the management of only one LUN.
- Using LUN groups to manage volumes of hosts and protection groups to manage volumes of applications

  This method is usually used to manage LUNs that belong to different applications which are deployed on the same host or host group.

  In such conditions, it is not appropriate to create the LUN group as a protection group, because data volumes of some applications may not need to be protected.

  Therefore, you can select the specified LUNs of the LUN group to create a protection group and perform protection operations for the protection group.
- Using LUN groups or protection groups separately

  If the data protection feature is not enabled, only LUN groups are available for you.

  If LUN group-based mappings are unnecessary, you can add LUNs of an application to a protection group to keep them consistent.

Capacity Expansion of LUN Groups or Protection Groups

For an application running out of storage space, you can add LUNs to expand the capacity. If the LUNs are managed in a LUN group or protection group, they inherit the mappings and protection settings of the group.

- Automatic creation of mappings

  After new LUNs are added to a LUN group, these LUNs share the mappings of the LUN group. Hosts accessible to the LUN group are also accessible to the LUNs.
- Automatic configuration of data protection

After new LUNs are added to a LUN group that is included in a protection group, the system will automatically append the protection settings of the protection group to the LUNs, such as creating clones, replication pairs, and HyperMetro pairs for the LUNs, and adding the created objects to their respective consistency groups.

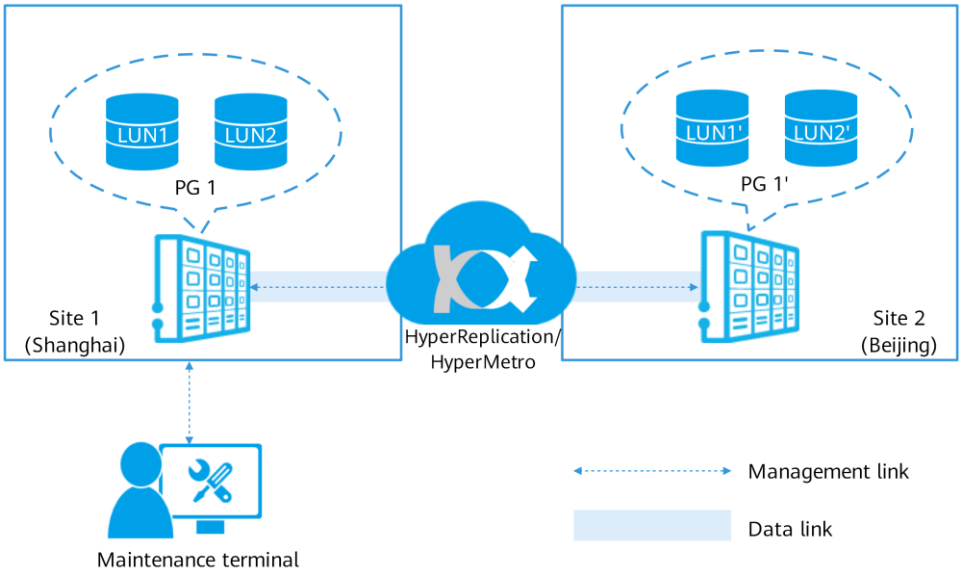
If new LUNs are added to a protection group, the system will append the protection settings of the protection group to the LUNs.

Similarly, if LUNs are removed from a LUN group or protection group, their mappings will be deleted, and they will be removed from consistency groups (their existing replication pairs and HyperMetro pairs will be retained and can be manually deleted).

Configuration on One Device for Cross-Device Data Protection

When cross-device data protection (such as replication, HyperMetro) is enabled, data protection configuration only need to be performed on one device. For example, as shown in Figure 10-3, if you want to replicate the protection group (composed of LUN 1 and LUN 2) at site 1 in Shanghai to site 2 in Beijing, you only need to configure protection settings at site 1, without logging in to site 2.

Figure 10-3 Configurations on one device



If you use one protection group to manage all LUNs of an application and have enabled protection for the protection group, the system automatically creates target volumes for all of the LUNs, pairs the sources with targets, and creates a consistency group. Currently, DeviceManager allows the following settings to be configured on one device:

- Remote replication

  After you select the protection group you want to replicate to a remote device, DeviceManager does the following automatically:

  - Creates a target protection group on the target device.
  - Creates target LUNs.
  - Pairs each source LUN with target LUN.
  - Creates a replication consistency group.

- Adds the pairs to the replication consistency group.

- HyperMetro

  After you select the protection group for which you want to enable active-active protection, DeviceManager does the following automatically:

  - Creates a target protection group on the target device.
  - Creates target LUNs.
  - Pairs each source LUN with target LUN.
  - Creates a HyperMetro consistency group.
  - Adds the pairs to the HyperMetro consistency group.

- DR Star

  After you select the protection group for which you want to enable DR Star protection, DeviceManager does the following automatically:

  - Creates target protection groups on the two target devices.
  - Creates target LUNs.
  - Pairs each source LUN with target LUN and form DR Star.
  - Creates three consistency groups to form DR Star.
  - Adds the replication or HyperMetro pairs to the corresponding consistency group.

DeviceManager uses data links between devices to transmit management commands (see Figure 10-3). You can directly use this feature with no need to configure network devices.

For security purposes, you must be authorized by the target device and enter the user name and password of the target device on the primary device.

10.1.1.2.1 Data Protection Based on Protection Groups

If an application has multiple LUNs, data protection for the application is to protect its LUNs simultaneously and ensure their data consistency. In this case, protection groups are introduced to protect the LUNs of an application.

A protection group consists of multiple LUNs. If you implement data protection on a protection group, such as creating snapshots, the operation applies to all LUNs in the protection group and ensures data consistency between LUNs. The following features support batch LUN data protection based on the protection group:

- Protection group-based snapshot

  When you create a snapshot for a protection group, a snapshot is created for each LUN in the protection group. A snapshot consistency group is automatically created and the snapshots are added to the snapshot consistency group.

- Protection group-based clone

  When you create a clone for a protection group, a clone is created for each LUN in the protection group. A clone consistency group is automatically created and the clone LUNs are added to the clone consistency group.

- Protection group-based remote replication

  When you create remote replication for a protection group, a remote replication pair relationship is established for each LUN in the protection group. A remote replication consistency group is automatically created and the remote replication pairs are added to the remote replication consistency group.

- Protection group-based HyperMetro

When you configure HyperMetro for a protection group, a HyperMetro pair relationship is established for each LUN in the protection group. A HyperMetro consistency group is automatically created and the HyperMetro pairs are added to the HyperMetro consistency group.

Like protecting an independent LUN, the protection group allows users to configure protection for multiple LUNs of an application. Users do not need to configure LUNs separately. Managing a batch of LUNs is as simple as managing one LUN and data consistency is ensured at the same time.

10.1.1.2.2 Flexible Use of LUN Groups and Protection Groups

As previously mentioned, LUN groups and protection groups are used to manage volumes of applications in batches. The following table describes the application scenarios.

Group	Application Scenario	Relationship with Applications
LUN group	Mapping	• Manages all volumes of an application. • Manages all volumes of a host (involving multiple applications).
Protection group	Snapshot, clone, replication, and active-active	Manages all volumes of an application.

- Using LUN groups to manage LUNs of an application

  This is the most recommended method. If LUN groups are used to manage LUNs of an application, LUNs are mapped and protected on the basis of LUN groups.

  This management model is simple. The system will automatically create a unique protection group and implement various types of data protection based on the protection group.

  This method makes the management of multiple LUNs as simple as the management of only one LUN.
- Using LUN groups to manage volumes of hosts and protection groups to manage volumes of applications

  This method is usually used to manage LUNs that belong to different applications which are deployed on the same host or host group.

  In such conditions, it is not appropriate to create the LUN group as a protection group, because data volumes of some applications may not need to be protected. Therefore, you can select the specified LUNs of the LUN group to create a protection group and perform protection operations for the protection group.
- Using LUN groups or protection groups separately

  If the data protection feature is not enabled, only LUN groups are available for you.

  If LUN group-based mappings are unnecessary, you can add LUNs of an application to a protection group to keep them consistent.

10.1.1.2.3 Capacity Expansion of LUN Groups or Protection Groups

For an application running out of storage space, you can add LUNs to expand the capacity. If the LUNs are managed in a LUN group or protection group, they inherit the mappings and protection settings of the group.

- Automatic creation of mappings

  After new LUNs are added to a LUN group, these LUNs share the mappings of the LUN group. Hosts accessible to the LUN group are also accessible to the LUNs.

- Automatic configuration of data protection

  After new LUNs are added to a LUN group that is included in a protection group, the system will automatically append the protection settings of the protection group to the LUNs, such as creating clones, replication pairs, and HyperMetro pairs for the LUNs, and adding the created objects to their respective consistency groups.

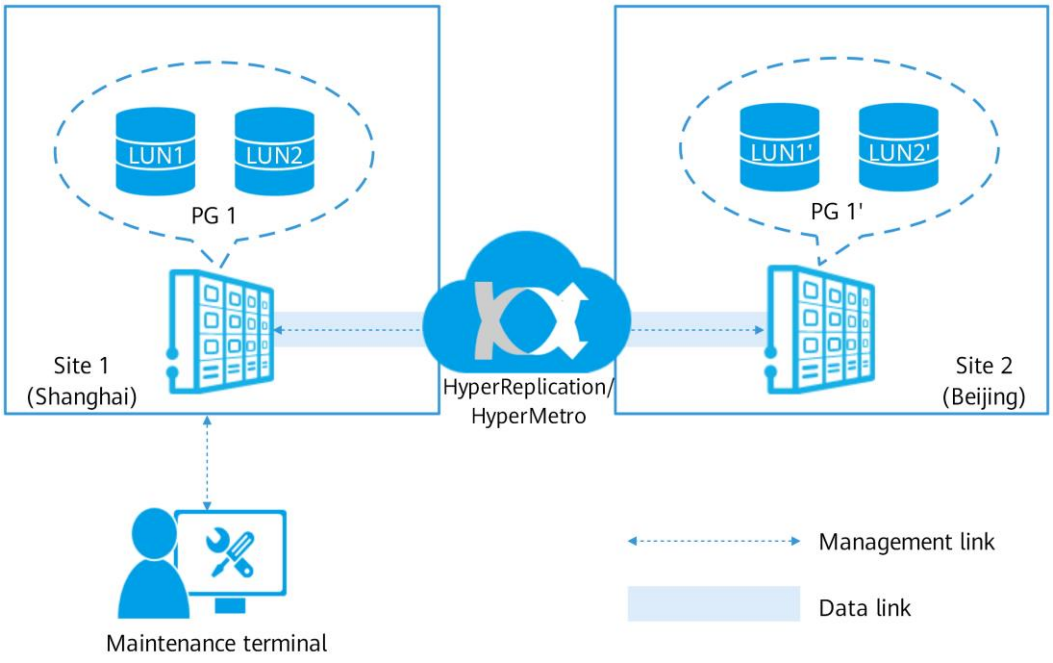
  If new LUNs are added to a protection group, the system will append the protection settings of the protection group to the LUNs.

  Similarly, if LUNs are removed from a LUN group or protection group, their mappings will be deleted, and they will be removed from consistency groups (their existing replication pairs and HyperMetro pairs will be retained and can be manually deleted).

10.1.1.2.4 Configuration on One Device for Cross-Device Data Protection

When cross-device data protection (such as replication and HyperMetro) is enabled, data protection configuration only needs to be performed on one device. For example, as shown in the following figure, if you want to replicate the protection group (composed of LUN 1 and LUN 2) at site 1 in Shanghai to site 2 in Beijing, you only need to configure protection settings at site 1, without logging in to site 2.

Figure 10-4 Configurations on one device



If you use one protection group to manage all LUNs of an application and have enabled protection for the protection group, the system automatically creates target volumes for all of the LUNs, pairs the sources with targets, and creates a consistency group. Currently, DeviceManager allows the following settings to be configured on one device:

- Remote replication

After you select the protection group you want to replicate to a remote device, DeviceManager does the following automatically:

a. Creates a target protection group on the target device.
b. Creates target LUNs.
c. Pairs each source LUN with the target LUN.
d. Creates a replication consistency group.
e. Adds the pairs to the replication consistency group.

- HyperMetro

  After you select the protection group for which you want to enable active-active protection, DeviceManager does the following automatically:

  a. Creates a target protection group on the target device.
  b. Creates target LUNs.
  c. Pairs each source LUN with the target LUN.
  d. Creates a HyperMetro consistency group.
  e. Adds the pairs to the HyperMetro consistency group.

- DR Star

  After you select the protection group for which you want to enable DR Star protection, DeviceManager does the following automatically:

  a. Creates target protection groups on the two target devices.
  b. Creates target LUNs.
  c. Pairs each source LUN with the target LUN and forms DR Star.
  d. Creates three consistency groups to form DR Star.
  e. Adds the replication or HyperMetro pairs to the corresponding consistency group.

DeviceManager uses data links between devices to transmit management commands (see Figure 10-4). You can directly use this feature with no need to configure network devices.

For security purposes, you must be authorized by the target device and enter the user name and password of the target device on the primary device.

10.1.1.3 Configuration Task

DeviceManager automatically enables a background configuration task after you submit a complex configuration operation. The task is running on the storage background. In this way, you can perform other operations while the previous one is still being processed.

Background configuration tasks also apply to LUN group-based mappings, protection group-based protection, and cross-device protection. Suppose that a user needs to configure protection for a LUN group that has hundreds of LUNs. It takes a long time and hundreds of operations to complete the configuration. However, the background configuration task helps the user complete the configuration in the background, freeing the user from long-term waiting.

- Task steps and progress

  A complex task usually contains multiple executable steps. You can use DeviceManager to check in which step the task is executed and view the overall task execution progress (%).

- Task execution in the background

After a configuration task is submitted, the task is executed in the background. You can close the DeviceManager page without waiting for the task to complete. You can also submit multiple configuration tasks. If the resources on which the tasks depend do not conflict, the tasks are automatically executed in sequence in the background.

- Resuming tasks from the breakpoint

  If the system is powered off unexpectedly during the task execution, the task will be resumed at the breakpoint after the system is restarted and all subsequent steps will be performed.

- Retrying failed tasks manually

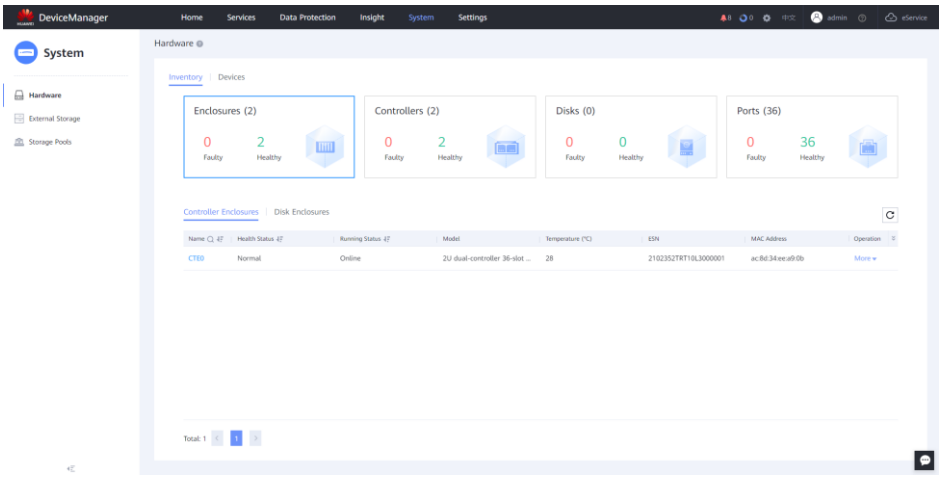
  If an exception occurs during the execution of a step, the task is automatically interrupted and the cause is displayed. For example, a task cannot be executed because the storage space is insufficient during LUN creation. In this case, you can manually rectify the fault that causes task interruption, and then manually restart the task. The task will then continue from the failure point.

10.1.1.4 Fault Management

10.1.1.4.1 Monitoring Status of Hardware Devices

This function provides hardware views in a what-you-see-is-what-you-get manner and uses colored digits to make status of different hardware more distinguishable. Figure 10-5 shows the status statistics of hardware components.

Figure 10-5 Inventory management



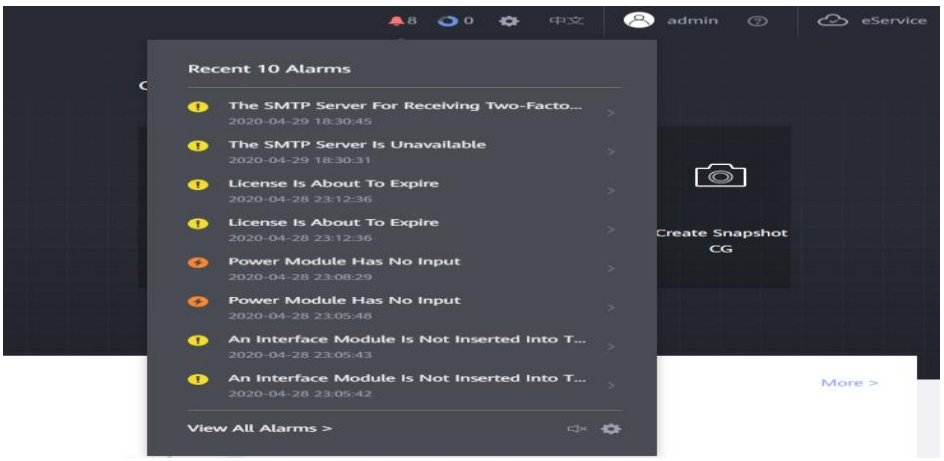
You can further navigate through a specific device frame and its hardware components, and view the device frame in the device hardware view. In the hardware view, you can monitor the real-time status of each hardware component and learn the physical locations of specific hardware components (such as ports and SSDs), facilitating hardware maintenance.

You can query the real-time health status of the SSDs, ports, interface modules, fans, power modules, BBUs, controllers, and disk enclosures. In addition, disks and ports support performance monitoring. You can refer to 10.1.1.5 Performance and Capacity Management for more details.

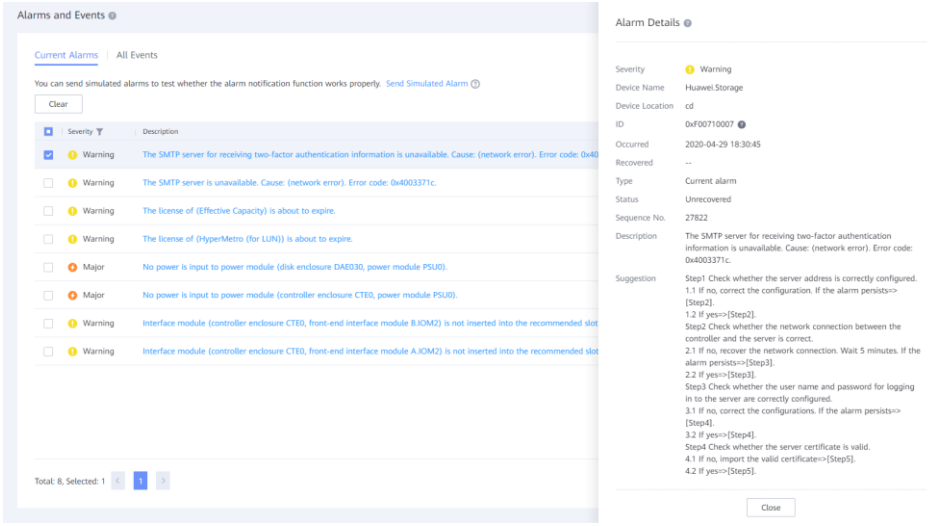
10.1.1.4.2 Alarm and Event Monitoring

This function provides you with real-time fault monitoring information. If a system fault occurs, the fault is pushed to the home page of DeviceManager in real time. Figure 10-6 shows an example.

Figure 10-6 Alarm notification



A dedicated page is available for you to view information about all alarms and events and also provides you with troubleshooting suggestions.



You can also receive alarm and event notifications through syslog, email, and SMS (a dedicated SMS modem is required). You can configure multiple email addresses or mobile phone numbers to receive notifications.

10.1.1.5 Performance and Capacity Management

Performance data collection and analysis are essential to daily device maintenance. Because the performance data volume is large and analyzing the data consumes many system resources, an extra server is often required for installing dedicated performance data collection and analysis software, making performance management complex.

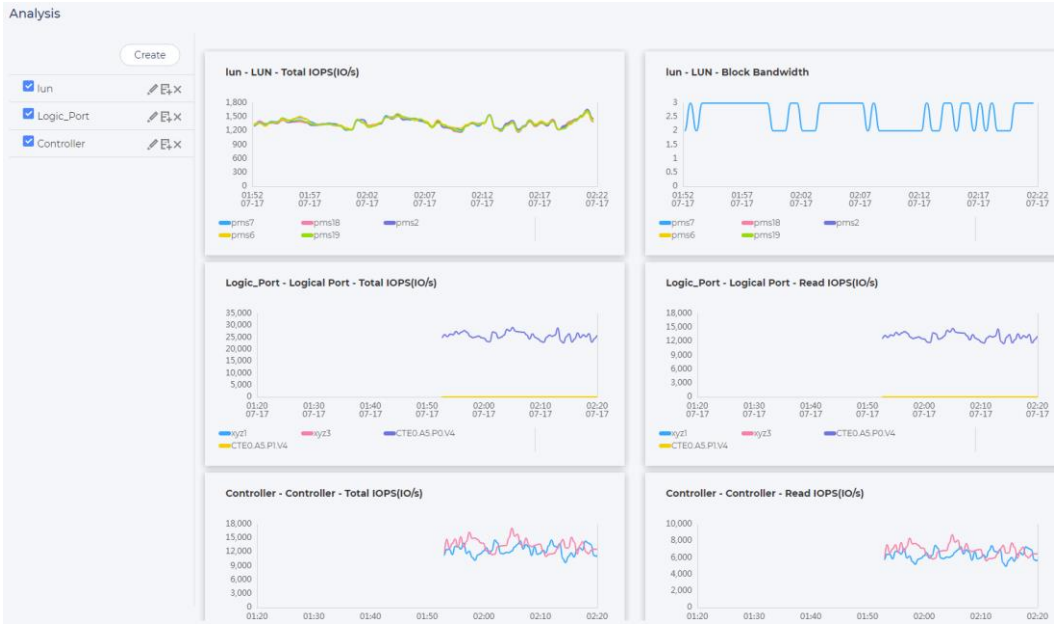
However, the storage system has a built-in performance and capacity data collection and analysis component that is ready for use. The component is specially designed to consume minimal system resources.

In addition, traditional O&M has problems such as delayed risk detection, ineffective communication, and difficulty in prediction. The storage system provides not only the basic performance and capacity collection and analysis functions, it also offers performance and capacity prediction (HyperInsight) based on machine learning, which effectively reduces device running risks and operation costs, and improves the overall O&M efficiency in the lifecycle.

10.1.1.5.1 Built-In Performance Data Collection and Analysis Capabilities

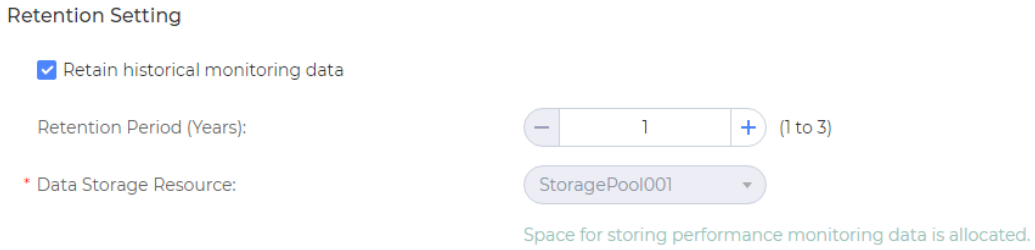
The storage system has built-in performance collection and analysis software. You do not need to install the software separately. You can collect, store, and query historical performance and capacity data of a maximum of the past three years. Alternatively, you can specify the period as required. Statistics and analysis on performance data of controllers, ports, disks, hosts, host groups, LUNs, LUN groups, remote replication, and replication links are displayed from the aspects of bandwidth, IOPS, average I/O response time, and usage.

Different objects and performance indicators can be displayed in the same view to help you analyze performance issues. You can specify the desired performance indicators to analyze the top or bottom objects, so that you can locate overloaded objects more efficiently and tune performance more precisely. The following figure shows performance monitoring. For details about monitoring objects, monitoring indicators, and performance analysis functions, see the online help.



10.1.1.5.2 Independent Data Storage Space

To store collected performance and capacity data, a dedicated storage space is required. DeviceManager provides a dedicated configuration page for users to select a storage pool for storing data.



You can also customize the data retention period. The maximum retention period is three years. DeviceManager automatically calculates the required storage space based on your selection and allocates the required storage space in the storage pool.

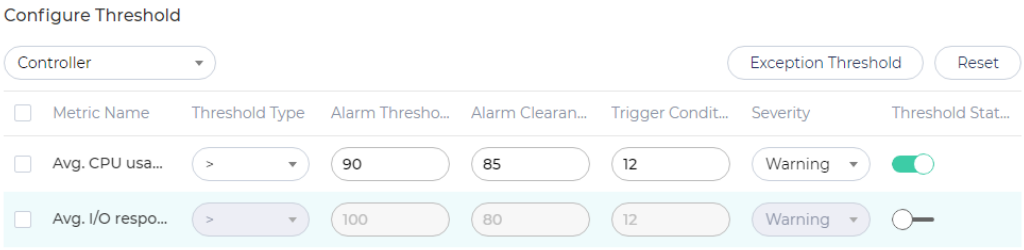
10.1.1.5.3 Capacity Estimation

DeviceManager provides the built-in HyperInsight to predict system and storage pool capacity usage in the next year. This function helps you make a more precise plan for further resource usage and allocate resources more effectively. For details, see Capacity Prediction.

10.1.1.5.4 Performance Threshold Alarm

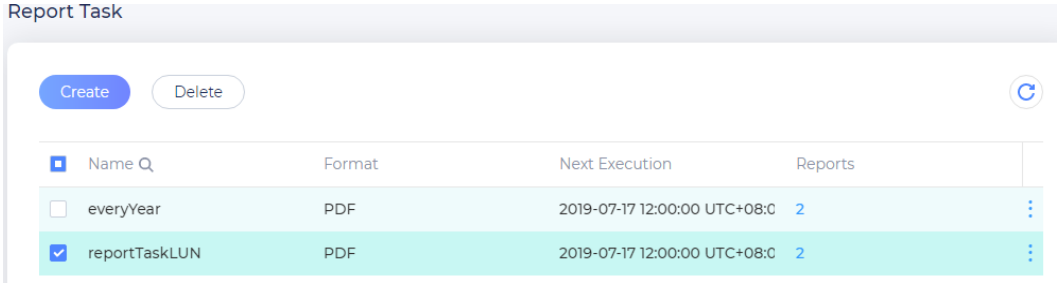
This function allows you to configure threshold alarms for objects such as controllers, ports, LUNs, and replication. The alarm threshold, flapping period, and alarm severity can be customized.

Different storage resources carry different types of services. Therefore, common threshold alarms may not meet requirements. Performance management allows you to set threshold rules for specified objects to ensure high accuracy for threshold alarms.



10.1.1.5.5 Scheduled Report

Performance and capacity reports for specific objects can be generated periodically. Users can learn about the performance and capacity usage of storage devices.



Reports can be generated by day, week, or month. You can set the time when the reports are generated, the time when the reports take effect, and the retention duration of the reports. You can select a report file format. Currently, the PDF and CSV formats are supported.

You can select the objects for which you want to generate a performance report. All the objects for which you want to collect performance statistics can be included in the report. You can also select the performance indicators to be displayed in the report. The capacity report collects statistics on the capacity of the entire system and storage pools.

You can create multiple report tasks. Each report task can be configured with its own parameters. The system automatically generates reports according to the task requirements.

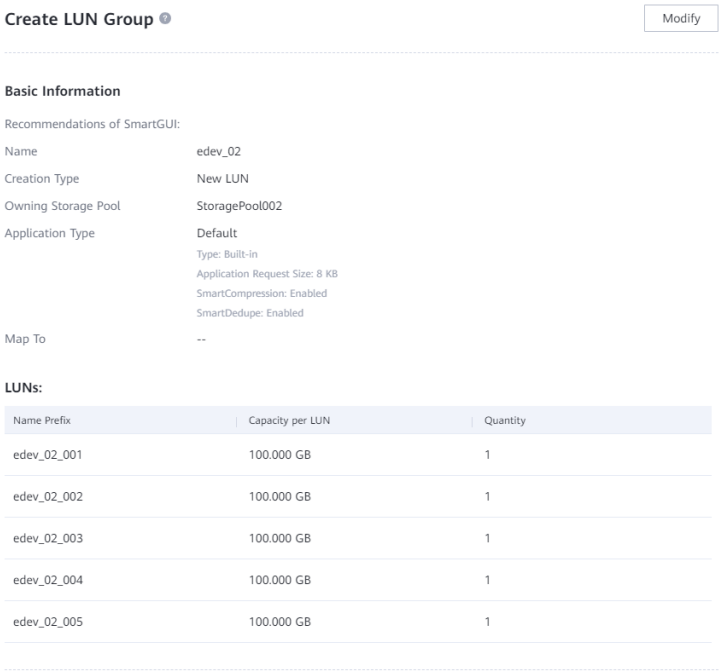
10.1.1.6 AIOps Intelligent O&M

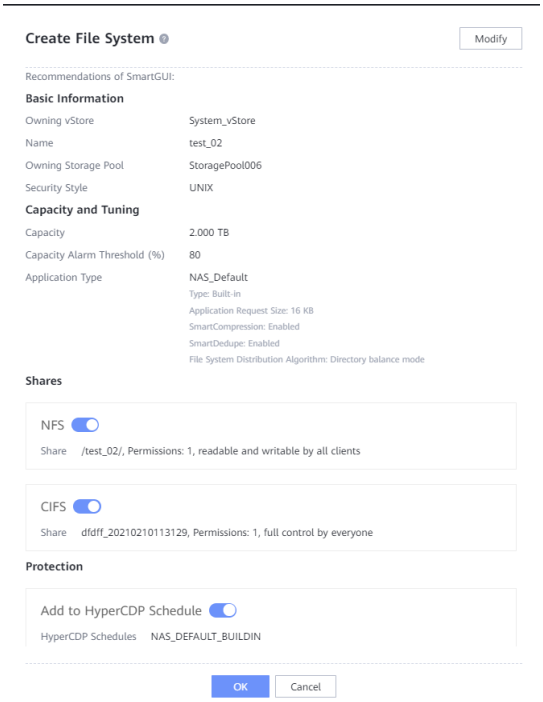
Artificial Intelligence for IT Operations (AIOps) aims to apply artificial intelligence technologies to O&M and build intelligent models based on historical data (such as logs, KPI data, and alarms) to solve the time-consuming and labor-intensive issues during automatic O&M.

OceanStor Dorado continuously builds intelligent O&M capabilities within storage devices. Compared with traditional O&M activities, intelligence technologies will further optimize O&M efficiency, efficient configuration delivery, proactive fault detection, and proper service placement. Compared with intelligent O&M tools on mobile terminals, the built-in intelligence capability of the storage system has advantages in data security. Users do not need to upload data to the cloud. The storage system automatically trains and builds intelligent models.

Intelligent GUI

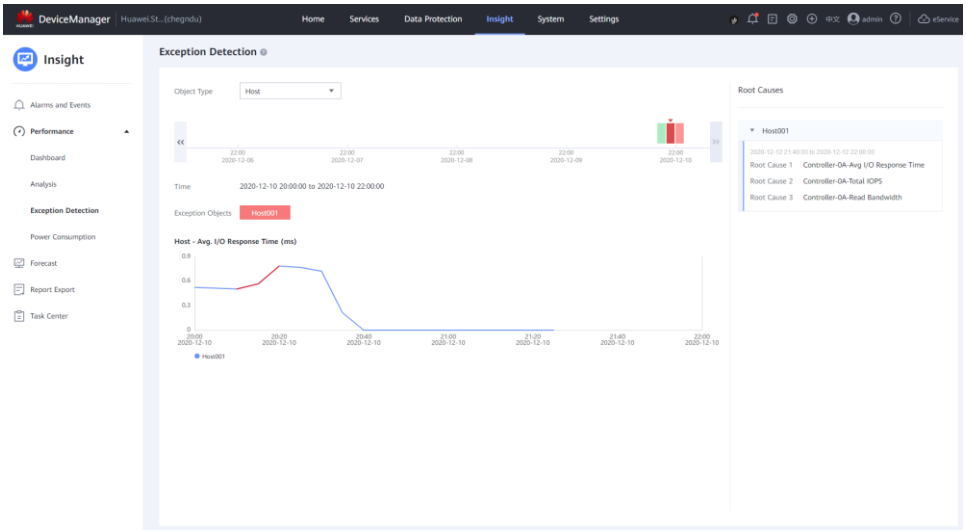
OceanStor Dorado provides intelligent parameter recommendation. Users can deliver SAN and NAS services on DeviceManager with zero parameter configuration. The parameter recommendation accuracy is ≥ 85%. To build the first recommendation feature in storage O&M, OceanStor Dorado breaks through key technologies such as the edge user behavior mining algorithm and recommendation algorithm-oriented performance improvement, and combines big data analysis and statistical learning methods to provide customers with more accurate configuration experience.





Performance Exception Identification

Performance issues are one of the fault scenarios that storage administrators are most concerned about. The sudden drop of IOPS and abnormal increase of latency directly affect service stability. OceanStor Dorado breaks through the unsupervised exception detection of host latency and the intelligent root cause locating algorithm based on hardware performance KPIs. It implements the incremental learning mechanism of the intelligent model. The entire process does not require manual intervention. In the lab environment, the exception identification accuracy is greater than 90%, and the root cause locating accuracy is ≥ 85%. The intelligence technology makes routine O&M more intelligent. The system provides built-in performance analysis to implement proactive prevention, simplifying performance exception analysis and ensuring the health of user devices 24/7.



Capacity Prediction

Capacity is the core asset of storage devices. Sufficient available capacity is the prerequisite for normal service running. Therefore, capacity expansion is one of the key activities in O&M. OceanStor Dorado uses the time series prediction algorithm to construct an intelligent model for storage pools and storage arrays to predict capacity bottlenecks and capacity trend in the next year, providing reference for customers' capacity expansion plans.

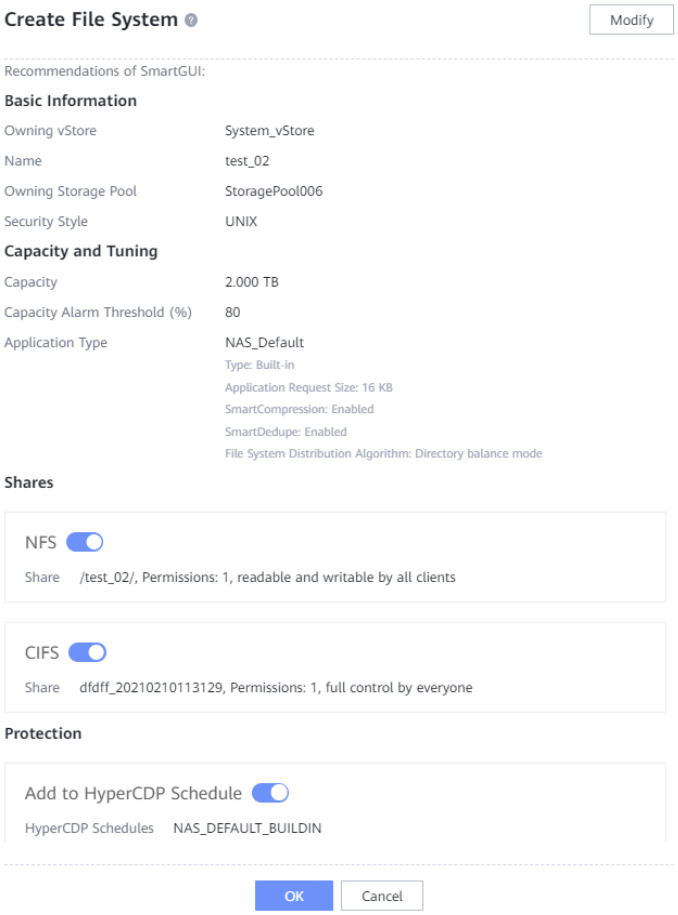
Performance Prediction

Performance indicators are a key reference for measuring the health status of storage devices and for storage administrators to properly provision services. The time series prediction algorithm is built in the storage system. Based on the storage arrays and controllers, the hyperparameter self-optimization mechanism is used to implement automatic learning of intelligent models in the device and predict the performance trend in the next week or month. The indicators include the average latency, total IOPS, and block bandwidth. This helps storage administrators predict performance bottlenecks in advance and properly place or migrate services to ensure stable device performance.

10.1.1.6.1 Intelligent GUI

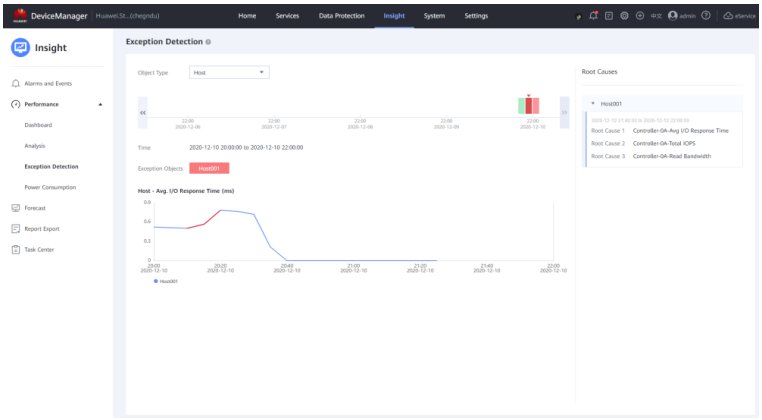
OceanStor Dorado provides intelligent parameter recommendation. Users can use HyperInsight on DeviceManager to deliver SAN and NAS services with zero parameter configuration. The parameter recommendation accuracy is ≥ 85%. To build the first recommendation feature in storage O&M, OceanStor Dorado breaks through key technologies such as the edge user behavior mining algorithm and recommendation algorithm-oriented

performance improvement, and combines big data analysis and statistical learning methods to provide customers with more accurate configuration experience.



10.1.1.6.2 Performance Exception Identification

Performance issues are one of the fault scenarios that storage administrators are most concerned about. The sudden drop of IOPS and abnormal increase of latency directly affect service stability. OceanStor Dorado provides built-in HyperInsight, which breaks through the unsupervised exception detection of host latency and the intelligent root cause locating algorithm based on hardware performance KPIs. It implements the incremental learning mechanism of the intelligent model. The entire process does not require manual intervention. In the lab environment, the exception identification accuracy is greater than 90%, and the root cause locating accuracy is greater than or equal to 85%. The intelligence technology makes routine O&M more intelligent. The system provides built-in performance analysis to implement proactive prevention, simplifying performance exception analysis and ensuring the health of user devices 24/7.

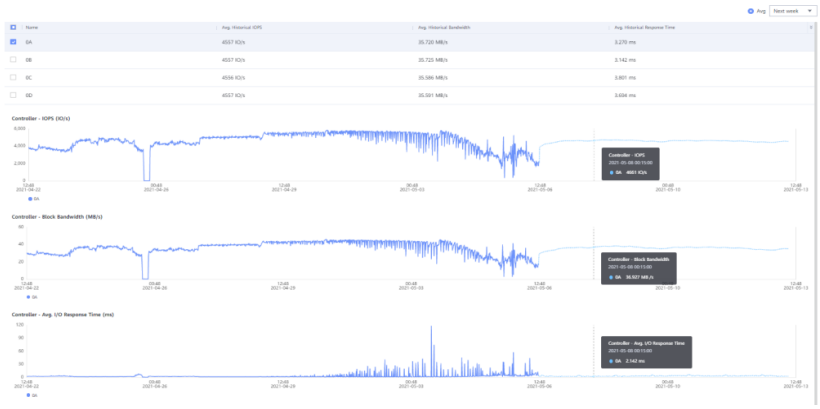


10.1.1.6.3 Capacity Prediction

Capacity is the core asset of storage devices. Sufficient available capacity is the prerequisite for normal service running. Therefore, capacity expansion is one of the key activities in O&M. OceanStor Dorado uses the time series forecasting algorithm to construct an intelligent capacity forecasting model for storage pools and storage arrays. This aims to predict capacity bottlenecks and the trend of the maximum used capacity in the next year, providing reference for customers' capacity expansion plans.

10.1.1.6.4 Performance Prediction

Performance indicators are a key reference for measuring the health status of storage devices and for storage administrators to properly provision services. DeviceManager provides built-in HyperInsight. The time series forecasting algorithm is built in the storage system. Based on the storage arrays and controllers, the hyperparameter self-optimization mechanism is used to implement automatic learning of intelligent models in the device and predict the performance trend in the next week or month. The indicators include the average latency, total IOPS, and block bandwidth. This helps storage administrators predict performance bottlenecks in advance and properly place or migrate services to ensure stable device performance.



10.1.2 CLI

The Command Line Interface (CLI) allows administrators and other system users to manage and maintain the storage system. It is based on the secure shell protocol (SSH) and supports key-based SSH access.

10.1.3 RESTful APIs

RESTful APIs of OceanStor Dorado 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 OceanStor Dorado.

10.1.4 SNMP

The storage system reports alarms and events through SNMP traps.

10.1.5 SMI-S

The Storage Management Initiative Specification (SMI-S) interface is a storage standard management interface developed and maintained by the Storage Networking Industry Association (SNIA). Many storage vendors participate in defining and implementing SMI-S. The SMI-S interface is used to configure storage hardware and services. Storage management software can use this interface to manage storage devices and perform standard management tasks, such as viewing storage hardware, storage resources, and alarm information. OceanStor Dorado supports SMI-S 1.5.0, 1.6.0, and 1.6.1.

10.1.6 Tools

OceanStor Dorado provides diversified tools for pre-sales assessment (eDesigner) and post-sales delivery (SmartKit). These tools effectively help deploy, monitor, analyze, and maintain the storage systems.

10.2 Non-Disruptive Upgrade (NDU)

The storage system upgrade usually requires controller restart and services need to be switched over between controllers to ensure service continuity. This not only greatly affects service performance but also requires a large amount of host information to be collected for evaluating potential upgrade compatibility risks. Host information collection requires host

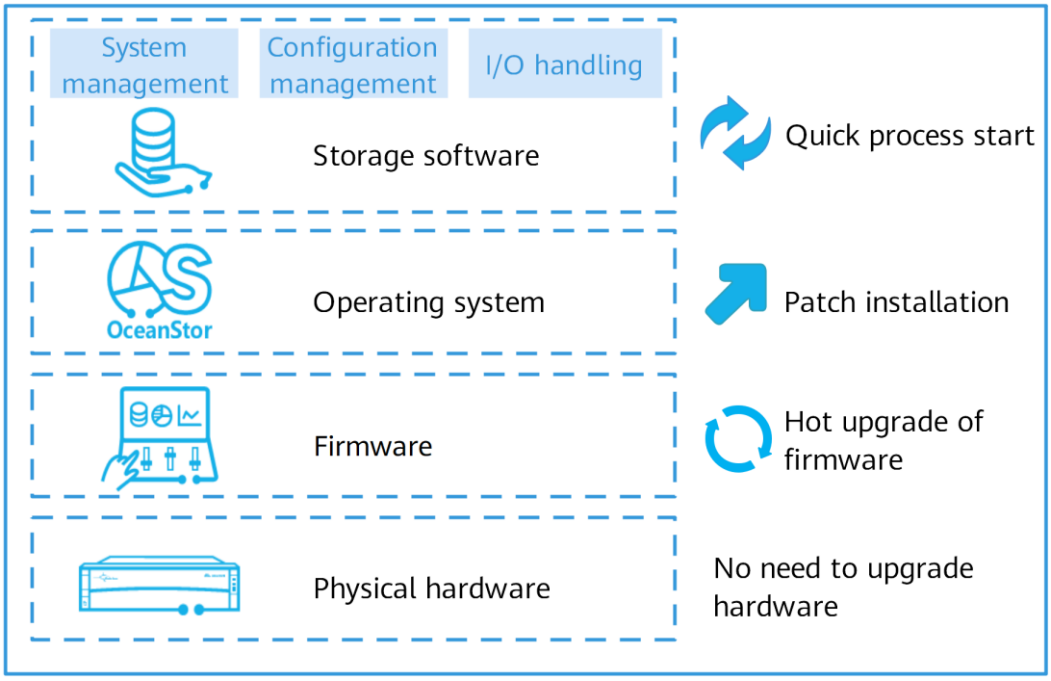
accounts and involves a large amount of information. Certain information cannot be automatically evaluated and demands manual intervention, adding to the complexity of an upgrade.

OceanStor Dorado provides an upgrade method that does not require controller restart. During an upgrade, controllers are not restarted, front-end links are not switched, and services are not affected. In addition, the performance is restored quickly after the upgrade. During an upgrade, the links between the storage system and host are not interrupted and no link switchover is performed, eliminating the need for collecting host information and avoiding compatibility risks caused by path switchover.

Component-based Upgrade

A storage system can be divided into four kinds of components: physical hardware, firmware, operating system, and storage software. Each kind is upgraded in different ways to complete the system upgrade in OceanStor Dorado, as shown in Figure 10-7.

Figure 10-7 Component-based upgrade



- Physical hardware does not need to be upgraded.
- Firmware, including the firmware of BIOS, CPLD, and interface modules, supports hot upgrade without restarting controllers.
- The operating system is upgraded by installing hot patches.
- Storage software is user-mode processes. Earlier processes are killed and new processes are started using upgraded codes, which are completed within seconds.

The component-based upgrade and front-end connection keepalive techniques eliminate the need for controller restart and front-end link switchover, leaving host services not affected.

Connection Keepalive

In scenarios where front-end interconnect I/O modules are used, links between the host and storage system are established over the I/O modules. After receiving service requests from the host, the I/O modules distribute the requests to the controller. Host I/Os received during the restart of service processes in the controller are stored in the I/O modules and then distributed to the controller after service processes run properly. The restart of service processes takes a very short time (1 to 2 seconds). In this way, I/Os issued by the host do not time out and the host is unaware of the restart.

In scenarios where front-end interconnect I/O modules are not used, the daemon process keeps the links between the controller and host connected during service process startup. The restart of service processes takes a very short time (1 to 2 seconds). In this way, I/Os issued by the host do not time out and the host is unaware of the restart.

Zero Performance Loss and Short Upgrade Time

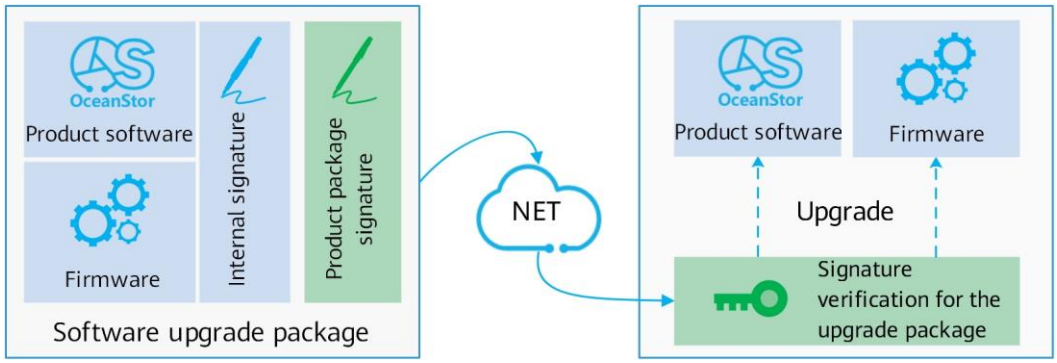
OceanStor Dorado does not involve service switchover. Therefore, an upgrade has little impact on host performance and service performance can restore to 100% within 2 seconds. The upgrade process does not demand controller restart or link switchover. You do not need to collect host information or perform compatibility evaluation. The end-to-end upgrade process (from importing packages to upgrade completion) takes less than 30 minutes, 10 minutes in general. The upgrade of the I/O handling process costs only 10 seconds, affecting services for only 10s. This ensures that the upgrade has the minimum impact on the storage system.

11 System Security Design

Storage security must be safeguarded by technical measures. Data integrity, confidentiality, and availability must be monitored. Secure boot and access permission control as well as security policies based on specific security threats of storage devices and networks further enhance system security. All these measures prevent unauthorized access to storage resources and data. Storage security consists of device security, network security, service security, and management security. This chapter describes the software integrity protection, secure boot, and data encryption capabilities related to system security. The digital signature technology ensures that the product package (including the upgrade package) developed by Huawei is not tampered with during device installation and upgrade. The secure boot technology guarantees that the startup components are verified during the startup of storage devices to prevent the startup files from being tampered with. The disk encryption feature is used to protect data stored on disks and prevent data loss caused by disk loss.

11.1 Software Integrity Protection

Figure 11-1 Software integrity protection

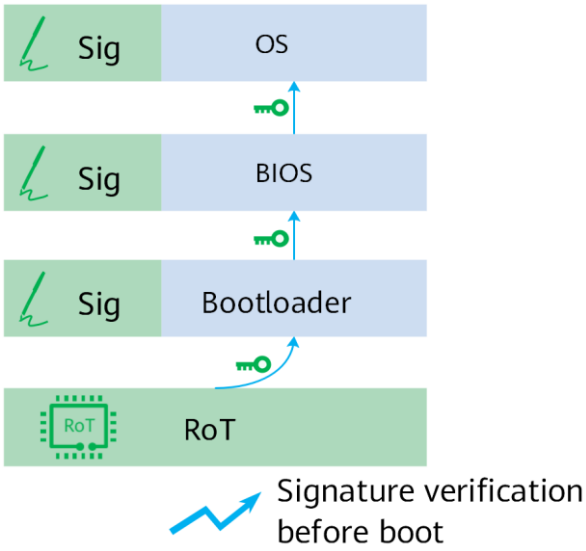


The product software package, upgrade package, and firmware package may be tampered with during the time of waiting for onsite R&D engineers. The digital signature of the product software package is used to protect the integrity of the upgrade package used during product development and onsite deployment. The software package uses an internal digital signature

and a product package digital signature. After the software package is sent to the customer over the network, the upgrade module of the storage system verifies the digital signature and performs the upgrade only after the verification is successful. This ensures the integrity and uniqueness of the upgrade package and internal software modules.

11.2 Secure Boot

Figure 11-2 Secure boot



After the device is powered on, the initial startup module starts and verification is performed level by level. If the verification is successful, the device starts. Digital signatures are used to verify firmware integrity to prevent firmware and operating systems from being tampered with. The related technologies are as follows:

- The root of trust (RoT) is integrated into the CPU to prevent software and physical attacks, providing the highest level of security in the industry.
- Software integrity is ensured by two levels of digital signatures (root key + level-2 key) and software uniqueness is ensured by digital certificates.
- The RSA 2048/4096 algorithm is used, which has the top security level in the industry.
- Level-2 keys (code signing keys) can be revoked.
- The built-in RoT of the CPU can prevent malicious tampering, such as tampering of flash firmware outside the ARM and replacement of the system disk.

12 Intelligent Storage Design

OceanStor Dorado uses intelligent chips to build an intelligent storage system and cloud system.

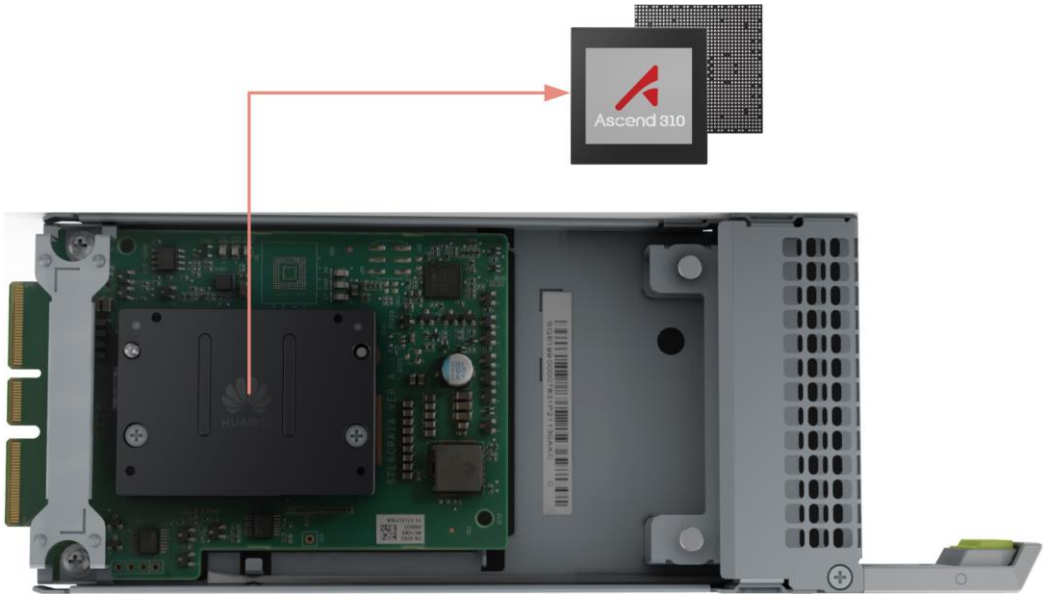
12.1 12.2 Intelligent Cloud ManagementIntelligent Storage

OceanStor Dorado relies on the powerful compute capability provided by intelligent accelerator cards. The deep learning algorithm is used to learn service load rules and predict service behavior, making the read cache and QoS control more intelligent.

12.1.1 Intelligent Cards

Currently, the mainstream intelligence technology is characterized as deep learning, and the mainstream deep learning model is various neural networks such as deep neural network (DNN), recurrent neural network (RNN), and convolutional neural network (CNN). The most basic mathematical operation in a neural network is matrix multiplication and addition. For example, if 64x64 is multiplied by an FP16 (floating point 16) matrix of 64x32, a single operation requires computing power of 64 x 64 x 32 = 131,072 FLOPS@FP16, and a common CPU core can perform floating point operations for only dozens of times in a clock cycle. In this way, an operation requires thousands of clock cycles. For the neural network model with teraFLOPS (TFLOPS), the efficiency of conventional CPUs cannot satisfy its requirements.

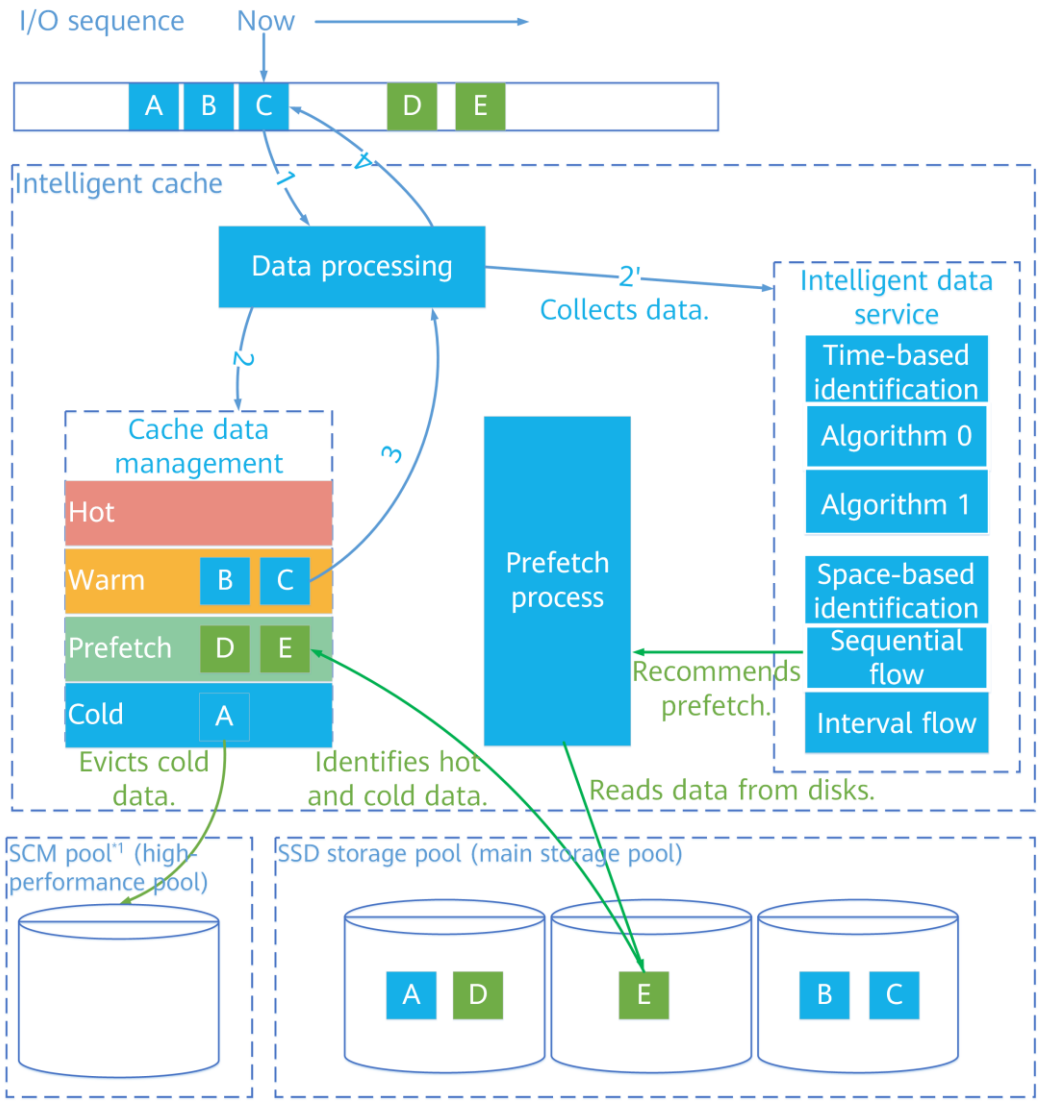
Figure 12-1 Intelligent accelerator card



12.1.2 Intelligent Cache and Tiering

OceanStor Dorado, used with intelligent accelerator cards and storage class memory (SCM) accelerator cards, implements intelligent read cache and data tiering, effectively mining users' data access models and optimizing system I/O stack processing. Figure 12-2 shows the intelligent read cache process.

Figure 12-2 Intelligent read cache process



- After a host delivers a read I/O (for data C) to a storage array, the process is as follows:
  a. The I/O enters the read process of the cache module to have concurrency and quotas processed.
  b. The read I/O searches the cache data management module for the data and sends the address information to the intelligent data service for background analysis.
  c. Data C is hit in the cache data management module and is returned to the process flow.
  d. After receiving data C, the read process assembles and returns it to the host.
- Intelligent data service process:
  a. After receiving the I/O metadata (C) from the foreground, the service identifies the corresponding prefetch mode.
  b. The intelligent prefetch algorithms analyze historical data (A, B, and C) and recommend prefetching data (D and E).
  c. The recommended prefetch data (D and E) is read from the main storage pool.

d. The read data (D and E) is stored in the cache data management module.

- Intelligent elimination process:
  a. In the cache, multi-level data management is implemented based on data importance. Cold data is automatically eliminated in case of memory insufficiency.
  b. Cold data is preferentially eliminated to and managed by the SCM storage pool of the underlying storage system.

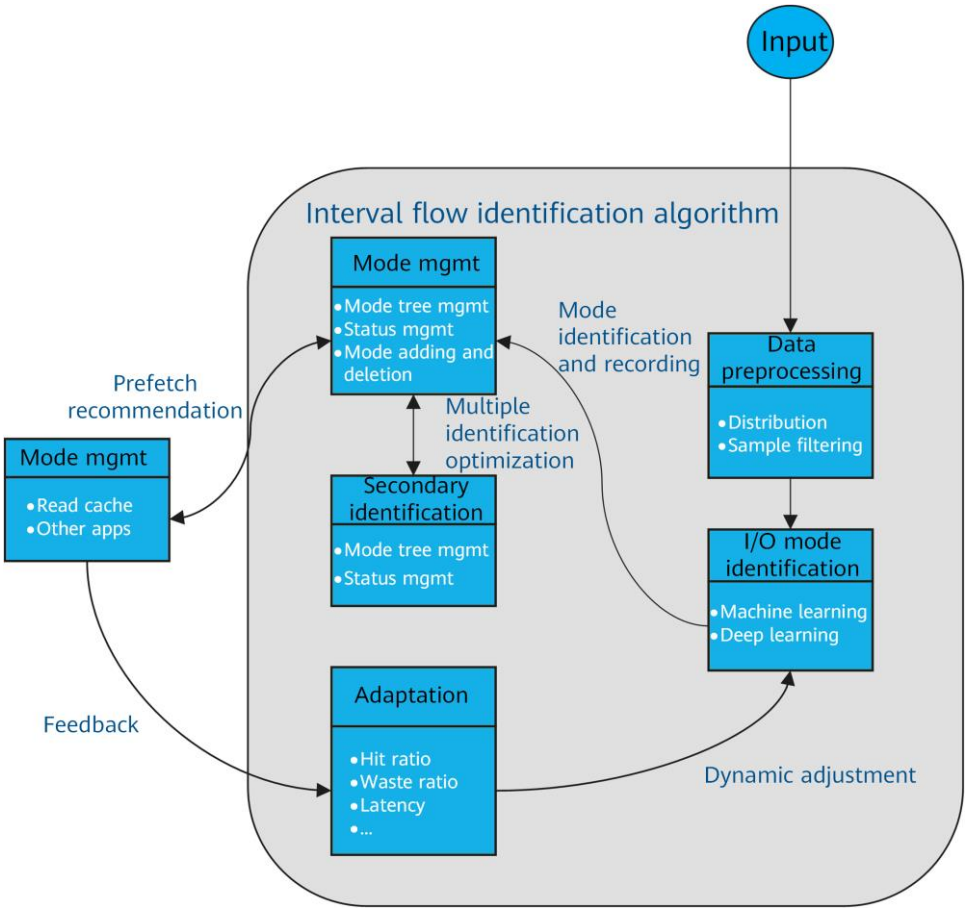
Spatial Prefetch

OceanStor Dorado enhances spatial association mining on data reading rules. On the basis of the more robust algorithm for interval flows, Huawei further researches and explores the method of reading the interval flows in certain batch services. Currently, the method can be accurately identified, and accurate prefetch can be performed in different complex interval flow scenarios such as multi-flow and multi-interval.

Identification of the multi-flow read mode: Services can access data in different address spaces. The read mode in different data spaces can be mined and distinguished by using the traffic distribution technology.

- Identification of the multi-interval read mode: The OceanStor Dorado prefetch algorithm can identify the data read modes of services and prefetch data, regardless of whether data is read from consecutive address spaces, from a single interval, or from multiple complex regular intervals.
- Identification of the multi-flow and multi-interval mode: Services can read data from different address spaces at different intervals. The OceanStor Dorado algorithm can effectively identify the data read modes of services and prefetch data.
- Mode management: The existing read modes of services are saved, reducing secondary learning costs and speeding up prefetch. When the read mode is changed, the original mode is replaced with the new one. Continuous spaces using the same mode can be combined.
- Dynamic mode adjustment: Prefetch effect feedback is provided based on key read performance indicators such as the prefetch hit ratio, waste rate, and latency to drive the deep learning algorithm to continue analyzing and mining new modes. This adjustment minimizes the impact of service mode changes on system performance.

Figure 12-3 Intelligent spatial prefetch process



Temporal Prefetch

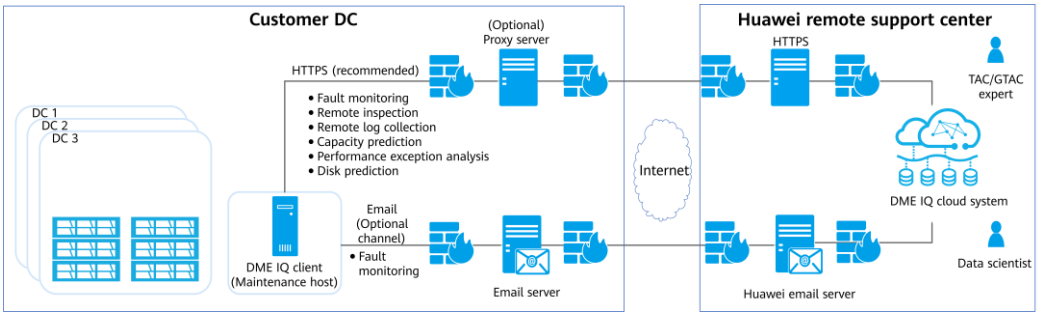
OceanStor Dorado enhances the hot data identification algorithm and mines data access time characteristics. At present, the semantic correlation feature of data is accurately mined, and hot data is effectively identified.

- I/O association mode mining: Association analysis is performed on the data read modes of services to identify the inherent read modes of the services and the semantic correlation of address spaces.
- I/O association mode management: The mined data association modes are saved in the memory and disks, and can be added, dynamically eliminated, or even deleted.
- Prefetch setting recommendation: Parameters such as the associated read threshold and recommended prefetch data quantity are set and modified based on service requirements.
- Dynamic mode adjustment: Prefetch effect feedback is provided based on key read performance indicators such as the prefetch hit ratio, waste rate, and latency to drive the deep learning algorithm to continue analyzing and mining new modes. This adjustment minimizes the impact of service mode changes on system performance.
- Hot data identification: Recent hot data is identified according to recent data access modes. In addition, data can be retained and evicted in the cache, facilitating quick access to hot data and reducing extra performance overhead caused by repeated prefetches.

12.2 Intelligent Cloud Management

In traditional service support mode, technical support personnel provide services manually. Faults may not be detected in a timely manner and information may not be delivered correctly. To resolve the preceding problems, Huawei provides the DME IQ cloud intelligent management system (DME IQ for short) based on the cloud native. With the customer's authorization, device alarms and logs are sent to DME IQ at a scheduled time every day. Based on artificial intelligence technologies, DME IQ implements intelligent fault reporting, real-time health analysis, and intelligent fault prevention to identify potential risks, automatically locate faults, and provide troubleshooting solutions. DME IQ remote assistance allows authorized technical service personnel to log in to the device remotely for technical support, minimizing device running risks and reducing operation costs.

Figure 12-4 Typical DME IQ networking



DME IQ enables the client to work with the cloud system.

- The client is deployed on the customer side.

  DME IQ client is used or DME IQ service is enabled on DeviceManager to connect to the DME IQ cloud system. Alarm information about customer devices is collected and sent to the Huawei cloud system in a timely manner.

- The cloud system is deployed in Huawei remote support center.

  DME IQ receives device alarms from the customer client all day long, automatically reports the problem to the Huawei remote support center, and creates the corresponding service request. Huawei service engineers will assist the customer to resolve the problem in time.

DME IQ has the following advantages:

- DME IQ provides a self-service O&M system for customers, aiming for precise personalized information services.
- Customers can use a web to access DME IQ to view device information anytime anywhere.
- High data security and reliability are ensured. Secure information transmission is provided and DME IQ can access the customer's system only after being authorized by the customer.
- DME IQ provides 24/7 secure, reliable, and proactive O&M services. SRs can be automatically created.
- Based on Huawei Cloud, the DME IQ cloud system drives O&M activities through big data analytics and artificial intelligence to identify faults in advance, reduce O&M difficulties, and improve O&M efficiency.

12.2.1 Scope of Information to Be Collected

With the authorization of customers, Huawei storage systems can be connected to DME IQ through a network to periodically collect their O&M data, helping fully understand storage O&M activities. The O&M data includes performance data, configuration information, alarm information, system logs, and disk information.

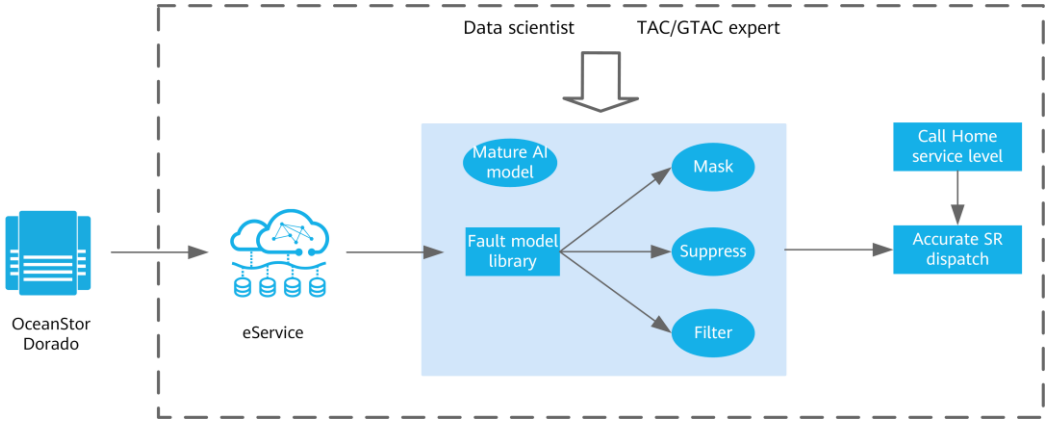
Table 12-1 Scope of Information to Be Collected

Data Type	Description	Interval of Data Upload
Performance data	A **.txt** file in the JSON format	Uploaded automatically. The new performance data is uploaded to the DME IQ cloud system every 5 minutes.
Configuration information	A **.txt** file	Uploaded automatically. The configuration is uploaded to the DME IQ cloud system once a day.
Alarm information	HTTPS message	Uploaded automatically. The new device alarm messages are uploaded to the DME IQ cloud system every 30 seconds.
System logs	Email	Uploaded automatically. The new alarm messages are uploaded to the DME IQ cloud system every 5 minutes.
	A **.tgz** file	Manually uploaded by Huawei technical support personnel on the DME IQ cloud system. In the current version, all system logs and system logs in the latest one hour, latest two hours, latest 24 hours, or a specific time period can be uploaded.
Hardware information	A **.txt** file	Uploaded automatically. The disk information is uploaded to the DME IQ cloud system once a day.

12.2.2 Intelligent Fault Reporting

DME IQ provides 24/7 health reporting. If a device fails, DME IQ is automatically notified. Traditional fault reporting mechanisms have difficulties in covering all scenarios and have problems such as false alarm reporting and alarm missing. DME IQ provides 24/7 active monitoring for customer device alarms. The alarms generated by the device are reported to DME IQ. Based on the fault feature model library of global devices, DME IQ performs automatic alarm masking to filter redundant alarms, improving the accuracy and efficiency of alarm handling. Based on the service level, DME IQ can automatically create an SR and send it to the corresponding Huawei engineer for problem processing. At the same time, DME IQ notifies the customer of the problem by using the pre-agreed method (email by default) to facilitate troubleshooting.

Figure 12-5 Working principles of intelligent fault reporting



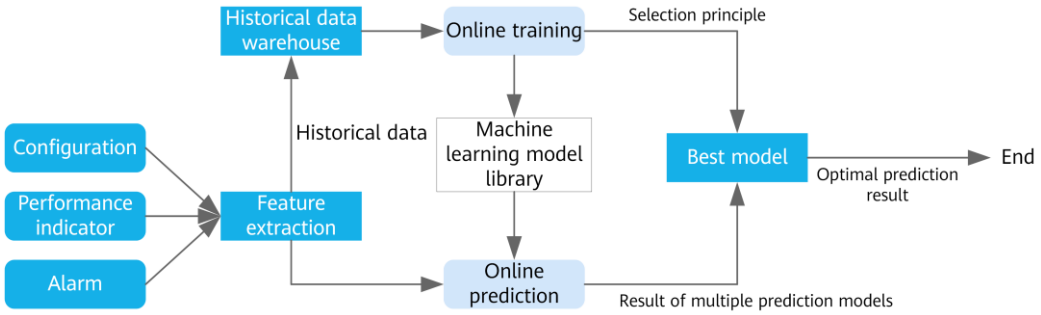
12.2.3 Capacity Prediction

System capacity changes are affected by multiple factors. The traditional single prediction algorithm cannot ensure the accuracy of prediction results. DME IQ ensures the rationality and accuracy of prediction results from the following aspects:

- DME IQ uses multiple prediction model clusters for online prediction, outputs the prediction results of multiple models, and then selects the best prediction results based on the selection rules recommended by online prediction. At the same time, based on the historical capacity data, DME IQ periodically trains and verifies itself in the light of linear prediction, recognizable trends, periods, and partial changes to optimize the model parameters, ensuring that the optimal prediction algorithm is selected.
- The DME IQ prediction algorithm model can accurately identify various factors that affect capacity changes, for example, sudden capacity increase and decrease caused by major events, irregular trend caused by capacity reclamation of existing services, and capacity hops caused by new service rollout. In this way, the system capacity consumption can be predicted more accurately.

DME IQ selects the best prediction model using the intelligent algorithm, and predicts the capacity consumption in the next 12 months. Based on the capacity prediction algorithm, DME IQ provides the overloaded resource warning, capacity expansion suggestions for existing services, and annual capacity planning functions for customers.

Figure 12-6 Working principles of capacity prediction



Responsibilities of each component:

- Data source collection: collects configurations, performance indicators, and alarm information to reduce the interference of multiple factors on machine learning and training results.
- Feature extraction: uses algorithms to transform and extract features automatically.
- Historical database warehouse: stores historical capacity data of the latest year.
- Online training
  - Uses a large number of samples for training to obtain the measurement indicator statistics predicted by each model and outputs the model selection rules.
  - For the current historical data, performs iteration for a limited number of times to optimize the model algorithm.
- Machine learning model library: includes ARIMA, fbprophet, and linear prediction models.
- Online prediction: performs prediction online using the optimized models and output the prediction results of multiple models and mean absolute percentage error (MAPE) values of measurement indexes.

$$\mathrm{M} = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

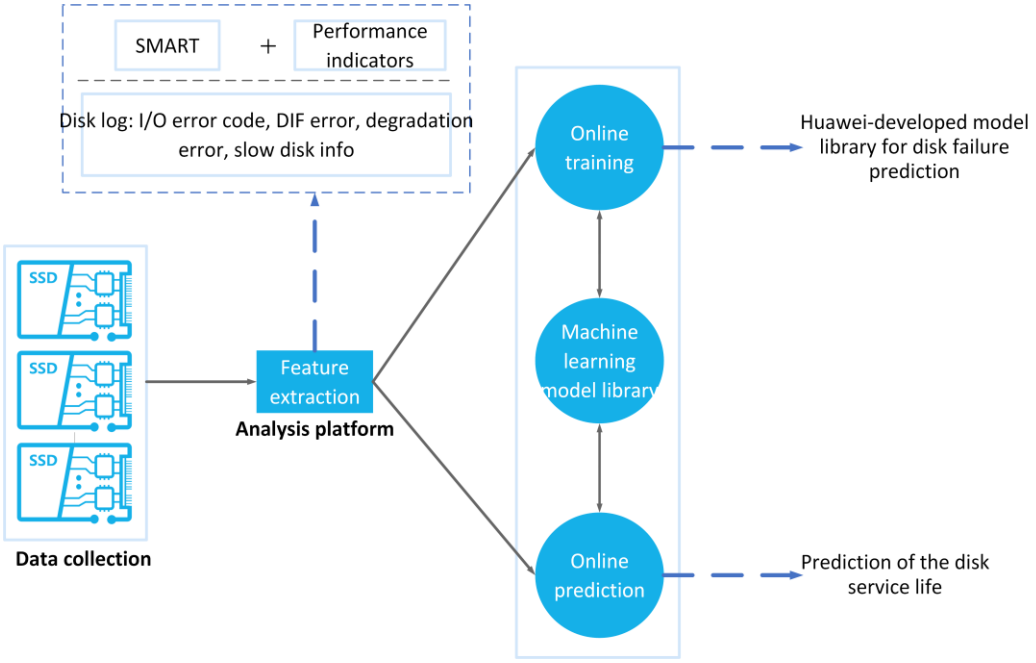
  In the preceding formula, **At** indicates the actual capacity and **Ft** indicates the predicted capacity.
- Best model selection: weights the model statistics of online training and results of online prediction, and selects the optimal prediction results.

12.2.4 Disk Health Prediction

Disks are the basis of a storage system. Although various redundancy technologies are used in storage systems, they allow the failure of a limited number of disks while ensuring service running. For example, RAID 5 allows only one disk to fail. When two disks fail, the storage system stops providing services to ensure data reliability. Disks are the largest consumables in a storage system. Therefore, disk service life is the most concerned topic for many users. SSDs are electronic components and their service life prediction indicators are limited. In addition, the number of read and write requests delivered by service varies every day, which further complicates disk service life prediction.

DME IQ collects the Self-Monitoring, Analysis and Reporting Technology (S.M.A.R.T.) information and I/O link information of SSDs, as well as reliability indicators of SSDs, and enters such information to hundreds of disk failure prediction models, implementing accurate SSD service life prediction. DME IQ uses intelligent algorithms to predict SSD risks to detect failed SSDs and replace risky SSDs in advance, preventing faults and improving system reliability.

Figure 12-7 Working principles of disk health prediction



- Data collection

  The disk vendors provide the S.M.A.R.T. data of disks. The S.M.A.R.T. data can indicate the running status of the disks and help predict risky disks in a certain range. However, it is difficult to ensure the accuracy of prediction results. DME IQ uses intelligence technologies to dynamically analyze S.M.A.R.T. changes of disks, performance indicator fluctuations, and disk logs, ensuring more accurate prediction results.

- S.M.A.R.T

  For SSDs, the interfaces provide SCSI log page information that records the current disk status and performance indicators, such as the grown defect list, non-medium error, and read/write/verify uncorrected errors.

- Performance indicators

  Workload information such as the average I/O size distribution per minute, IOPS, bandwidth, and number of bytes processed per day, and performance indicators such as the latency and average service time

- Disk log

  I/O error codes collected by Huawei storage systems, DIF errors, degradation errors, slow disk information, slow disk cycles, and disk service life

- Feature extraction

  Based on massive amounts of historical big data, feature transformation and feature extraction are automatically performed by using the algorithm.

- Analysis platform
  - Online training: performs trainings based on the model algorithms, and perform iteration for a limited number of times to optimize the model algorithms.
  - Machine learning model library: disk failure prediction model.
  - Online prediction: uses the optimized training models to predict disk failures.

- Prediction result

DME IQ tests and verifies massive amounts of SSD data and accurately predicts SSD service life based on the disk failure prediction model.

12.2.5 Device Health Evaluation

Generally, after a device goes online, the customer performs inspections to prevent device risks, which has two disadvantages:

- Inspection frequency

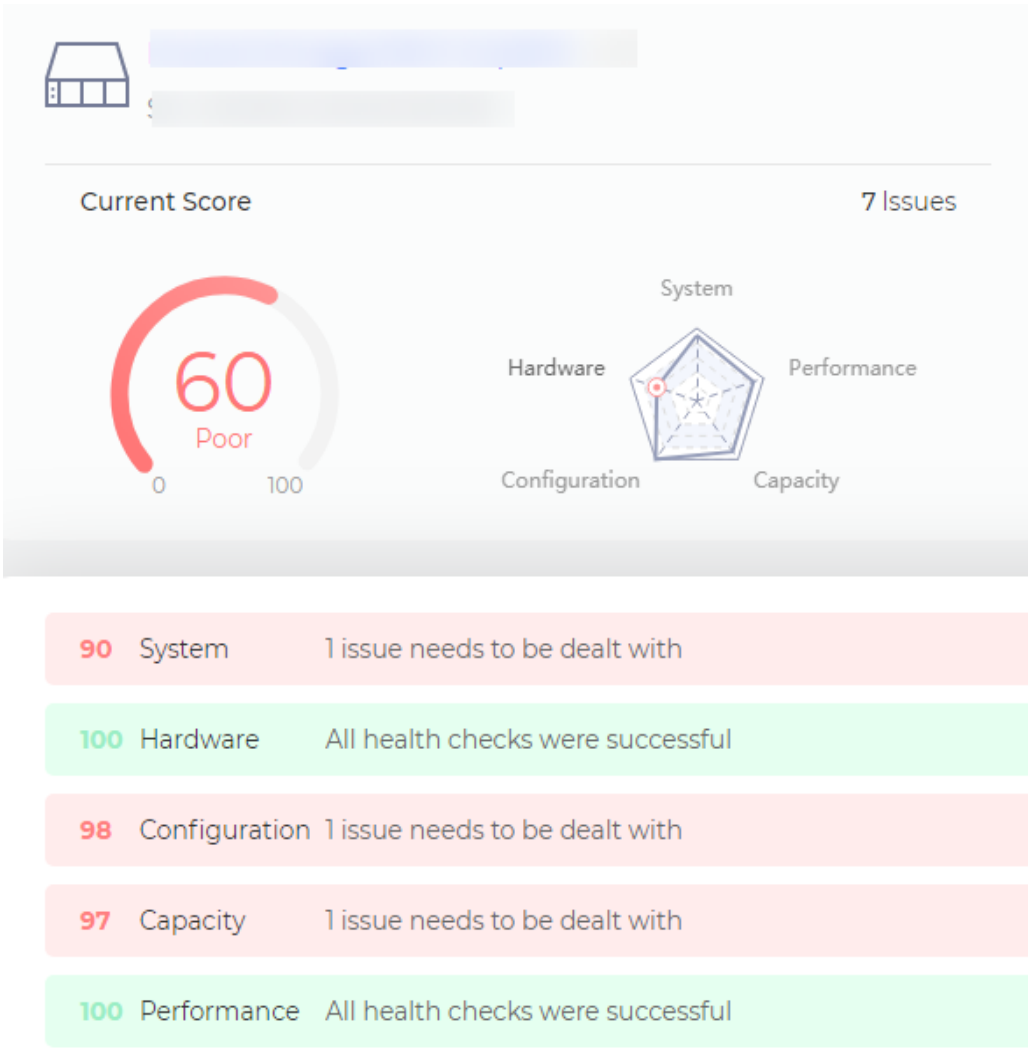
  Inspections are performed monthly or quarterly. As a result, customers cannot detect problems in a timely manner.

- Inspection depth

  The customer can only check whether the current device is faulty. The system, hardware, configuration, capacity, and performance risks are not analyzed.

DME IQ evaluates device health in real time from the system, hardware, configuration, capacity, and performance dimensions, detects potential risks, and displays device running status based on health scores. In addition, DME IQ provides solutions for customers to prevent risks.

Figure 12-8 Device health evaluation details

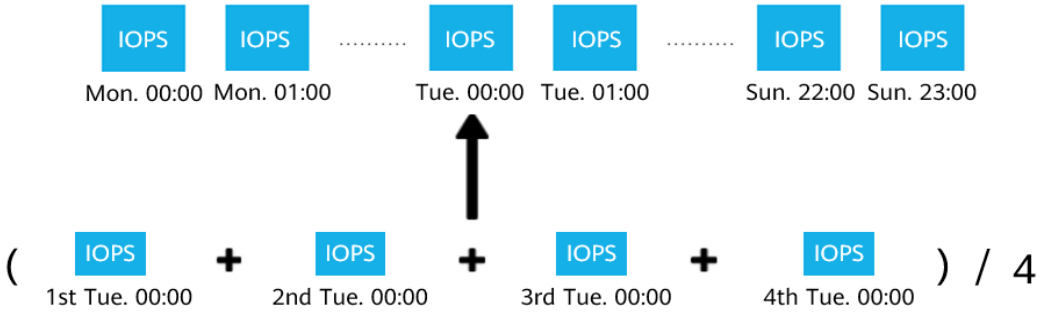


12.2.6 Performance Fluctuation Analysis

Periodic service operations (such as scheduled snapshot or SmartTier) or temporary changes (such as online upgrade, capacity expansion, and spare parts replacement) should be performed during off-peak hours to avoid affecting online services. In the past, O&M personnel estimate the proper time window according to experiences or performance indicators of a past period of time.

Based on historical device performance data, DME IQ analyzes performance fluctuations from the dimensions of load, IOPS, bandwidth, and latency. Users can view the service period patterns from the four dimensions and select a proper time window to perform periodic operations on services (such as scheduled snapshot) or temporary service changes (such as online upgrade, capacity expansion, and spare parts replacement) to prevent impacts on services during peak hours.

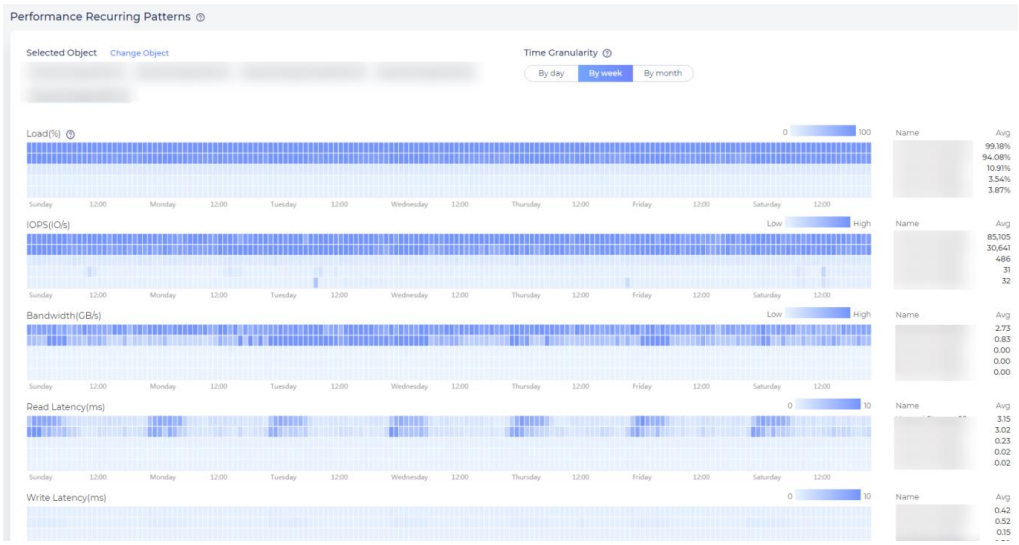
Figure 12-9 Working principles of weekly performance fluctuation analysis



DME IQ calculates the performance indicator values of each hour from Monday to Sunday based on the device performance data in the past four weeks. For example, the IOPS is calculated as follows: Calculate the sum of the IOPS on each hour in the past four weeks and then divide the sum by 4 to obtain the weekly performance fluctuation patters. For daily and monthly calculation, the methods are similar.

Users can view the service performance statistics by day, week, or month as required, as shown in the following figure.

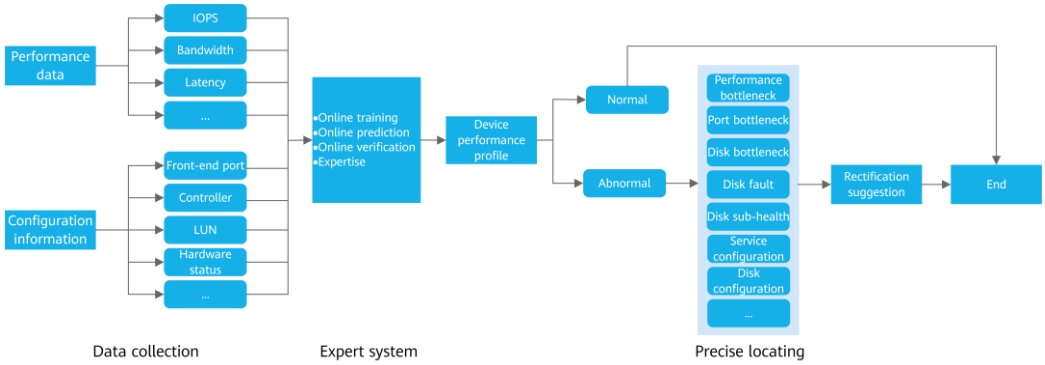
Figure 12-10 Weekly performance fluctuation



12.2.7 Performance Exception Detection

The biggest concern of enterprises is whether services can run smoothly. However, because the performance problems are complicated and difficult to identify and solve in advance, the problems are detected until they get worse, affecting the services and causing losses to enterprises. The performance exception detection function is provided. For service latency, the deep learning algorithm is used to learn service characteristics from historical performance data. Combining service characteristics with the industry and Huawei expertise, DME IQ obtains the device performance profiles which show real-time exceptions, precisely locate faults, and provide rectification suggestions.

Figure 12-11 Working principles of performance exception detection

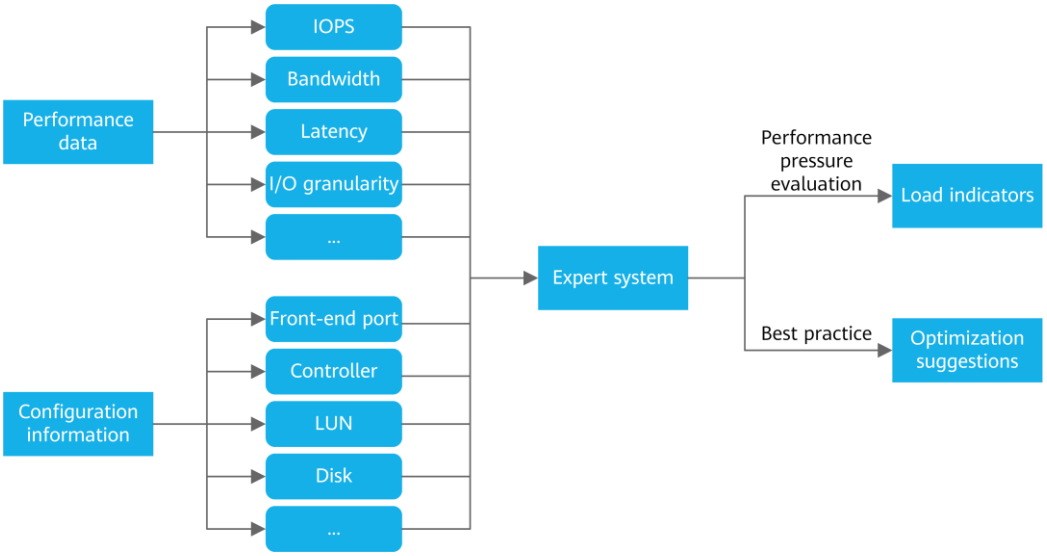


12.2.8 Performance Bottleneck Analysis

After a service goes online, O&M personnel are concerned about whether the device has high performance pressure and whether the service can run properly. Due to complicated factors that affect device performance, such as hardware configuration, software configuration, service type, and performance data, multiple performance indicators need to be compared and analyzed concurrently and manually. Therefore, performance pressure evaluation and performance optimization become big challenges.

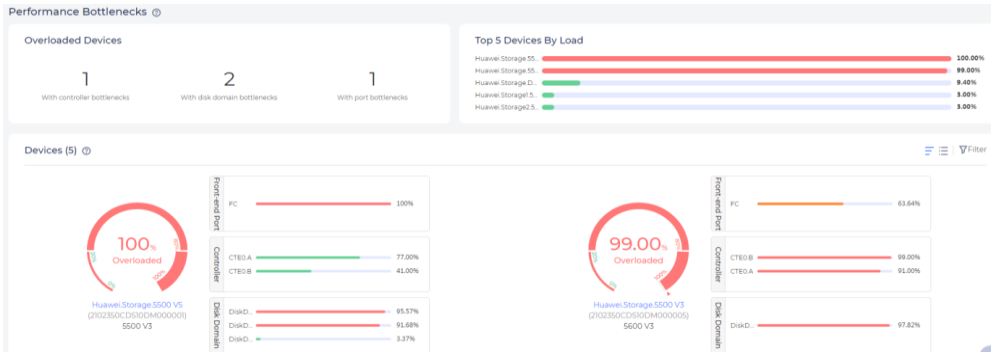
DME IQ performance bottleneck analysis covers device configurations and performance data. DME IQ evaluates device performance pressure, provides clear overall device loads and loads on each component based on Huawei expertise, identifies performance bottlenecks, and provides optimization suggestions. Customers can make adjustment based on suggestions for the performance bottleneck to ensure stable service running.

Figure 12-12 Working principles of performance bottleneck analysis



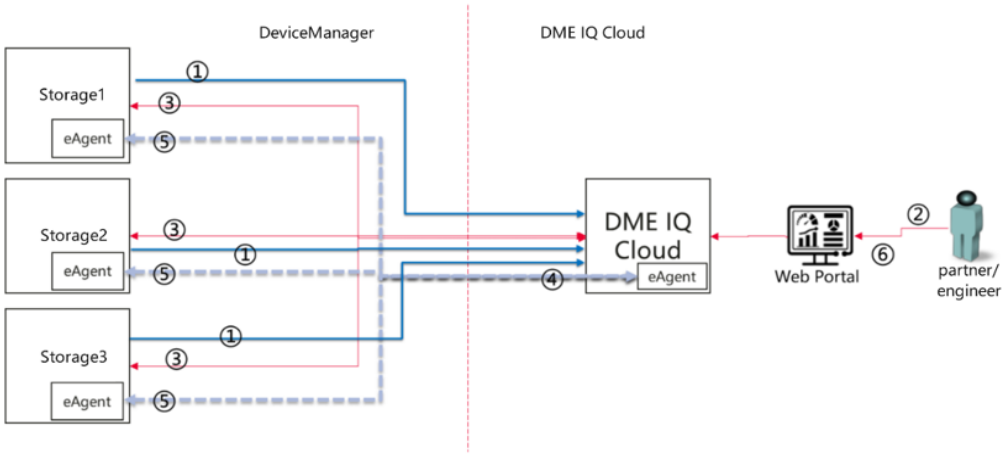
Users can view the overall device loads and loads on each component, as shown in Figure 12-13.

Figure 12-13 Performance bottleneck analysis



12.2.9 Remote Assistance

Service personnel can use the DME IQ remote assistance function to help customers troubleshoot simple device faults, analyze performance, configure storage services, and expand disk or storage pool capacity without visiting sites.



Based on the existing solution of directly connecting the storage system to DME IQ, a persistent bidirectional channel can be established between the storage system and DME IQ cloud to implement remote operations.

1. Complete DME IQ setting on the storage side and enable the remote assistant function.

2. A service personnel initiates a remote assistance request from the DME IQ cloud.

4. The storage system creates a persistent connection channel with the DME IQ cloud for real-time communication.

5. The DeviceManager page is extended to the cloud based on the persistent connection channel.

6. The service personnel performs authorized remote operations on the cloud.

13 Ecosystem Compatibility

IT ecosystem infrastructure includes hardware (servers, network devices, and storage devices) and software (virtualization systems, operating systems, cluster software, and management and control software) and different software and hardware products must be compatible with each other. As core infrastructure, OceanStor Dorado all-flash storage is compatible with IT software and hardware products of different vendors, types, and versions.

OceanStor Dorado all-flash storage supports a large number of scenarios where different software and hardware products are combined. For details, visit https://info.support.huawei.com/storage/comp/#/oceanstor-dorado.

13.1 Data Plane Ecosystem Compatibility

13.1.1 Host Operating System

OceanStor Dorado is compatible with mainstream host operating systems in the industry (including IBM AIX, HP-UX, Solaris, Red Hat Enterprise Linux, SuSE Linux Enterprise Server, Oracle Enterprise Linux, Windows Server, NeoKylin, Kylin (Tianjin), Hunan Kylin, Deepin, Linx-TECH Rocky, and Redflag Linux) and multipathing software for operating systems (including embedded multipathing software of the host operating system, Huawei UltraPath, and third-party multipathing software Veritas DMP). In addition, OceanStor Dorado provides active-active storage solutions for mainstream host operating systems.

13.1.2 Host Virtualization System

OceanStor Dorado is compatible with mainstream host virtualization systems in the industry (including VMware ESXi, Microsoft Hyper-V, XenServer, Red Hat RHV, IBM PowerVM (VIOS), Huawei FusionCompute, and NeoKylin advanced server operating system (OS) (virtualization version)) and multipathing software (including embedded multipathing software of host virtualization systems and Huawei UltraPath). In addition, it provides active-active storage solutions.

OceanStor Dorado also supports various VMware features, including VAAI, VASA, SRM, vSphere Web Client Plug-in, vRealize Operations, and vRealize Orchestrator. It is deeply

integrated with VMware, providing customers with comprehensive storage services in VMware virtualization environments.

13.1.3 Host Cluster Software

OceanStor Dorado supports various host cluster software, including IBM PowerHA/HACMP, IBM GPFS, IBM DB2 PureScale, HPE ServiceGuard, Oracle SUN Cluster, Oracle RAC, Windows Server Failover Clustering, and Red Hat Cluster Suite, providing reliable shared storage services for host services in cluster scenarios.

13.1.4 Database Software

OceanStor Dorado supports various database software to meet customers' requirements for different service applications, including Oracle, DB2, SQL Server, SAP, GaussDB, Dameng, GBase, and Kingbase.

13.1.5 Storage Gateway

OceanStor Dorado can be taken over by third-party storage gateways, including third-party hardware storage gateways (IBM SVC and EMC Vplex), third-party software storage gateways (DataCore SANsymphony-V and FalconStor CDP/NSS), and gateways of third-party storage product HDS VSP series that support heterogeneous storage.

13.1.6 Heterogeneous Storage

Heterogeneous virtualization feature SmartVirtualization of OceanStor Dorado supports a wide range of storage products from other vendors, including EMC, IBM, HPE, HDS, and NetApp, helping customers protect their historical storage investments, upgrade storage devices, and migrate data.

13.1.7 Storage Network

OceanStor Dorado supports protocols such as FC, iSCSI, and TCP/IP, and is compatible with mainstream FC switches and directors (including Brocade and Cisco), standard Ethernet switch, and mainstream FC HBAs and standard Ethernet NICs.

13.2 Management and Control Plane Ecosystem Compatibility

13.2.1 Backup Software

OceanStor Dorado supports mainstream backup software in the industry and snapshot-based backup solutions, improving backup efficiency, saving host resources, and ensuring data security. Mainstream third-party backup software includes IBM TSM, Veeam, Veritas NBU, EISOO, and SCUTECH.

13.2.2 Network Management Software

OceanStor Dorado supports mainstream network management protocols and standards (including SNMP, RESTful, and SMI-S), as well as management software (including IBM Spectrum Control, SolarWinds Storage Resource Monitor, Microsoft System Center Visual

Machine Manager (SCVMM), HPE Operations Manager, and BMC Atrium Discovery) in the industry. In addition, it provides unified O&M management for customer data centers, reducing O&M costs.

13.2.3 OpenStack Integration

OceanStor Dorado 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. In addition, OceanStor Dorado supports commercial versions of OpenStack such as Huawei FusionSphere OpenStack, Red Hat OpenStack Platform, Mirantis OpenStack, and EasyStack.

13.2.4 Container Platform Integration

OceanStor Dorado releases CSI Plugin and FlexVolume Plugin, and supports mainstream container management platforms such as Kubernetes and Red Hat OpenShift.

OceanStor Dorado releases the CDR plugin to provide backup and recovery capabilities for containerized applications, ensuring data security of mission-critical services.

14 More Information

You can obtain more information about OceanStor Dorado at the following site:

https://support-it.huawei.com/dorado-v6/#/home

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 https://storage.huawei.com/index.html.

You can also visit Huawei's official website to obtain more information about Huawei storage:

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

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

https://support.huawei.com/enterprise

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:

https://e.huawei.com/en/branch-office-query

15 Feedback

Huawei welcomes your suggestions for improving our documentation.

If you have comments, send your feedback to the following mailbox:

storagedoc@huawei.com.

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

16 Acronyms and Abbreviations

Acronym or Abbreviation	Full Spelling
CIFS	Common Internet File System
CK	Chunk
CKG	Chunk Group
CLI	Command-line Interface
DIF	Data Integrity Field
DC	Data Center
DCL	Data Change Log
DTOE	Direct TCP/IP Offloading Engine
eDevLUN	External Device LUN
FC	Fiber Channel
FTL	Flash Translate Layer
FIM	Front-end Interconnect I/O Module
FRU	Field Replaceable Unit
GC	Garbage Collection
GUI	Graphical User Interface
LUN	Logical Unit Number
NAS	Network Attached Storage
NFS	Network File System
NUMA	Non-uniform Memory Access
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
OP	Over-Provisioning

Acronym or Abbreviation	Full Spelling
PID	Proportional-Integral-Differential
RAID	Redundant Array of Independent Disks
RAID-TP	Redundant Array of Independent Disks-Triple Parity
RDMA	Remote Direct Memory Access
ROW	Redirect-on-Write
SAS	Serial Attached SCSI
SSD	Solid-State Drive
SCSI	Small Computer System Interface
SCM	Storage Class Memory
SMB	Server Message Block
SMP	Symmetric Multiprocessing
TCO	Total Cost of Ownership
T10 PI	T10 Protection Information
VDI	Virtual Desktop Infrastructure
VSI	Virtual Server Infrastructure

