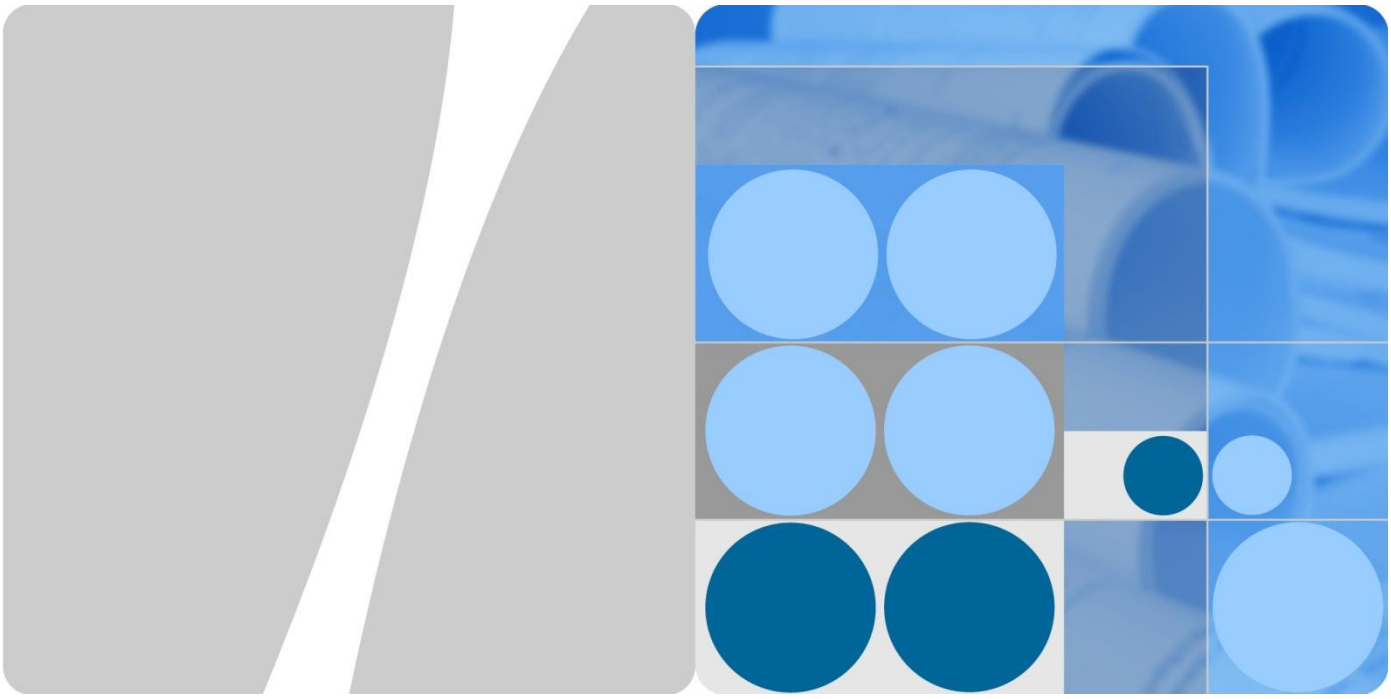


Huawei OceanStor Dorado V3 All Flash Storage

Technical White Paper

Issue 1.0

Date 2017-03-30

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

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

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

Notice

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

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

Huawei Technologies Co., Ltd.

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

Website: http://e.huawei.com

Email: support@huawei.com

Contents

Figures

1 About This Document

This document describes the architecture and key features and technologies of Huawei OceanStor Dorado V3 all-flash storage systems (OceanStor Dorado V3 for short), highlighting the unique advantages and customer benefits.

2 Overview

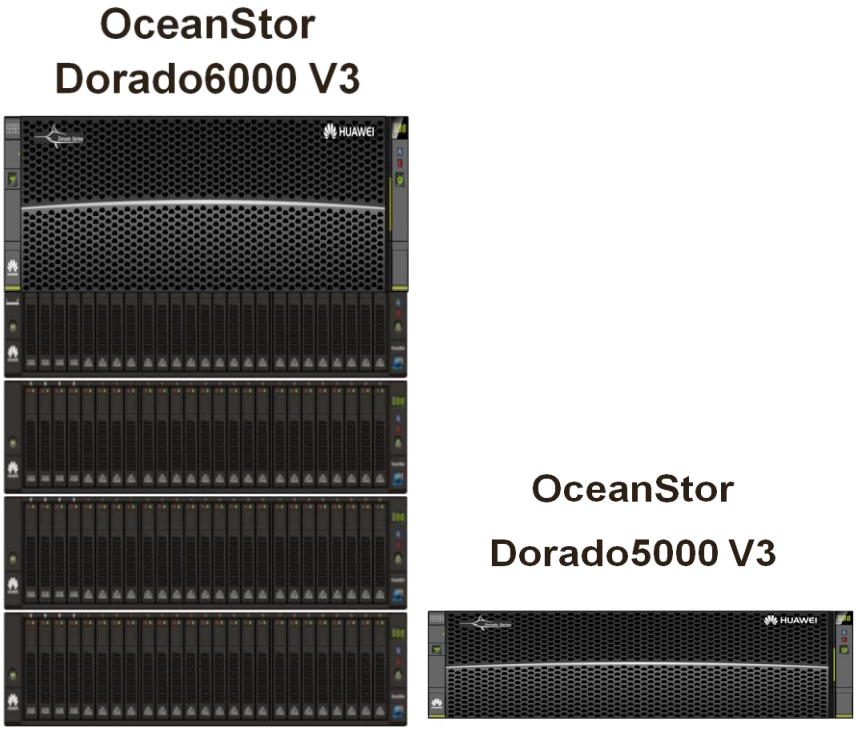
To survive in an increasingly fierce competition environment and shorten the rollout time of new services, enterprises' IT systems need to transform from a traditional cost center to a powerful weapon with the ability to help enterprises improve their competitiveness and achieve business success. In addition to providing high performance and robust reliability for mission-critical services, storage systems must address service growth needs, enhance service agility, and help services flexibly adapt to an increasingly fierce competition environment.

Storage technologies have developed rapidly over the past two decades, with many reliability enhancing technologies being introduced, such as RAID, RAID2.0+, remote replication, and active-active arrays. CPU performance that represents the computing capability has been improved by nearly 580 times. I/O channel performance is also almost 1000 times higher than before. However, improvement in storage media is only 20 times. Disks have become a roadblock to the improvement of storage systems. Therefore, the overall IT system performance cannot meet fast-growing service requirements.

To address this challenge, a major change occurs in the storage industry. Solid state disks (SSDs) are gradually replacing hard disk drives (HDDs). SSDs have absolute advantages over HDDs in terms of performance, reliability, and power consumption. However, they bring about new problems, for example, the amount of written data is limited, and the I/O stack of traditional storage system software is designed for HDDs and cannot bring SSD performance into full play, resulting in non-optimal total cost of ownership (TCO). To solve these problems, all-flash storage systems specially designed for SSDs are introduced. The emerging all-flash storage systems do not inherit many enterprise network features from traditional enterprise storage, so users cannot achieve an optimal IT solution that balances reliability, user habits, and performance.

To address this issue, Huawei launches the new-generation all-flash storage system OceanStor Dorado V3.

Figure 2-1 Appearance of OceanStor Dorado V3



Based on an industry-leading architecture, OceanStor Dorado V3 provides enterprise users with high-performance and efficient storage services, supports advanced data backup and disaster recovery technologies, ensuring secure and smooth operation of data services and meeting requirements of various applications such as online transaction processing (OLTP)/online analytical processing (OLAP), virtual server infrastructure (VSI), and virtual desktop infrastructure (VDI). In addition, OceanStor Dorado V3 offers various methods for easy-to-use management and convenient local and remote maintenance, remarkably reducing management and maintenance costs.

3 Solution

Huawei OceanStor Dorado V3 all-flash storage systems (the OceanStor Dorado V3 for short) are dedicated to setting a new benchmark for the enterprise storage field and providing data services of the highest level for enterprises' mission-critical businesses. With the advanced all-flash architecture and rich data protection solutions, the OceanStor Dorado V3 delivers world-leading performance, efficiency and reliability that meet the storage needs of various applications such as large-scale database OLTP/OLAP, VDI, and VSI. Applicable to sectors such as government, finance, telecommunications, energy, transportation, and manufacturing, the OceanStor Dorado V3 is the best choice for mission-critical applications.

3.1 System Architecture

The OceanStor Dorado V3 adopts a hardware architecture that has been proven highly reliable on live networks and uses mature products and Huawei SSDs (HSSDs), providing end-to-end reliable hardware. As to software, the OceanStor Dorado V3 employs OceanStor OS of which a brand-new architecture is designed for SSDs, ensuring stable low latency and high performance. In addition, it inherits rich enterprise network features of Huawei enterprise storage and supports convergence and interconnection with products using OceanStor OS, providing strong combined solutions. Thanks to the advanced system architecture, the OceanStor Dorado V3 supports scale-out and scale-up. Scale-out enables stable low latency as well as linear increase of capacity and performance. Scale-up helps achieve capacity expansion. When capacity is increased, the system's back-end data balancing mechanism automatically distributes data onto all the disks of the storage array evenly. The following details the OceanStor Dorado V3 architecture from the aspects of hardware and software.

3.1.1 Hardware Architecture

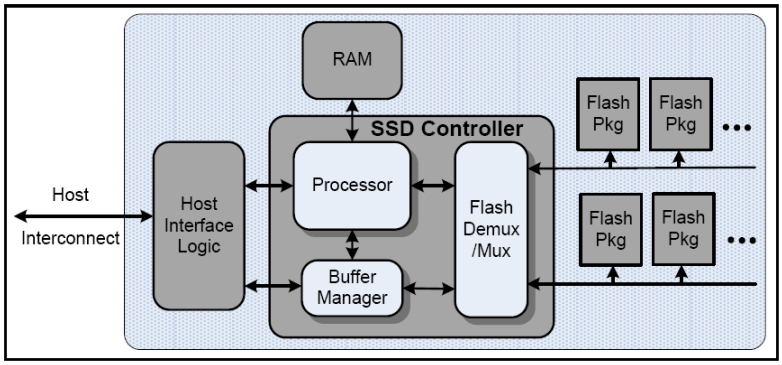
3.1.1.1 HSSD Architecture

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

- Host interface: the protocol and physical interface used by a host to access an SSD. NVMe, SAS, NoF (NVMe over Fabric) are supported.
- SSD controller: a core SSD component responsible for read and write access from a host to the back-end media and for protocol conversion, table entry management, data caching, and data checking.

- RAM: a component responsible for the Flash Translation Layer (FTL) table and data caching to provide fast data access.
- NAND FLASH: physical entity for data storage

Figure 3-1 SSD hardware architecture



3.1.1.2 Storage Device Architecture

OceanStor Dorado V3 is classified into OceanStor Dorado6000 V3 and OceanStor Dorado5000 V3.

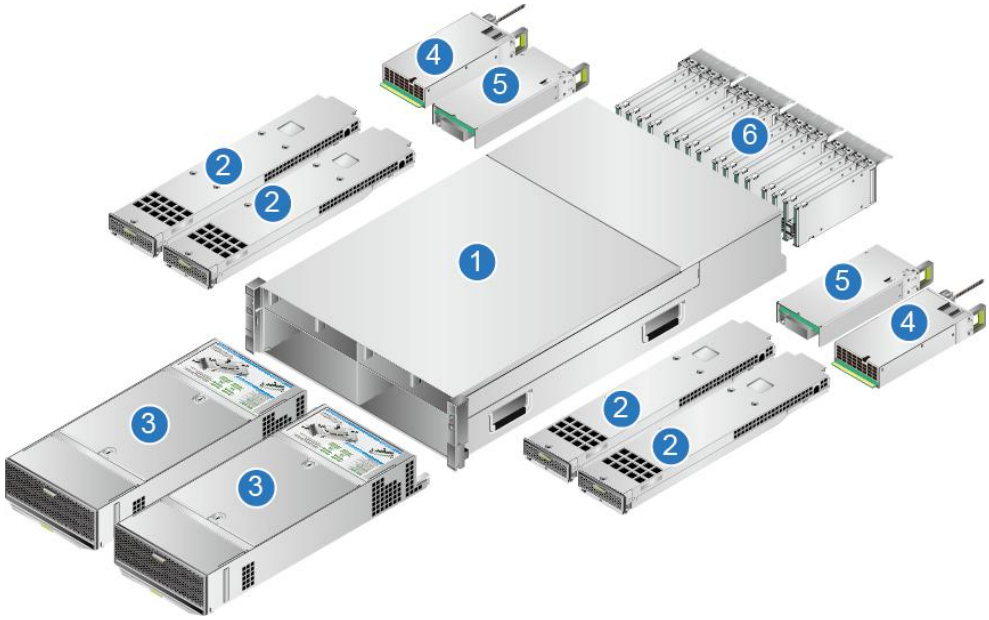
In OceanStor Dorado6000 V3, the controller enclosure and disk enclosure are independent, allowing flexible scale-out and scale-up. The controller enclosure is a 3 U device housing dual controllers interconnected by midplanes. The disk enclosure is a 2 U device based on redundant architecture through midplane interconnection, and supports a maximum of 25 2.5-inch SSDs.

In OceanStor Dorado5000 V3, the controller enclosure and disk enclosure are integrated to achieve high-density performance and capacity. The controller enclosure is a 2 U device housing dual controllers interconnected by midplanes, and supports a maximum of 25 2.5-inch SSDs.

An OceanStor Dorado V3 all-flash storage system consists of controller enclosures, disk enclosures, and PCIe switches.

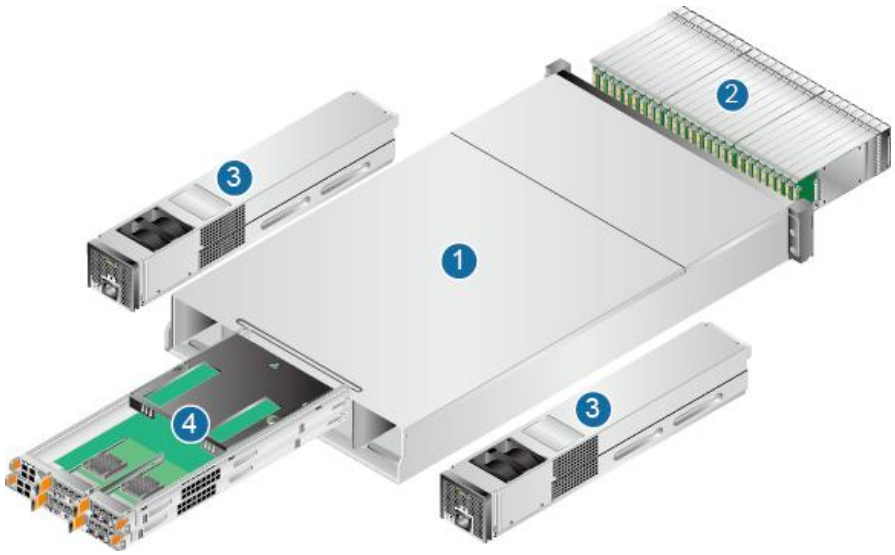
- Controller enclosure: a core component responsible for processing all service logic and allowing host access and device management. It consists of a system subrack, a control board, interface modules, power modules, BBUs, and management modules.

Figure 3-2 Hardware architecture of a 3 U controller enclosure



1	System subrack	2	BBU module
3	Controller	4	Power module
5	Management module	6	Interface module

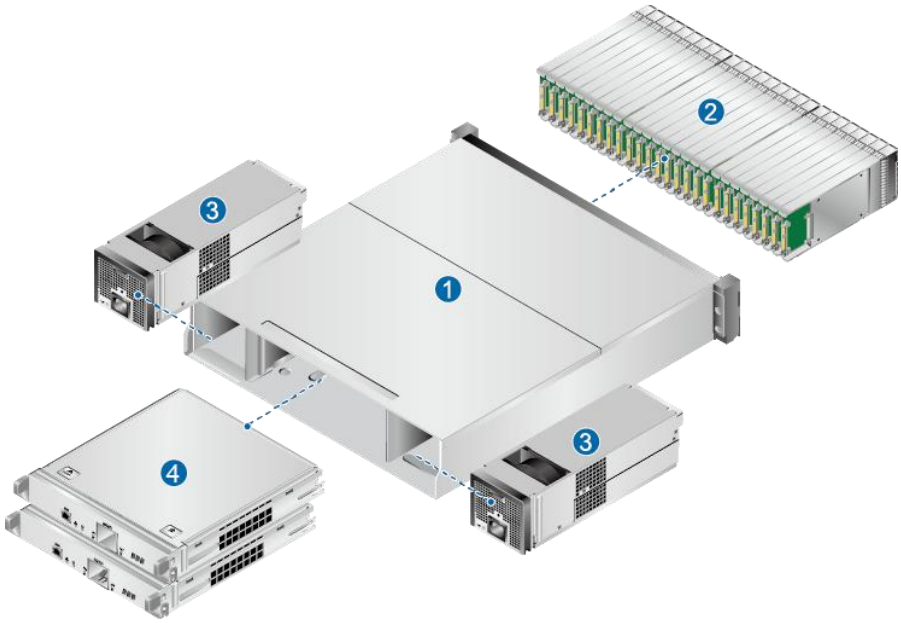
Figure 3-3 Hardware architecture of a 2 U controller enclosure



1	System subrack	2	Disk module
3	Power-BBU module	4	Controller

- Disk enclosure: contains SSDs, manages disks, and interconnects and provisions service access. It consists of a subrack, expansion modules, and SSDs.

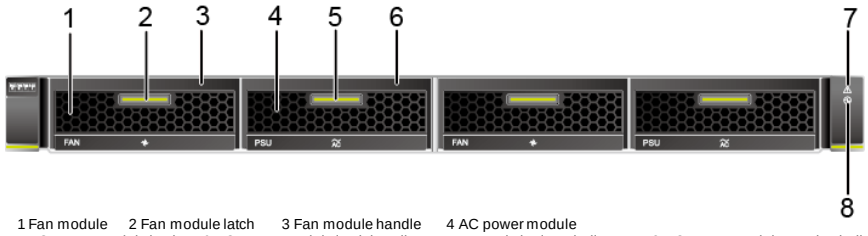
Figure 3-4 Hardware architecture of a disk enclosure



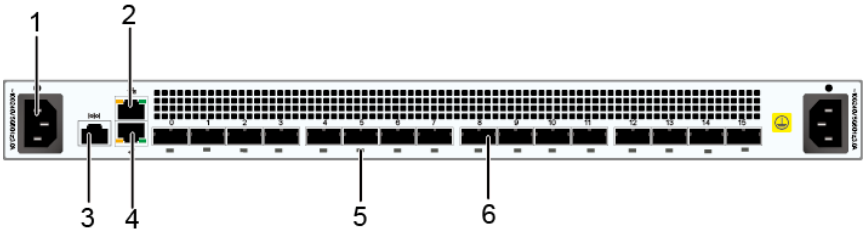
1	System subrack	2	Disk module
3	Power module	4	Expansion module

- PCIe switch: used for data transmission between controllers

Figure 3-5 Hardware architecture of a PCIe switch



1 Fan module 2 Fan module latch 3 Fan module handle 4 AC power module
5 AC power module latch 6 AC power module latch handle 7 Data switch alarm indicator 8 AC power module running indicator

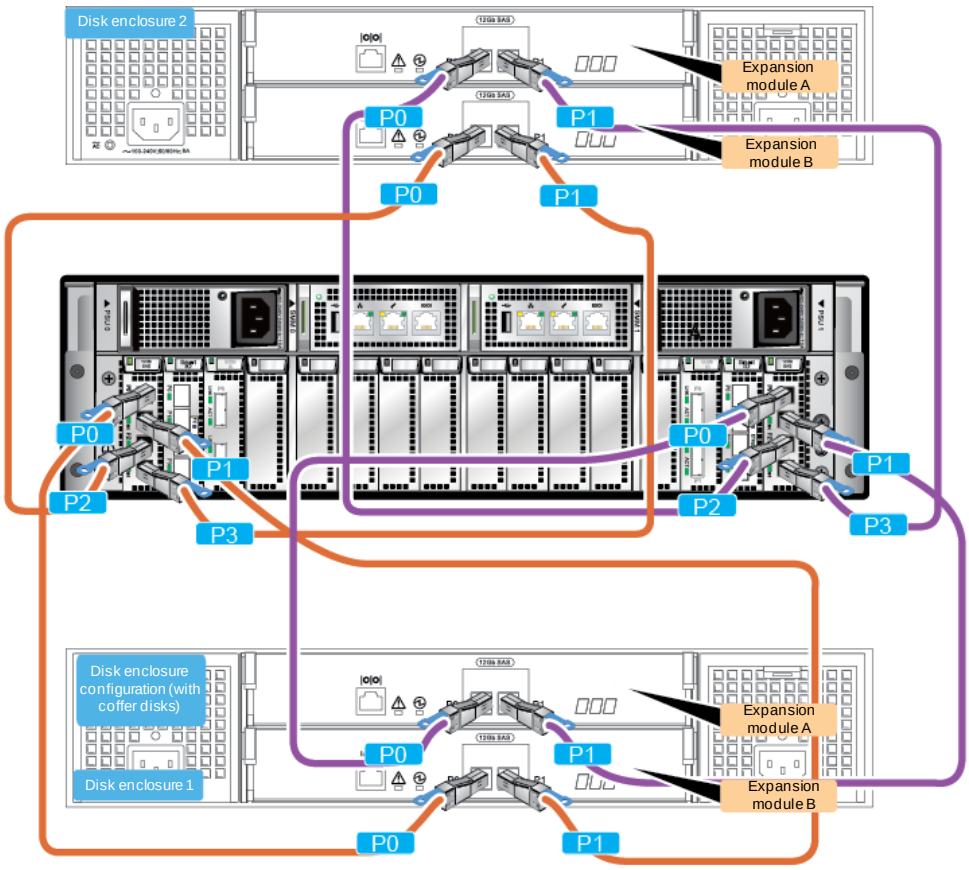


1 Power socket 2 Management network port 3 Serial port 4 Management network port
5 PCIe port link/speed indicator 6 PCIe port

3.1.1.3 Hardware Expansion Capability

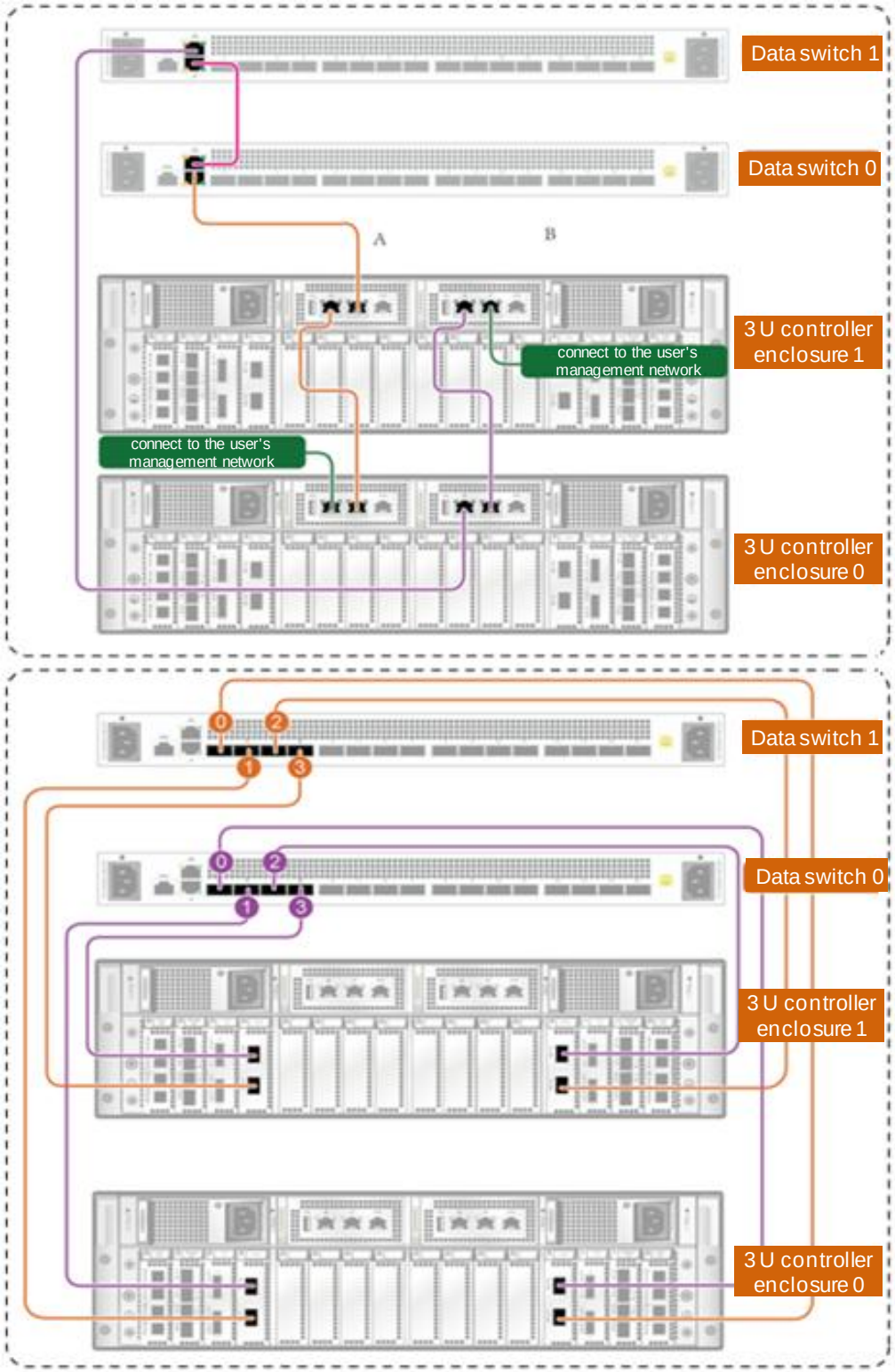
Scale-up: The controller enclosure and disk enclosures of OceanStor Dorado V3 are redundantly interconnected using SAS3.0. Two 48 Gbit/s SAS ports are available, providing level-1 expansion capability to make full use of SSD performance.

Figure 3-6 Scale-up networking



Scale-out: Controllers of OceanStor Dorado V3 are redundantly connected using PCIe 3.0 for data transmission, while redundant GE networks enable scale-out management.

Figure 3-7 Scale-out networking



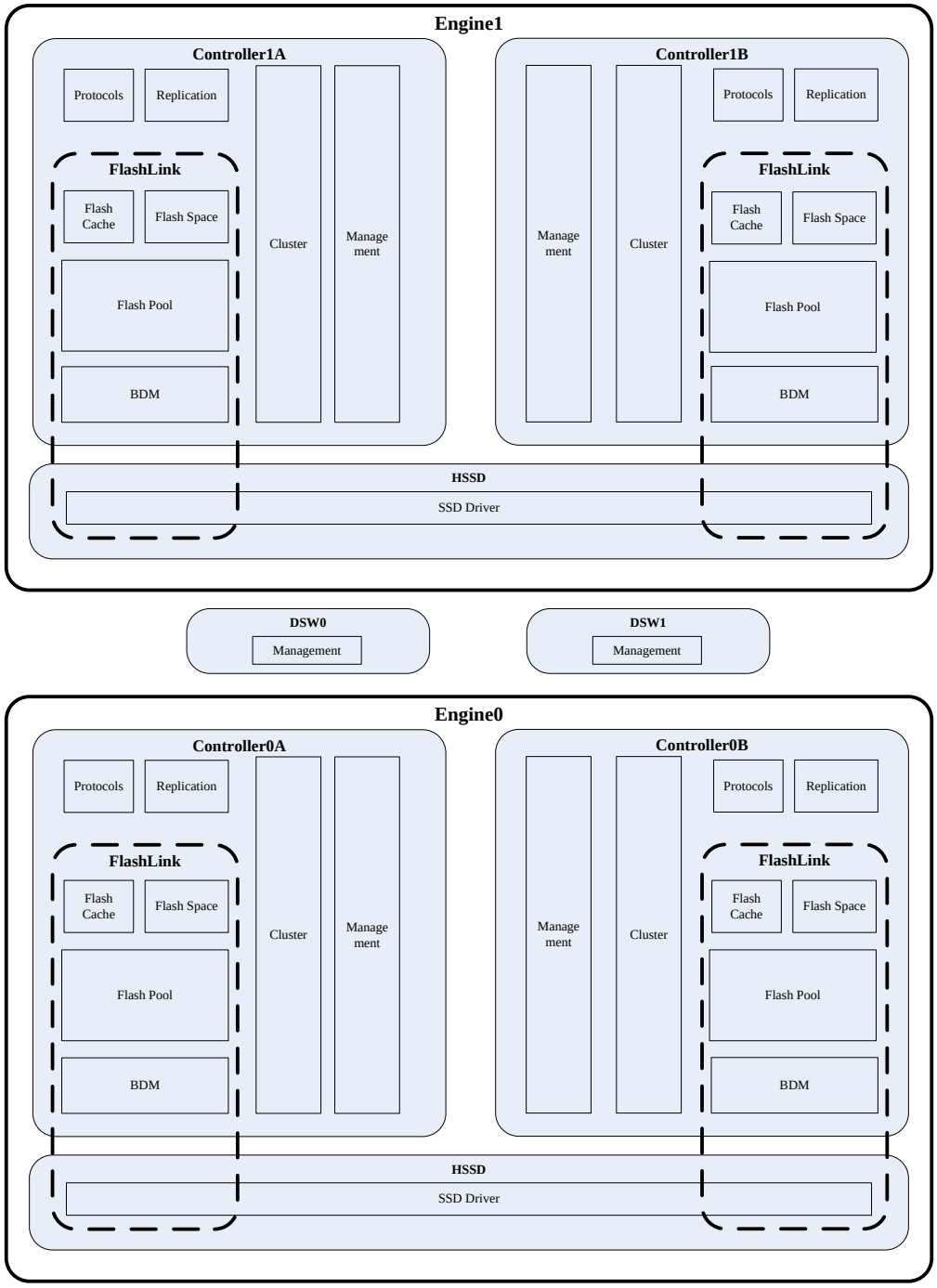
3.1.1.4 Hardware Architecture Characteristics

1. Optimal performance: end-to-end high-speed architecture, PCIe 3.0 bus, SAS 3.0 disk ports, 16 Gbit/s Fibre Channel/10GE/56 Gbit/s InfiniBand host interfaces, high-performance HSSDs
2. Stable and reliable: mature hardware products, fully redundant hardware architecture proven by tens of thousands sets of systems on live networks
3. Efficient: OceanStor Dorado V3 supports both Scale-out and Scale-up and its controllers and disks can be expanded online. Based on modular design, I/O modules are hot swappable. Front-end and back-end ports can be configured on demand.

3.1.2 Software Architecture

OceanStor Dorado V3 uses OceanStor OS that is especially designed for SSDs and employs the FlashLink technology (see section 3.2 "FlashLink") to provide users excellent performance, robust reliability, and high efficiency.

Figure 3-8 Software architecture of OceanStor Dorado V3



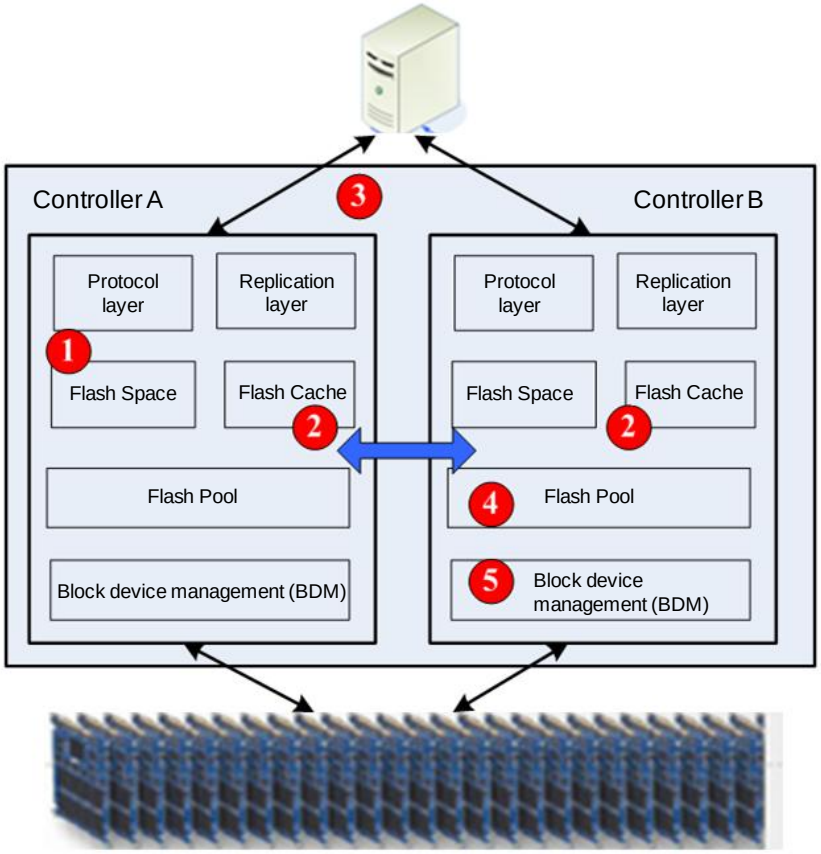
- The storage controller software architecture mainly consists of the Cluster & Management plane and service plane.
  - The Cluster & Management plane provides a basic environment for system running, controls multi-controller Scale-out logic, and manages alarms, performance, and user operations.
  - The service plane schedules storage service I/Os, realizes data Scale-out capabilities, and implements controller software–related functions of the FlashLink technology

such as deduplication and compression, full-stripe sequential write, cold and hot data separation, global wear leveling, and anti-wear leveling.

- HSSD software architecture mainly includes Huawei-developed disk drives, realizes basic SSD functions and hardware-related functions of the FlashLink technology such as I/O priority and multi-channel data flows.

3.1.2.1 Write Process

Figure 3-9 Write I/O process



1. Write I/Os enter Flash Space after passing the protocol layer. The system checks whether the I/Os belong to this controller. If no, the I/Os are forwarded to the peer controller.
2. If yes, the I/Os are written to local Flash Cache and mirrored to peer-end Flash Cache.
3. A write success is returned to the host.
4. Flash Cache flushes data to Flash Pool where the data will be deduplicated and compressed.
   a. Flash Pool divides the received data into data blocks with a fixed length (8 KB).
   b. Flash Pool calculates the fingerprint value of each data block and forwards the data block to the owning controller based on the fingerprint value.
   c. After the local controller receives data blocks, Flash Pool checks the fingerprint table.
   d. If the same fingerprints exist in the fingerprint table, obtain the locations where related data is stored, and compare the data with data blocks byte by byte. If they

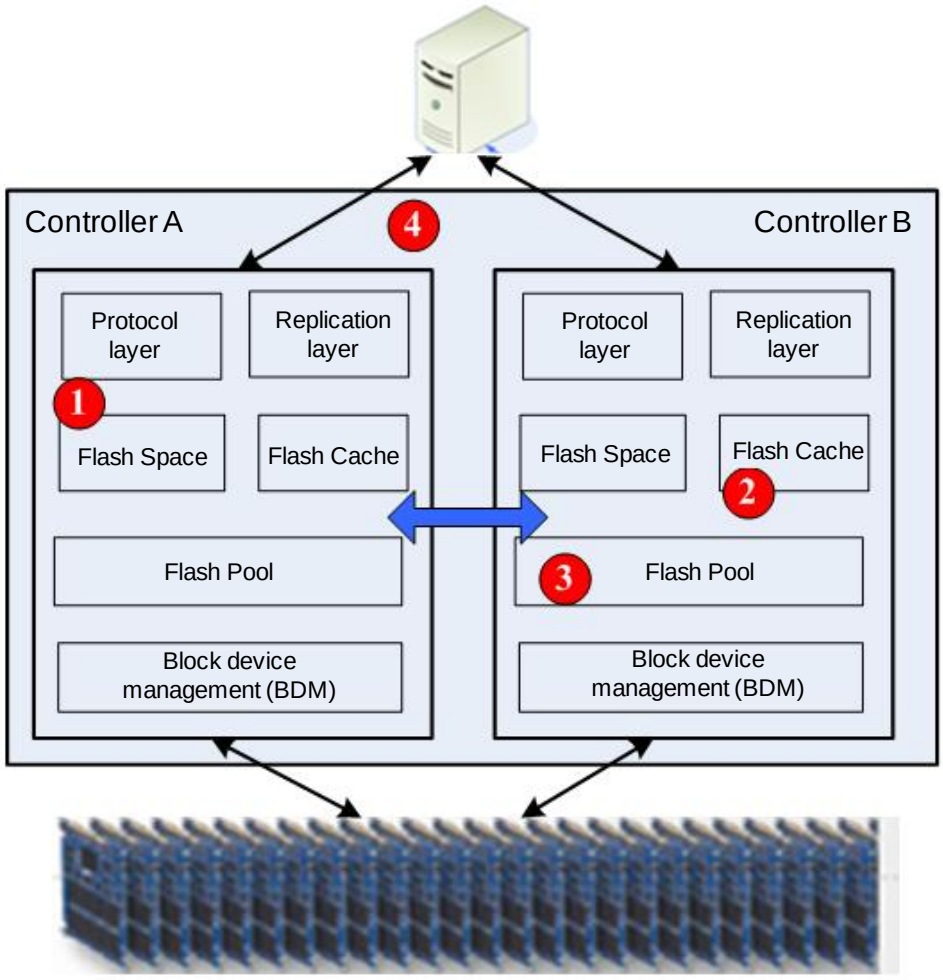
are the same, the system increases the reference count of the fingerprints and does not write the data blocks to SSDs.

e. If the fingerprint table does not contain the same fingerprints or the data and data blocks are not consistent, compress the data blocks (based on the size of 1 KB).

5. Flash Pool combines the data into full stripes and writes it to SSDs.
   a. Compressed I/Os are merged into write stripes of which the size is an integer multiple of 8 KB.
   b. If the I/Os are merged into full-write stripes, calculate the checksum and write the data and checksum to disks.
   c. If the I/Os are not merged into full-write stripes, add 0s to the tail before data is written to disks (the 0s will be cleared in subsequent garbage collection).
   d. Data is written to a new location every time and metadata mapping relationships are updated.
   e. After a message is returned indicating that I/Os are successfully written to disks, Flash Cache deletes the corresponding data pages.

3.1.2.2 Read Process

Figure 3-10 Read I/O process



1. Flash Space analyzes received I/Os and judges whether the I/Os belong to the local controller. If no, it forwards the I/Os to the peer controller.
2. The owning controller searches for the desired data in Flash Cache and returns the data to the host. If it cannot find the data, Flash Pool will continue the process.
3. Flash Pool divides the I/Os into data blocks with a fixed size (8 KB), determines the owning controller of each data block based on the LBA, and forwards the data block to the owning controller.
    a. On the owning controller of a data block, check the LBA-fingerprint mapping table and obtain the fingerprint.
    b. Forward the data block read request to the fingerprint owning controller according to fingerprint forward rules.
    c. On the fingerprint owning controller, find the fingerprint-storage location mapping table and read the data at the storage location.
4. Decompress data on the fingerprint owning controller and return data to the host.

3.1.2.3 Hardware Architecture Characteristics

1. Excellent performance: The FlashLink technology realizes efficient I/O scheduling, providing high performance on condition that the service system latency is steadily low.
2. Stable and reliable: SSD drives, controller software, and multi-level reliable solutions provide users 99.9999% reliability and ensure 24/7 stable service system running.
3. Efficient: Multiple efficiency-improving features are available, such as online deduplication and compression and heterogeneous virtualization, to protect customer investments.

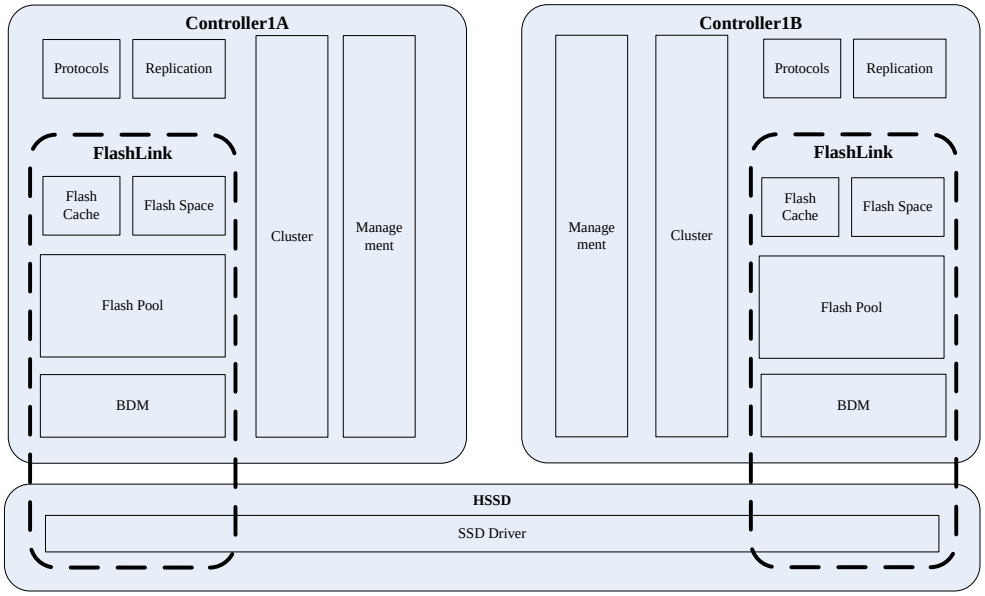
3.2 FlashLink

3.2.1 Introduction

As stated in Chapter 2 "Overview", while SSDs have absolute advantages over HDDs in terms of performance, reliability, and power consumption, they bring about new problems.

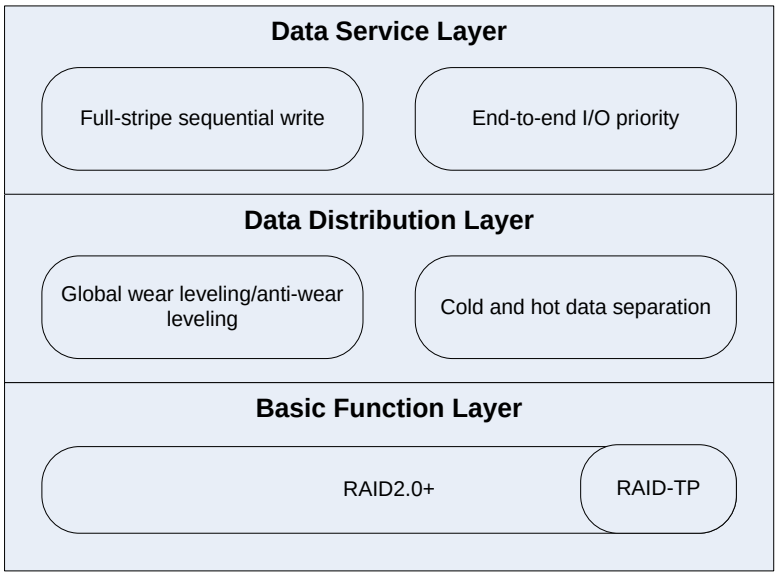
As the only vendor capable of developing both storage arrays and SSDs, Huawei adopts innovative FlashLink in OceanStor Dorado V3. This most cost-effective all-flash storage system has resolved these problems. The following figure shows FlashLink in the system architecture.

Figure 3-11 FlashLink illustration



As shown in the preceding figure, FlashLink contains the major I/O modules of controller software and disk drivers. The following figure shows functions provided by FlashLink.

Figure 3-12 FlashLink functions



FlashLink provides functions at three layers:

- Basic function layer: Provides efficient and reliable RAID functions based on the advanced RAID 2.0+ technology and adds the RAID-TP function that supports three-disk redundancy. In scenarios of large-capacity disks, reliability is ensured and impacts on services are reduced at the same time.
- Data distribution layer: Allocates data to disks evenly and adjusts the allocation algorithm dynamically based on disk status using SSD drives and controller software.

This layer stores hot and cold data separately based on the characteristics of SSDs to obtain better performance and reliability.

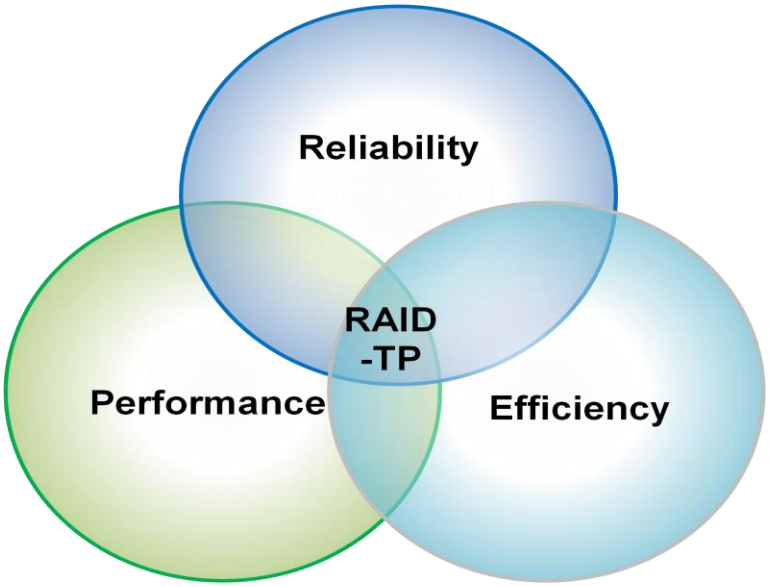
- Data service layer: Merges all writes to one full stripe and then sends the data to disks, completely eliminating write penalty. Efficient I/O scheduling is realized and RAID usage is improved. You can configure priorities from host interfaces to disks to grant the highest priority to latency-sensitive I/Os.

3.2.2 RAID-TP

The growing capacity of single disks requires longer reconstruction time if a disk fails. To ensure system reliability, you can improve reconstruction speed to reduce the time but this affects service performance. You can also increase redundancy to ensure system reliability during the reconstruction process at the cost of disk utilization.

Based on the redirect-on-write (ROW) technology, OceanStor Dorado V3 adopts full-stripe writes and eliminates write penalties completely. Write performance has no loss in large-stripe RAID configurations, where the cost of adding one redundant disk can be ignored. For example, OceanStor Dorado V3 supports 23+3 RAID configuration at most. Compared with 23+2 RAID configuration, capacity usage is reduced by only 3.5% but reliability improves by two orders of magnitude. In summary, OceanStor Dorado V3 ensures reliability, performance, and efficiency in scenarios of large-capacity disks by using the RAID-TP function.

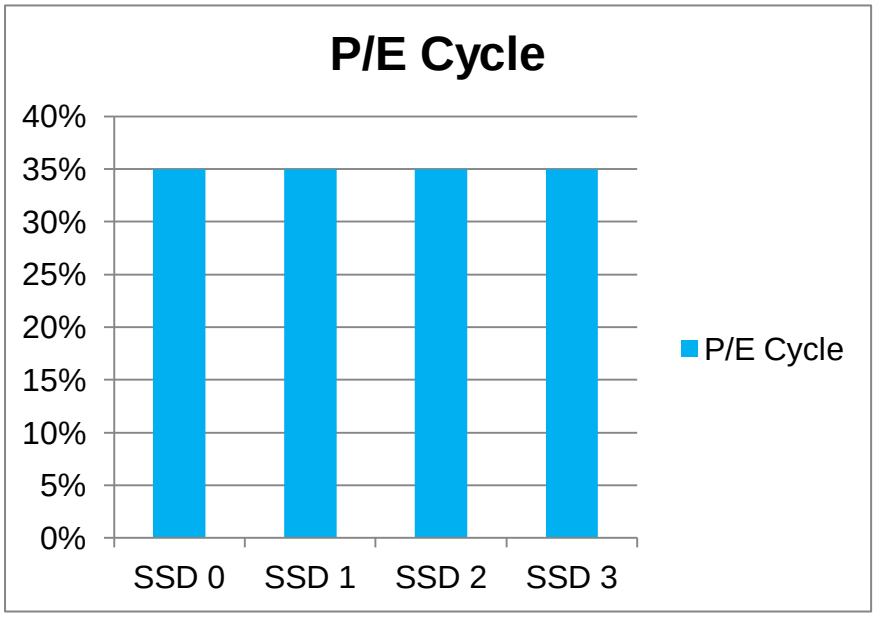
Figure 3-13 Customer benefits from RAID-TP



3.2.3 Global Wear Leveling and Anti-Wear Leveling

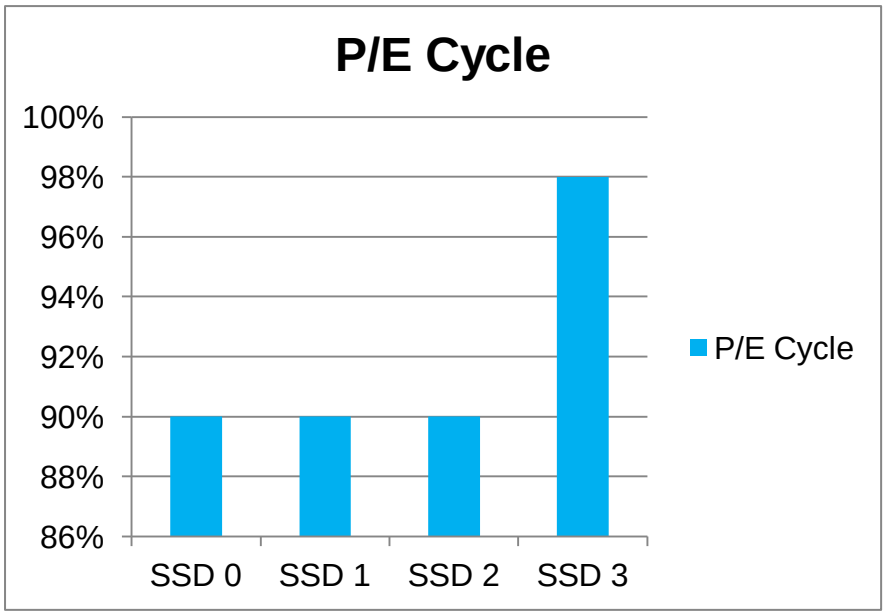
The biggest difference of an SSD and an HDD is that the amount of data written to the SSD is limited and the SSD lifespan is inversely related to this amount. Therefore, an all-flash storage system requires load balancing between disks to prevent some disks from failures due to frequent writes. FlashLink uses controller software and disk drives to regularly query the SSD controller about disk wearing level, which is used as one basis for space allocation.

Figure 3-14 Global wear leveling



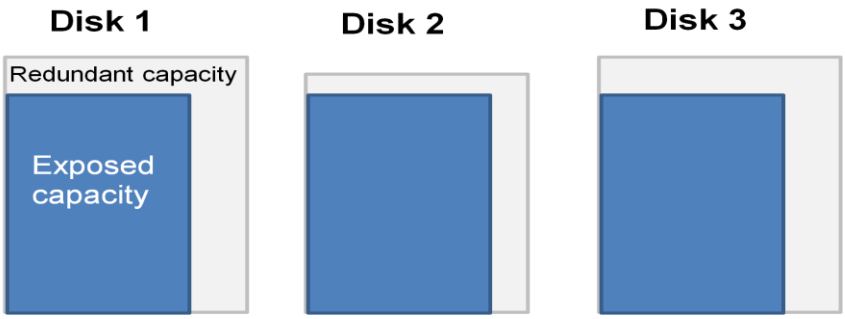
However, if OceanStor Dorado V3 is approaching the end of its life, for example, the disk wearing level reaches over 80%, multiple disks may fail at the same time and data may be lost if global wear leveling is still used. In this case, the system enables anti-global wear leveling to avoid failures of SSDs in batch. The system selects the SSD that is most severely worn and writes data onto this disk as long as it has idle space. As a result, this SSD runs out of its service life faster than other disks. Users will be prompted to replace this disk. In this way, SSDs will not fail in batch.

Figure 3-15 Global anti-wear leveling



Every SSD reserves certain space for garbage collection. During system running, one disk may have more bad blocks than others, consuming more redundant space. As a result, performance and reliability of the disk are affected. If the system keeps using the policy of global wear leveling, the performance of the entire system will be affected and this disk will be more vulnerable to damage. FlashLink obtains partial redundant space from each disk as shared redundant space, called global capacity redundancy. Based on the remaining redundant space on each SSD, FlashLink dynamically adjusts the data allocation algorithm to reduce data written to the SSD that is most severely worn and scatter data to other disks, improving SSD lifespan and storage performance.

Figure 3-16 Global capacity redundancy

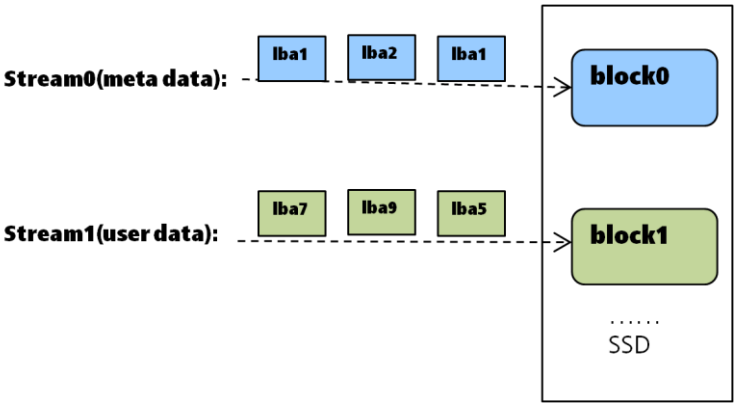


3.2.4 Hot/Cold Data Separation

HDDs use magnetic medium for storage, which can be directly overwritten. SSDs use NAND FLASH that must be erased before data writes. Page is the basic write unit and block is the smallest erasure unit for NAND FLASH. A block consists of several pages. Before writing data on a page, the system must erase the block where the page resides. New data is written to SSDs constantly but the storage space of NAND FLASH is fixed. To solve this problem, SSDs introduce garbage collection (GC) to release blocks containing invalid data to store new data.

In an ideal situation, if all the data in one block is invalid, the SSD can directly erase the whole block instead of migrating valid data. FlashLink adds labels to data with different change frequencies in the controller software and then sends the data to SSDs, as shown in the following figure. This technology is called separation of multiple channels of data.

Figure 3-17 Separation of multiple channels of data



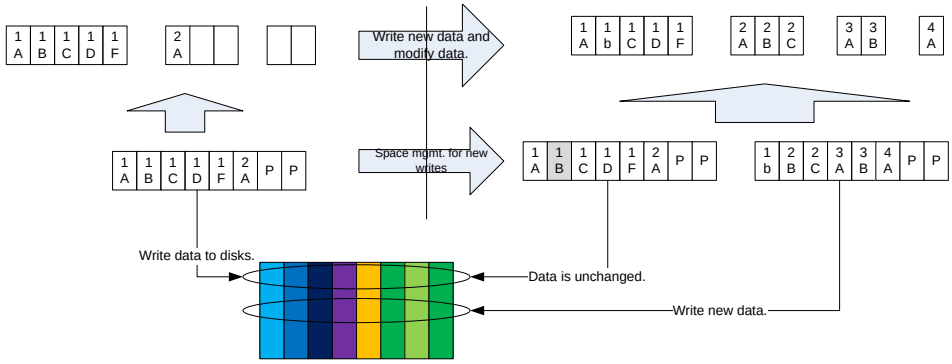
SSDs save data with the same labels to the same block. In this way, hot and cold data are stored in different blocks, reducing the amount of data migration for garbage collection, and improving performance and reliability of SSDs.

Figure 3-18 Hot and cold data separation

3.2.5 Full-Stripe Sequential Write

As discussed above, wear leveling must be achieved among SSDs. However, if a user has to frequently change data in one block (such as log volumes of a database), FlashLink uses redirect-on-write (ROW) to ensure wear leveling. The ROW technology re-allocates one space to new data or updating existing data for each write operation. All written data can be evenly distributed to different disks no matter what service model is adopted by the users. Different service data is merged in to one full stripe and then written to SSDs, eliminating the read penalty during the write process required by traditional RAID, as shown in the following figure.

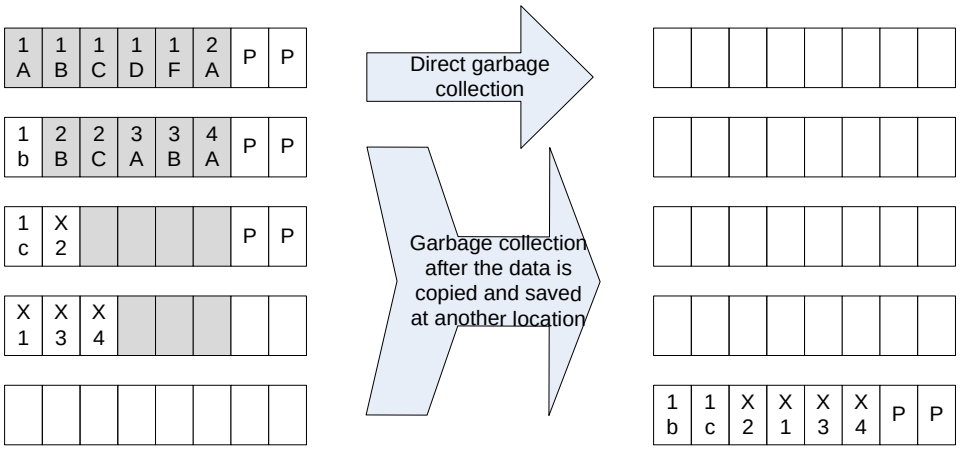
Figure 3-19 Full-stripe sequential write



When new data is written to SSDs, the system merges data of LUN 1 and LUN 2 as a full stripe and then writes data to disks, as shown in the left part of the figure. If some data is modified as shown in the right part of the figure, the user modifies **1B** to **1b** in LUN 1, writes new data **3A** and **3B** to LUN 3, and new data **4A** to LUN 4. The system merges **1b**, **2B**, **2C**, **3A**, **3B**, and **4A** as a new stripe and then writes data to disks. The system marks **1B** data block as junk data.

After a long time of running, a large amount of junk data will be generated and the system cannot find any space for full-stripe write. When the amount of junk data reaches a certain value, FlashLink uses global garbage collection to clear storage space, ensuring enough space for full-stripe write at any capacity usage rate.

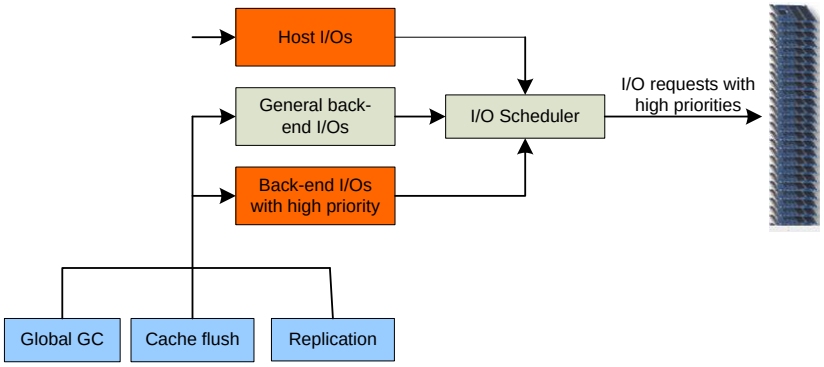
Figure 3-20 Global garbage collection



3.2.6 End-to-End I/O Priority

To ensure stable latency, Dorado V3 controllers mark the priorities of I/Os. For example, the priority of a read request from a host is higher than that of a cache flushing request, and a cache flushing request is prior to a back-end I/O copy in asynchronous replication. I/O priorities are sent to SSDs together with read and write requests. Upon receiving I/Os, SSDs check the priority labels of I/Os and SSD controllers process I/Os with high priorities first to realize end-to-end I/O priority control.

Figure 3-21 End-to-end I/O priority



3.3 Key Features

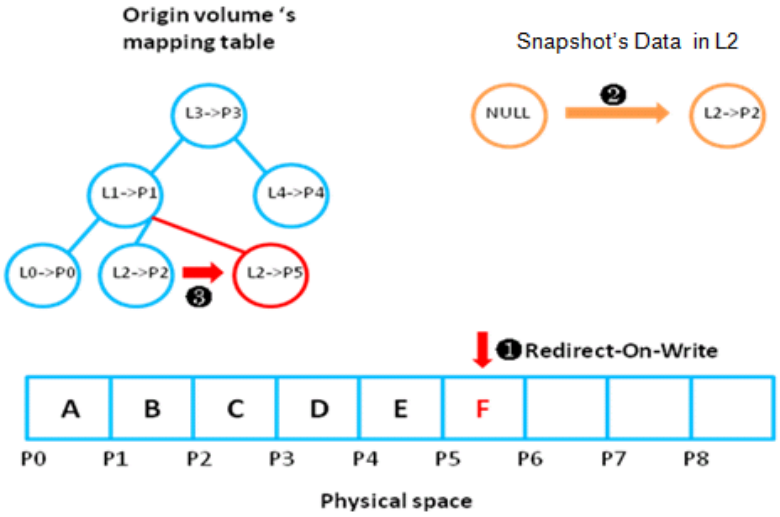
Dorado V3 provides two series of software:

- Hyper series for data protection: HyperSnap, HyperReplication, and HyperMetro, providing disaster recovery and data backup for users
- Smart series for efficiency improvement: SmartDedupe, SmartCompression, SmartThin, SmartVirtualization, and SmartMigration to improve storage efficiency and reduce user TCO

3.3.1 Snapshot (HyperSnap)

Snapshot is implemented using COW and ROW technologies. COW allows data to be copied when it is initially written. Data copy affects write performance of hosts. ROW does not involve data copy. However, after data is overwritten frequently, data distribution on the source LUN will be damaged and data of the source LUN is distributed randomly. This is not a problem if SSDs are used. If HDDs are used, read performance deteriorates dramatically due to the mechanical disk seek time. Lossless snapshots realized using the ROW mechanism can solve the problem of sharp performance drop in traditional arrays.

Figure 3-22 Basic principle of the snapshot (ROW) technology

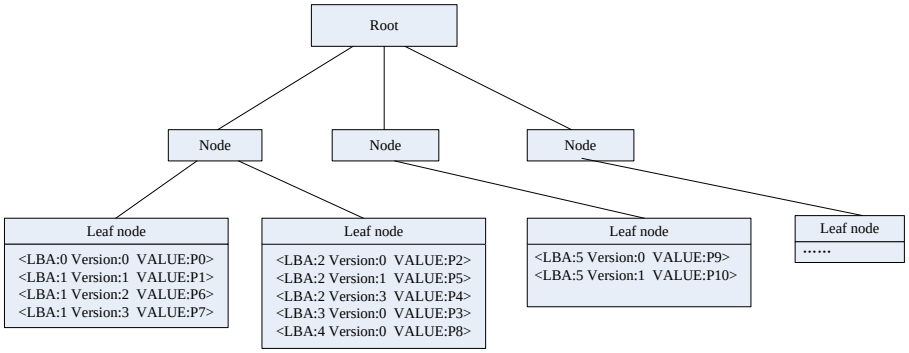


The original volume (source LUN) and snapshot use a mapping table to access physical space. The initial data of the origin volume is **ABCDE** and is saved in sequence in terms of physical space. The metadata of the snapshot is null. All read requests to the snapshot are redirected to the origin volume.

- When the origin volume receives a write request in which **C** is changed to **F**, the data is directly written into new physical space P5 instead of being overwritten into physical space P2, as shown in step 1 in the preceding figure.
- After the data is written into the new physical space, mapping item **L2->P2** is inserted into the metadata of the snapshot. In this way, accesses to logical address L2 of the snapshot are not redirected to the origin volume and data is directly read from physical space P2, as shown in step 2 in the preceding figure.
- **L2->P2** in the mapping metadata of the origin volume is changed to **L2->P5**, as shown in step 3 in the preceding figure.

Data in the origin volume is changed to **ABFDE** and data in the snapshot is still **ABCDE**.

Figure 3-23 Distribution of the LUN data and metadata of the source LUN



The system has a unique metadata organization method and supports high-speed query, deletion, insertion, and update of metadata. Snapshots using this organization method have no performance loss.

- When a snapshot is being created, IOPS and latency of the source LUN remain unchanged.
- When a snapshot is being deleted, IOPS and latency of the source LUN remain unchanged.
- When a snapshot is mapped to the host for read and write, IOPS and latency of the source LUN remain unchanged.

The following shows how to delete snapshot TP1 to explain the implementation principles of lossless snapshot.

Deleting the data exclusively occupied by a snapshot: Check whether mapping data exists at the time one time point greater than the snapshot activation time. If mapping data exists, the data is exclusively occupied by a snapshot. If not, the data is shared. The user has created three snapshots at time points TP0, TP1, and TP2. The latest time point is TP3. In the first figure on the left, LBA0 only has data in VERSION0 that is shared by VERSION0, VERSION1, VERSION2. LBA1 has no data in VERSION0. Therefore, VERSION1, VERSION2, and VERSION3 occupy data exclusively. <LBA:1 VERSION:1 VALUE:P1> needs to be deleted.

Lossless snapshots are realized through the metadata management mechanism with time points and high-speed random access of SSDs.

3.3.2 Remote Replication (HyperReplication)

- **Working principle**

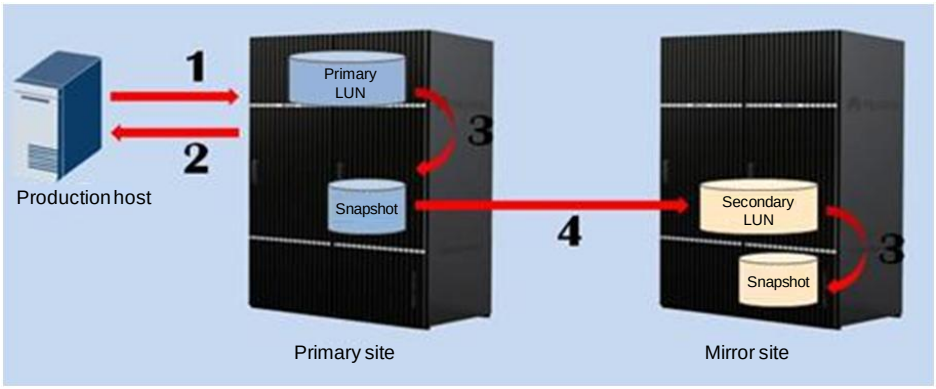
  After an asynchronous remote replication relationship is set up between a primary LUN at the primary site and a secondary LUN at the secondary site, initial synchronization is implemented. After initial synchronization is completed, the data status of the secondary LUN becomes **Synchronized** or **Consistent**. Then, I/Os are processed as follows:

  1. The primary LUN receives a write request from a production host.
  2. After data is written to the primary LUN, a write completion response is immediately returned to the host.
  3. Incremental data is automatically synchronized from the primary LUN to the secondary LUN based on the user-defined synchronization period that ranges from 1 to 1,440 minutes. (If the synchronization type is **Manual**, the user needs to trigger

the synchronization manually.) Before synchronization starts, a snapshot is generated for each of the primary LUN and the secondary LUN. The snapshot of the primary LUN ensures that the data read from the primary LUN during the synchronization keeps unchanged. The snapshot of the secondary LUN backs up the secondary LUN's data in case that the data becomes unavailable if an exception occurs during the synchronization.

4. During the synchronization, data is read from the snapshot of the primary LUN and copied to the secondary LUN. After the synchronization is complete, the snapshot of the primary LUN and that of the secondary LUN are canceled, and the next synchronization period starts.

Figure 3-24 Working principle of asynchronous replication



- **Technical advantages**
  - Data compression

    Data compression is supported specific to iSCSI links and data compression ratio varies with the service data type. The maximum ratio for database services can be 4:1.
  - Quick response to host requests

    After a host writes data to the primary LUN at the primary site, the primary site immediately returns a write success to the host before the data is written to the secondary LUN. In addition, data is synchronized from the primary LUN to the secondary LUN in the background and does not affect the 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 lost data is determined by the synchronization period (ranging from 3 to 1440 minutes, 30s by default) that is specified by the user based on site requirements.
  - Splitting, switchover of primary and secondary LUNs, and rapid fault recovery

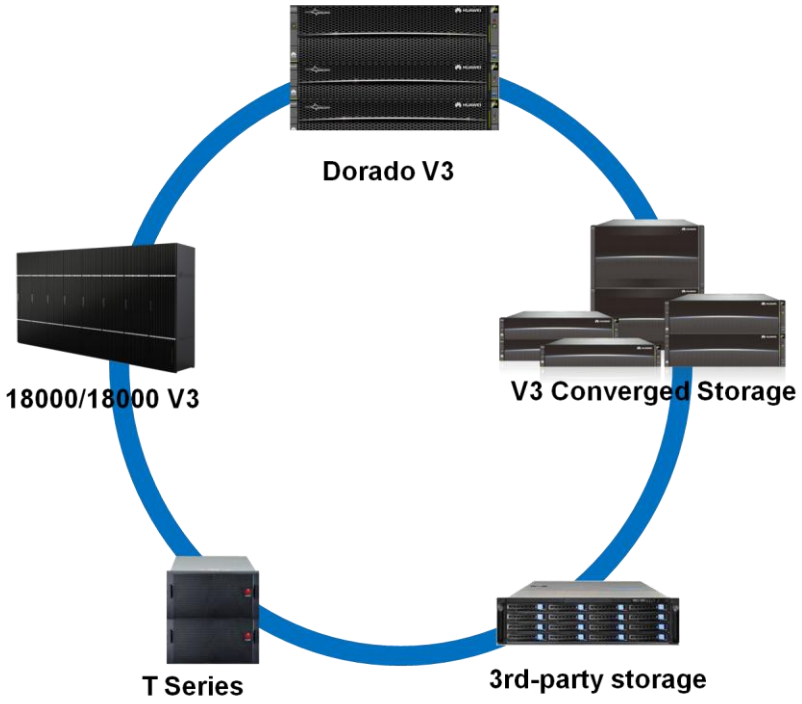
    The asynchronous remote replication supports splitting, synchronization, primary/secondary switchover, and recovery after disconnection.
  - Consistency group

    Consistency group functions are available, such as creating and deleting consistency groups, creating and deleting member LUNs, splitting LUNs, synchronization, primary/secondary switchover, and forcible primary/secondary switchover.
  - Interoperability between high-end, mid-range, and entry-level storage

    Developed on the OceanStor OS unified storage software platform, OceanStor Dorado V3 is completely compatible with the replication protocols of Huawei high-end, mid-range, and entry-level storage products. Remote replication can be

created among different types of products to construct a highly flexible disaster recovery solution.

Figure 3-25 Interoperability between high-end, mid-range, and entry-level storage



- **Application scenarios**

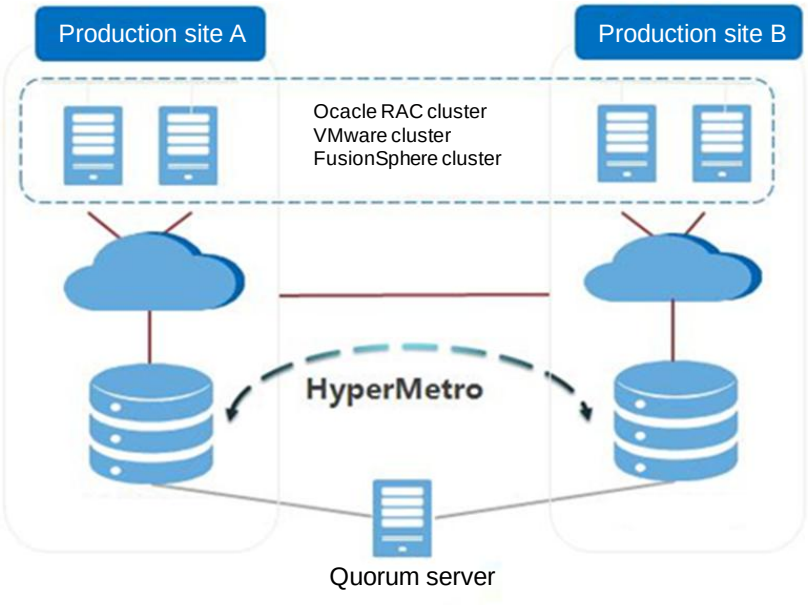
  Remote data disaster recovery and backup: For HyperReplication/A, the write latency of foreground applications is independent from the distance between the primary and secondary sites. Therefore, HyperReplication/A applies to disaster recovery scenarios where the primary and secondary sites are far away from each other, or the network bandwidth is limited.

3.3.3 Active-Active Arrays (HyperMetro)

HyperMetro, an array-level active-active technique provided by OceanStor Dorado V3, enables two storage systems to work in active-active mode in the same equipment room, in the same city, or in two places that are 100 km away from each other. HyperMetro allows two LUNs from two storage arrays to maintain real-time data consistency and to be accessible to hosts. If one storage array fails, hosts will automatically choose the path to the other storage array for service access. If only one storage array can be accessed by hosts due to failures of the links between storage arrays, the arbitration mechanism determines which storage array continues to provide services. The quorum server is deployed at the third-place site.

HyperMetro supports both Fibre Channel and IP networking (GE/10GE).

Figure 3-26 Active-active arrays



Technical features of HyperMetro:

1. Gateway-free A/A solution: The networking is simple, and the deployment is easy. Without the use of a gateway, the reliability and performance are better because there is one less possible failure point and 0.5 ms of latency caused by a gateway is avoided.
2. A/A mode: Storage arrays in both data centers support data read and write. The upper-layer application system can make full use of such a service capability to implement load balancing across data centers.
3. Site access optimization: The UltraPath is optimized specific to the A/A scenario. It can identify region information to reduce cross-site access, thereby reducing latency. The UltraPath can read data from the local or remote storage array. However, when the local storage array is working properly, the UltraPath preferentially reads data from and writes data to the local storage array, preventing data read and write across data centers.
4. FastWrite: In a common SCSI write process, a write request goes back and forth between two data centers twice to complete two interactions, namely Write Alloc and Write Data. FastWrite optimizes the storage transmission protocol and reserves cache space on the destination array for receiving write requests. Write Alloc is omitted and only one interaction is required. FastWrite reduces the time for data synchronization between two arrays by 50%, improving the overall performance of the HyperMetro solution.
5. Service granularity–based arbitration: If links between two sites fail, HyperMetro can enable some services to run in data center A and other services run in data center B based on service configurations. Compared with traditional arbitration where only one data center provides services, HyperMetro improves resource usage of hosts and storage systems and balances service loads. Arbitration granularity can be LUNs or consistency groups.
6. Automatic link quality adaptation: If multiple links exist between two data centers, HyperMetro automatically balances loads among links based on quality of each link. The system dynamically monitors link quality and adjusts load sharing ratio on two links to reduce retransmission ratio and improve network performance.

7. Compatibility with other features: HyperMetro can work with existing features such as HyperSnap, SmartThin, SmartDedupe, and SmartCompression.

3.3.4 Inline Deduplication (SmartDedupe)

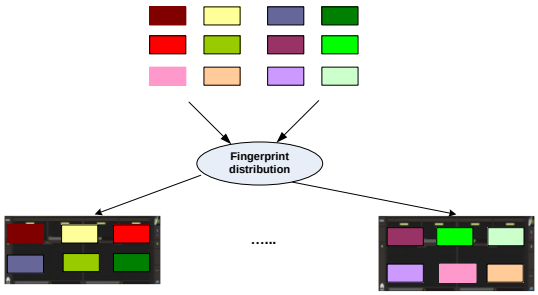
Inline deduplication deletes duplicate data online before writing data to flash media. With this function enabled, data is first cached after entering the storage system. The background automatically deletes duplicate data before writing dirty data into the flash media. Deduplication is performed in real time and is not handled in post-processing.

To ensure data correctness, if the fingerprint of the new data is the same as that of the existing data, the system compares the new data and existing data byte by byte. This method avoids data inconsistency if the system misjudges that the new data and existing data are the same only by checking their fingerprints.

Identification of zero data blocks that do not occupy the storage space of any data or metadata: When an application is reading data, zero is returned if a mapping relationship does not exist between LBA and the fingerprint. If the write data has zero data blocks, an internal zero page is used to replace the zero database, requiring no space allocation and storage, as well as saving storage space and improving performance.

Written data is sent to system controllers based on the fingerprint values and processed. Data is then evenly distributed to all SSDs, as shown in the following figure.

Figure 3-27 Working principle of deduplication



The efficiency of deduplication varies with data type. In VDI applications, deduplication ratio can reach 10 times while in database scenarios, deduplication ratio is smaller than two times. Deduplication can be disabled based on LUNs. In scenarios where higher performance and a small deduplication ratio are required, for example, databases, you can disable deduplication.

3.3.5 Inline Compression (SmartCompression)

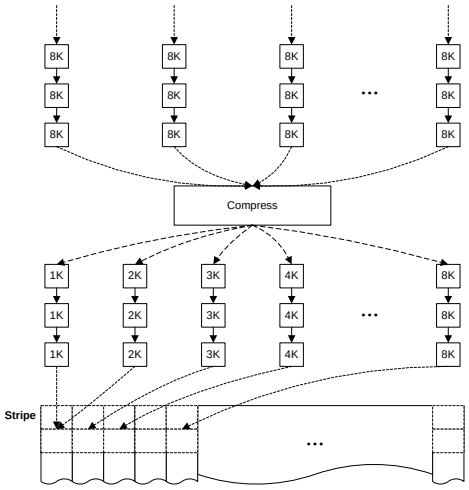
Inline compression compresses data online before writing data to flash media. In addition, compression is performed after deduplication, ensuring that no duplicate data is compressed and improving compression efficiency.

Compression is performed in real time and is not handled in post-processing. The overall compression ratio is determined by the nature of data sets. The compressed data blocks are stored in arrays but need fewer SSDs. SmartCompression minimizes write amplification (WA) of SSDs and improves durability of flash arrays.

Dorado V3 adopts the enhanced LZ4 rapid compression algorithm. The unit for storing compressed data is 1 KB, which significantly increases the data compression ratio. In the

following figure, 8 KB data blocks are compressed, converged into full stripes, and then written to disks.

Figure 3-28 Working principle of compression



The efficiency of compression varies with the data type and the compression ratio is generally 2 to 3.5 times. Compression can be disabled based on LUNs. In scenarios where higher performance is required, you can disable compression.

3.3.6 Intelligent Thin Provisioning (SmartThin)

Dorado V3 supports thin provisioning, which enables the storage system to allocate storage resources on demand. SmartThin does not allocate all capacity in advance, but presents a virtual storage capacity larger than the physical storage capacity to users. In this way, users see a larger storage capacity than the actual storage capacity. When users begin to use the storage capacity, SmartThin provides only required space to users. If the storage space is about to be consumed up, SmartThin triggers the storage resource pool expansion to add system storage capacity. The whole expansion process is transparent to users and causes no system downtime.

Application scenarios:

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

OceanStor Dorado V3 identifies zero data and its zero page does not occupy any storage space or generate any metadata.

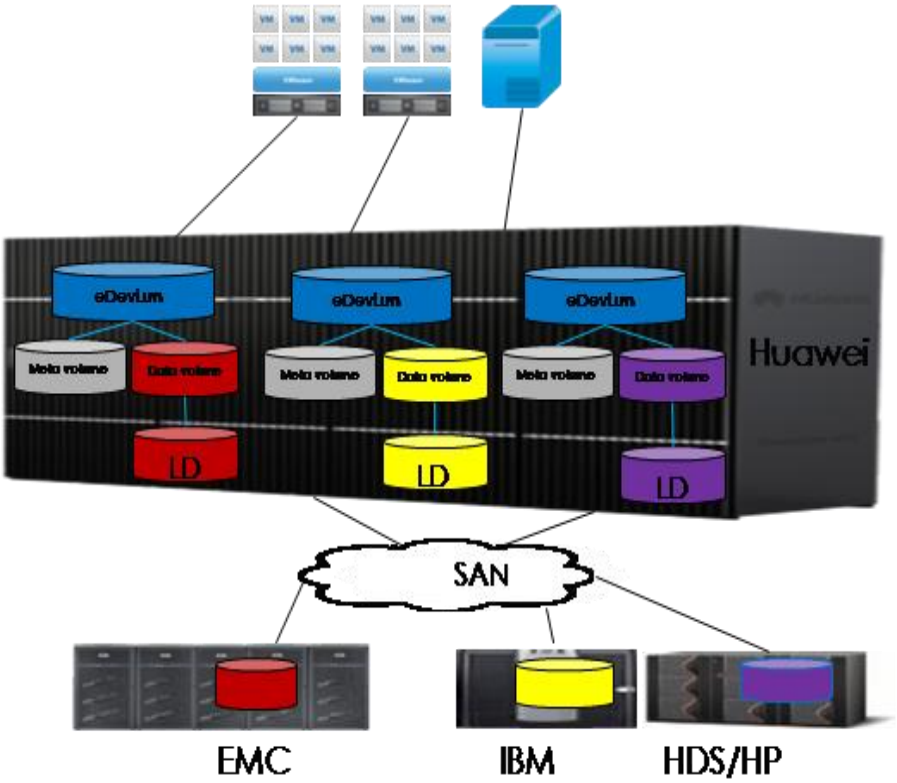
3.3.7 Heterogeneous Virtualization (SmartVirtualization)

OceanStor Dorado V3 aims at providing rich virtualization functions for heterogeneous storage systems of customers. The heterogeneous online migration function allows data to be smoothly migrated among heterogeneous LUNs without interrupting services. The heterogeneous remote replication function implements disaster recovery for heterogeneous LUNs. The heterogeneous snapshot function implements rapid backup for heterogeneous LUNs.

SmartVirtualization uses LUNs mapped from heterogeneous storage systems to the local storage system as logical disks (LDs) that can provide storage space for the local storage system and create eDevLUNs that can be mapped to the host on LDs. LDs provide data storage space for data volumes of eDevLUNs, and the local storage system provides storage space for metadata volumes of eDevLUNs.

SmartVirtualization ensures the data integrity of external LUNs.

Figure 3-29 Heterogeneous storage virtualization



eDevLUNs and local LUNs have the same properties. For this reason, SmartMigration is used to provide online migration for heterogeneous LUNs.

SmartVirtualization applies to:

1. Heterogeneous array takeover

   As customers build their data center over time, their storage arrays in the data center may come from different vendors. How to efficiently manage and use storage arrays from different vendors is a technical challenge that storage administrators must tackle. Storage administrators can leverage the heterogeneous virtualization takeover function of

Huawei OceanStor to simplify the management of heterogeneous arrays. To manage all heterogeneous arrays, now they only need to manage Huawei storage arrays and remarkably reduce their workloads. This scenario is characterized by simplified system management.

2. Heterogeneous data migration

   A customer data center may have a large number of third-party storage systems whose warranty periods are about to expire or whose performance cannot meet service requirements. After purchasing OceanStor Dorado V3, the customer wants to migrate services from the legacy storage systems to the new ones. In this case, the customer can use the online migration function of SmartVirtualization to migrate data on LUNs of the legacy storage systems to the new ones. The migration process will not affect ongoing host services, but the third-party LUNs must have been taken over before the migration. Characteristic of this scenario: Ongoing host services are uninterrupted when data on LUNs of the third-party storage systems is being migrated.

3.3.8 Intelligent Data Migration (SmartMigration)

OceanStor Dorado V3 provides intelligent data migration based on LUNs. Services on a source LUN can be completely migrated to the target LUN without interrupting ongoing services. In addition to service migration within a storage system, LUN migration also supports service migration between a Huawei storage system and a compatible third-party storage system.

SmartMigration replicates all data from a source LUN to a target LUN and uses the target LUN to completely replace the source LUN after the replication is complete. Specifically, all internal operations and requests from external interfaces are transferred from the source LUN to the target LUN transparently.

Implementation of SmartMigration has two stages:

1. Service data synchronization

   Ensures that data is consistent between the source LUN and target LUN after service migration.

2. LUN information exchange

   Enables the target LUN to inherit the WWN of the source LUN without affecting host services.

SmartMigration applies to:

1. Storage system upgrade by working with SmartVirtualization

   SmartMigration works with SmartVirtualization to migrate data from legacy storage systems (storage systems from Huawei or other vendors) to new Huawei storage systems to improve service performance and data reliability.

2. Service performance adjustment

   SmartMigration can be used to improve or reduce service performance. When the performance of a legacy storage system fails to meet service requirements, you can migrate services to a storage system that provides higher performance. Conversely, if services on a legacy storage system do not need high storage performance, you can migrate those services to a low-performance storage system. For example, cold data can be migrated to entry-level storage systems without interrupting host services to reduce operating expense (OPEX).

3.4 System Management

OceanStor Dorado V3 provides device management interfaces and integrated northbound management interfaces. Device management interfaces include a graphic management interface DeviceManager and a command-line interface (CLI). Northbound interfaces are RESTful interfaces, supporting SMI-S, SNMP, evaluation tools, and third-party network management plug-ins. For details, refer to the compatibility list of OceanStor Dorado V3.

3.4.1 Device Management

3.4.1.1 DeviceManager

DeviceManager is a common graphic management system for Huawei OceanStor systems and accessed through a web page. The GUI client uses standard HTTP protocol to communicate with Dorado V3. Most system operations can be executed on DeviceManager, but certain operations must be run in the CLI.

3.4.1.2 CLI

The CLI allows administrators and other system users to perform supported operations. You can define key-based SSH user access permission to enable users to compile scripts on a remote host. You are not required to save the passwords in the scripts and log in to the CLI remotely.

3.4.2 Northbound Management

3.4.2.1 Restful API

Restful API of Dorado V3 allows system automation, development, query, and allocation based on HTTPS interfaces. With Restful API, you can use third-party applications to control and manage arrays and develop flexible management solutions for Dorado V3.

3.4.2.2 SNMP

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

3.4.2.3 SMI-S

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

3.4.2.4 Tools

OceanStor Dorado V3 provides diversified tools for pre-sales assessment and post-sales delivery. These tools can be accessed through WEB, ToolKit, DeviceManager, SystemReporter, and CloudService and effectively help users deploy, monitor, analyze, and maintain OceanStor Dorado V3.

3.4.3 OpenStack Integration

OceanStor Dorado V3 launches the latest OpenStack Cinder Driver for itself in the OpenStack community as the community is updated. Vendors of commercial OpenStack versions can

obtain OpenStack Cinder Driver and integrate it to OpenStack so that their products support OceanStor Dorado V3.

OceanStor Dorado V3 provides four versions of OpenStack Cinder Driver: OpenStack Juno, Kilo, Liberty, and Mitaka. In addition, OceanStor Dorado V3 supports commercial versions of OpenStack such as Huawei FusionSphere OpenStack, Red Hat OpenStack Platform, and Mirantis OpenStack.

3.4.4 Virtual Machine Plug-ins

OceanStor Dorado V3 supports:

1. VMware vSphere
   - VMware VAAI (vStorage APIs for Array Integration)
   - VASA (vStorage APIs for Storage Awareness)
   - SRM (Site Recovery Manager) feature
   - vCenter plug-in for unified management on the vCenter interface
2. Windows Hyper-V
   - Windows Thin space reclamation technology
   - Windows Offload Data Transfer (ODX) technology
   - System Center plug-in which can be managed by Microsoft System Center Operations Manager (SCOM) and System Center Virtual Machine Manager (SCVMM)

3.4.5 Host Compatibility

OceanStor Dorado V3 supports mainstream host components, including operating systems, virtualization software, HBAs, volume management, and cluster software. OceanStor Dorado V3 supports a wider range of operating systems and VM platforms for mainstream database software.

- Host operating system:

  Windows Server, Red Hat Enterprise Server, SuSE Enterprise Linux Server, CentOS, Oracle Linux, Red Flag, Kylin, Neokylin, Rocky, HP-UX, AIX, Solaris, Mac OS X
- Mainstream virtualization software:

  Huawei FusionSphere, VMware, XenServer, Windows Hyper-V, Oracle VM, IBM VIOS, Red Hat Enterprise Virtualization
- HBA:

  QLogic, Emulex, Brocade
- Volume management software:

  Oracle ASRU, Symantec VxVM
- Cluster software:

  Windows WSFC, Symantec Veritas Cluster Server, IBM PowerHA, RoseHA, VMwareHA
- Host multipathing software:

OceanStor Dorado V3 supports multipathing software built in host operating systems for virtualization and self-developed multipathing software UltraPath to better work with Windows, Linux, AIX, Solaris, and VMware systems and ensure service continuity and high reliability.

NOTE

This section lists part of host compatibility information. For more information about OceanStor DoradoV3 compatibility, visit http://support-open.huawei.com/ready/index.jsf;jsessionid=D6185C2E6A2B671741F928B671E4B098.

4 Best Practices

Huawei is continuously collecting requirements of important customers in major industries and summarizes the typical high-performance storage applications and the challenges that face these customers. Based on the information, Huawei provides best practices which are tested and verified together with application suppliers.

Application software	Best Practice Deliverable	Customer Benefit
Oracle	Dorado 6000 V3–based Oracle database reference architecture	Guides operations in Dorado V3 Oracle application scenarios.
	Benchmark (typical configuration, basic and value-added performance)	Guides configurations in Oracle scenarios and provides performance comparison.
	Backup: Oracle uses Dorado 6000 V3 lossless snapshots	Lossless snapshots are used for Oracle data backup and rapid recovery.
	Disaster recovery: Dorado 6000 V3 asynchronous replication is used to protect Oracle databases	Asynchronous replication is used to ensure Oracle service continuity.
	Test and development: The Oracle test and development system uses Dorado 6000 V3	Guides customers to select optimal configurations for Oracle services.
	Data migration: Use the heterogeneous takeover and migration functions of Dorado 6000 V3 to migrate Oracle data	Multiple migration solutions are used to improve migration efficiency and reduce services on impact.
VMware	Dorado 6000 V3–based VMware VDI reference architecture	Guides operations in Dorado V3 VMware VDI application scenarios.
	Dorado 6000 V3–based VMware VSI reference architecture	Guides operations in Dorado V3 VMware VSI application scenarios.
	Benchmark-VDI (typical configuration, basic and	Guides configurations in VMware VDI scenarios and

Application software	Best Practice Deliverable	Customer Benefit
	value-added performance)	provides performance comparison.
	Benchmark-VSI (typical configuration, basic and value-added performance)	Guides configurations in VMware VSI scenarios and provides performance comparison.
	Test and development: The VMware test and development system uses Dorado 6000 V3	Guides customers to select optimal configurations for VMware scenarios.
	Joint certification: VMware Horizon Fast Track Proven Solution	Joint certification improves cognition of Dorado in VMware VDI scenarios.
SQL Server	Dorado 6000 V3–based SQL Server database reference architecture	Guides operations in Dorado V3 SQL Server application scenarios.
	Benchmark (typical configuration, basic and value-added performance)	Guides configurations in SQL Server scenarios and provides performance comparison.
	Test and development: The SQL Server test and development system uses Dorado 6000 V3	Guides customers to select optimal configurations for SQL Server services.
SAP HANA	Dorado 6000 V3–based SAP HANA reference architecture	Guides operations in Dorado V3 SAP HANA application scenarios.
	Benchmark (typical configuration, basic and value-added performance)	Guides configurations in SAP HANA scenarios and provides performance comparison.
	Efficiency optimization: The snapshot function of Dorado 6000 V3 simplifies enterprises' SAP HANA Landscape management	Improves SAP HANA service processing efficiency.
	Joint certification: Dorado 6000 V3 series SAP HANA TDI certification	Joint certification improves cognition of Dorado in SAP HANA scenarios, promoting sales.
FusionSphere	Dorado 6000 V3–based FusionSphere VDI reference architecture	Guides operations in Dorado V3 FusionSphere application scenarios.
	Benchmark (typical configuration, basic and value-added performance)	Guides configurations in FusionSphere scenarios and provides performance comparison.

For more information about best practices, visit http://e.huawei.com/en/marketing-material.

5 Conclusion

OceanStor Dorado V3, an all-flash storage array specially designed for critical enterprise services, adopts the multi-controller architecture dedicated to flash storage and disk-controller coordination FlashLink technology to meet the requirements of unexpected service growth. Furthermore, 1 ms gateway-free active-active design ensures always-on businesses. Inline deduplication and compression technologies improve efficiency and cut TCO by 90%.

OceanStor Dorado V3 provides high-performance, reliable, and efficient storage for enterprise applications such as databases and virtualization, helping the financial industry, governments, enterprises, and carriers smoothly evolve to the flash memory age.

6 Acronyms and Abbreviations

Table 6-1 Acronyms and abbreviations

Acronym and Abbreviation	Full Spelling
FlashLink	FlashLink
CK	Chunk
CKG	Chunk Group
DIF	Data Integrity Field
FC	Fiber Channel
FTL	FLASH Translation Layer
GC	Garbage Collection
SSD	Solid State Disk
LUN	Logical Unit Number
OLAP	On-Line Analytical Processing
OLTP	On-Line Transaction Processing
OP	Over-Provisioning
RAID	Redundant Array of Independent Disks
SAS	Serial Attached SCSI
SCSI	Small Computer System Interface
SSD	Solid State Disk
T10 PI	T10 Protection Information
VDI	Virtual Desktop Infrastructure
VSI	Virtual Server Infrastructure
WA	Write amplification

Acronym and Abbreviation	Full Spelling
Wear Leveling	Wear Leveling
TCO	Total Cost of Ownership