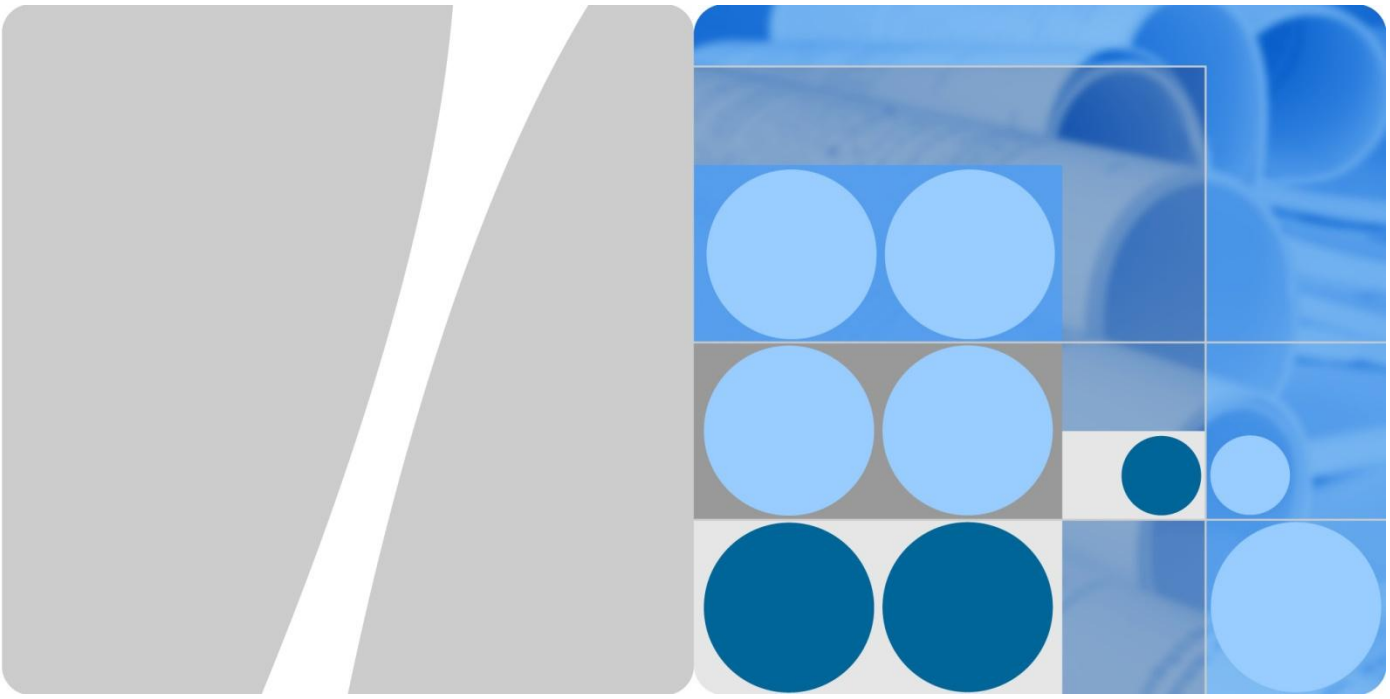


HUAWEI OceanStor 6800 V3 and 6900 V3 Converged Storage Systems

Technical White Paper

Issue 5.0

Date 2014-08

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://enterprise.huawei.com

Contents

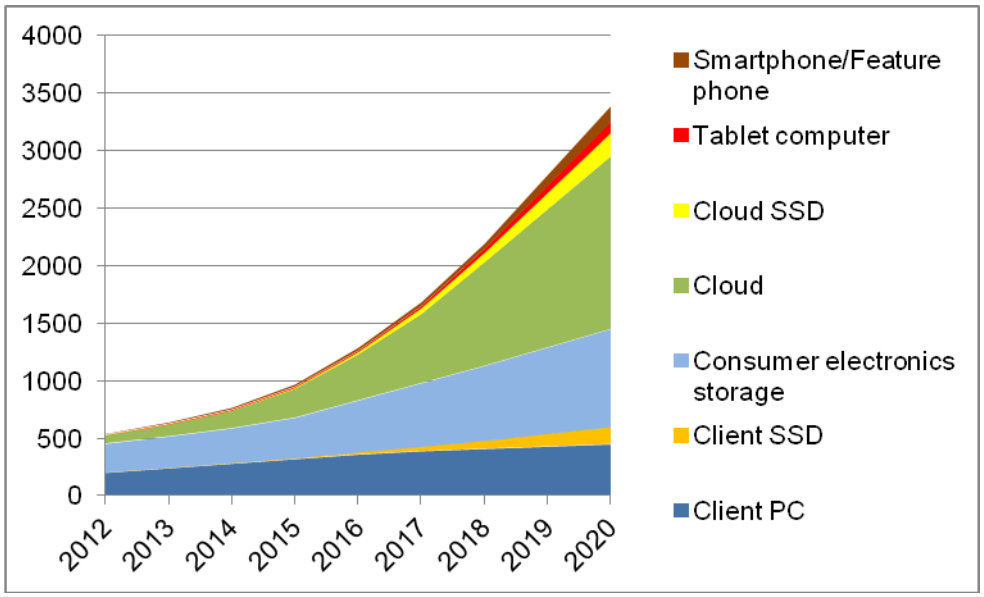
1 Overview **1**
2 Definition of "Converged & Compact, Intelligent&Efficient" **3**
3 Storage Architecture **5**
4 Converged: Heterogeneous Resource Management **7**
5 Converged: SAN&NAS Converged Architecture **10**
6 Compact: High-Density Disk Enclosures **12**
7 Compact: SmartDedupe& SmartCompression **21**
8 Intelligent: SmartCache **23**
9 Intelligent: SmartTier **25**
10 Intelligent: SmartThin **28**
11 Intelligent: SmartMigration **31**
12 Efficient: SmartQoS **33**
13 Efficient: SmartPartition **36**
14 Summary **38**
15 Acronyms and Abbreviations **39**

1 Overview

Evolving from mainframe servers to midrange computers, PCs, and desktop Internet, the information technology (IT) is penetrating into all walks of life. Nowadays, we are embracing the mobile Internet era. The change of application environments hastens data explosion. According to Gartner's statistics, about 2.6 EB of data was generated around the world in the era of midrange computers and 15.8 EB of data when PCs were popular. In the era of desktop Internet, the amount of data was almost quadrupled, reaching 54.5 EB. Up to 1800 EB of data may be generated in the era of mobile Internet. The skyrocketing amount of data not only requires superlarge storage capacities but also imposes demanding requirements on other features of storage products.

Since data sources are increasingly diversified, clouds will gradually become the largest data sources, replacing PCs and consumer electronics (CE). The following figure shows predicted rankings of data sources.

Figure 1-1 Predicted rankings of data sources



Since data sources are changing constantly, data types change accordingly. Although the amount of critical service data, such as databases, increases continuously, it accounts for a decreasing percentage of the total data amount; whereas enterprise office data, such as emails

and large media files, once accounted for the highest percentage of the total data amount. In recent years, since the amount of personal data increases sharply, media and entertainment data replaces enterprise office data as the largest data sources. In 1993, critical service data and enterprise office data accounted for 50% of the total data amount respectively, and the amount of personal data could be ignored. In 2002, 70% of data was enterprise office data, and 20% was critical service data. Since 2010, personal data accounts for 50% of the total data volume, whereas enterprise office data accounts for 40%, and critical service data accounts for only 10%.

Different types of data from diversified sources have different requirements on the performance, reliability, and costs of storage media. Critical service data requires high-performance and robust-reliability storage devices, whereas personal entertainment data requires inexpensive storage devices. The reality is that critical service data and personal entertainment data usually need to be stored in a single set of storage device. Such contradicting requirements impose new challenges. To keep with IT development, next-generation mid-range storage products must have:

- Integrated, simple, intelligent, and cost-effective system architecture
- High flexibility, meeting diverse storage needs
- Agile data planning and management
- Rich and practical functions

2 Definition of "Converged & Compact, Intelligent&Efficient"

HUAWEI OceanStor 6800 V3 and 6900 V3 Converged storage systems (the V3 Converged storage systems for short) are the next-generation unified storage products specifically designed for enterprise-class applications. Leveraging a storage operating system built on a cloud-oriented architecture, a powerful new hardware platform, and suites of intelligent management software, the V3 converged storage systems deliver industry-leading functions, performance, efficiency, reliability, and ease-of-use. They provide data storage for applications such as large-database Online Transaction Processing (OLTP)/Online Analytical Processing (OLAP), file sharing, and cloud computing, and can be widely applied to industries ranging from government, finance, telecommunication, energy, to media and entertainment (M and E). Meanwhile, the V3 converged storage systems can provide a wide range of efficient and flexible backup and disaster recovery solutions to ensure business continuity and data security, delivering excellent storage services.

Converged:

- Convergence of heterogeneous storage systems

  Thanks to the heterogeneous virtualization function, the V3 converged storage systems can efficiently manage storage systems from other mainstream vendors and unify resource pools for central and flexible resource allocation.

- Convergence of SAN and NAS

  SAN and NAS are converged on one platform that provides file and block access services, enabling flexible configurations and protecting customer investment.

Compact:

- High-density disk enclosures

  High-density disk enclosures fully utilize their 4 U space. Vertical instead of horizontal disk slots are provided, and the size of cascading boards is reduced by more than half, providing as much space as possible for disks.

  The delicate air duct design, optimized fan speed adjustment policy, and enhanced fan usage resolve the heat dissipation issue and meet the noise and power consumption requirements.

- SmartDedupe and SmartCompression

  SmartDedupe and SmartCompression deduplicate and compress data before storage, reducing space for data storage, lowering the storage cost per GB, and improving the data storage efficiency.

Intelligent:

- SmartTier

SmartTier automatically analyzes data access frequencies per unit time and migrates data to disks of different performance levels based on the analysis result. (High-performance disks store most frequently accessed data, performance disks store less frequently accessed data, and large-capacity disks store seldom accessed data.) In this way, the optimal overall performance is achieved, and the IOPS cost is reduced.

- SmartThin

  SmartThin allocates storage space on demand rather than pre-allocating all storage space at the initial stage. It is more cost-effective because customers can start business with a few disks and add disks based on site requirements. In this way, the initial purchase cost and TCO are reduced.

- SmartMigration

  SmartMigration migrates host services from a source LUN to a target LUN without interrupting these services and then enables the target LUN to take over services from the source LUN without being noticed by the hosts. After the service migration is complete, all service-related data has been replicated from the source LUN to the target LUN.

- SmartCache

  SmartCache uses SSDs to compose the performance tier of storage systems and employs SSDs' powerful random read and write capabilities to identify and accelerate hotspot data. In addition, SmartCache can accelerate metadata access, achieving system performance optimization and improvement.

Efficient:

- SmartQoS

  SmartQoS categorizes service data based on data characteristics (each category represents a type of application) and sets a priority and performance objective for each category. In this way, resources are allocated to services properly, fully utilizing system resources.

- SmartPartition

  The core idea of SmartPartition is to ensure the performance of mission-critical applications by partitioning core system resources. Users can configure cache partitions of different sizes. The V3 converged storage systems ensure the number of cache partitions occupied by service applications. Based on the actual service condition, the V3 converged storage systems dynamically adjust the number of concurrent access requests from hosts to different cache partitions, ensuring the service application performance of each partition.

3 Storage Architecture

The V3 Converged storage systems employ a 6 U four-controller compact design, improving the I/O port density per unit space and simplifying controller networking. The four controllers are fully interconnected based on PCIe 3.0. Figure 3-1 shows the logical connections. PCIe switching chips are built in each controller. The chips enable high-speed interconnections among the four controllers and support 80-lane PCIe 3.0 connections, providing 80 GB/s interconnection bandwidth. The V3 Converged storage systems support a maximum of eight controllers. The four controllers in one controller enclosure are interconnected with the four controllers in the other controller enclosure over 10GE switches for service exchange. Each controller is connected to two switches through at least two 10GE ports. The global array design eliminates single points of failure, supports the IP address expansion of the four controllers in one controller enclosure, and provides 16 or more controllers.

Figure 3-1 6 U four-controller full interconnection architecture

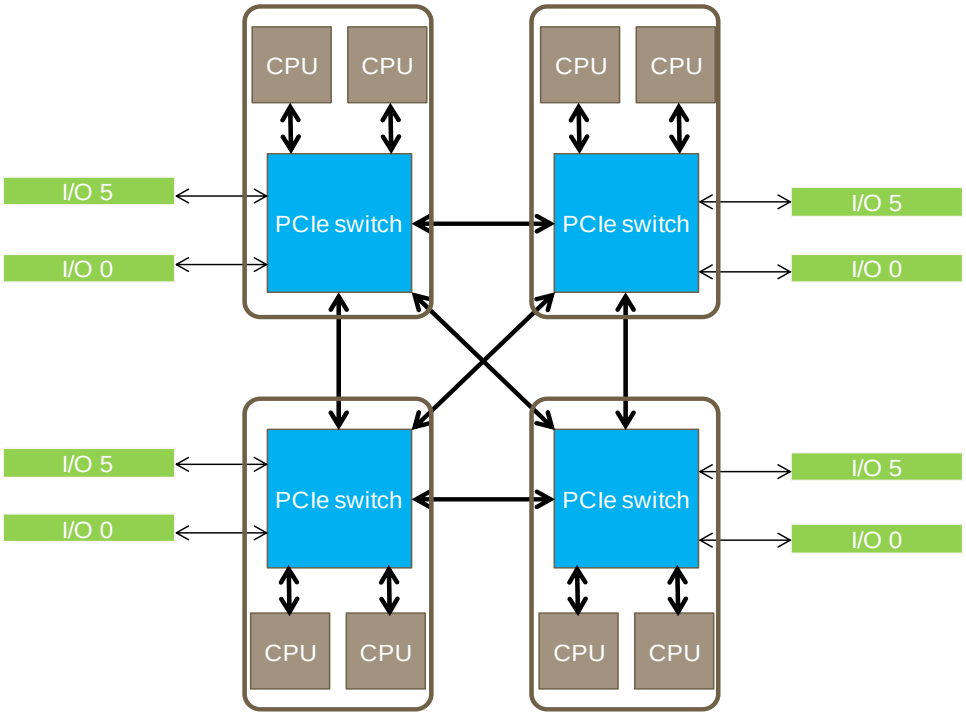
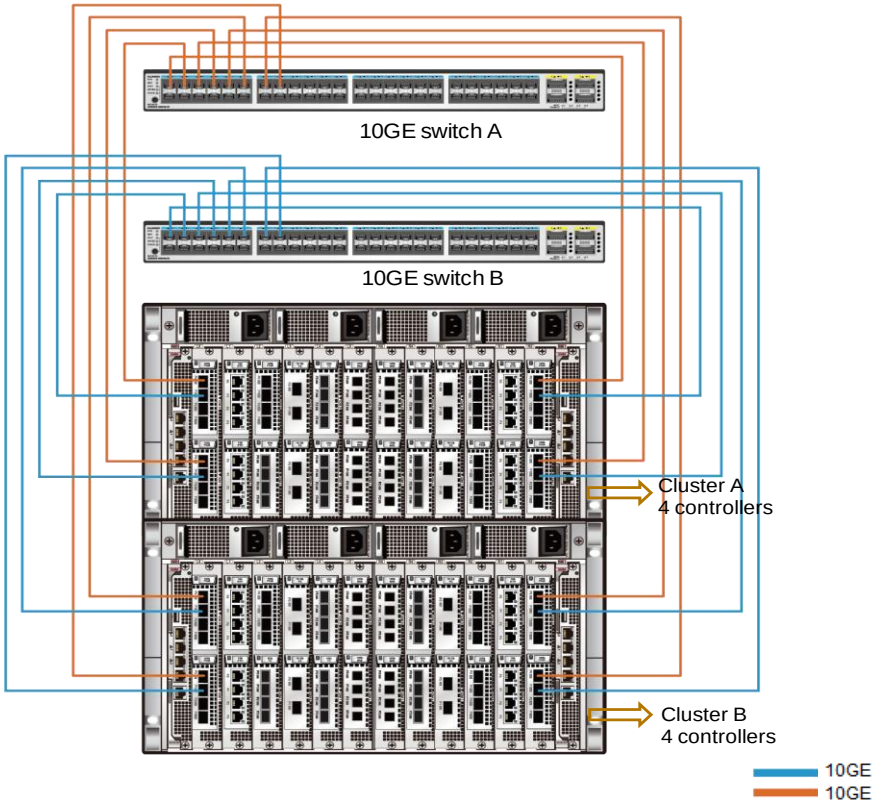


Figure 3-2 shows the connections among multiple controllers of the V3 Converged storage systems. A full-redundancy networking mode is adopted, eliminating single points of failure. Built-in backup battery units (BBUs) and coffer disks help users easily cope with incidents.

Two four-controller clusters are interconnected using 10GE switches. Each controller in a cluster is connected to two switches by two 10GE cables, ensuring that the cluster works correctly even when one link is down. In the mean time, 10GE switches enable future cluster expansion.

Figure 3-2 Network diagram of a storage system with eight controllers



The V3 Converged storage systems boast all-PCIe 3.0 interconnection, back-end SAS 3.0, and high-speed channels and powerful computing capability of Intel Ivy Bridge CPUs, meeting increasingly demanding performance requirements.

4 Converged: Heterogeneous Resource Management

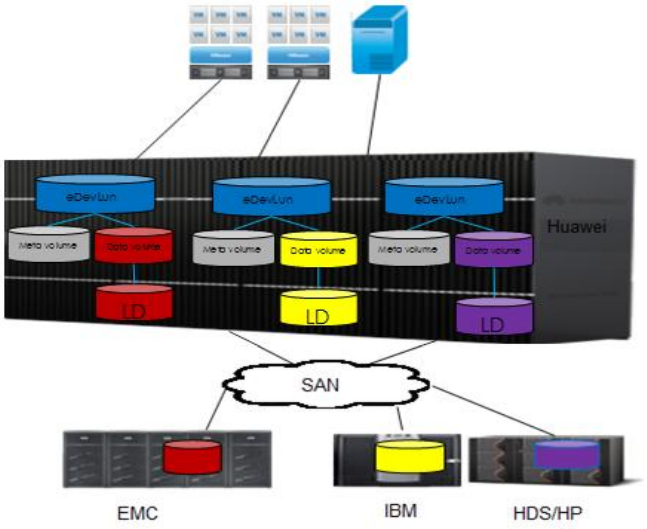
The V3 Converged storage systems aim at providing rich virtualization functions for heterogeneous storage systems of customers:

- The heterogeneous takeover function reduces complexity in managing heterogeneous storage systems and improves LUN performance.
- The heterogeneous online migration function allows data to be smoothly migrated among LUNs of heterogeneous storage systems without interrupting services.
- The heterogeneous remote replication function implements disaster recovery for LUNs of heterogeneous storage systems.
- The heterogeneous snapshot function implements rapid backup for LUNs of heterogeneous storage systems.

The heterogeneous virtualization feature provided by the V3 converged storage systems is called SmartVirtualization.

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, and the local storage system provides storage space for meta volumes of eDevLUNs. SmartVirtualization ensures the data integrity of external LUNs.

Figure 4-1 SmartVirtualization



eDevLUNs and local LUNs have the same properties. For this reason, SmartMigration, HyperReplication/S, HyperReplication/A, and HyperSnap are used to provide online migration, synchronous remote replication, asynchronous remote replication, and snapshot functions respectively for LUNs of heterogeneous storage systems. Meanwhile, SmartQoS, SmartPartition, and cache write back are used to improve the LUN performance of heterogeneous storage systems.

SmartVirtualization applies to:

- Heterogeneous array takeover

  As users' data centers develop, storage systems in the data centers may come from different vendors. How to efficiently manage and apply storage systems from different vendors is a challenge that storage administrators must tackle. Storage administrators can leverage the takeover function of SmartVirtualization to simplify heterogeneous array management. They need only to manage Huawei storage systems, and their workloads are remarkably reduced. In such a scenario, SmartVirtualization simplifies system management.

- Heterogeneous data migration

  A large number of heterogeneous storage systems whose warranty periods are about to expire or whose performance cannot meet service requirements may exist in a customer's data center. After purchasing the V3 converged storage systems, the customer wants to migrate services from the existing storage systems to the new storage systems. The customer can leverage the online migration function of SmartMigration to migrate data on LUNs of heterogeneous storage systems to the new storage systems. The migration process has no adverse impact on ongoing host services, but the LUNs must be taken over before the migration. In such a scenario, SmartVirtualization ensures ongoing host services when data on LUNs of heterogeneous storage systems is migrated.

- Heterogeneous disaster recovery

  If service data is scattered at different sites and there are demanding requirements for service continuity, the service sites need to serve as backup sites mutually, and service switchovers can be performed between sites. When a disaster occurs, a functional service site takes over services from the failed service site and recovers data. However, as storage systems at the data site come from different vendors, data on the storage systems cannot be backed up mutually. The synchronous and asynchronous replication functions

of SmartVirtualization enable data on LUNs of heterogeneous storage systems to be backed up mutually, achieving data disaster recovery among sites.

- Heterogeneous data protection

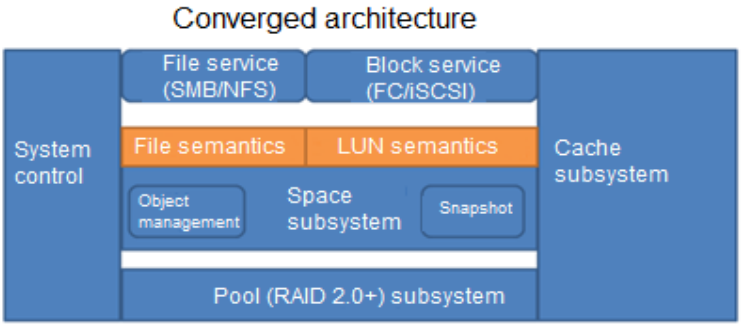
  Data on LUNs of heterogeneous storage systems may be attacked by viruses or damaged. SmartVirtualization leverages the heterogeneous snapshot function to create snapshots for LUNs of heterogeneous storage systems instantly, and rapidly restores data at a specific point in time using the snapshots if data is damaged.

5 Converged: SAN&NAS Converged Architecture

Built on the Huawei OceanStor OS architecture, the V3 Converged storage systems are Huawei's first-generation Converged unified storage systems that converge SAN and NAS. The storage systems provide file system and block access ports and apply to scenarios where file systems, virtualization applications, and databases are deployed.

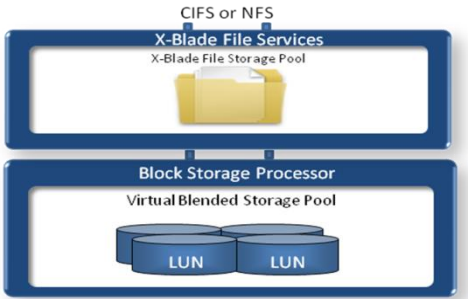
Figure 5-1 shows the converged architecture of the V3 converged storage system. On this architecture, file systems and LUNs work above the space subsystem, whereas the storage pool subsystem based on RAID 2.0+ works under the space subsystem. File systems and LUNs directly interact with the space subsystem. The file system architecture is based on objects. Each file or directory serves as an object, and each file system is a collection of objects. Generally, LUNs are divided into thin LUNs and thick LUNs. Both the thin LUNs and thick LUNs are from the storage pool system and space system instead of file systems. In this way, this converged architecture delivers a simplified software stack and provides a higher storage efficiency than the traditional unified storage architecture shown in Figure 5-2. In addition, LUNs and file systems are independent from each other.

Figure 5-1 OceanStor OS architecture



For the NAS function of EMC VNX, X-Blade (a NAS gateway) is required to provide file sharing services. File systems and block services work on different operation platforms, complicating the architecture and software stack. EMC VMAX also employs NAS gateways to implement file access semantics of Converged storage systems. For NetApp FAS series, although the unified storage system works on a unified architecture, the block semantics is based on Write Anywhere File Layout (WAFL), and its software layer is more complex than OceanStor OS. The V3 converged storage systems built on OceanStor OS are more efficient than other storage systems in terms of software stack.

Figure 5-2 Traditional unified storage architecture



OceanStor OS provides file and block access services, significantly improving the flexibility and access efficiency of Converged storage systems.

6 Compact: High-Density Disk Enclosures

High-density disk enclosures of the V3 Converged storage systems employ a dual-channel architecture for redundancy. Each of the two cascading boards provides four 6 Gbit/s mini SAS HD ports (four PHY per port). A channel can provide 4 GB/s bandwidth. A high-density disk enclosure can be cascaded to other disk enclosures.

Figure 6-1 Topology of a high-density disk enclosure of the V3 converged storage systems

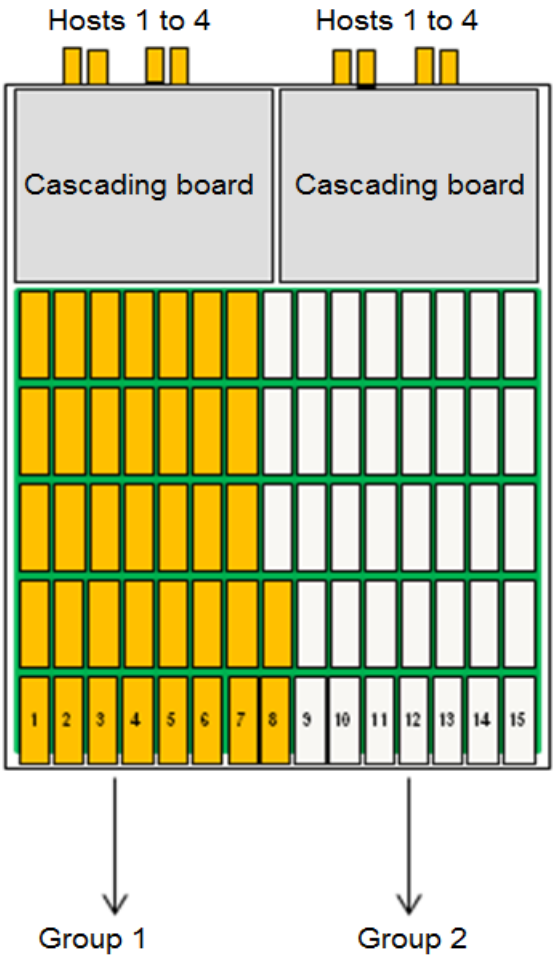
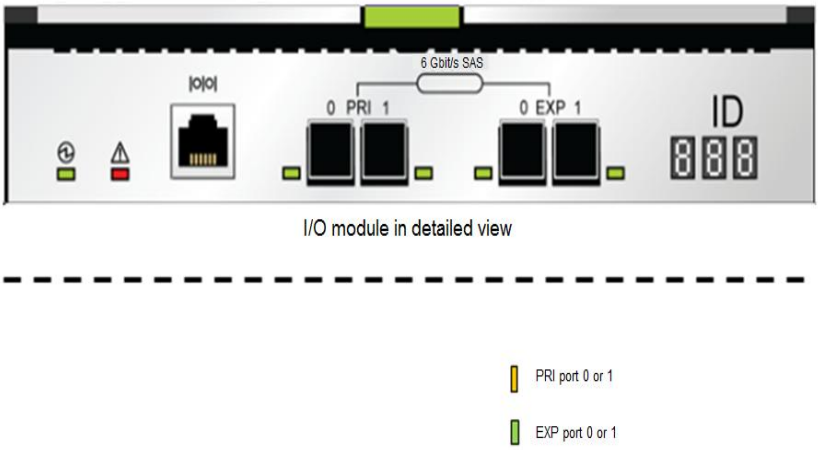


Figure 6-2 Panel of a cascading board



The following figures show the networking diagram of a high-density disk enclosure of the V3 converged storage systems from different views.

Figure 6-3 Top view of the network

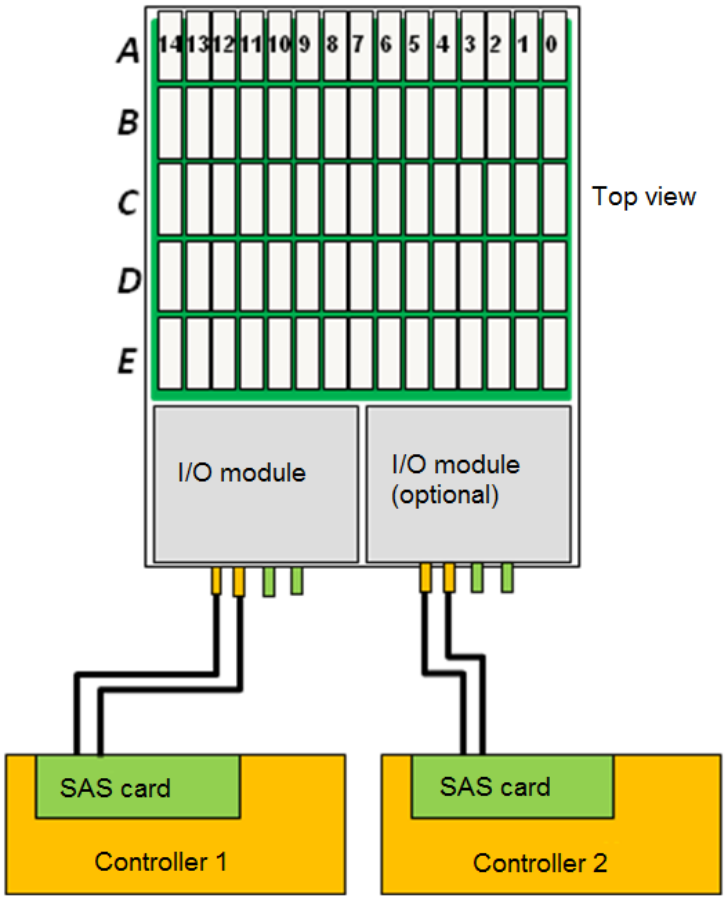
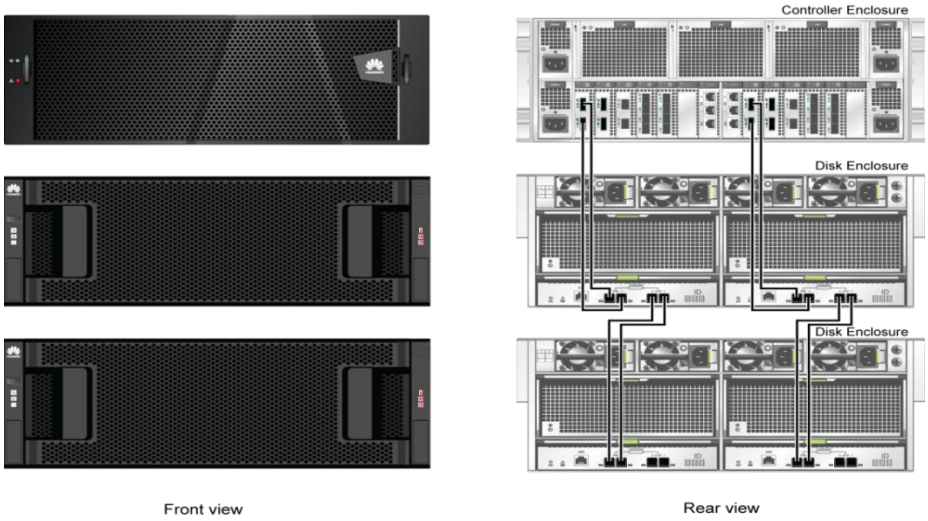


Figure 6-4 Front and rear views of the network



At least one 6 or 12 Gbit/s SAS card must be configured for each channel of a controller.

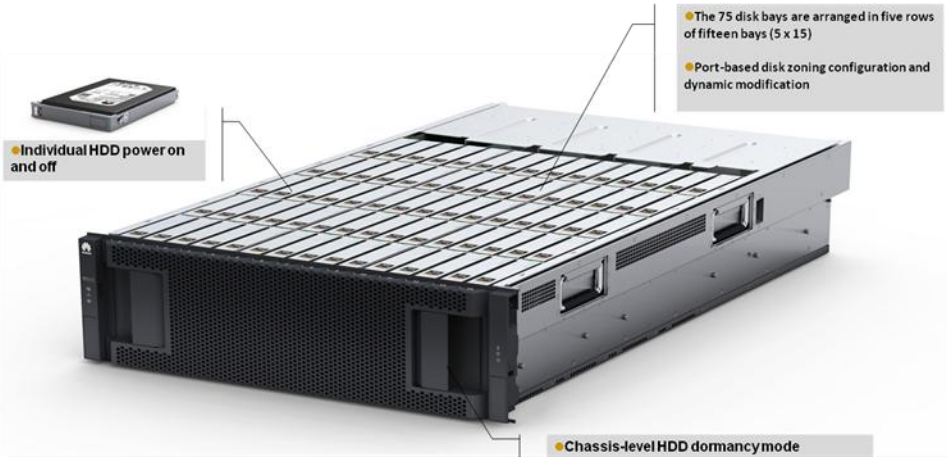
A 6 Gbit/s SAS card provides two mini SAS ports. Mini SAS to mini SAS HD cables must be used to connect high-density disk enclosures. The two ports are horizontally connected to ports HOST 1 and HOST 2 of a high-density disk enclosure, and ports HOST 3 and HOST 4 serve as EXP ports and are connected to another high-density disk enclosure.

A 12 Gbit/s SAS card provides four 12 Gbit/s mini SAS HD ports. Mini SAS HD cables must be used to connect to high-density disk enclosures. The four ports are divided into two channels: ports 0 and 1, ports 2 and 3. Ports 0 and 1 are horizontally connected to ports HOST 1 and HOST 2 of a high-density disk enclosure, and ports HOST 3 and HOST 4 serve as EXP ports and are horizontally connected to another high-density disk enclosure.

- **High-density architecture**

  High-density disk enclosures fully utilize their 4 U space. Vertical instead of horizontal disk slots are provided, and the size of cascading boards is reduced by more than half. In this way, more space is provided to disks.

Figure 6-5 Disk layout in a high-density disk enclosure

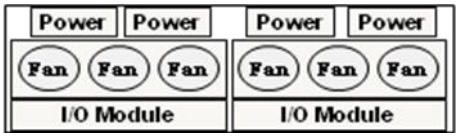


- **Heat dissipation**

  Higher density requires a more powerful heat dissipation capability. Heat generated by 75 disks must be dissipated in a timely manner. Otherwise, the storage device is overheated and damaged.

  The delicate air duct design, optimized fan speed adjustment policy, and enhanced fan usage resolve the heat dissipation issue and meet the noise and power consumption requirements.

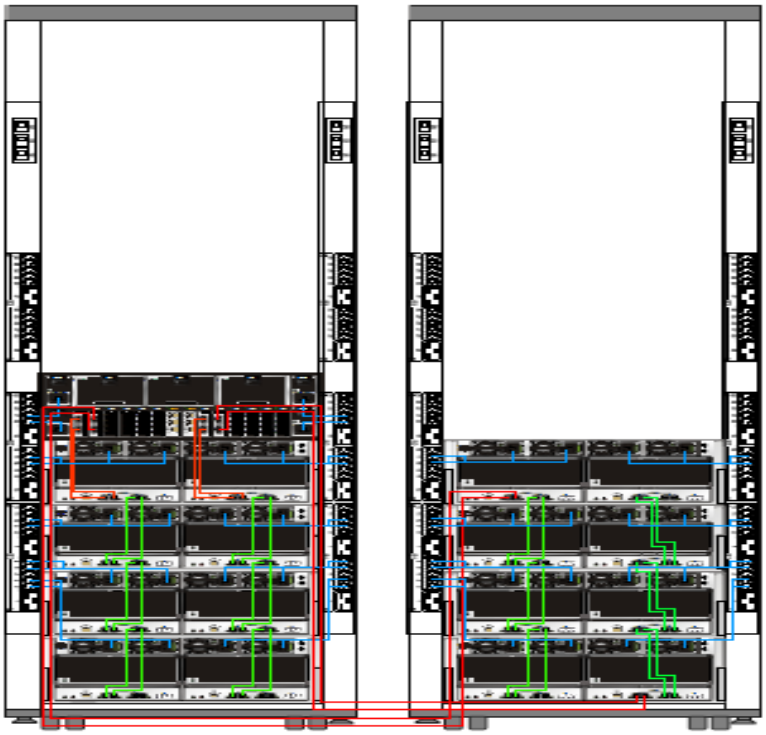
Figure 6-6 Heat dissipation



- **Weight**

  Another challenge that high-density disk enclosures must tackle is weight. Higher density means a heavier disk enclosure. Therefore, the cabinet cannot be fully configured. Recommend cabinet configurations are as follows:

Figure 6-7 Recommended configurations of high-density disk enclosures



The following table lists recommended configurations of high-density disk enclosures.

Table 6-1 Recommended configurations of high-density disk enclosures

Item	Value
Rack 1 weight	471.8 kg
Rack 2 weight	426.8 kg
Rack 1 power consumption	5.65 kW
Rack 2 power consumption	4.8 kW
Ground load bearing capacity	Single unit: 1169.4 kg per square meter
	Average: 612.5 kg per square meter

You can adjust the configurations based on site requirements.

- **Specifications**

  The following table lists the specifications of high-density disk enclosures of the V3 Converged storage systems.

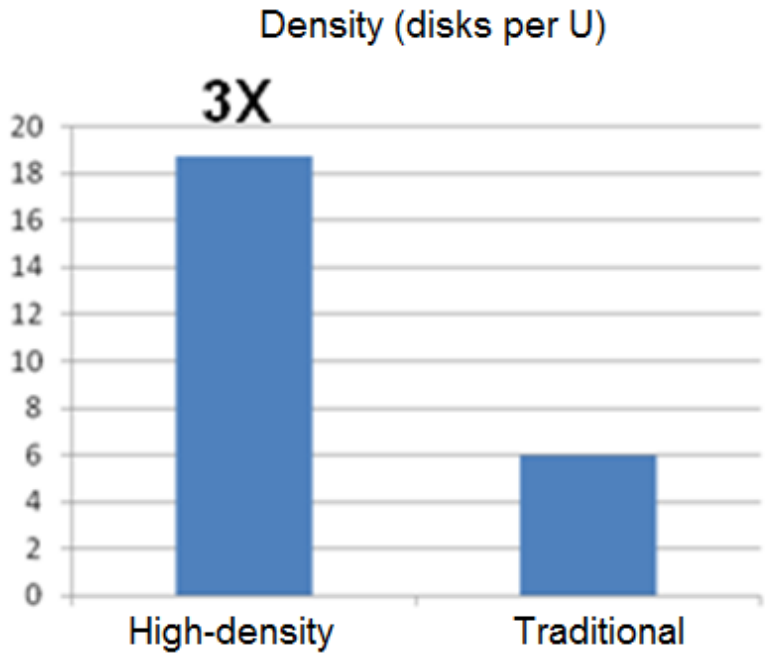
Table 6-2 High-density disk enclosure specifications

Item	Description
Host interface	Mini SAS HD (four 6 Gbit/s per I/O module)

Item	Description
Disk interface	NL-SAS 6 Gbit/s or 3 Gbit/s Up to 75 disks per enclosure
Redundant components	Power module Cooling module I/O module Disk module
Chassis dimensions	Height: 176.5 mm (4 U) Width: 446 mm Depth: 790 mm (without cable management assembly) 1070 mm (with cable management assembly)
Maximum chassis weight	46.7 kg without disks 106.7 kg with disks (4 TB NL-SAS)
Total power	Maximum: 1200 W
Input voltage	90 V to 264 V
Input frequency	45 Hz to 65 Hz (typical 50 Hz or 60 Hz)
Input current	10 A
Maximum average output power	800 W or 1200 W: 110/220 V input: four 800 W for a chassis 220 V input: two 1200 W for a chassis

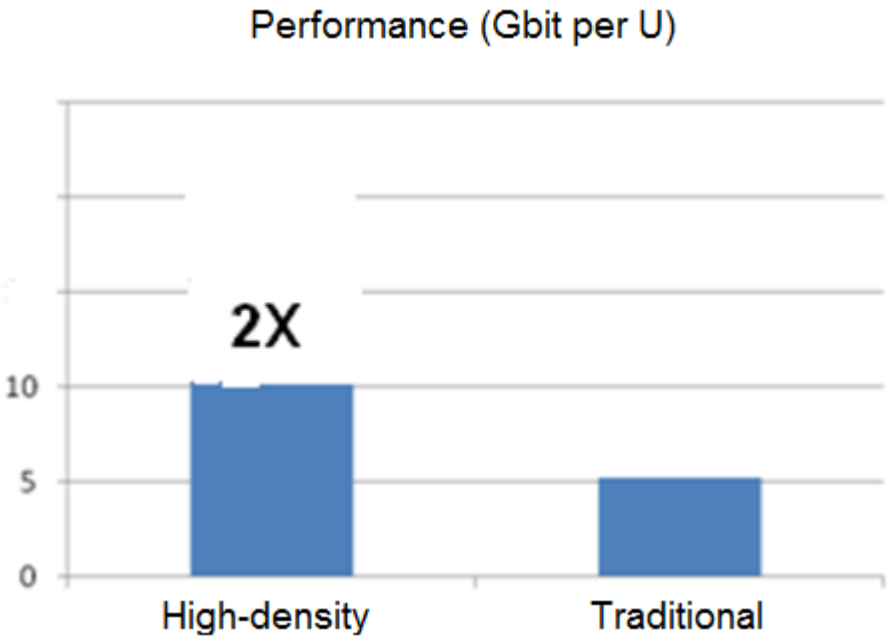
- **Disk density**

  The storage density of high-density enclosures is three times that of traditional disk enclosures.



- **Performance**

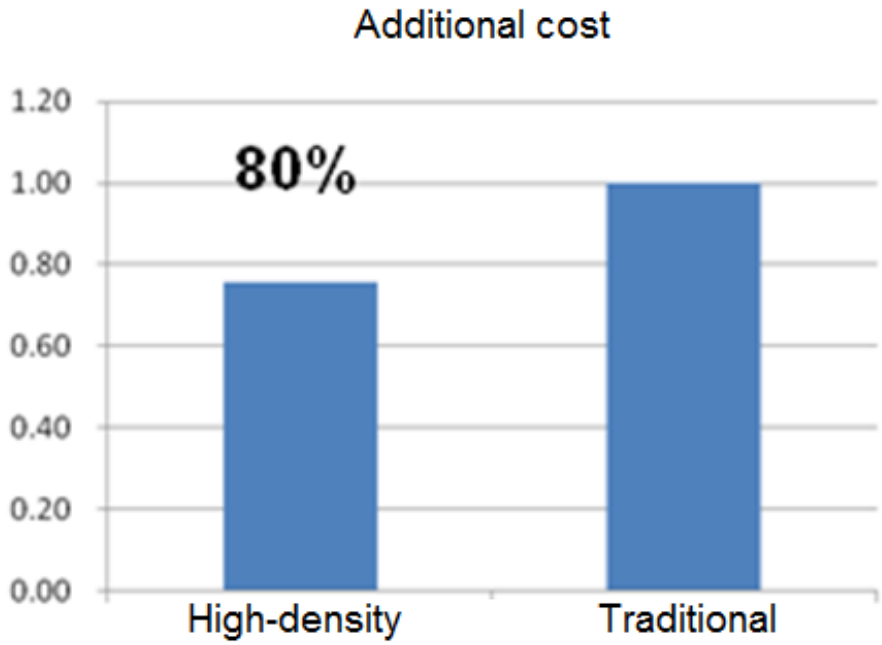
  The performance per U of high-density disk enclosures is two times that of traditional disk enclosures.



  As the SAS channel width of high-density disk enclosures is limited, you are advised to ensure that the layers of cascaded high-density disk enclosures do not exceed four to ensure storage performance.

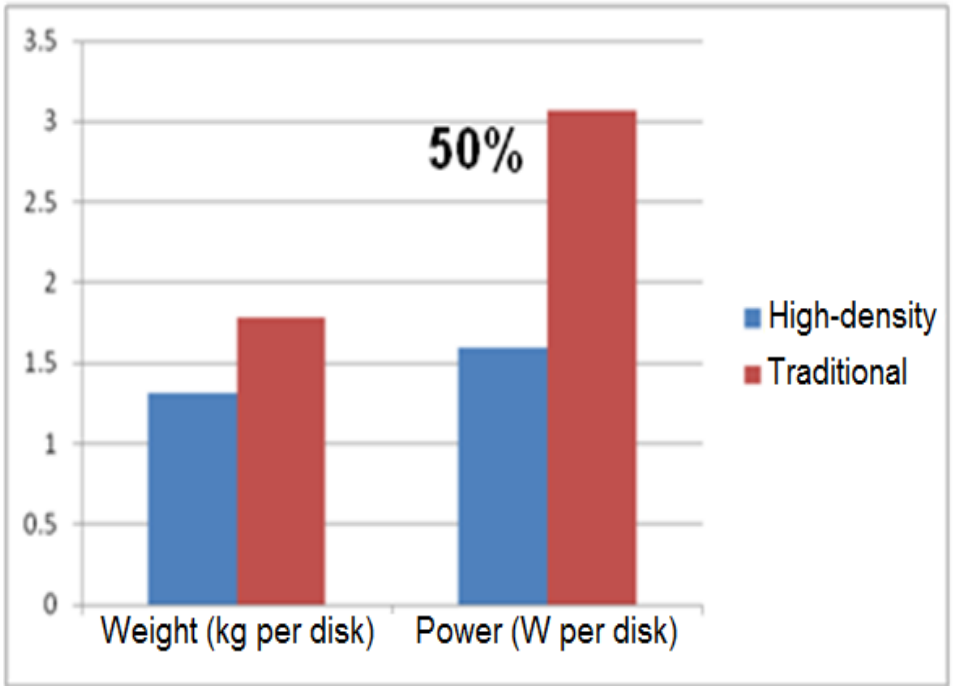
- **OPEX**

  The OPEX per disk of high-density disk enclosures is 20% lower than that of traditional disk enclosures.



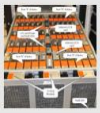
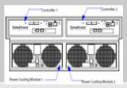
- **Power consumption**

  The power consumption per disk of high-density disk enclosures is 50% lower than that of traditional disk enclosures.



- **Competitor comparison**

Figure 6-8 Specifications comparison

Vendor	Huawei	DDN SFA12K*	EMC	Cray 1300	IBM EXP5060
Status	Release soon	Released	Released	Released	Released
PI					
Panel					
ARCH	Planting	Planting	Planting	Locking	Locking
Disks	75	84	60	84	60
Disk Type	3.5 NL-SAS	3.5SAS/SSD/NL	2.5SAS/SSD,3.5 NL	3.5SAS/SSD/NL	3.5SATA/NL,2.5 SAS
Size	4U*19"*790mm	4U*19"*1030mm	4U*19"*889mm	5U*19"*880mm	4U*19"*866mm
Weight	106.7Kg	136Kg	117.94Kg	130Kg	102.1Kg
Power	1200W	2181W	1300W	2300W	1280W
Environment	5~35℃	10~35℃	10~40℃*	5~35℃	10~35℃
Port	6G SAS*8	6G SAS*4	6G SAS*8	6G SAS*4	4G FC*4
Rack	1070mm	Customize	1100mm	1100mm	1100mm

7 Compact: SmartDedupe& SmartCompression

The V3 Converged storage systems not only deliver the SAN and NAS converged architecture, but also provide data deduplication and compression functions to shrink data for file systems and LUNs. As one of the data storage efficiency improvement methods, the data deduplication and compression functions have extended from the backup area to the primary storage area. They play a critical role in tiered storage with the SSD tier and all-flash arrays because they can save storage space and reduce the TCO of enterprise IT architectures.

The V3 Converged storage systems implement data deduplication based on file systems and thin LUNs in in-line mode. In the storage systems, the data deduplication granularity is consistent with the minimum data read and write unit (grain) of file systems or thin LUNs. Meanwhile, as users can specify the grain size (4 KB to 64 KB) when creating file systems or thin LUNs, the V3 Converged storage systems can implement data deduplication based on different granularities.

When the data deduplication function is enabled, user data is delivered to the deduplication module in grains. The deduplication module first calculates data fingerprints and then checks whether duplicate fingerprints exist. If yes, the data block is a duplicate one and will not be saved. If no, the data block is a new one and will be delivered to disks for storage. In addition, byte-by-byte comparison can be enabled or disabled. If byte-by-byte comparison is enabled, the deduplication module compares the data with the fingerprint byte by byte.

The V3 Converged storage systems implement data compression based on file systems and thin LUNs in in-line mode. When data compression is enabled, user data is delivered to the compression module in grains and is stored after being compressed. The compression module combines multiple data blocks that belong to the same compression object type and have continuous logical block addresses (LBAs) and compresses these data blocks at a time to improve the compression ratio. Tests show that compression performance is the beast when the compression granularity is 32 KB. For this reason, data blocks whose size is smaller than 32 KB are compressed together, whereas data blocks whose size is larger than 32 KB are compressed directly.

To reduce the impact of decompression on host read performance in low-compression ratio scenarios, the compression module checks whether the compression effect reaches the preset threshold. If the compression effect does not reach the threshold, the data is considered low-compression ratio data and will be stored as decompressed data. In this way, the data can be read without decompression, reducing the impact on read performance.

After SmartDedupe and SmartCompression are enabled simultaneously, data is deduplicated and then compressed before being stored onto disks. The following describes the process for

processing a data write request when SmartDedupe and SmartCompression are enabled simultaneously:

- Calculates the data fingerprint using the SHA1 algorithm.
- Checks whether a duplicate fingerprint exists in the fingerprint library of the file system or thin LUN.
- (Optional) Compares the data with the fingerprint byte by byte if byte-by-byte compression is enabled.
- Returns duplicate information if a duplicate data block exists and indicates the data block is unique if no duplicate data block exists.
- Compresses the unique data block.
- Writes the compressed data block to the disks and updates data block information in the fingerprint.
- Returns data block deduplication and compression information, such as deduplication and compression flags and physical address, to the file system or thin LUN.

The following describes the process for processing a data read request when SmartDedupe and SmartCompression are enabled simultaneously:

- Reads data from the disk based on the deduplication and compression information, such as physical address, delivered by the file system or thin LUN.
- Determines whether the data block is compressed based on the deduplication and compression flags delivered by the file system or thin LUN and directly returns the data block to the upper-layer application if the data block is not compressed.
- Decompresses and returns the data block to the upper-layer application if the data block is compressed.

In the meantime, the V3 Converged storage systems provide deduplication/compression acceleration cards. Customers can configure hardware acceleration cards based on service requirements. Hardware acceleration cards carry operations, such as fingerprint calculation, compression, and decompression, reducing CPU resource consumption. SmartDedupe and SmartCompression can work with SmartCache to employ SSDs to accelerate deduplication metadata, improving metadata query efficiency and index speed and significantly reducing the impact of data deduplication on system performance.

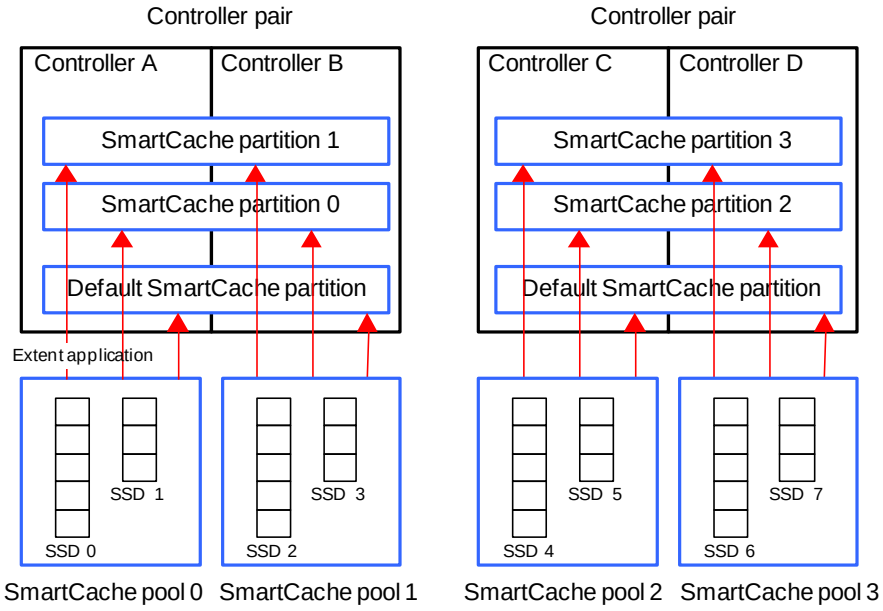
Thanks to the SAN and NAS converged architecture, SmartDedupe and SmartCompression can be enabled or disabled for file systems and thin LUNs, achieving block-based online data shrinking. As SmartDedupe and SmartCompression are independent from each other, they can be enabled or disabled independently. In addition, enabling and disabling these functions do not compromise system performance. SmartDedupe and SmartCompression provided by the V3 Converged storage systems work in in-line mode. When the functions are enabled, new data is deduplicated and compressed. When the functions are disabled, deduplicated data cannot be restored.

8 Intelligent: SmartCache

SmartCache employs high-performance SSDs to compose an independent performance tier to accelerate data. The SSD tier can serve as cache expansion to cache hotspot data that the upper-tier cache cannot store or as read cache for internal metadata. For example, SmartCache can be directly used to accelerate the read of deduplicated metadata.

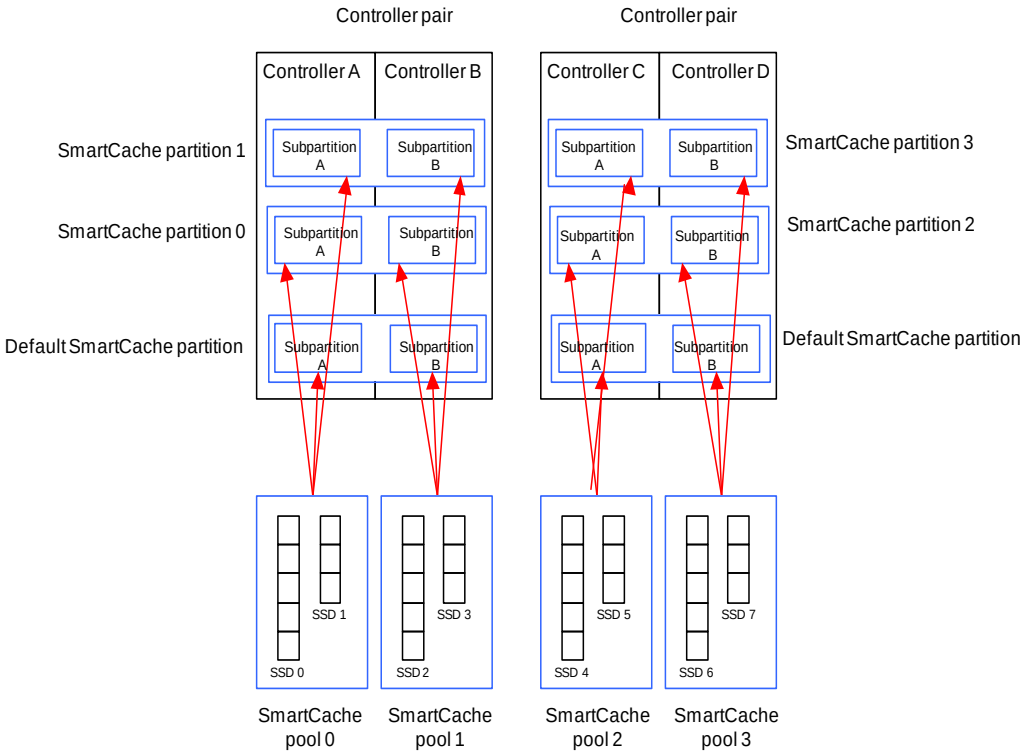
SmartCache employs SmartCache pools and SmartCache partitions to manage SSD resources. By default, one SmartCache pool is generated on each controller. The SmartCache pool manages all SSDs in the controller and provides fine-grained local SSD resource application and release functions. Resources in each partition of the SmartCache pool must come from different SSDs to ensure SSD load balancing.

Figure 8-1 Relationships between SmartCache pools and SmartCache partitions



By default, one default SmartCache partition is generated on each controller pair. Users can create multiple SmartCache partitions based on the controller pair. SmartCache partitions on a controller pair consist of subpartitions on the two controllers. After obtaining fine-grained, discontinuous SSD resources, SmartCache subpartitions complete more fine-gained resource applications and release. After obtaining cache resources, SmartCache subpartitions independently cache and evict data.

Figure 8-2 Relationship between SmartCache pools and subpartitions



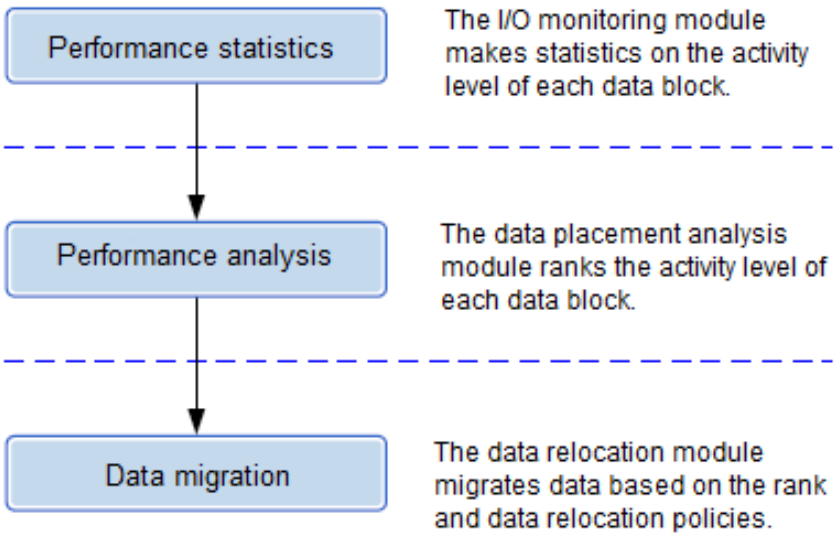
For file systems, three types of data are involved: user data (object data on disks), indirect blocks, and file metadata (stored in the object management structure of disks). When a user data or file property operation is complete in level-1 cache, the file system returns an operation success acknowledgement to users. The subsequent operations can be performed asynchronously. File system access is accelerated by modifying file system metadata and combing small I/Os.

9 Intelligent: SmartTier

The V3 Converged storage systems support Huawei's self-developed SmartTier feature. This feature is used to implement automatic storage tiering. SmartTier stores right data onto right media at right time. SmartTier improves storage system performance and reduces storage costs to meet enterprises' requirements on both performance and capacities. By preventing historical data from occupying expensive storage media, SmartTier ensures effective investment and eliminates energy consumption caused by useless capacities, reducing TCO and optimizing cost-effectiveness.

SmartTier performs intelligent data storage based on LUNs and identifies LUNs based on a data migration granularity from 512 KB to 64 MB. The data migration granularity is called extent. SmartTier collects statistics on and analyzes the activity levels of data based on extents and matches data of various activity levels with storage media. Data that is more active will be promoted to higher-performance storage media (such as SSDs), whereas data that is less active will be demoted to more cost-effective storage media with larger capacities (such as NL-SAS disks). The data migration process of SmartTier consists of performance statistics collection, performance analysis, and data migration, as shown in the following figure:

Figure 9-1 Three phases in data processing by SmartTier



Performance statistics collection and performance analysis are automated by the storage system based on users' configuration, and data migration is initiated manually or by a user-defined scheduled policy.

The I/O monitoring module collects performance statistics.

SmartTier allows user-defined I/O monitoring periods. During the scheduled periods, it collects statistics on data reads and writes. Activity levels of data change throughout the data life cycle. By comparing the activity level of one data block with that of another, the storage system determines which data block is more frequently accessed or less frequently accessed. The activity level of each extent is obtained based on the performance indicator statistics of data blocks.

The working principle is as follows:

- During scheduled I/O monitoring periods, each I/O is recorded to serve as data sources for performance analysis and forecasting. The following information is recorded based on extents: data access frequency, I/O size, and I/O sequence.
- The I/O monitoring module records the I/Os of each extent based on memories. Each controller can monitor a maximum of 512 TB storage space.
- The I/O monitoring module performs weighting for I/O statistics on a daily basis to weaken the impact of historical services on current services.

The data placement analysis module implements performance analysis.

The collected performance statistics are analyzed. This analysis produces rankings of extents within the storage pool. The ranking progresses from the most frequently accessed extents to the least frequently accessed extents in the same storage pool. Note that only extents in the same storage pool are ranked. Then a data migration solution is created. Before data migration, SmartTier determines the direction of relocating extents according to the latest data migration solution.

The working principle is as follows:

- The data placement analysis module determines the I/O thresholds of extents on each tier based on the performance statistics of each extent, the capacity of each tier, and the access frequency of each data block. The hottest data blocks are stored to the tier of the highest performance.
- Extents that exceed the thresholds are ranked. The hottest extents are migrated first.
- During data placement, a policy is made specifically for SSDs and another policy is made to proactively migrate sequence-degraded extents from SSDs to HDDs.

The data relocation module migrates data.

Frequently accessed data (hotspot data) and seldom accessed data (cold data) are redistributed after data migration. Random hotspot data is migrated to the high-performance tier and performance tier, and non-hotspot data and high-sequence data are migrated to the capacity tier, meeting service performance requirements. In addition, the TCO of the storage system is minimized and the costs of users are reduced.

SmartTier has two migration triggering modes: manual and automatic. The manual triggering mode has a higher priority than the automatic one. In manual triggering mode, data migration can be triggered immediately when necessary. In automatic triggering mode, data migration is automatically triggered based on a preset migration start time and duration. The start time and duration of data migration are user-definable.

In addition, SmartTier supports three levels of data migration speeds: high, medium, and low. The upper limits of the low-level, medium-level, and high-level data migration rates are 10 MB/s, 20 MB/s, and 100 MB/s respectively.

The working principle is as follows:

- The data relocation module migrates data based on migration policies. In the user-defined migration period, data is automatically migrated.
- The data relocation module migrates data among various storage tiers based on migration granularities and the data migration solution generated by the data placement analysis module. In this way, data is migrated based on activity levels and access sequences.
- The data relocation module dynamically controls data migration based on the current load of a storage pool and the preset data migration speed.
- The minimum unit for data migration is extent. Service data can be correctly accessed during migration. Relocating an extent is to read data from the source extent and write the data to the target extent. During data migration, read I/Os read data from the source extent while write I/Os write data to both the source and target extents. After data migration, the metadata of the source and target extents is modified. Then read and write I/Os access the target extent. The source extent is released.

10 Intelligent: SmartThin

The V3 converged storage systems support Huawei's self-developed SmartThin feature. This feature is used to implement thin provisioning. SmartThin allows users to allocate desired capacities to LUNs when creating LUNs. When LUNs are being used, storage capacities are allocated on demand to improve storage resource utilization and meet the requirements of growing services. SmartThin does not allocate all space out, but presents users a virtual storage space larger than the physical storage space. In this way, users see larger storage space than the actual storage space. When users begin to use storage space, SmartThin provides only required space to users. If the storage space is insufficient, SmartThin expands the capacity of the back-end storage unit. The whole expansion process is transparent to users and causes no system downtime.

If the actual amount of data is larger than expected, LUN space can be adjusted dynamically. Free space can be allocated to any LUN that needs space. In this way, storage space utilization and effectiveness are improved. In addition, LUN space can be adjusted online without affecting services.

SmartThin creates thin LUNs based on RAID 2.0+ virtual storage resource pools. Thin LUNs and thick LUNs coexist in a same storage resource pool. Thin LUNs are logical units created in a thin pool. They can be mapped and then accessed by hosts. The capacity of a thin LUN is not its actual physical space, but only a virtual value. Only when the thin LUN starts to process an I/O request, it applies for physical space from the storage resource pool based on the COW policy.

SmartThin allows a host to detect a capacity larger than the actual capacity of a thin LUN. The capacity detected by a host is the capacity that a user can allocate to the thin LUN, namely the volume capacity (virtual space) displayed on the host after a thin LUN is created and mapped to the host. The actual capacity of a thin LUN refers to the physical space actually occupied by a thin LUN. SmartThin hides the actual capacity of the thin LUN from the host and provides only the nominal capacity of the thin LUN.

In addition, SmartThin allows users to create a thin LUN whose capacity is larger than the maximum available physical capacity of a storage resource pool. For example, if the maximum physical capacity of a storage resource pool is 2 TB, SmartThin allows users to create a thin LUN larger than 10 TB.

SmartThin uses the capacity-on-write and direct-on-time technologies to respond to read and write requests from hosts to thin LUNs. Capacity-on-write is used to allocate space upon writes, and direct-on-time is used to redirect data.

- Capacity-on-write

  When a thin LUN receives a write request from a host, the thin LUN uses direct-on-time to determine whether the logical storage location of the request is allocated with an actual storage location. If the actual storage location is not allocated, a space allocation

task is triggered with a minimum grain of 64 KB. Then data is written to the newly allocated actual storage location.

- Direct-on-time

  Because capacity-on-write is used, the relationship between the actual storage location and logical storage location of data is not calculated using the formulas, but is determined by mappings based on capacity-on-write. Therefore, when a thin LUN is read or written, the relationship between the actual storage location and logical storage location must be updated based on the mapping table. A mapping table is used to record mappings between actual storage locations and logical storage locations. A mapping table is dynamically updated in the write process and is queried during the read process. Therefore, direct-on-time is divided into read direct-on-time and write direct-on-time.

  - Read direct-on-time: After a thin LUN receives a read request from a host, it queries the mapping table. If the logical storage location of the read request is assigned an actual storage location, the thin LUN redirects the logical storage location to the actual storage location, reads data from the actual storage location, and returns the read data to the host. If the logical storage location of the read request is not assigned an actual storage location, the thin LUN sets data at the logical storage location to all zeros and returns all zeros to the host.
  - Write direct-on-time: After a thin LUN receives a write request from a host, it queries the mapping table. If the logical storage location of the write request is assigned an actual storage location, the thin LUN redirects the logical storage location to the actual storage location, writes data to the actual storage location, and returns an acknowledgement to the host indicating a successful data write. If the logical storage location of the write request is not assigned an actual storage location, the thin LUN performs operations based on capacity-on-write.

SmartThin supports the online expansion of a single thin LUN and a single storage resource pool. The two expansion methods do not affect services running on a host.

- The expansion of a single thin LUN is to expand the nominal storage space of the thin LUN. After the nominal storage space of a thin LUN is modified, SmartThin provides the new nominal storage space of the thin LUN to the host. Therefore, the volume capacity (virtual space) displayed on the host is the capacity after expansion. In the expansion process, the original storage location is not adjusted. If new data needs to be written to the newly added thin LUN storage space, the thin LUN applies for physical space from the storage resource pool based on capacity-on-write.
- The expansion of a storage resource pool is a capability provided by RAID 2.0+ storage virtualization. Storage capacities are expanded without affecting services running on hosts. In addition, SmartMotion balances data among all the disks including newly added disks in the storage resource pool.

  SmartThin provides two methods of space reclamation: standard SCSI command (**unmap**) reclamation and all-zero data space reclamation. The working principles of these two methods are described as follows:

- Standard SCSI command reclamation: When a virtual machine is deleted, a host issues the **unmap** command using the SCSI protocol. After receiving this command, SmartThin uses direct-on-time to search for the actual storage location that corresponds to the logical storage location to be released on a thin LUN, releases the actual storage location on the thin LUN to a storage resource pool, and removes the mapping from the mapping table. To use this space reclamation method, applications on hosts must be able to issue the **unmap** command. VMware, SF, and Windows 2012 support the **unmap** command.
- All-zero data space reclamation: When receiving the write request from a host, SmartThin determines whether data blocks contained in the write request are all zeros. If the logical storage location that issues the all-zero data space is not allocated with an

actual storage location, SmartThin returns a message indicating a successful data write to the host without space allocation. If the logical storage location that issues the all-zero data space is allocated with an actual storage location, SmartThin releases the actual storage location from the thin LUN to the storage resource pool, removes the mapping from the mapping table, and returns a message indicating a successful data write to the host. This space reclamation method does not require any special commands from hosts.

11 Intelligent: SmartMigration

The V3 converged storage systems employ LUN migration to provide intelligent data migration. 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 heterogeneous storage system. The LUN migration feature provided by the V3 converged storage systems is called SmartMigration.

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:

- 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.

- Service performance adjustment

  SmartMigration can be used to improve or reduce service performance. It can migrate services either between two LUNs that have different performances within a storage system or between two storage systems that have different configurations.

  - Service migration within a storage system

    When the performance of a LUN that is carrying services is unsatisfactory, you can migrate the services to another LUN that provides higher performance to boost service performance. For example, if a user requires quick read/write capabilities, the user can migrate services from a LUN created on low-speed storage media to a LUN created on high-speed storage media. Conversely, if the priority of a type of services decreases, you can migrate the services to a low-performance LUN to release the

high-performance LUN resources for other high-priority services to improve storage system serviceability.

- Service migration between storage systems

  When the performance of an existing storage system fails to meet service requirements, you can migrate services to a storage system that provides higher performance. Conversely, if services on an existing 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).

- Service reliability adjustment

  SmartMigration can be used to adjust service reliability of a storage system.

  - To enhance the reliability of services on a LUN with a low-reliability RAID level, you can migrate the services to a LUN with a high-reliability RAID level. If services do not need high reliability, you can migrate them to a low-reliability LUN.
  - Storage media offer different reliabilities even when configured with the same RAID level. For example, when the same RAID level is configured, SAS disks provide higher reliability than NL-SAS disks and are more often used to carry important services.

- LUN type adjustment to meet changing service requirements

  Thin LUNs and thick LUNs can be flexibly converted without interrupting host services.

12 Efficient: SmartQoS

The V3 converged storage systems support Huawei's self-developed SmartQoS feature. This feature is used to ensure the QoS. SmartQoS intelligently schedules and allocates computing resources, cache resources, concurrent resources, and disk resources of a storage system, meeting the QoS requirements of services that have different priorities.

SmartQoS uses the following technologies to ensure the quality of data services:

- **I/O priority scheduling:** Service response priorities are divided based on the importance levels of different services. When allocating system resources, a storage system gives priority to the resource allocation requests initiated by services that have the high priority. If resources are in shortage, more resources are allocated to services that have the high priority to maximize their QoS. Currently, three priorities are available: high, medium, and low.
- **I/O traffic control**: Based on a user-defined performance control goal (IOPS or bandwidth), the traditional token bucket mechanism is used to control traffic. I/O traffic control prevents specific services from generating excessive large traffic that affects other services.
- **I/O performance assurance**: Based on traffic suppression, a user is allowed to specify the lowest performance goal (minimum IOPS/bandwidth or maximum latency) for a service that has a high priority. If the minimum performance of the service cannot be ensured, the storage system gradually increases the I/O latency of low-priority services, thereby restricting the traffic of low-priority services and ensuring the lowest performance goal of high-priority services.

The I/O priority scheduling is implemented based on storage resource scheduling and allocation. In different application scenarios, the performance of a storage system is determined by the consumption of storage resources. Therefore, the system performance is optimized as long as resources, especially bottleneck resources, are properly scheduled and allocated. The I/O priority scheduling technique monitors the usage of computing resources, cache resources, concurrent resources, and disk resources. If a resource bottleneck occurs, resources are scheduled to meet the needs of high-priority services to the maximum. In this way, the QoS of mission-critical services is ensured in different scenarios.

The I/O priority scheduling technique employed by SmartQoS schedules critical bottleneck resources on I/O paths. Those resources include **computing resources**, **cache resources**, **concurrency resources**, and **disk resources**. Scheduling policies are implemented based on user-defined LUN priorities. The priority of a LUN is determined by the importance of applications served by the LUN. There are three LUN priorities available: **high**, **medium**, and **low**.

The I/O priority scheduling technique controls the allocation of front-end concurrency resources, CPU resources, cache resources, and back-end disk resources to control the response time of each schedule object.

- Priority scheduling of **front-end concurrency resources** is implemented at the front end of the storage system to control concurrent access requests from hosts. A storage system's capability to process concurrent host access requests is limited. Therefore, when the maximum number of concurrent host accesses allowed by a storage system is reached, SmartQoS restricts the maximum number of concurrent accesses for each priority based on the number of LUNs of each priority running under each controller. The restriction principle is as follows: High-priority services and large-traffic services are allocated a larger number of concurrent access resources.
- Priority scheduling of **computing resources** is implemented by controlling the allocation of CPU runtime resources. Based on the weight of each of high, medium, and low priorities, SmartQoS allocates CPU runtime resources to services of each priority. When CPU resources become a performance bottleneck, priority scheduling ensures that high-priority services are allocated more CPU runtime resources.
- Priority scheduling of **cache resources** is implemented by controlling the allocation of cache page resources. Based on the weight of each priority, SmartQoS first processes page allocation requests initiated by high-priority services.
- Priority scheduling of **disk resources** is implemented by controlling the I/O delivery sequence. Based on the priorities of I/Os, SmartQoS enables high-priority I/Os to access disks first. If most I/Os are queuing on the disk side, priority scheduling of disk resources reduces the queuing time of high-priority I/Os. In this way, the overall latency of high-priority I/Os is reduced.

The priority scheduling technique employed by SmartQoS is implemented based on LUN or file system priorities. Each LUN or file system has a priority property, which is configured by a user and saved in a database. When a host (SCSI target) sends an I/O request to a disk array, the disk array gives a priority to the I/O request based on the priority of the LUN or file system that will process the I/O request. Then the I/O carries the priority throughout its processing procedure.

The I/O traffic control technique of SmartQoS:

- Restricts the performance of some applications in the system by limiting the total IOPS or bandwidth of one or multiple LUNs in the storage system. This technology prevents those applications from generating high burst traffic that may affect the performance of other services in the system.
- Limits data processing resources available for data services on specific LUNs. The objects of traffic control consist of the I/O class (read, write, or read and write) and traffic class (IOPS or bandwidth). Based on the two classes, a 2-tuple (I/O class and traffic class) is obtained for traffic control specific to a certain LUN.
- Controls traffic based on the I/O class and the obtained 2-tuple. Each I/O class corresponds to one traffic control group. Each traffic control group contains a certain number of LUNs and LUN groups whose maximum traffic is restricted. The I/O class-based traffic control function is implemented based on the I/O class queue management, token allocation, and dequeue control.

The I/O latency control technique of SmartQoS ensures the minimum performance requirements of some critical services by restricting low-priority services. Users can set minimum performance requirements for high-priority services. If those requirements cannot be met, the storage system restricts low- and medium-priority services in sequence to ensure the minimum performance set for the high-priority services.

SmartQoS restricts the performance of low- and medium-priority services by gradually increasing their latency. To prevent performance instability, SmartQoS does not prolong the latency after the latency reaches the upper limit. If the actual lowest service performance

becomes 1.2 times of the preset lowest performance indicator, the storage system gradually cancels the increased latency of low- and medium-priority services.

13 Efficient: SmartPartition

The V3 converged storage systems support Huawei's self-developed SmartPartition feature. This feature is used to optimize cache partitioning. The core idea of SmartPartition is to ensure the performance of mission-critical applications by partitioning core system resources. An administrator can allocate a cache partition of a specific size to an application. The storage system ensures that the application uses the allocated cache resources exclusively. Based on the actual service condition, the storage system dynamically adjusts the front- and back-end concurrent accesses to different cache partitions, ensuring the application performance of each partition. SmartPartition can be used with other QoS technologies (such as SmartQoS) to achieve better QoS effects.

Caches are classified into read caches and write caches. The read cache pre-fetches and retains data to improve the hit ratio of read I/Os. The write cache improves the disk access performance by means of combination, hitting, and sequencing. Different services need read and write caches of different sizes. SmartPartition allows users to specify read and write cache sizes for a partition, meeting cache requirements of different services.

The read cache configuration and the write cache configuration affect the I/O procedure differently. The impact on the write I/Os shows up in the phase of cache resource allocation. In this phase, the host concurrency and write cache size of a partition are determined. The reason for determining the two items in this phase is that it is the initial phase of a write procedure and the cache of the storage system is actually not occupied in this phase.

The impact on read I/Os involves two aspects. The first aspect is similar to the write I/O scenario. Specifically, the storage system determines whether the host concurrency meets the requirement. If not, the storage system returns the I/Os. The read cache is intended to control the size of the cache occupied by read data. The size of a read cache is controlled by the read cache knockout procedure. Therefore, the second aspect of the impact shows up in the read cache knockout procedure. If the read cache size of the partition does not reach the threshold, read cache resources are knocked out extremely slowly. Otherwise, read cache resources are knocked out quickly to ensure that the read cache size is below the threshold.

Compared with host applications, the processing resources of a storage system are limited. Therefore, a storage system must restrict the total host concurrency amount. For each partition, the concurrency is also restricted to ensure the QoS.

Regarding SmartPartition, the host concurrency of a partition is not fixed but calculated based on the priority weighted algorithm with the following factors taken into account:

- Number of active LUNs in the partition in the last statistics period
- Priorities of active LUNs in the partition in the last statistics period
- Number of I/Os completed by each LUN in the last statistics period

- Number of I/Os returned to hosts because the partition concurrency reaches the threshold in the last statistics period

Weighting the preceding factors not only fully uses the host concurrency capability but also ensures the QoS of a partition.

After one statistics period elapses, the concurrency capability of a partition may need to be adjusted based on the latest statistical result. The SmartPartition logic adjusts the concurrency capability based on a specific step to ensure a smooth adjustment, minimizing host performance fluctuation.

Similar to host concurrency control, back-end concurrency control is also intended to fully use system resources while ensuring the QoS of a partition. The back-end concurrency is calculated based on the priority weighted algorithm with the following factors taken into account:

- Amount of dirty data on high-priority LUNs in a partition in the last statistics period
- Disk flushing latency of LUNs in a partition in the last statistics period
- Actual disk flushing concurrency of LUNs in a partition in the last statistics period

The adjustment period and approach are similar to those of host concurrency control.

14 Summary

The V3 Converged storage systems are brand-new converged storage systems that employ a cloud-oriented architecture. Built based on block-level virtualization and SAN and NAS convergence, the storage systems support online deduplication and compression and heterogeneous virtualization. In addition, SmartPartition and SmartCache accelerate and improve the performance of file system sharing and block-based access, and SmartDedupe and SmartThin improve the storage efficiency. The V3 Converged storage systems can easily cope with performance and management challenges in virtualization scenarios. Thanks to heterogeneous virtualization and the convergence of SAN and NAS, SSDs and HDDs, the V3 Converged storage systems deliver robust performance, enhanced reliability, high availability, and improved cost-effectiveness. Huawei is dedicated to providing high-quality storage products and user-friendly services for customers. Guiding by this concept, the V3 Converged storage systems fully meet customers' requirements for high performance, functions, and efficiency, and maximize customers' benefits.

15 Acronyms and Abbreviations

Table 15-1 Acronyms and abbreviations

Acronym and Abbreviation	Full Spelling
BBU	Backup Battery Unit
BP	Block Point
DIF	Data Integrity Field
eDevLUN	External Device LUN
HDD	hard disk drive
IOPS	Input and Output Per Second
KV_DB	KeyValue-DataBase
LUN	Logical Unit Number
NL-SAS	Nearline Serial Attached SCSI
OLAP	On Line Analysis Process
OLTP	On Line Transaction Process
QoS	Quality of Service
RAID	Redundant Array of Independent Disks
SAS	Serial Attached SCSI
SCSI	Small Computer System Interface
SSD	Solid State Disk
TCO	Total Cost Ownership
WAFL	Write Anywhere File Layout
WWN	World Wide Number